Article

An Interactive Scholarly Collaborative Network Based on Academic Relationships and Research Collaborations

Abrar A. Almuhanna *, Wael M. S. Yafooz * and Abdullah Alsaeedi

Department of Computer Science, College of Computer Science and Engineering, Taibah University, Medina 20012, Saudi Arabia; aasaeedi@taibahu.edu.sa

* Correspondence: abrar_ali_m@hotmail.com (A.A.A.); waelmohammed@hotmail.com (W.M.S.Y.)

Abstract: In this era of digital transformation, when the amount of scholarly literature is rapidly growing, hundreds of papers are published online daily with regard to different fields, especially in relation to academic subjects. Therefore, it difficult to find an expert/author to collaborate with from a specific research area. This is thought to be one of the most challenging activities in academia, and few people have considered authors' multi-factors as an enhanced method to find potential collaborators or to identify the expert among them; consequently, this research aims to propose a novel model to improve the process of recommending authors. This is based on the authors' similarity measurements by extracting their explicit and implicit topics of interest from their academic literature. The proposed model mainly consists of three factors: author-selected keywords, the extraction of a topic's distribution from their publications, and their publication-based statistics. Furthermore, an enhanced approach for identifying expert authors by extracting evidence of expertise has been proposed based on the topic-modeling principle. Subsequently, an interactive network has been constructed that represents the predicted authors' collaborative relationship, including the top-k potential collaborators for each individual. Three experiments have been conducted on the collected data; they demonstrated that the most influential factor for accurately recommending a collaborator was the topic's distribution, which had an accuracy rate of 88.4%. Future work could involve building a heterogeneous co-collaboration network that includes both the authors with their affiliations and computing their similarities. In addition, the recommendations would be improved if potential and real collaborations were combined in a single network.

Keywords: scholarly big data; scholar similarity; collaborator recommendation; expert finding; expertise evidence; academic social networking

check for updates

Citation: Almuhanna, A.A.; Yafooz, W.M.S.; Alsaeedi, A. An Interactive Scholarly Collaborative Network Based on Academic Relationships and Research Collaborations. *Appl. Sci.* **2022**, *12*, 915. https://doi.org/10.3390/app12020915

Academic Editors: Gaoang Wang and Jenq-Neng Hwang

Received: 10 December 2021
Accepted: 14 January 2022
Published: 17 January 2022

1. Introduction

In recent times, global collaboration between researchers has majorly expanded the level of cooperation between them and universities with regard to the exchange of knowledge and resources. Academics are now transmitting their knowledge and publications throughout the world. Digital technology has the power to bridge distances and promote cross-disciplinary and cross-border collaborations, which are an inspiration for scholarly networks. Such collaborations will consequently improve the experience of social activities at conferences and will create larger professional networks. Scholarly networks also help with shared learning and the transfer of technologies to poor countries. Such partnerships between scientists increase the number of citations and open up new and improved opportunities for further research and collaboration. Today, the transfer of knowledge has become extremely easy, and the number of citations has increased tremendously [1]. In addition, scholarly communities have revolutionized the methods of information publication and research sharing. Indeed, many of these groups have emerged and created interrelations that may affect the structure of the research community itself. For instance, the website ResearchGate is designed to facilitate access to academic research and increase collaboration between researchers. People from all around the world register there and obtain an instant online presence, and it also provides them with a convenient platform for showcasing their research, even before the publication stage. Today, researchers commonly use such platforms to discuss their research with the academic community and create professional partnerships. One of the largest examples is Google Scholar, which supports new and inexperienced scientists in finding support and advice from people who are more experienced in their field. Despite this easy access, scholarly communication and publishing still face shortcomings in terms of maximum accessibility and usability, support for an expandable range of contributions, open infrastructure, and, most importantly, community building [2].

In recent years, social-network (SN) analysis has become a widely used technique across diverse fields. SN analysis is not only applied to analyze social media applications, such as Facebook and Twitter, but it is also more broadly utilized in relation to scientific research, where it provides integrated services. Moreover, this form of analysis has been increasingly used to aid research on subjects by facilitating socialization and collaboration as SNs bring together interrelated organizations and individuals with common interests or objectives. Academic social networks (ASNs) contain academic entities, such as publications, authors, academic institutions, keywords, or content, which are represented by nodes (Figure 1), whereas connections between entities usually indicate relationships such as co-citations and co-collaboration. These different types of relationships form different networks, which create diverse perspectives in terms of scholarly communication and research interaction. For researchers with a vast amount of scholarly data, extracting valuable collaborations is often a highly time-consuming and complex task. Through ASNs, looking deeper into scholarly relationships may aid scholars in determining possible co-authorship. Finding collaborators for researchers is a critical problem that needs to be addressed; therefore, a need exists to address relevant k-collaborators for a selected author [3].

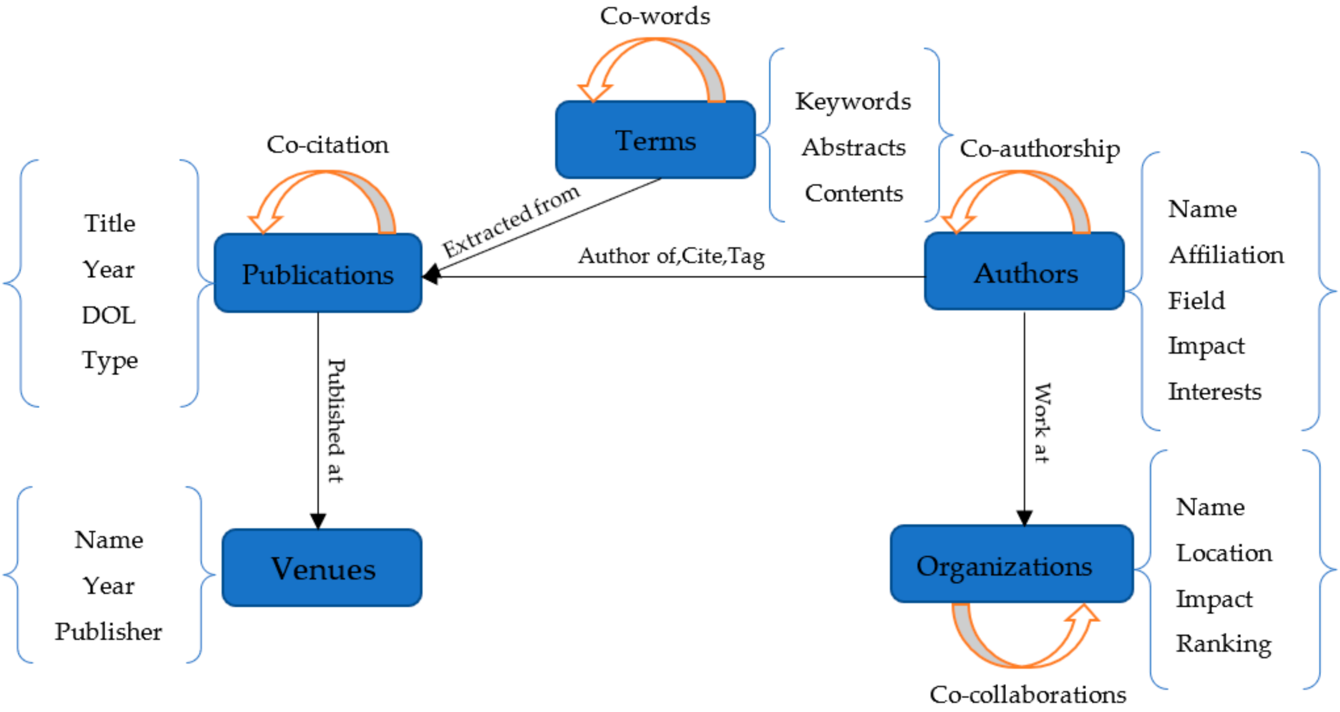


Figure 1. Academic social networks' entities [3].

Furthermore, scholarly data is expanding every day, creating a knowledge repository. This heterogeneous network contains various types of information, such as papers, journals, and authors. At present, the extraction of relevant experts in a specified field is one of the most critical research topics, and experts are considered the most crucial entity in the vast knowledge society. Indeed, in many scenarios, the recognition of researchers' skills is of considerable importance; for example, appreciating the qualitative attributes of an expert can be vital to a project. In every field of life, experts with professional skills, knowledge, and experience are required. From social institutions to research institutes, they play an integral role in team guidance, team management, and productivity and efficiency improvements. Experts are diverse in nature and continually evolve; thus, acquiring relevant experts from a wide range of information networks is critical. In some instances, authors have similar areas of expertise but differ in their specialized fields. For instance, many papers have been written on computer science; identifying an author who specializes in computer science is crucial because he or she could be a specialist in artificial intelligence, networks, web development, or data science.

Similarly, this logic can be applied to all other domains of science, technology, the humanities, and anywhere research work is applied [4]. However, to keep track of the expertise in an academic community, researchers are diligently trying to automate this problem within this domain. By processing the content of published research papers, several theories have been proposed to extract the expertise of authors. We have reviewed the existing approaches in [5], where a few have focused on query matching using keywords [6], while others have focused on the number of citations [7] or extracting and categorizing scholarly articles using their abstracts [8]. On the other hand, some research has concentrated on mining the relationships of authors with the aim of finding the most suitable collaboration based on the similarities between them; for example, studies have inspected likenesses in terms of venues and topics [9], co-word and co-citation similarity [10], the topics used and the academic rank of collaborators [11], or the demographical data of their topics [12]. Consequently, information visualization is a valuable methodology for organizing data with higher volumes and providing accurate computations, especially for a scholarly network. However, this field is still progressing, and researchers are attempting to make the task more efficient, less time-consuming, and less complex.

The overall aim of this research is to build a model for academic researchers to recommend the most significantly influential collaborations based on similarities in their publication records while recognizing the most significant factor with regard to the outcome similarity patterns. Additionally, another goal is to identify the expert authors who have a significant role in each research area and to facilitate the representation of this information into a network graph by grouping them together based on their research areas. Furthermore, studying the influence of collaborations with universities where these institutions are affiliated with authors' publications is also a priority.

In particular, we examine the following research questions:

- Q1: How do the selected factors affect the quality of the collaboration recommendation?
- Q2: How can the identification of experts be improved in a certain research area?
- Q3: Does the representation of the relationship and collaboration between experts and authors enhance the quality of the academic network based on their similarities?

This research considers the field of academic networks as the main subject of its analysis. The main contribution of this research can be seen from four perspectives:

- Building a novel model for improving the co-collaboration recommendation for academic authors based on multi-factors: author-selected keywords, total number of citations, total number of published papers, year ranges (from the year of the first publications to the year of the last publication), and the weight of topic distribution within their publications.
- Identifying the expert author in a specific research area relying on the multi-factor indicators for a topic of interest: total number of publications in the expertise area, total number of their citations, and total topic weight in terms of an author's publication-based topic distribution.

- Interactive network representation of the predicted authors' collaboration: utilizing a topic-clustering model based on their main research area, and for each author, there is a personalized ranked list of the top k-potential collaborators.
- Building a dataset: includes the multi-factors of individual authors' publications in order to mine their academic research area.

The rest of the paper is organized as follows. Section 2 begins with definitions of the main concepts discussed in this research and then goes on to review related studies. Section 3 elaborates on the methods and materials used in this research, while Section 4 discusses the experiments along with their results; a discussion of these results will be presented in Section 5. Finally, Section 6 concludes the paper with a summary.

2. Related Studies

This section reviews the literature relating to this research in order to find valuable collaborators and extract the author's expertise in the scholarly network.

2.1. Preliminaries

This section presents some definitions that are relevant to the current research for a better understanding of the common phrases.

1. **Co-citation Networks** are networks constructed according to the citation relationships between articles. If two publications have been cited on the same paper, this can be called co-citation. Since the two papers cannot cite each other at the same time, these networks are directed.
2. **Co-word Networks** reflect the frequency of co-words by exploring which papers have two keywords that appear simultaneously, which is also known as the keyword co-occurrence frequency. Co-word analysis is the process of discovering and visualising the interactions between keywords by studying the strengths of two keywords co-occurrence in two different publications. Keywords are especially powerful, as they are used by scholars and researchers to communicate the essence of an article; co-word analysis can thus investigate the network of ideas, scholarly topics, and research trends in a given discipline.
3. **Co-authorship Networks** as a commonly used ASN, each node in these networks refers to an author, while every edge denotes a co-authored relationship. Co-author networks can be studied from multiple perspectives, which is relevant to the vast majority of disciplines. Moreover, the rapid evolution and adoption of technology means that scientists can collaborate more easily across geographical boundaries and no longer need to be physically co-located at the same institution. Collaboration behaviors based on co-author networks can thus be studied by scholars, while collaboration teamwork has emerged as a new research pattern.
4. **Scientific Collaboration Network** is a form of social network in which the nodes represent scholars, and the links represent co-authorships. As such, if two authors have co-authored at least once, then there will be a link between them [12].
5. **Scholar Similarity Measurement** is based on the publication of two scholars and their behavior. The key concept is that if two scholars publish at same venues, or behave similarly in some aspect, they are more likely to be similar [12].
6. **Scientific Collaborator Recommendation** is the process of suggesting the suitable collaborators to a specific author. Potential collaborators are chosen based on their similarity and ranked according to how similar they are to the target authors [12].

2.2. Scholarly Data Analysis

This section discusses several approaches to multi-type correlations among diverse academic entities within scholarly big-data environments where industries and academic societies have attracted attention towards the phenomenon of scholarly big data itself [13,14].

Kwiek [15] investigated international academic collaboration in order to find a collaborator from two perspectives: the behavioral aspect was examined in relation to international research collaboration (IRC), and the attitudinal aspect was looked into with regard to international research orientation (IRO). The results of the experiment found that in terms of the academic discipline, individuals in hard academic fields consider their research to be less internationally concerned (low IRO), while they collaborate intensely with international communities (high IRC). The opposite situation was observed among individuals in soft academic fields. Leading on from this, Zhou [16] proposed the use of the multi-dimensional network analysis method (AIMN) in this area of research. AIMN is a computational method that describes multi-type correlations among basic academic entities. Additionally, their model highlights the relationships between academic activities through big-data-based scholarly environments, which includes underlining the connection between article recommendations, co-authors, etc. Moreover, they used an improved algorithm known as the random walk with restart (RWR) model in order to calculate the similarity relevance of two nodes. AIMN outperformed three other methods: MVCWalker, CCRec, and basic RWR without AIMN. Meanwhile, Batagelj and Maltseva [17] explored bibliographic networks using the longitudinal approach and introduced the concept of time by using temporal quantities. They also employed the authorship network in order to obtain the author with the largest number of works and the citation network to ascertain the most cited paper by implementing the multiplication of networks.

2.3. Author Similarity Identification

Scientific collaboration in different research areas becomes very popular and complex. For researchers having big scholarly data, it is often a very time-consuming and complex task to extract valuable collaboration [3]. Kumar et al. [18] proposed a model for co-citation analysis, where citing sentences was used by authors to increase the accuracy of the similarity analysis. Compared to the traditional approach, which only examines the first author, this technique considers all the authors within the reference section. In relation to the content-based approach, the lower level of graph density and the clustering coefficient makes the network sparse; hence, it is very useful for finding the detailed sub-structure of domain analysis.

Al-Sultany [19] proposed a novel method for recommending collaborators using four important connections: author–author, co-author–author, author–paper, and author–conference relationships. The model was implemented on a total of 100 target authors whose total number of collaborators range from 0 to 30. The PageRank algorithm was also utilized for addressing relevant k-collaborators for the selected authors while considering three factors: time of collaboration, times of collaboration, and co-author order where they estimate that the first and second authors in the publication have a higher level of contribution. The comparison of the proposed model with the MVSWalker indicates that the recommendation system provides highly accurate outcomes. Meanwhile, [20] proposed a similarity metric based on how likely it is that two authors will work in the same domain. Then, clusters can be classified that comprise researchers from both institutions who work on related research topics; subsequently, joint research projects can be suggested using the Louvain clustering method, and they can build a profile for each scholar that is represented by a set of tuples. Then, the average similarity matrix can be computed for the topic discovered based on the scholars' profile. They used a micro-evaluation in relation to the topic; as a result, they identified 48 predicted collaborations with a recall of 0.86.

Furthermore, Alshareef et al. [9] proposed a model to compute the similarity score between scholars in a scientific community based on their publications' metadata, such as the title, keywords, the abstract, and the place of publication. Additionally, they computed the topic similarity, which was extracted using the latent semantic analysis (LSA) model, which employed three types of similarities: author–author articles' metadata, venue–venue, and topic similarity. This was followed by ranking the combined author–venue similarity where the exploration of topical similarity will maintain the same subject as the scientific collaboration. As result, their hybrid approach, which consists of content-based and collaborative filtering approaches, has a positive impact on making more effective recommendations. Sun et al. [12] presented a novel system for the recommendation of personalized collaborators, which included demographical characteristics as an academic factor based on the fact that researchers have different collaboration strategies at different career ages. The model is based on three parts: the extraction of authorship from the DBLP dataset, the extraction of a topic based on articles' abstracts/titles, and random walk for career-age awareness to determine the significance of the collaboration between nodes x and y. Experimental results show that compared to six baseline methods, the proposed model outperformed.

2.4. Expert Recommendation Systems

Many researchers use search engines to locate an expert to work with, yet in various scenarios, they do not receive optimal outcomes during the initial phases because of their keyword-based searches. Several recommendation approaches have been proposed in order to rank the authors based on their expertise in a specific area [21–24].

Gao et al. [25] proposed a novel technique for the analysis and extraction of relevant experts by integrating a topic model and a paper cooperation network. Additionally, with the help of latent Dirichlet allocation (LDA), expert cooperation networks play a vital role in the extraction process of expert topics; for instance, if two experts often publish their research work together, they are likely to be associated with the same field. The experimental evaluation indicates that the perplexity of the model decreased first and then increased when they expanded the number of topics. The generalizability of the model was also achieved with the implementation of an optimal number of topics.

Furthermore, Berger et al. [26] proposed a contextualized embedding method (document-centric) for expertise data extraction rather than non-contextual techniques (word or team-centric), which investigated the effects of retrofitting on contextualized embeddings. In addition, they compared two different ways of merging paper titles and abstracts by embedding what was used to extract the semantic similarity. They also tested four innovative strategies to rank the authors of co-authored papers. Today, most papers are written by multiple authors; therefore, the ranking systems used in this study for scoring authors based on the strength of their connection to their paper's topic was a novel idea. Meanwhile, ref. [27] focused on finding experts in the field of computer science, specifically with regard to faculties of computer science at Indonesian universities, based on the abstracts of respective scholars' theses from students at Fasilkom. The model was designed based on word and vector representation by incorporating doc2vec and word2vec. Thus, suitable experts were identified to supervise theses subject to their expertise by using the cosine similarity to assign them. However, the results showed that using word2vec to represent the expertise performs was more productive than using doc2vec.

Li et al. [28] proposed a network approach for expertise retrieval using credit allocation for co-authors through analyses of the degree of each co-author's contribution to collaborative work, and it assigns expertise level according to their contribution. Thereby, the proposed methodology worked on the base of credit allocation, specific author papers, and medical subject heading (MeSH) connections using a topic-modeling LDA method. The weighted version of HeteAlloc uncovered the concentration of authors on a particular topic and classified the papers with smaller MeSh terms. The unweighted version is very helpful when the range of topics is too widely spread and most papers have multiple MeSh terms. Nevertheless, the unweighted version is often employed because the MEDLINE dataset mostly carries publications with numerous MeSh terms. Moreover, Javadi [29] proposed the expert finding system (EFS) where the quality of the results for a team of expert researchers will outperform the results for individual work. To guage the similarity, they used two measurements: co-occurrence and shared neighbor. As a result, the product value of the two weights gives an estimate that can indicate the relationship between the expertise of an author and the actual work of the user. Consequently, it provides a list of names for collaboration with an expert in a specific field, with an accuracy rate of 71.50%.

2.5. Scholarly Data Visualisation

Data visualization is a leading field of data mapping, which helps with predictions with regard to the structure, mapping, trends, and graphical representation of the data. The visualization is an important mapping feature that enables the user to find out the important trends within bulky amounts of data and their organization [30,31]. Mokhtari et al. [32] presented a bibliometric visualization and overview of the *Journal of Documentation* (*JDoc*) from its initiation in 1945 to 2018 with regard to 2056 papers. It is worth noting here that *JDoc* is one of the most pioneering and influential journals in the domain of library and information science (LIS). The analysis was performed in relation to different aspects. Overall, the research was very helpful in terms of creating an effective bibliometric visualization and examination of *JDoc*, and it helped to add some interesting topics to LIS to facilitate better connections between contributors. Liu et al., [33] also proposed a new system that used mining techniques called the web of scholars. The model uses methods such as searching, mining, and other complex network visualizations that correspond to scholars from the field of computer science. Where it mainly relies on the knowledge graph, the approach can effectively be used as an interoperable tool that can provide in-depth analysis. Meanwhile, [34] proposed a model based on evidence-based librarianship (EBL), which employs the integration of social network analysis by using keywords to identify collaboration networks and visualizations of the same elements in EBL research.

Although studies have been conducted by many authors, the problem of finding potential collaborators or expert scholars is still insufficiently explored. Thus, a more systematic and theoretical analysis is required in that regard. However, Table 1 presents a review of the literature, showing that most of the studies focused on the topic but neglected the extraction of the main topic relating to the authors. Additionally, the authors' publication histories are rarely included as a factor of similarity, even though they reveals his/her topics of interest over the years.

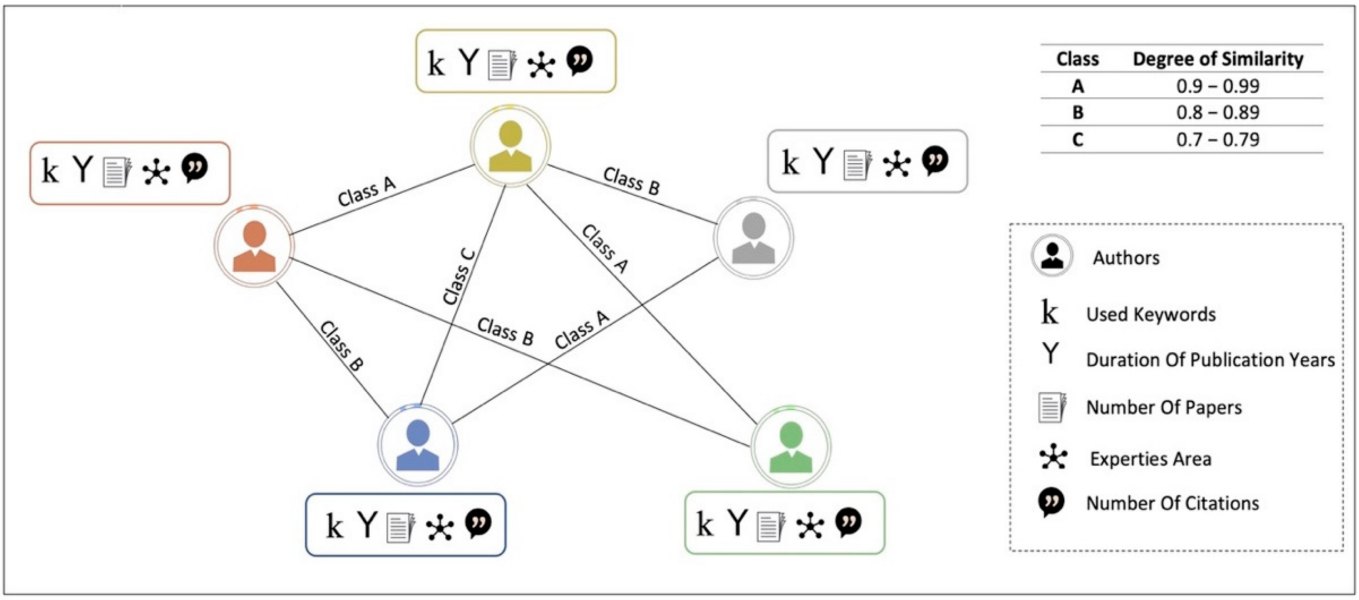
Table 1. Review of the used factors in the literature.

Author(s)	Type	Network	Metadate			Topic	Citation	Other Factors
			Keywords	Abstract	Venue			
[18]	Collaboration	Yes	No	No	No	No	Yes	-
[34]	Collaboration	No	No	Yes	No	No	No	Times of collaboration
[20]	Collaboration	Yes	No	Yes	No	Yes	No	-
[9]	Collaboration	No	Yes	Yes	Yes	Yes	No	-
[12]	Collaboration	Yes	No	Yes	No	Yes	No	Career age
[25]	Expert	Yes	No	No	No	Yes	No	Co-authored
[26]	Expert	No	No	Yes	No	Yes	No	Title
[27]	Expert	No	No	Yes	No	No	No	Supervisor's name
[28]	Expert	No	Yes	No	No	Yes	No	Number of publications
[29]	Expert	No	Yes	No	Yes	No	No	Publication year

3. Methods and Materials

This research employed a quantitative method that was based on the collection of quantifiable data and the subsequent application of computational and statistical techniques; thus, the correlation between the authors using different features was calculated where the impact of using one such feature on the relationships is observed along with how it changes if the number of the considered features also alters. This research builds on the existing literature within the field of academic informatics and expands on prior academic studies by scholars that examines the capabilities of the proposed networking platform to facilitate the prediction of suitable co-collaboration by providing an enhanced recommendation quality. In particular, the research is composed of two main parts: enhanced metrics and network representation. The first part aims to develop a new metric of similarity between authors based on multiple factors when attempting to predict the most similar authors to collaborate with; for each research area, the extraction of expert authors can also be undertaken. Meanwhile, the second part aims to use the knowledge gained from the analysis of the matrix to build an interactive academic network that represents the predictable collaborative relationships. Additionally, the goal is to build an interactive university network that aims to mine the relationships between these institutions. The underlying idea of this network is that if two publications from these affiliations share common keywords, they will have a higher probability of cooperating. The research concludes by elucidating how various academic disciplines might utilize such a predictive model to estimate a potential beneficial collaboration.

Figure 2 presents a conceptual view of the proposed network based on the aforementioned matrix where all published papers were collected for each author, including the citation number for each publication. This was followed by an extraction of the metadata for each paper (title, keywords, year). While researchers often act in various ways across multiple domains of interest, which may cause topic drift in general recommendation systems, the main research area for each author was extracted in order to determine the expertise area. Hence, while having all these factors, the similarity between the authors has been calculated based on five different factors: keywords used in published papers, publication span (based on the time from the first to the last publication), number of the published papers, total number of citations, and the degree of expertise in each area.



Class	Degree of Similarity
A	0.9 – 0.99
B	0.8 – 0.89
C	0.7 – 0.79

Figure 2. Conceptual view of the proposed approach.

3.1. Model Architecture

This section presents the model's design and details about the data gathering approach that was chosen, which are required to fully develop the taxonomy that is necessary to format and structure the research idea, including the sample orientation selection and the selection of similarity measures that are employed in order to construct an effective interactive network. As shown in Figure 3, the proposed model for this research can be divided into four phases: data acquisition; data pre-processing; extracting, linking, and mapping; and interactive network visualization. The first two phases are concerned with data collection and pre-processing, while the remaining two are the main components for the proposed approach. The third phase will extract the similarity between the authors, whereas the construction of the interactive network will be carried out in the last section.

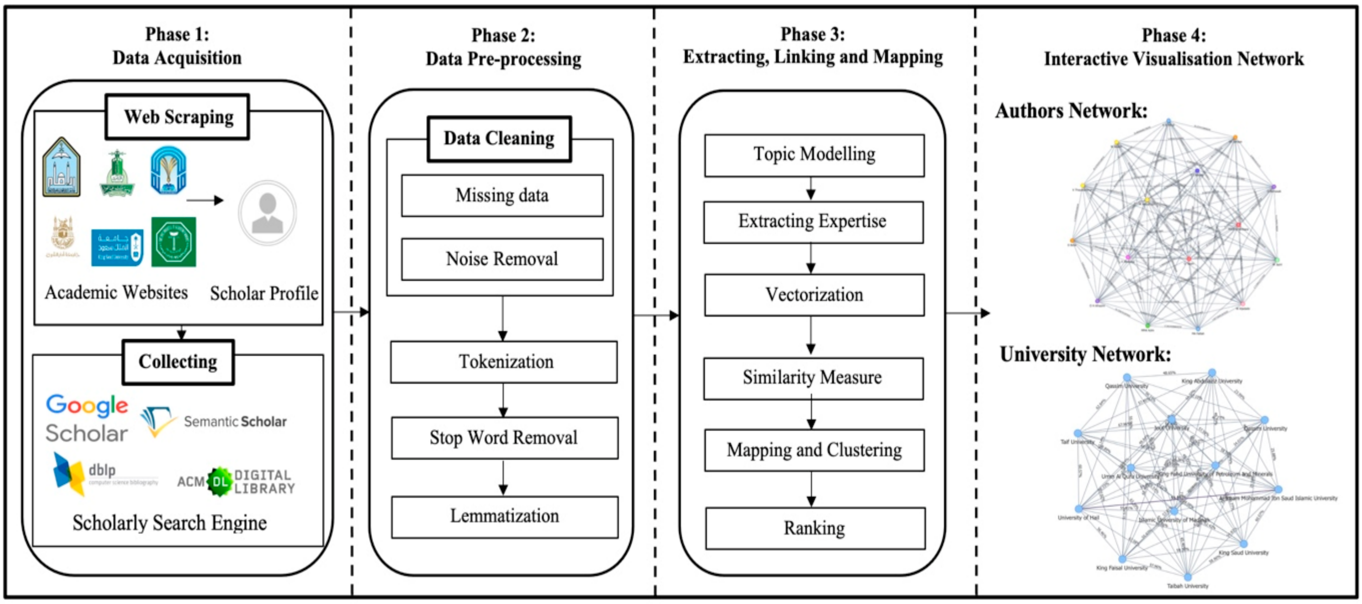


Figure 3. The proposed approach of the academic network.

3.1.1. Data Acquisition

This research adopted a web-scraping methodology to collect its initial dataset. The primary goal of this method is to automatically extract the data from a variety of unstructured websites that retrieve relevant content based on the query. Moreover, this study focused on academic publications from the field of computer science, which is our own research area, meaning we have a clear grasp of the publishing community; thus, the Publish or Perish (PoP) tool was used to gather the required data, which encompasses raw metadata from a variety of data sources such as Google Scholar, DBLP, and Microsoft Academic Search [35].

In addition, the collected publications were further restricted to authors from twelve Saudi Arabian universities, making it a more focused case study. Specifically, King Abdul Aziz University (KAU), King Fahd University of Petroleum and Minerals, King Saud University, Umm Al-Qura University, King Faisal University, Imam Abdulrahman Bin Faisal University (IAU), Islamic University of Madinah, Qassim University, Taibah University, Taif University, Jouf University, and the University of Hail were concentrated [36]. Hence, the authors' names were manually selected from the official university websites in relation to three different academic ranks: professor, associate professor, and assistant professor. Each university had 10 selected authors with 100 or fewer publications each. This narrower focus also allowed for the consideration of the specifics of the authors' publication time ranges. However, while the quality of the similarity result can be enhanced by providing additional data for the author, the data extracted using the PoP tool alone was insufficient and missing the keywords that the scholars had assigned to their publications, which is one of the important features of this study. Hence, we integrated the data from the PoP tool with the keywords and affiliations related to these authors' publications. Figure 4 illustrates the full process of collecting the required data.

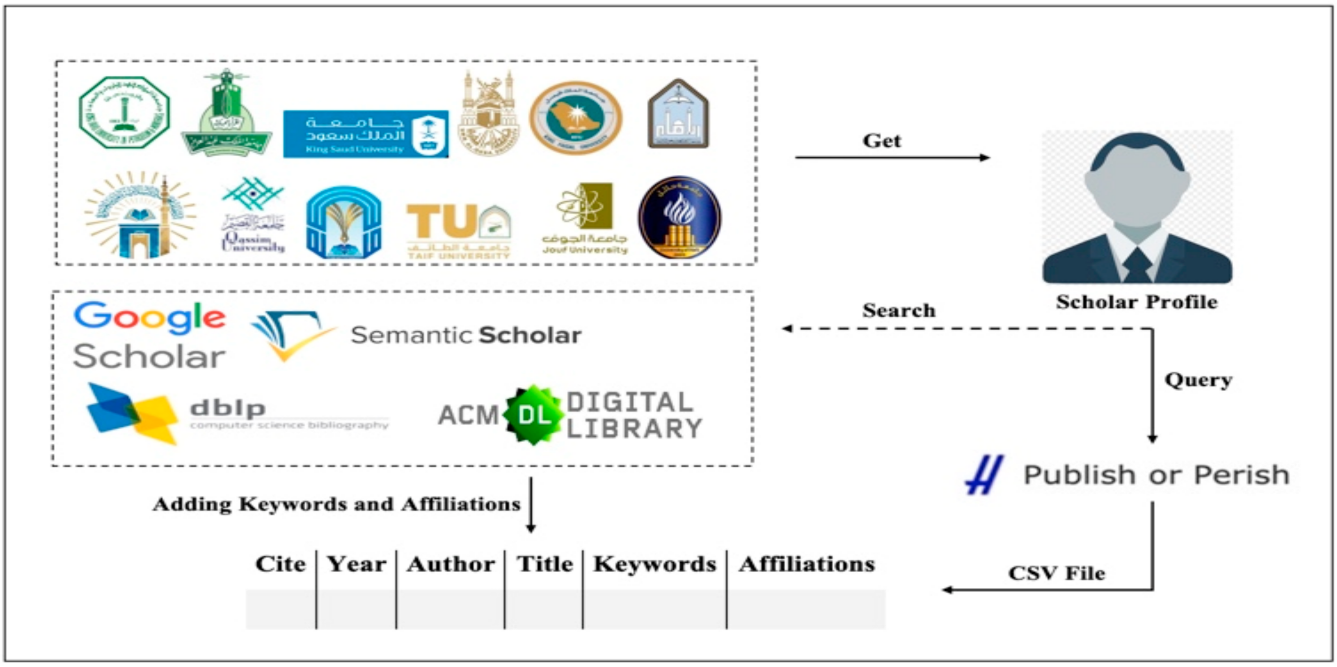


Figure 4. Process of collecting the academic publications.

3.1.2. Data Pre-Processing

At this stage, preparing input text for subsequent algorithmic analysis is a critical step in any text-processing procedure. According to the collected data, the pre-processing steps in this research were data cleaning, tokenization, stop word removal, and lemmatization, including part-of-speech (POS) tagging.

Data Cleaning: where all uppercase characters are converted to lowercase characters, and all extra characters, numbers, and whitespaces were removed. Additionally, any word with less than three characters and high-frequency words was removed. In addition, if the paper was in a language other than English or had no keywords, the full paper was omitted only from the keywords analysis [37].

Tokenization: to translate the human-readable text into machine-readable format for the purposes of text analysis; these were defined as meaningful units (tokens), hence breaking a stream of textual content up into words and thus forming a set of individual words separated by a line break or punctuation characters [38].

Stop words removal: each language contains words that are frequently used, also known as noise words, which are the words that provide little information and are rarely needed. The removal of stop words is mainly motivated by the need to enhance execution speed and accuracy. Without the removal of stop words, the similarity may be overloaded with junk words and would render inaccurate results [39].

Lemmatization: The method of identifying the dictionary forms of words (e.g., learn) given one of its inflected variants (e.g., learning, learns). It is generally followed by a labelling of words using a specific part-of-speech (POS) tagging [40].

3.1.3. Extracting, Linking, and Mapping

This phase consists of three main parts: topic modeling, similarity measures, and extracting expertise. In terms of topic modelling, this involved the clustering of authors and the extraction of the main area for each of them, which determined the field of expertise. This is followed by similarity measures based on three factors: keywords, publications statistics, and the research topic. The sections below will elaborate on these processes in detail.

Topic Modeling

In the academic community, publications are used to express scientific progress. Additionally, most academic papers adopt common formats and structures; consequently, research projects have similar keywords that can be defined using a uniform topic-event structure. In order to find the similarities between authors whose writings engage with the same area, topic modeling was applied with the aim of clustering in order to identify relatedness among these individuals. Subsequently, in order to determine the author's expertise, the topic has been extracted from each paper and the mean value for each topic has been calculated to determine the most relevant topic for each author based on their publications. However, as keywords are considered to be short texts, this may induce a sparsity problem; thus, the non-negative matrix factorization (NMF) model is the most suitable and popular method for overcoming this problem, as it offers an unsupervised means of reducing the dimensionality of non-negative matrices [41]. Furthermore, in order to represent the matrix, it uses inverse document frequency (IDF) as a statistical measure of the terms present within a dataset, giving more weight to the words or terms that appear less frequently across the whole dataset or documents and assigning less relevance to more frequent terms. On the other hand, term frequency (TF) offers more weight to terms appearing in a single document, that is, the number of times a term appears in a document. The TF-IDF is the product of these two statistical measures. The TF is calculated as follows:

$$\mathrm{TF} \ = \ \mathrm{f}_{\mathrm{t_i d}} / (\text{number of words in d}) \tag{1}$$

where d is the document under consideration. The IDF is calculated in the following manner [42]:

$$\mathrm{idf}(\mathrm{t_i}) = \log \frac{\mathrm{N}}{\mathrm{n_i}} \tag{2}$$

where $\mathrm{n_i}$ is the frequency of term $\mathrm{t_i}$ in the total N documents.

Additionally, NMF will view each paper as a mixture of various topics that are represented by a vector of weights over topics. Nevertheless, we only considered the topic with the highest weight as the main topic for that publication, which indicates the degree of the topic's relatedness to the author's publication. Figure 5 denotes the hierarchy of topic layers [43]. Moreover, in order to determine the similarity of the topics between the authors, the cosine similarity is employed, given that $\mathrm{T_i} = \{ \mathrm{t_0}, \mathrm{t_1}, \ldots, \mathrm{t_5} \}$ [18]:

$$\mathrm{Topic}_{\mathrm{sim}}\left(\vec{\mathrm{u}}.\vec{\mathrm{v}}\right) = \frac{\vec{\mathrm{u}}.\vec{\mathrm{v}}}{\left|\vec{\mathrm{u}}\right| \times \left|\vec{\mathrm{v}}\right|} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{u_i} \times \ \mathrm{v_i}}{\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{u_i^2} \times \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{v_i^2}}} \tag{3}$$

Similarity Measures

This research proposes a novel matrix that can be used to determine similarity between authors based on three factors: keywords, the statistics of publication records, and research area topics. However, the author-selected keywords are the main topic indicator for the published papers, as they offer many specifications to consider, such as keyword diversity, the number of used keywords, and their degree of relevance to the main topic. Hence, with regard to the keywords factor, CountVectorizer has been employed in relation to these variables. The keywords presented in the vector model using CountVectorizer in order to produce a matrix to calculate the cosine similarity for the authors from the sklearn.metrics.pairwise library. It treats each word as a vector by counting the frequency of each word's appearance and replacing it with the corresponding frequency. Among the existing metrics, in order to disclose the relatedness of words, using the cosine similarity is the most accurate way to determine the similarity degree between two authors. Given that two vectors $\vec{u}$, $\vec{v}$ exist with N dimensions, the cosine similarity between them is computed as follows [18]:

$$\mathrm{Word}_{\mathrm{sim}}\left(\vec{\mathrm{x}}.\vec{\mathrm{y}}\right) = \frac{\vec{\mathrm{x}}.\vec{\mathrm{y}}}{\left|\vec{\mathrm{x}}\right| \times \left|\vec{\mathrm{y}}\right|} = \frac{\sum_{\mathrm{i}\,=1}^{\mathrm{n}} \mathrm{x_i} \times\ \mathrm{y_i}}{\sqrt{\sum_{\mathrm{i}\,=1}^{\mathrm{n}} \mathrm{x_i^2} \times \sum_{\mathrm{i}\,=1}^{\mathrm{n}} \mathrm{y_i^2}}} \tag{4}$$

wherein $\vec{\mathrm{x}}$ and $\vec{\mathrm{y}}$ are the components of vector $\vec{u}$ and $\vec{v}$, respectively.

On the other hand, to ascertain the similarity in the statistics between publication records, including paper, citation, and year numbers (PCYnum) that consist of the total number of papers published by an author, the number of citations and their publication year ranges were assessed for each author.

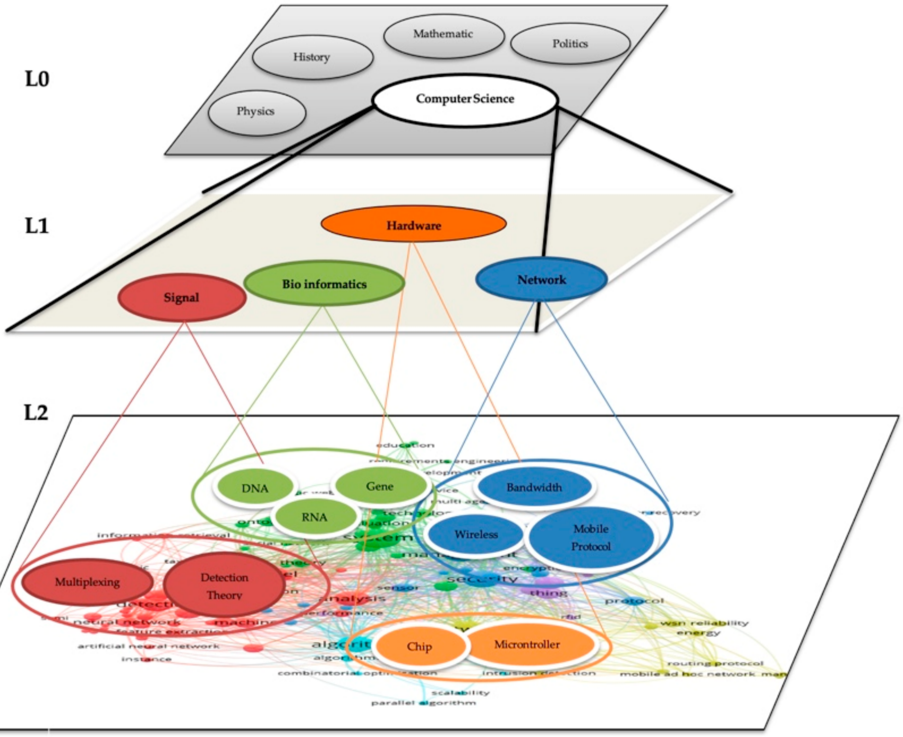


Figure 5. The topics' hierarchy layers [43].

- Number of Papers

In practice, academics who publish more papers tend to be thought of as more knowledgeable in a field than others; therefore, we computed the number of all the published papers by each author [44]. $PNum(x_j)$ is the total number of papers for an author x_j, where $j = \{j_1,\ j_2,\ j_3,\ \ldots\ldots,\ j_n\}$, and n is the total number of authors.

- Citation Count

Citation analysis is one of the primary methods used to evaluate papers in an academic social network. Indeed, citations certainly act as a popular indicator (weighting) for research outcomes; hence, they can index the beneficial nature and magnitude of the paper's contribution [45]. This study thus considered the number of citations as a factor that reflects the contribution of an author to scholarly communities, so we calculated the total number of citations per author by totaling up all of the citations for all of the author's papers. This was given that $\mathrm{i} = \left\{\mathrm{i}_1,\ \mathrm{i}_2,\ \mathrm{i}_3,\ \ldots\ldots,\ \mathrm{i}_{\mathrm{p}_\mathrm{j}}\right\}$:

$$\mathrm{Cite}\left(\mathrm{x_j}\right) = \sum_{\mathrm{i}=1}^{\mathrm{p_j}} \mathrm{C_i}\left(\mathrm{x_j}\right) \tag{5}$$

wherein $\mathrm{p_j}$ is the total number of papers published by author $\mathrm{x_j}$, and $\mathrm{C_i}\left(\mathrm{x_j}\right)$ is the citation count of the ith paper for user $\mathrm{x_j}$.

- Range of Publication Years

One of the most well-known facts about an author's research activity within academic social networks is that it usually evolves on the basis of previous publication areas. Therefore, we relied on the author's publication duration, from their first piece of research being published to their latest example. Given that $\mathrm{y_j}\left(\mathrm{x_j}\right) = \{\mathrm{y_1}\left(\mathrm{x_j}\right),\ \mathrm{y_2}\left(\mathrm{x_j}\right),\ \mathrm{y_3}\left(\mathrm{x_j}\right), \ldots\ \mathrm{y}_{\mathrm{p_j}}\left(\mathrm{x_j}\right)\}$, the range was calculated as follows:

$$\mathrm{Years}\left(\mathrm{x_j}\right) = \mathrm{max}_{1\leq \mathrm{i} \leq \mathrm{p_j}} \left\{\mathrm{y_j}\left(\mathrm{x_j}\right)\right\} - \mathrm{min}_{1\ \leq\ \mathrm{i}\ \leq\ \mathrm{p_j}} \left\{\mathrm{y_j}\left(\mathrm{x_j}\right)\right\}, \mathrm{max_i} \neq \mathrm{min_i} \tag{6}$$

wherein $\mathrm{min_i}$ represents the year of the first publication for author $\mathrm{x_j}$, while $\mathrm{max_i}$ represents the year of most recent publication. Algorithm A1 (Appendix A) illustrates the processes prior to calculating the similarity when computing the number of papers for each author (lines 4–10), the total number of author citations (lines 11–13), and the range of the publication years (lines 14–19).

Furthermore, as a consequence of $\mathrm{PNum}\left(\mathrm{x_j}\right)$, $\mathrm{Cite}\left(\mathrm{x_j}\right)$, and $\mathrm{Years}(\mathrm{x})$, they are converted to vectors. Before applying the cosine similarity equation to PCYnum, the vectors were normalized to unit length in order to be more suitable and accurate for similarity calculation by rescaling and subtracting the minimum value for each number in PCYnum and then dividing the result by the range (the difference between the maximum and the minimum value) [37].

$$\mathrm{X_{norm}} = \frac{\mathrm{x} - \mathrm{x_{min}}}{\mathrm{x_{max}} - \mathrm{x_{min}}} \tag{7}$$

Additionally, all of these explanations and calculations lead to the proposed enhanced authors' similarity measure, which can be written as follows:

$$\mathrm{Author_{sim}}\left(\vec{\mathrm{x}}, \vec{\mathrm{y}}\right) = \alpha \times \mathrm{Topic_{sim}}\left(\vec{\mathrm{x}}, \vec{\mathrm{y}}\right) + (1 - \alpha) \times \left(\mathrm{Word_{sim}}\left(\vec{\mathrm{x}}, \vec{\mathrm{y}}\right) + \mathrm{PCYnum_{sim}}\left(\vec{\mathrm{x}}, \vec{\mathrm{y}}\right)\right) \tag{8}$$

wherein α is the weight parameter; $\mathrm{Word_{sim}}$ represents the keywords similarity; $\mathrm{PCYnum_{sim}}$ represents the paper, citation, and year similarities; and $\mathrm{Topic_{sim}}$ represents the topics similarity.

Meanwhile, Algorithm A2 (Appendix A) illustrates the processes that were used to determine the similarity between authors using three factors: the author-selected keywords (lines 4–5), the authors' publication statistics (lines 6–9), and publication topic distributions (lines 10–14). Additionally, combining the computed similarities gives greater weight to the topics distribution due to the significant impact they have on the similarity (lines 15–18).

Extracting Expertise

In this phase, the expertise of an author in a specific topic is extracted by relying on the multi-factor indicators for a topic of interest, such as the number of papers written about that topic, the number of citations they have attracted, and the total topic weight for the authors' papers; hence, a researcher's publication list presents all of their contributions to academia. In addition, exploring the data on the list of a single researcher will give an accurate and realistic evaluation of their expertise. It reflects the level of knowledge they have on the topics they have worked on and may also indicate their main research area and their field of expertise. Consequently, the papers' keywords convey an expert's research areas perfectly. In order to identify the expert author for a topic, first, the extracted topics must be converted into the vector format (1,0) for each publication in order to recognize the author with the highest number of publications relating to that topic. For each publication, the most common topic should be taken as its main topic. Instead of the weighted number, it converts to the vector format by assigning the main area of interest with 1 and 0 for other topics, while the citation number can be added to the corresponding topic. Additionally, for each topic, the author with the highest weight in that topic should be identified based on their papers and keywords, which reflect the depth of their publications on that subject area. Therefore, each author will have the number of published papers on each topic with their citation count, including the extracted weight for that topic; hence, for each topic, the authors should have three factors: a publication number, a citation number, and the topic weight from the topics' distribution. Therefore, for each author, the weighted average can be ascertained by combining the publication and citation numbers with the corresponding topic weight while assigning greater relevance to the topic weight:

$$\mathrm{Avg}_{\mathrm{Weight}_i} = \alpha \times \mathrm{TopicAVG}(t_i) + (1 - \alpha) \times (\mathrm{PNum}_{t_i} + \mathrm{Cite}_{t_i}) \tag{9}$$

where α is weight parameter and $\mathrm{Avg}_{\mathrm{Weight}_i}$ represents the expertise of an author on a certain topic t_i, $\mathrm{TopicAVG}(t_i)$ is the total weight from the topic distribution, and PNum_{t_i} is the count of papers written on that topic, while Cite_{t_i} is the total number of their citations.

Finally, for each topic, the expertise results need to be ranked, which means that the author assigned with the highest average weight in that topic will be considered the most knowledgeable expert in that grouping. Figure 6 illustrates the process of assigning an expert to a topic, and the process for extracting the authors' expertise is shown in Algorithm A3 (Appendix A). First, this is conducted by verifying the topic's distribution in relation to each paper (lines 5–10), while assigning the highest topic weight as the main topic for the paper and changing the topic distribution into a vector format is the next step; this is where 1 stands in for the main topic and 0 represents the other topics (lines 11–18). Therefore, for each author, the number of papers in each topic is accumulated and their citation number is also aggregated (lines 20–32). While the total weight of the topics for each author is given, the three factors are computed while giving the topic weight the highest priority (lines 24–28). Finally, in relation to each topic, the author with the highest weighted average is highlighted (lines 29–32).

3.1.3.4. Interactive Visualization Network

Static graphs are useful for presenting network structures, yet they restrict the amount of information that can be expressed and thus view the network from the same perspective all the time. Owing to that, in this research, we employed an interactive network to present the predicted collaboration with more flexibility and functionality. The following subsections will illustrate two interactive networks: the authors' network and the universities' network. The purpose of this is to present the predicted co-collaborative relationships but with different factors in each instance, as both the authors' relationships and the universities' co-collaboration will be analyzed.

Authors' Network

This section presents the author network, which consists of the authors as the nodes and the edges between them as reflections of their predicted co-collaboration, while the weight of the edge represents the degree of their similarities. Additionally, the authors have been clustered based on their research area, including the expert author in each area, which are the individuals with the highest topic weight. In order to construct the collaborative academic network, we built an undirected weighted graph reflecting the predicted degree of collaboration based on their similarity. This is given that $G = (V, E)$, wherein V represents the vertex (node) for each author in the network ($v_i \in V$), while the edge $\left(e_{ij} \in E\right)$ is the evidence of the degree of similarity between two authors' nodes $\left(v_i \text{ and } v_j\right)$. While the edge ($e_{ij}$) and weight ($w_{ij}$) reflect the number of their similarities based on their similar research counts and citations, their keywords and research areas are considered in relation to their active time period.

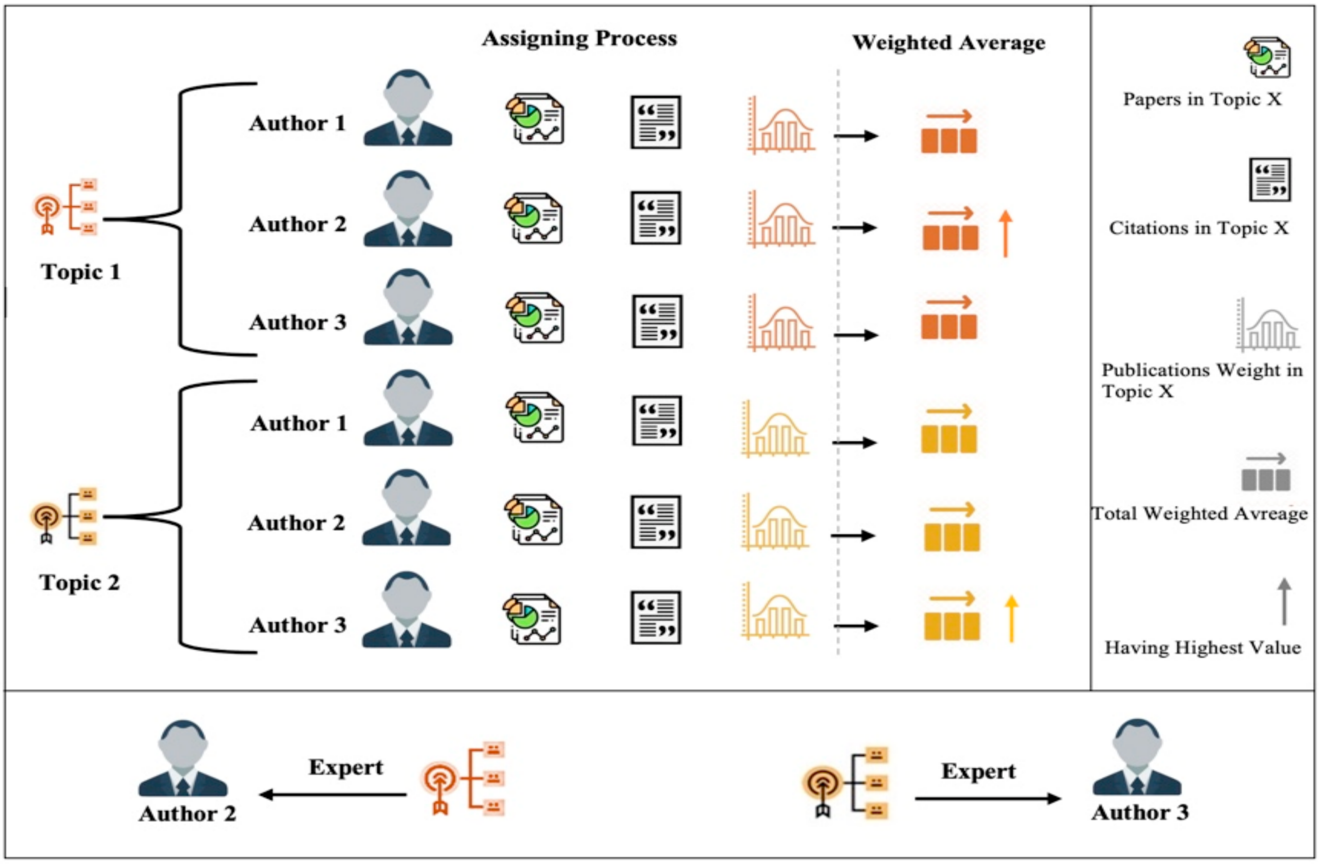


Figure 6. The process of extracting expertise.

Furthermore, the nodes in the network were clustered based on the topic distribution among the data where every author node belongs to one cluster by analyzing the areas of all authors' papers and then taking the mean value for each topic distribution. Hence, by distinguishing the influence of the papers on that topic, it is possible identify the topic with the highest impact as the main research area. Furthermore, in order to measure the impact of a topic t_i, we measure the impact as TopicAVG given that $P_i = \{p_0, p_1, \ldots, p_n\}$:

$$\text{TopicAVG}(t_i) = \frac{\sum_{p\ \epsilon P,T} P_{\text{weight}}\ (p_i)}{N_p} \tag{10}$$

where N_p is the total number of the author's papers on this topic, and $P_{\text{weight}}\ (p_i)$ is the topic weight number of paper p_i. However, TopicAVG not only reflects the average weight of all papers linked to that topic, but it also discloses its impact. Therefore, the highest TopicAVG among all of the other topics for that author will be considered the main research area for that individual and will be clustered based on it. Figure 7 illustrates the process of assigning each author to a topic. Moreover, to represent the expert author among their peers in the same topic within the network, a larger node size was employed as an indicator of importance in terms of topic representation; thus, each topic has one node that is larger than the others.

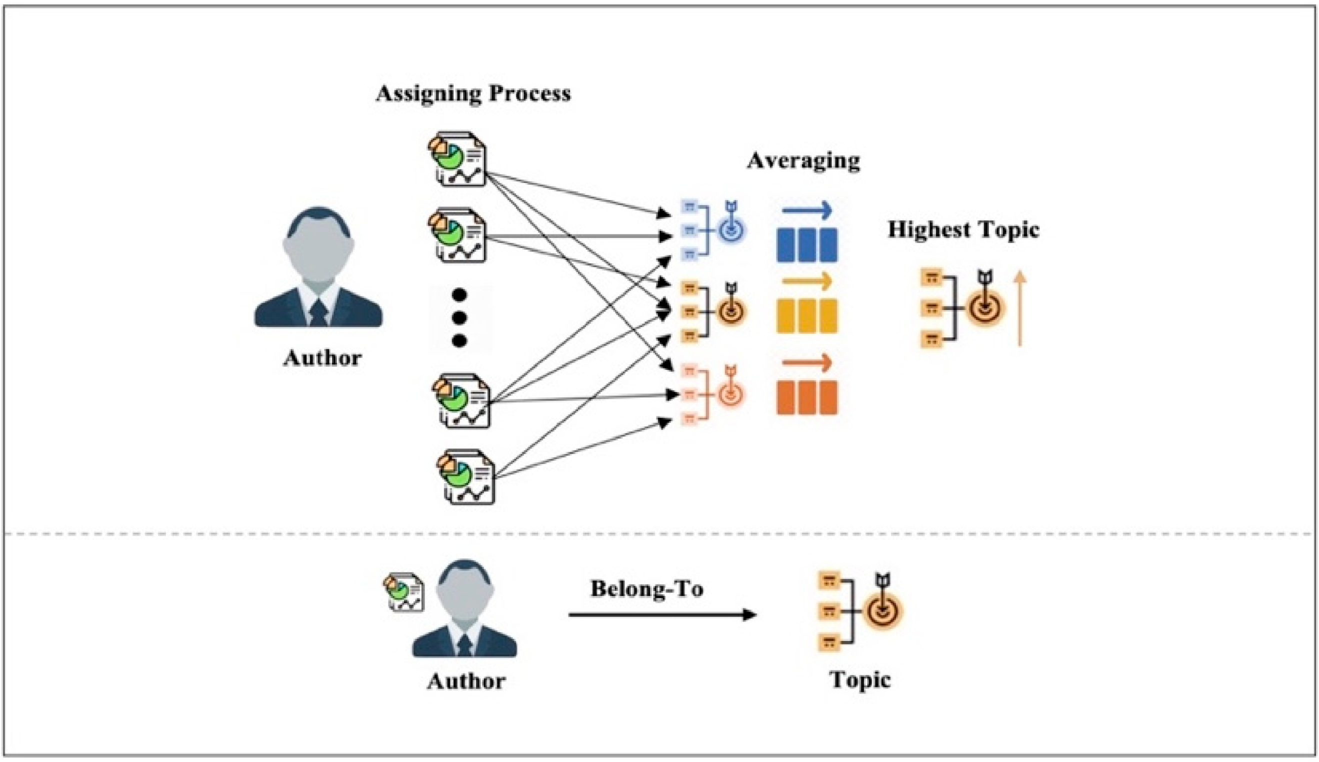


Figure 7. Assigning process for the author's main research area.

Furthermore, a threshold set-up was applied to the edge values to remove the influence of insignificant probabilities on similarity edges between author nodes, meaning they will not add noise to the similarity calculation; this threshold was set to neglect all probabilities below 67%. Additionally, useful features were also added to the network; when a node is hovered over or clicked on, a new window appears containing specific information about that author. This includes the following elements:

- Their total number of published papers;
- Their total number of citations;
- The range of their publication years;
- Their main field of study;
- The name of their university;
- A ranked list of similar authors' names and their similarity percentages.

The process for building the authors' network is shown in Algorithm A4 (Appendix A). In order to cluster the authors based on their main research area, they are represented according to the highest topic weight in the topic's distribution for each of them (lines 5–10). Moreover, a node has been added in the graph for each author; its color should correspond to the author's research area, while there is an increase in the size of the node in order to represent the expert in that topic (lines 11–19). Lastly, edges have been added between the authors if their similarity is higher than 66% (lines 20–28).

Additionally, when the node is clicked on or hovered over, all edges connected to that node are displayed as thicker edge lines in a color that highlights to distinguish them from others; this allows for the easy identification of directly connected nodes. Hovering the cursor around an edge also shows the similarity percentage of all connected nodes. Algorithm A5 (Appendix A) illustrates the steps taken to rank the authors' similarity list, while Figure 8 offers a representation of the proposed interactive network visualization.

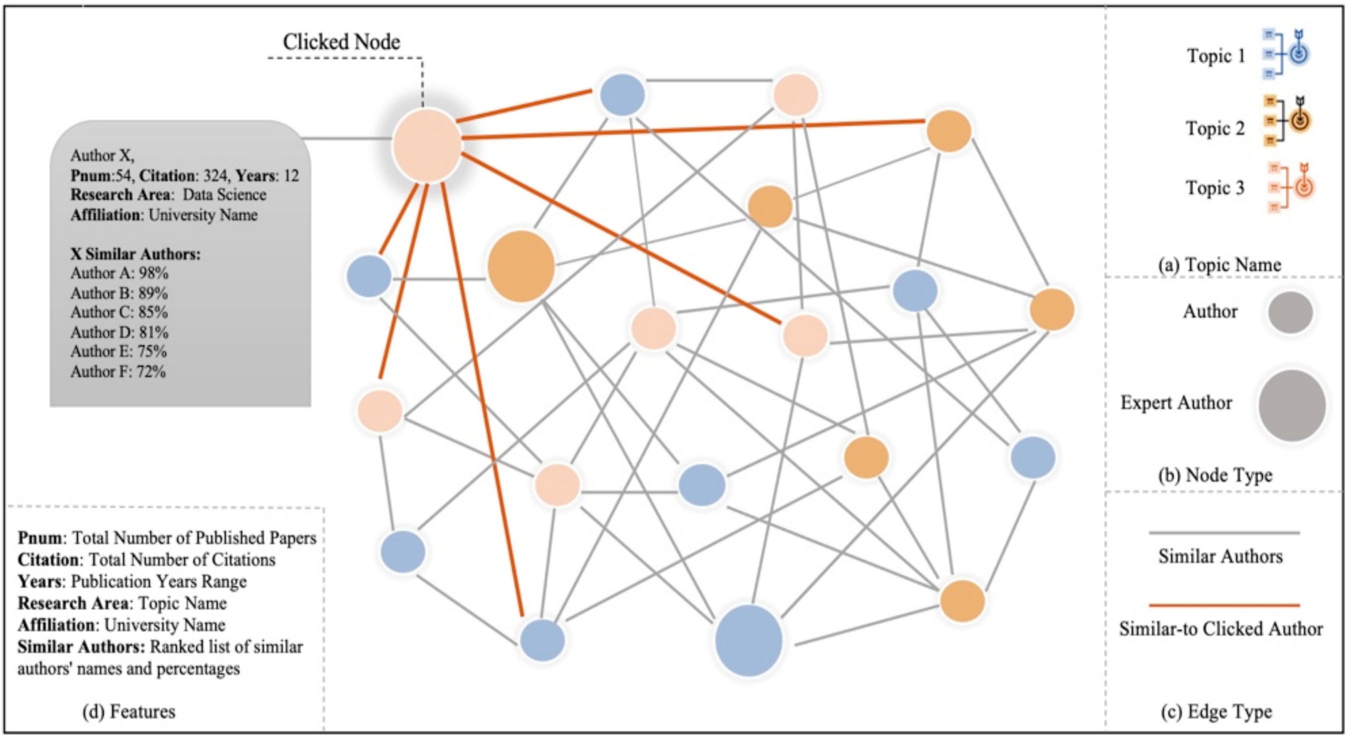

Clicked Node
Author X,
Pnum:54, Citation: 324, Years: 12
Research Area: Data Science
Affiliation: University Name
X Similar Authors:
Author A: 98%
Author B: 89%
Author C: 85%
Author D: 81%
Author E: 75%
Author F: 72%
Pnum: Total Number of Published Papers
Citation: Total Number of Citations
Years: Publication Years Range
Research Area: Topic Name
Affiliation: University Name
Similar Authors: Ranked list of similar authors' names and percentages
(d) Features
Topic 1
Topic 2
Topic 3
(a) Topic Name
Author
Expert Author
(b) Node Type
Similar Authors
Similar-to Clicked Author
(c) Edge Type

Figure 8. Informative representations of the proposed interactive network.

The link analysis uses two measurements: relatedness and significance. However, the relationship between two nodes in the same graph is measured in accordance with their relatedness. Subsequently, after building the network graph with a threshold of the edges, to evaluate the importance of a node in the network, the PageRank algorithm was implemented. It was introduced originally by Google Search during the process of grading websites in their search engine results in order to rank pages that have been indexed by a search engine. Furthermore, it measures the likelihood that a person that will arrive at any particular page [31]. In relation to this, we implemented the PageRank on the authors' network, yet as it is an undirected graph, we converted it into a directed graph by only adding another edge between two nodes if there was an originally existing undirected edge (e_i, e_j). Hence, one edge will be pointing into (in) v_i and the other out of (out) v_j. As shown in Figure 9, node A has a higher number of connected edges, so it should have higher edge scores, which indicates its importance.

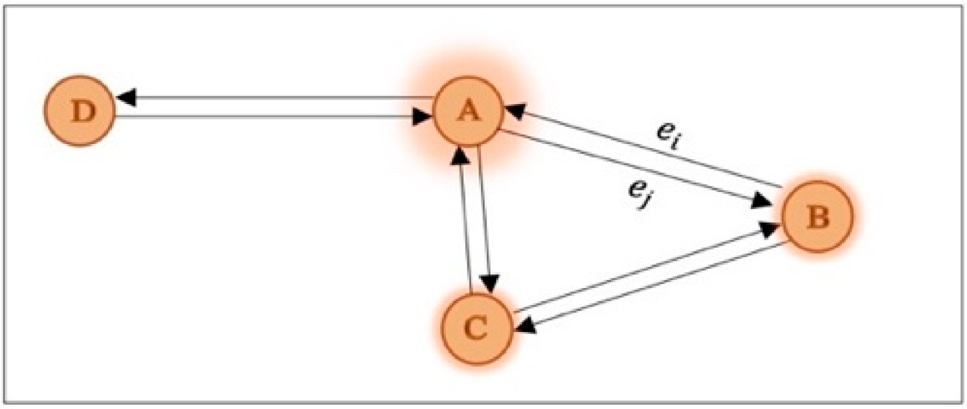


Figure 9. Undirected graph converted into a directed graph.

Universities' Network

This section presents the universities' network, which consists of nodes that represent the university and weighted edges that depict their degree of similarity. With the goal of finding the strongest possible influence for one university on the another in mind, the universities' network was constructed to predict the co-collaboration between them by looking at their publications in order to quantify the level of similarity. For instance, if two universities have authors who usually write papers in the same areas, this suggests that these two universities are somehow intellectually connected by the topic or methodology of their authors' papers. The more often their authors write about the same areas, the stronger the link is between these universities; therefore, it is possible to obtain a connected network of universities based on their authors' papers are often present in the same research areas. After applying CountVectorizer to the keywords, the similarity of the publication's keywords between the two universities i and j can be calculated using cosine similarity (Equation (4)). Additionally, we included the hover and clicked node features as we did in the authors' network, and the threshold was also set; thus, all of the insignificant probabilities will not be counted as a similarity edge between nodes.

4. Experiment and Results

This section elaborates on the findings of the proposed approach and its implementation, which is based on three experiments in relation to the authors' advanced similarity measure, assessing the authors' expertise, and building the authors' co-collaboration network. The first experiment was based on evaluating the similarity between author-selected keywords, the publication topic distributions, and publication statistics, including the number of papers, years of publication activity, and how many citations each individual has. Meanwhile, the second experiment focused on finding an expert author in each topic. The third experiment summarizes the outcome of its two predecessors by building a network that consists of authors and their similarities assembled based on their research area, while some interactive functionality was also added to facilitate acquiring the required result. In addition, the universities network was built based on the authors' selected-keywords similarities in order to predict the influence of co-collaboration in terms of universities.

4.1. Dataset

The main objective of this section is to discuss the collected representative dataset for each author to compile the final results for the proposed model.

4.1.1. Dataset Description

One key challenge when analyzing an author's individual publication information was that, to the best of our knowledge, there has not been any prior work that examines such information about the author as an individual unit, which can be referred to. This research therefore focuses on individual authors in terms of their publication activity, research areas, and affiliations; hence, our work is based on a new dataset that is comprised of the authors' publication details, including the range from their first publication to their last one. After collecting the required data for this research using the PoP tool in a CSV form, it is then further enriched by adding keywords and affiliations. The collected data shows that there are a few records with three different languages other than English, which are Arabic, French, and Chinese. However, since this research work is carried out in English, only English publications were considered during the data analysis. The key attributes of this dataset are illustrated in Table 2. Moreover, the collected data includes 120 authors and 3898 publication records for 12 universities from up until 23 February, 2021. The statistics of the gathered data are shown in Table 3.

Table 2. The key attributes of the used dataset.

Attribute Name	Attribute Description
Cites	Number of citations per paper
Authors	Authors' initial first name, last name
Title	Title of the paper
Year	Paper publication year
Keywords	Paper keywords
Affiliations	The academic institute that the author belongs to

Table 3. Statistics of the collected data.

Authors	Publications	Universities	Years Range	Citations
120	3898	12	43	39,363

4.1.2. Analyzing Publication Records

This section discusses data pre-processing and the analysis of the publication from two perspectives: the temporal evolution of authors' publications and keywords analysis. In text mining research, data pre-processing is a common task for enhancing the efficiency of the final results. In this research, the main objective of this process is to obtain the required dataset, including the author's publication specifications; hence, the processing strategy may differ depending on the type of data. Data cleaning was the first step to be carried out and included lowercase and punctuation removal as well as the removal of duplicates and papers that did not originally include the relevant keyword phrases (e.g., book chapter). Data lemmatization and tokenization were then conducted, including stop word removal. As a result, the final data set contained a total of 3628 publication records.

Temporal analysis: This is one of the most effective strategies for analysis, since it helps to examine large amounts of data and to highlight their similarities and differences. Therefore, analyzing the authors' publications leads to an understanding of their significance and influence on scholarly communities. The data was examined from different perspectives, such as the number of publications, the continence range of publishing years, and the significance of the citation. As shown in Figure 10, the distribution of the publication years is scattered, which indicates that some of the authors started publishing much earlier than others. Additionally, the data shows that there were significant differences in the citation importance and number of publications in the same year; the highest number of publications was in 2020, whereas the publications in 2013 had the highest citation count, even though there was a lower overall number of publications over the course of that year. However, it is worth noting that a 2020 publication has not been published for a long period of time and is therefore likely to have a lower volume of citations. At the same time, this statistic indicates the significance of the papers that were published in 2013 and their distinguished content.

Figure 11 shows the total number of publications for each author in the data set, which can be used to assess the statistical significance. M AlRodhaan has the highest number of papers, while S AlThanyyan has the lowest number. Along with that, Figure 12 illustrates the publication and citation distribution over the years for both authors, and it can be seen that even though S AlThanyyan has a low number of publications, the number of citations was relatively high. Therefore, the number of publications alone does not reflect the significance of a publication. Furthermore, the number of citations cannot be considered in isolation. Instead, all three factors (year distributions, publications number, citations number) should be considered equally important in the exploration of the significance of a publication.

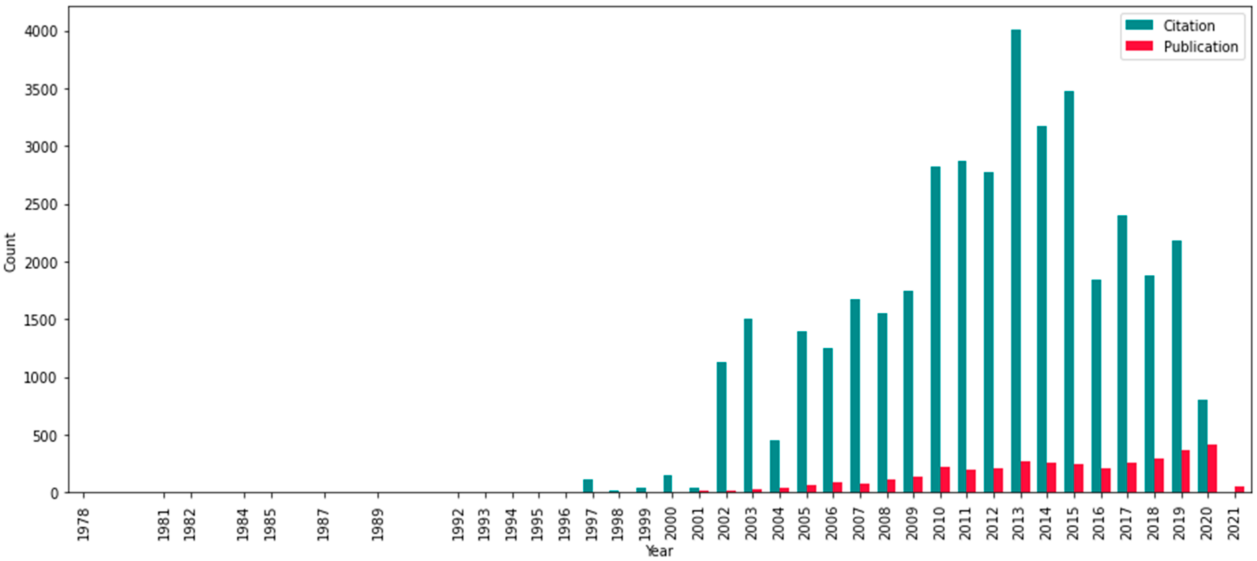


Figure 10. Number of citations and publications in each year.

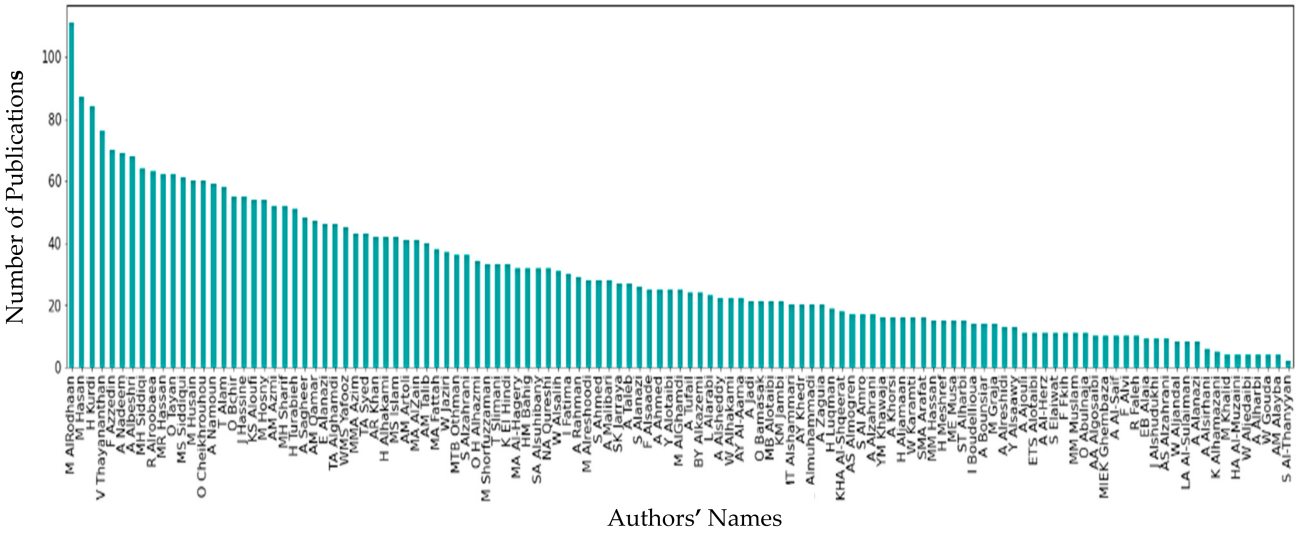


Figure 11. Number of publications for each author in descending order.

Figure 13 combines all three factors and depicts them individually for each author within the drop-down list that contains all of the authors' names. Additionally, for the year ranges, Start Year reflects the first publication year, while End Year is the year of the last publication, and they can be adjusted to show a specific range of years.

Keyword Co-occurrence Analysis: This represents the pattern of connections between co-keywords and their significant role in this network; thus, it clusters the words based on their co-occurrence and their relevance to each other. This is conducted using the VOSviewer clustering technique and by mapping the relationship between those words [46]. However, in order to ignore the words with low relevance to the main cluster, a threshold was applied to restrict the occurrence of the term to 10 times. As shown in Figure 14, the keywords have been grouped into six clusters based on a relevance score of co-occurrence, which reflects the main topics that the authors publish on.

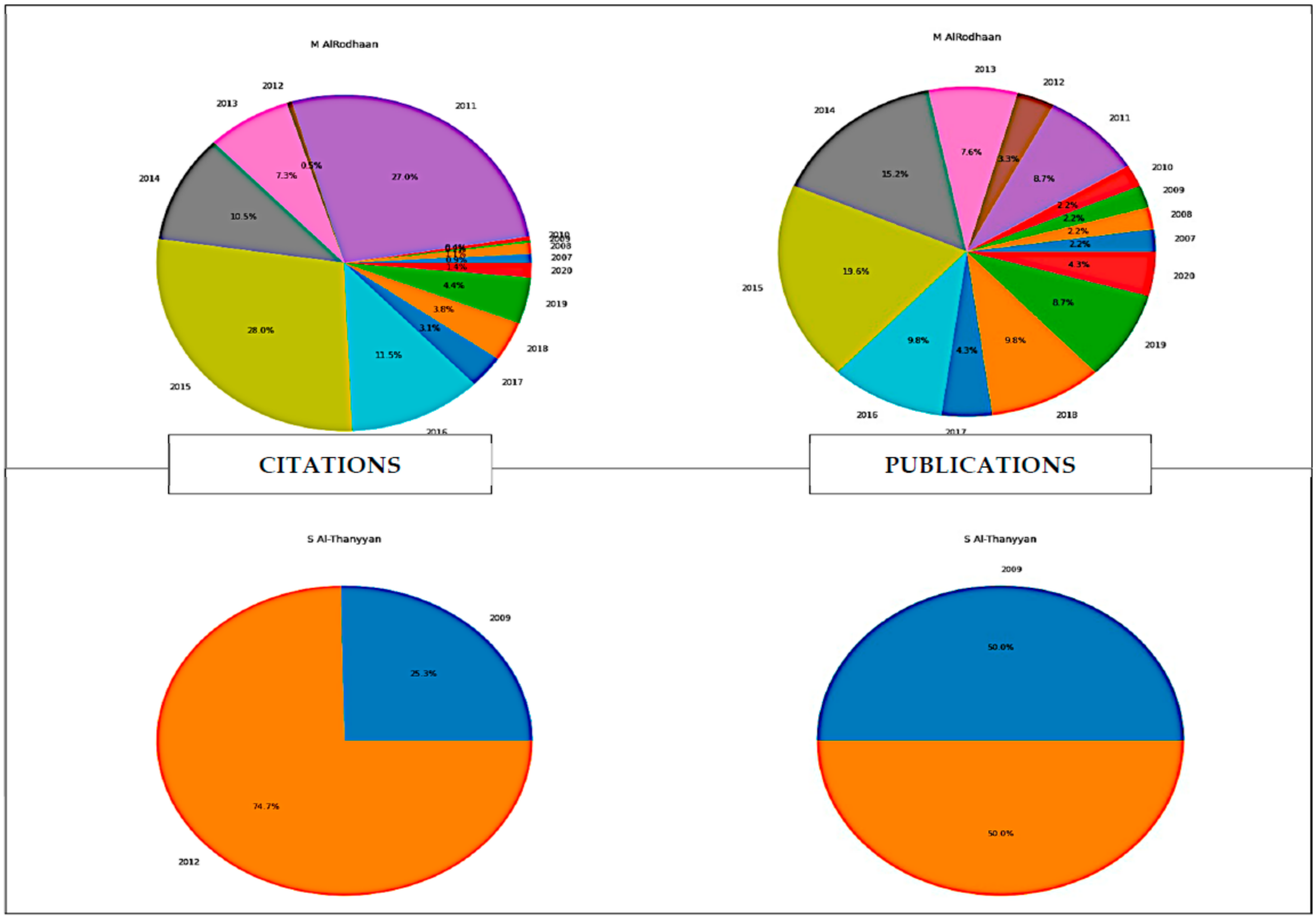


Figure 12. Two authors' publication and citation distributions over the years.

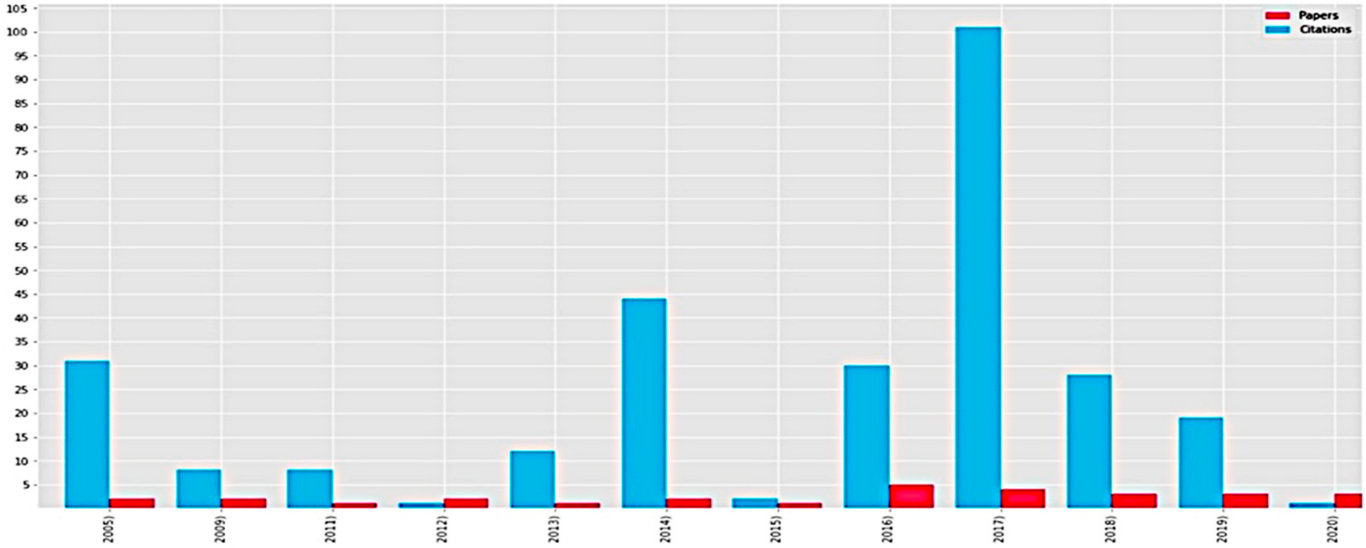


Figure 13. Publications' statistical analysis of year ranges.

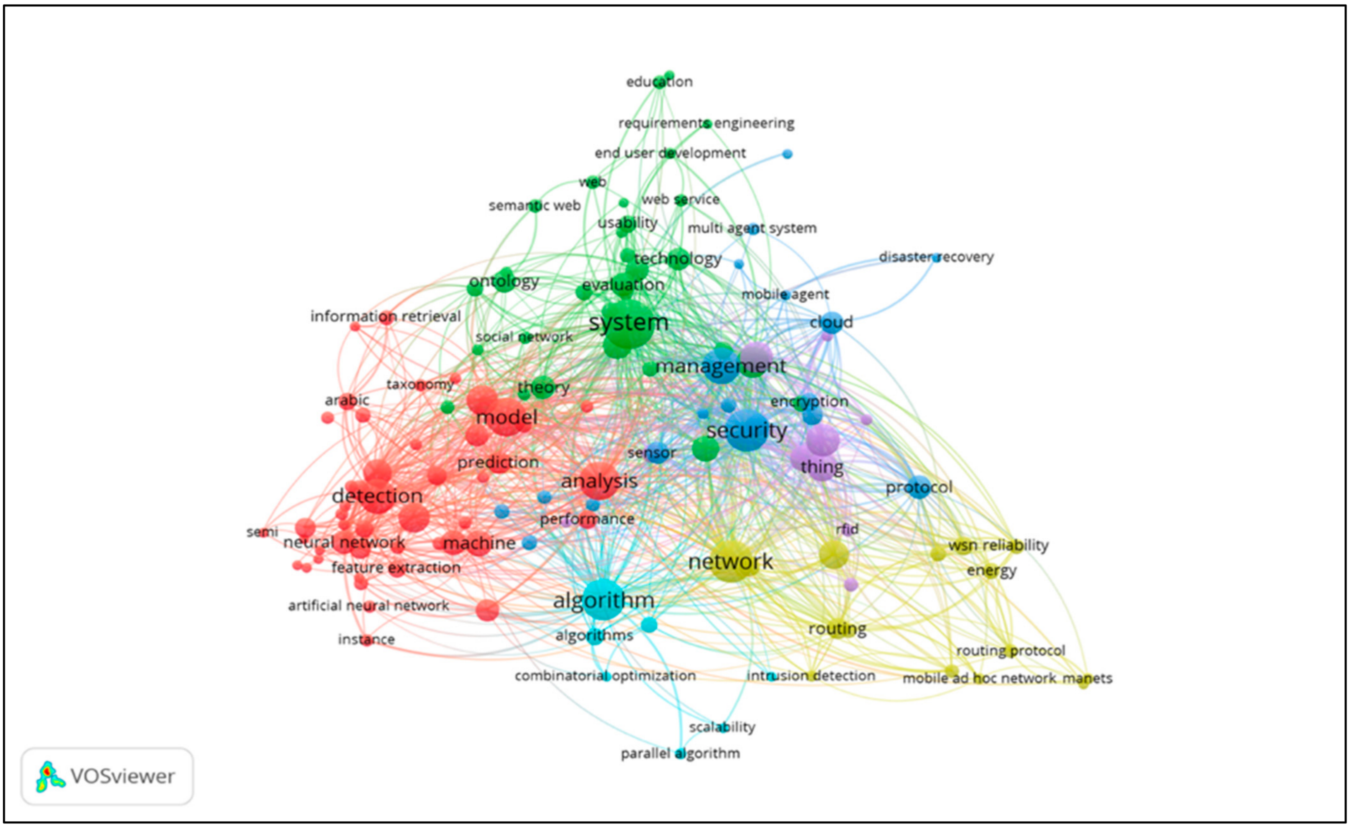


Figure 14. Co-occurrence keywords' network.

4.2. Experiment Settings

This section provides a detailed description of the experiments that were carried out in order to meet the research objectives, including how they were setup and what actions were taken. This research was performed using a MacBook Pro that was running macOS Big Sur 11.2.3 while using Jupiter Notebook from Anaconda Navigator with Python 3.9. Furthermore, all of the researchers' contributions to science are set out in the publication lists. This study focused on individual authors' publication activity to find similar peers who may be suitable for future collaborations, and three experiments based on the authors' publications were explored with regard to the authors' similarities, their levels of expertise, and constructing the recommended co-collaboration network.

E1–Authors Similarity: The main objective of this experiment was to find the similarities between the authors based on three factors. First, this was completed by finding the similarities based on the keywords used in their publications. Second, their publication records statistics were utilized, which consist of the number of their published pieces of work, how many citations they have, and the range of years where they had papers published. Third, an extraction of the topics they worked on using their keywords was undertaken, meaning the topic of each of the author's publications could be determined. The topic distribution was then aggregated from all of the author's publications by taking the average weight of each topic in order to find the similarities based on the topics' averaged weight. Finally, the similarities between the authors were computed by combining all of the resulting similarities.

E2–Assessing Authors' Expertise: This experiment aimed to determine the expert author for each research area by considering three factors for each topic: the number of published papers on the topic; the citations count; and the topic weight, which had extracted from the previous experiment. The topic distribution for each paper was subsequently transformed into a vector format, which involved highlighting the main topic for each paper alongside their corresponding citation count. Then, this was also completed in relation to each author's topic aggregate with regard to all of their papers alongside their corresponding citation count. After that, for each author, the weighted average including the three factors was computed. Therefore, the authors were ranked in terms of each topic, meaning that those individuals with the highest value became the expert in that area.

E3–Constructing the Network: The objective of this experiment was to build an interactive network that represents the recommended co-collaboration between authors based on their similarity from the first experiment. Hence, the first step was to extract the main research area for each author, and the average weight for each topic was taken from the second experiment. Each author's topic with the highest score was considered to be their main research area. The academics were then clustered according to their main topic. A network was subsequently built using the authors as nodes and the similarity values as the edges. Finally, further interactive features were added to the network.

4.3. Authors' Similarity Extraction

This section focuses on finding the similarities between the authors using three factors: the used keywords in their publications, the statistics relating to their publication records, and their publication topics. This was conducted by computing the similarities for each factor separately and then combining the resulting similarities into an enhanced similarity matrix. The following sections will illustrate the three similarities measures in detail as well as underlining their results.

4.3.1. Similarity of Publications' Keywords

The author-selected keywords are the main topic indicator for the published papers; therefore, they were one of the most important factors in being able to measure the similarities. First, in order to compute the cosine similarity between the authors' keywords, all of each author's keywords that were used in their published papers were treated as one document. Additionally, the CountVectorizer was utilized to compute an author's keywords frequency matrix; based on this, the cosine similarity was calculated. Table 4 sets out a sample of these results.

Table 4. Similarity matrix between authors based on keywords.

Authors	**Author 1**	**Author 2**	**Author 3**	**Author 4**	**Author 5**	**Author 6**
Author 1	0.00000000	18.5865564	71.7619352	17.909209	21.7499681	6.14723873
Author 2	18.5865564	0.00000000	23.9411005	61.7959224	56.3382012	63.7912953
Author 3	71.7619352	23.9411005	0.00000000	19.4128249	26.1079481	8.802141
Author 4	17.909209	61.7959224	19.4128249	0.00000000	65.8127783	61.4811754
Author 5	21.7499681	56.3382012	26.1079481	65.8127783	0.00000000	55.4024585
Author 6	6.14723873	63.7912953	8.802141	61.4811754	55.4024585	0.00000000

Figure 15 illustrates the similarity matrix in a simple network and dismisses the similarity values that are lower than 60. Moreover, it indicates that Authors 2, 4, and 6 are relatively similar in terms of the keywords used.

4.3.2. Similarities of Statistical Publication Records

Prior to computing the authors' similarities for their statistical publication records, some of the operations were ascertained according to the generation process of the statistics, including the total number of publications for each author, the total number of citations, and the continence range of publishing years, which is from the year of the first publication until the year of their latest publication. Table 5 depicts the authors' statistics after computing the year range, publication number, and number of citations.

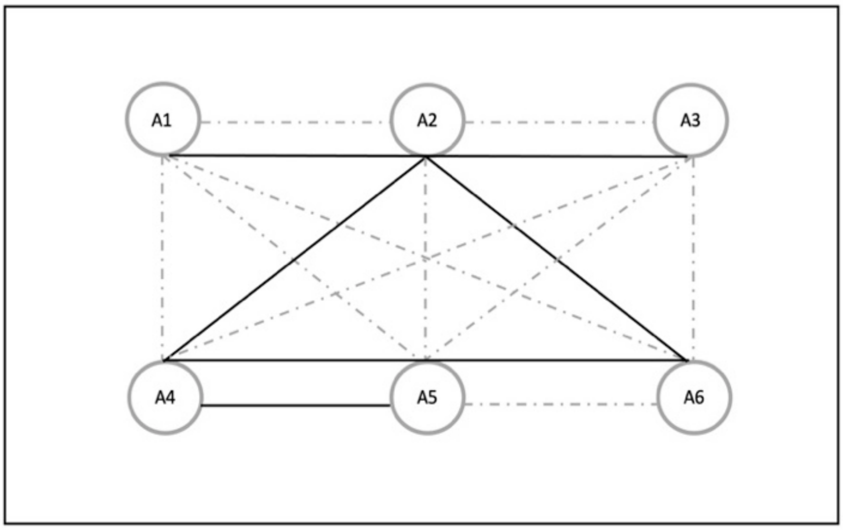


Figure 15. Similarity between authors based on keywords.

Table 5. Authors' statistics in relation to publication records.

Authors	Year	PapersNum	Citation
Author 1	10	36	382
Author 2	13	111	2298
Author 3	9	33	989

Furthermore, vector normalization was implemented before computing the cosine similarity between the authors; Table 6 illustrates the resulting similarities.

Table 6. Authors' similarity matrix based on the statistics of their publication records.

Authors	Author 1	Author 2	Author 3	Author 4	Author 5	Author 6
Author 1	1.00000000	0.90032947	0.88305555	0.83770483	0.94676091	0.94255144
Author 2	0.90032947	1.00000000	0.96932289	0.57778796	0.88569069	0.95651244
Author 3	0.88305555	0.96932289	1.00000000	0.66693209	0.9431802	0.98638411

4.3.3. Similarities between Statistical Publication Records

An author's publication list reflects their cumulative formal contributions to science. They not only contain their research activity but also implicitly expose the topics the authors have worked on. It is therefore of great importance to measure the similarity factors between the authors. Furthermore, in order to extract the topics from the authors' publications, TfidfVectorizer was applied to the keywords to employ the NMF model that will produce the required variables. Using 3627 features that were extracted from the keywords, six different topics were extracted. Additionally, in order to assign a research-area name for each topic, the appropriate name was inferred from the top 10 words in each area. Table 7 presents all of the topics along with their top keywords and the assigned research area.

Table 7. Topics' specifications.

Topics	Topic Related Words	Research Area
Topic 0	2.340073*"cloud" + 1.798998*"computing" + 0.499167*"service" + 0.492035*"security" + 0.406478*"internet" + 0.389438*"thing" + 0.385234*"storage" + 0.382136*"smart" + 0.301981*"multi" + 0.289075*"fog"	Computer Security
Topic 1	1.420433*"network" + 0.887853*"sensor" + 0.783095*"wireless" + 0.306670*"hoc" + 0.277345*"protocol" + 0.224257*"energy" + 0.222750*"security" + 0.219170*"rout" + 0.212544*"internet" + 0.200105*"mobile"	Networks
Topic 2	1.238230*"learning" + 0.942592*"machine" + 0.407995*"analysis" + 0.385140*"feature" + 0.379130*"classification" + 0.378085*"deep" + 0.377475*"detection" + 0.367168*"recognition" +0.366274*"image" + 0.326451*"arabic"	Artificial Intellqsigence (AI)
Topic 3	1.369969*"system" + 0.618557*"information" + 0.599722*"software" + 0.451404*"management" +0.346359*"engineering" + 0.273435*"model" + 0.270433*"requirement" + 0.231994*"process" + 0.215679*"application" + 0.197406*"service"	Software Engineering
Topic 4	1.748337*"data" + 0.592130*"mining" + 0.329126*"big" + 0.256031*"security" + 0.243313*"web" + 0.179450*"performance" + 0.161883*"database" + 0.145166*"privacy" + 0.142380*"cluster" + 0.123463*"integrity"	Data Science
Topic 5	1.553971*"algorithm" + 0.562206*"genetic" + 0.521833*"optimization" + 0.433186*"problem" + 0.379081*"search" + 0.214585*"parallel" + 0.192708*"combinatorial" + 0.177214*"approximation" + 0.155371*"multi" + 0.154928*"heuristic"	Algorithms And Theory

Moreover, as shown in Figure 16a, the NFM features predicted the topic distribution for each publication by using the keyword vectors where the highest topic weight is the main topic for that paper and the total number of papers for each topic. Furthermore, an aggregate of all the papers' predictions was calculated using the mean value for each topic with regard to the authors' topic distribution. Whereas the highest topic weight represents the main research area for the author, Figure 16b illustrates the total number of authors after computing their main topic. Additionally, Table 8 offers an example of the topics' distribution.

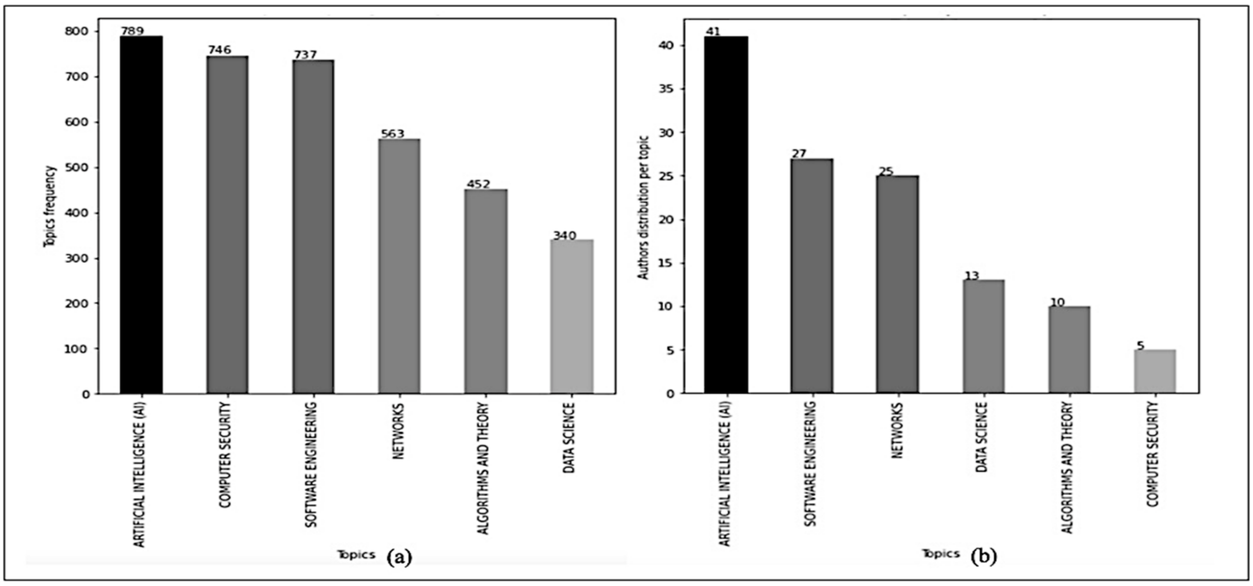


Figure 16. Topics' distribution. (**a**) The total number of publications on each topic; (**b**) The total number of authors in each topic.

Table 8. Topics' distribution for authors.

Authors	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Author 1	0.08461012	0.00082923	0.00450022	0.07453312	0.03191502	0.00169824
Author 2	0.00674379	0.05268865	0.01595955	0.01202155	0.00831521	0.02183269
Author 3	0.03732404	0.0053186	0.00979788	0.01809716	0.04586964	0.00260079

Figure 17 is an animated representation of the topics' distribution over the range of years for a particular author, and there is a drop-down list to select a specific author.

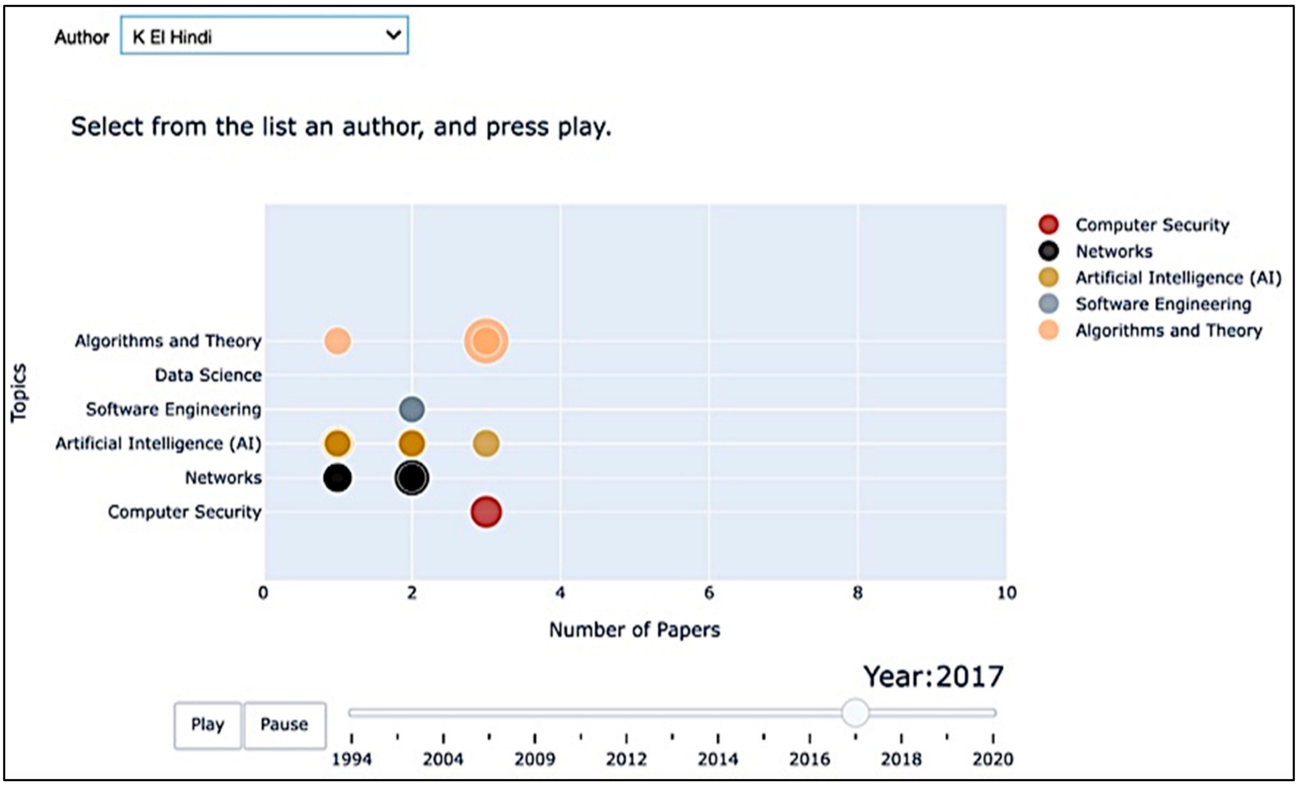


Figure 17. An animated representation of the topics' distribution.

The topics' vector distribution was normalized prior to computing the cosine similarity between the authors; Table 9 highlights the resulting similarity.

Table 9. Similarity matrix between authors based on topics' distribution.

Authors	Author 1	Author 2	Author 3	Author 4	Author 5	Author 6
Author 1	1.00000000	0.26180808	0.81658032	0.28527477	0.270762	0.13779837
Author 2	0.26180808	1.00000000	0.34820367	0.93831064	0.97558835	0.95986175
Author 3	0.81658032	0.34820367	1.00000000	0.36345333	0.3013472	0.22413259
Author 4	0.28527477	0.93831064	0.36345333	1.00000000	0.98166036	0.95456905
Author 5	0.270762	0.97558835	0.3013472	0.98166036	1.00000000	0.97278061
Author 6	0.13779837	0.95986175	0.22413259	0.95456905	0.97278061	1.00000000

4.3.4. The Advanced Similarity Measure

After calculating all three of the similarity measures, they were combined while a greater weight was given to the topics' distribution similarities; thus, a new matrix was therefore produced. Table 10 represents the final enhanced matrix results.

Table 10. Similarity matrix between authors based on three factors.

Authors	Author 1	Author 2	Author 3	Author 4	Author 5	Author 6
Author 1	1.00000000	0.3667296	0.80018708	0.36413612	0.39002971	0.27585119
Author 2	0.3667296	1.00000000	0.43978972	0.77089477	0.83406769	0.86260725
Author 3	0.80018708	0.43978972	1.00000000	0.37360102	0.41768945	0.33574954

4.4. Assessing Authors' Expertise

This section discusses the results of identifying the expert author in each topic. The publication records are vital when evaluating the authors' expertise and knowledge; therefore, this research utilized all of each author's publications in order to analyze their expertise in their extracted research area. The publications were examined first by assessing their relevance and importance to the extracted topics. Each publication was then assigned to a main topic from the previous experiment, employing the vector format of topic distribution by assigning the main topic with 1 and 0 otherwise, including the corresponding citation number of each publication, as shown in Table 11.

Table 11. Distribution of topics and citations in vector format.

Papers	Distribution	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Paper 1	Main Topic	0	0	0	0	0	1
	Citations	0	0	0	0	0	17
Paper 2	Main Topic	0	1	0	0	0	0
	Citations	0	18	0	0	0	0
Paper 3	Main Topic	0	0	1	0	0	0
	Citations	0	0	22	0	0	0

Moreover, Figure 18 illustrates the assignment process for an author; the size of the node represents the number of citations for that paper, while the color represents its topic. All of the publications and their features in the selected year will be highlighted.

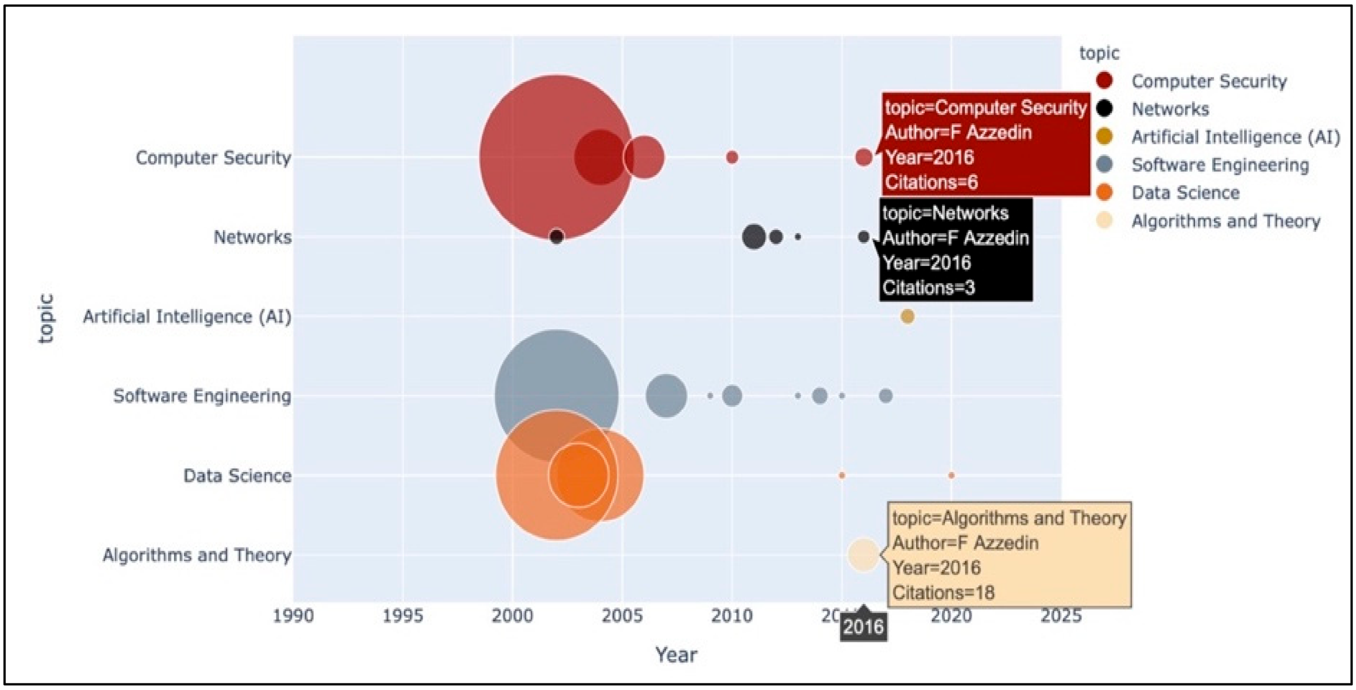


Figure 18. Topics distribution of publications over the years.

All of the distributed topics were accumulated by summing the total citations and counting the number of papers that addressed that topic. Additionally, the extracted weight was used, which is the average number of the topic's distribution extracted from the previous experiment; each author has the total number of papers on each topic, the count of their citations, and the topic weight accounted for. Table 12 represents the authors' factors in each topic; in Topic 0, Author 1 had the highest number of papers, while the highest number of citations was attributed to Author 2, yet the greatest topic weight was attributed to Author 3. For Topic 1, the largest number of publications and the topic weight were attributed to the same author. However, the weighted average calculated for each author in the topic is represented as the expertise number, which is determined by totaling the number of publications, citations, and topic weight given that the topic weight gets a higher priority weight in the process of extracting expertise.

Table 12. Extracting authors' expertise.

Topic	Authors	Papers	Citations	Topic Weight	Expertise
Topic 0	Author 1	21	282	2.65666089	62.19399653
	Author 2	7	517	0.89587796	105.337527
	Author 3	19	263	3.38440497	58.43064298
Topic 1	Author 1	45	927	5.84844025	197.9090642
	Author 2	29	1009	4.03656339	210.021938

4.5. Generating Interactive Visual Networks

This section underlines the steps that were made to build an interactive network. The research adopted academic social network analysis, as it has become a widely-used technique across a diverse range of fields.

4.5.1. Authors' Interactive Network

The authors' collaboration network is one of the important networks that are used to visualize and evaluate the implicit relationships that need to be uncovered. This research focuses on finding the most similar authors based on their publications in order to generate a list of valuable possible collaborators for the target author. The network is therefore based on the similarities that have been computed, the extraction of the authors' main research areas and, their expertise.

Network Construction: The underlying idea of building the co-collaboration network is that authors can be clustered together based on their main research areas while identifying the expert.

Link Weight Calculation: This is based on the selected factors adding a link between the authors, while its weight reflects their corresponding similarities.

Consequently, the authors were clustered based on their main research area, and one author in each cluster was assigned as the role of expert (represented by a larger node size). All of the nodes represent authors, while the edges (links) are the predicted collaboration based on their similarities. Figure 19 visually depicts the collaboration network of authors without constraining the edges with a threshold; hence, the network is very dense and can be pruned by setting the threshold and removing irrelevant edges.

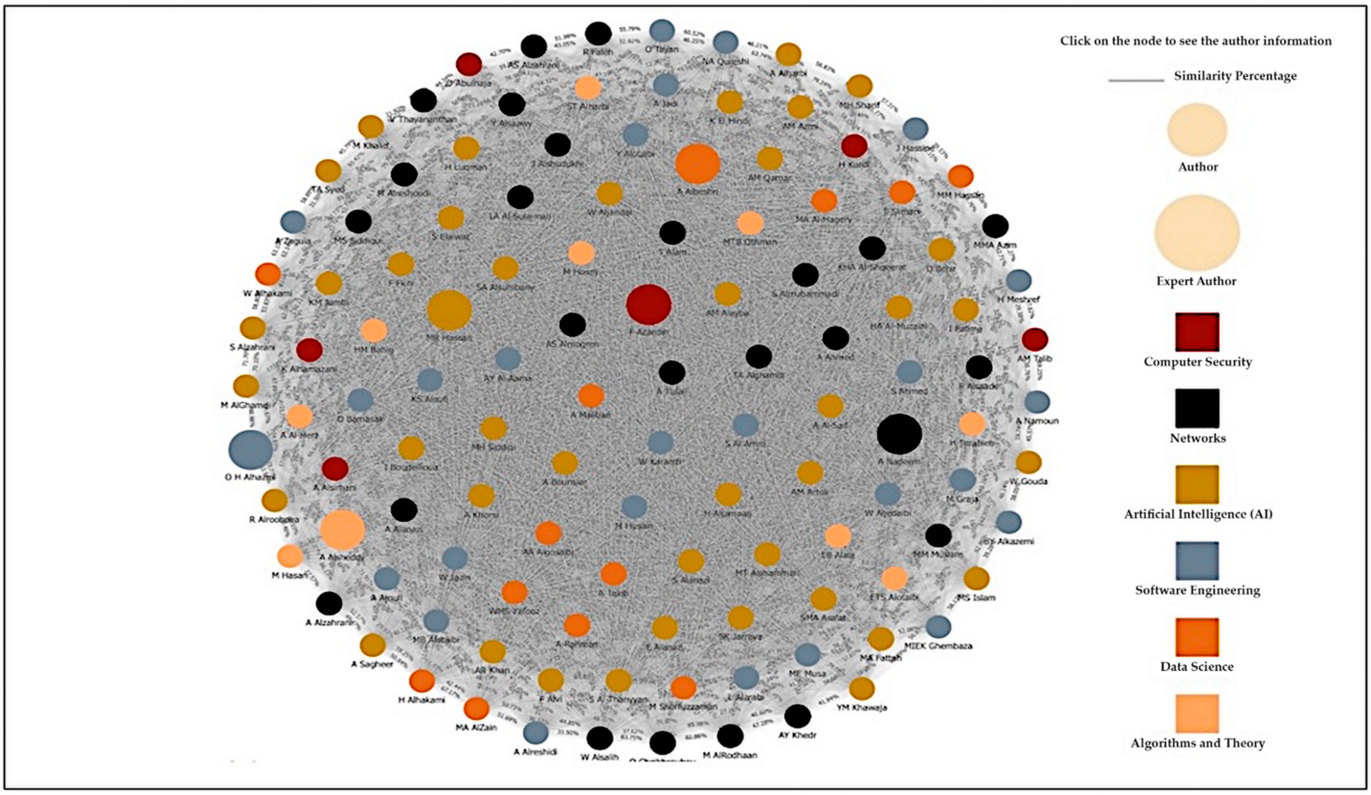
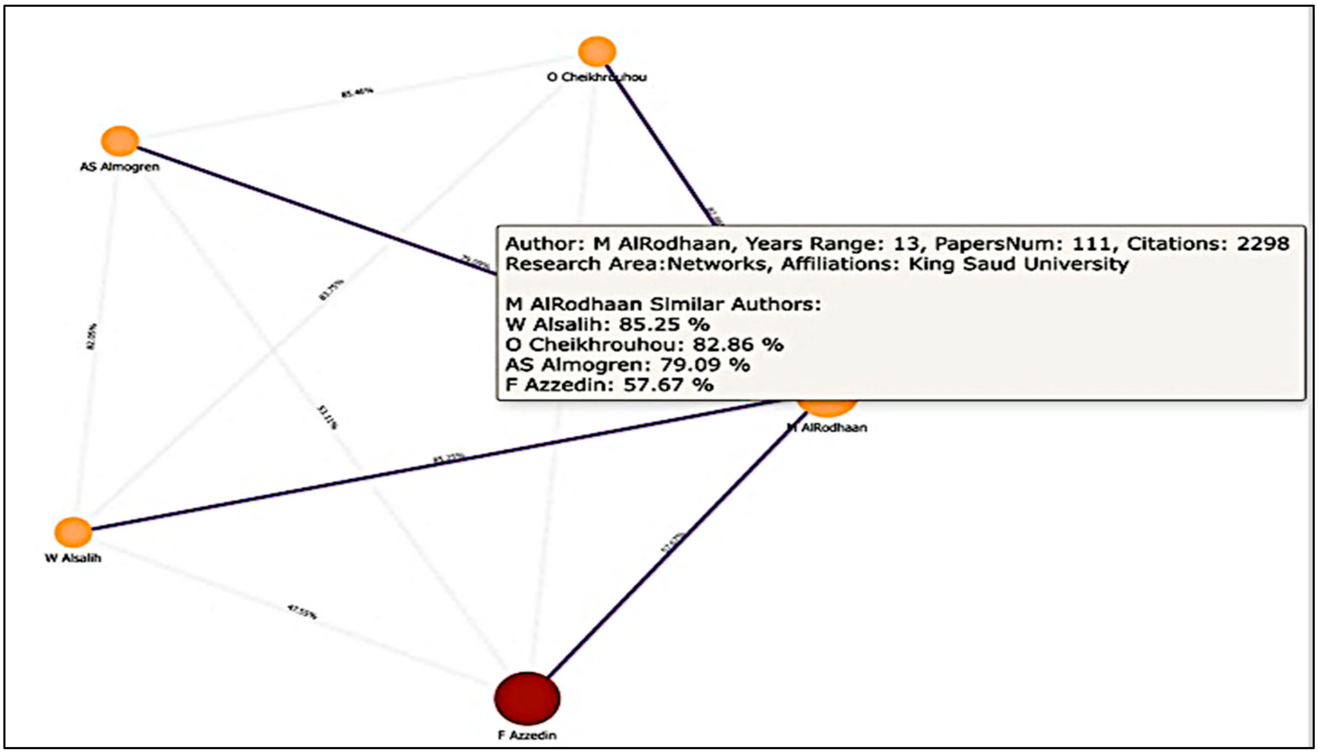


Figure 19. The authors' collaboration network.

Figure 20 represents a sample of the interactive network's functionality where the nodes represent the authors, and the edges are their similarity percentage. If a node is clicked on (or hovered over), all of the connected nodes will be highlighted, and the features window will be displayed. The features consist of the name of the author with their number of papers and citations as well as the range of years they have been active, their main research area, and their corresponding affiliation. Moreover, this information is accompanied by a ranked list of connected authors based on their similarities. Although it is not possible to unambiguously assign an author to another author, the ranked list contains the similarity levels between the scholars.

In order to only present the significant similarities between the authors, a threshold was set for the edge value; to be more specific, a network hyperlink between two nodes denotes their cooperation relationship. Discovering those nodes with high values is preferable when evaluating the magnitude of the predicted relationship, so all of the insignificant similarity values are ignored, as no edges lower than 67% will be added. Figure 21 demonstrates the final representation for the authors' interactive network where the colors of the nodes represent the authors' main topics (Topics names presented in Figure 19). Table 13 illustrates the network entities' descriptions.

Moreover, the PageRank algorithm was employed to extract the author with the highest number of similarities. The author with the highest degree of centrality is considered to be an influential member of the group, and they are represented by the node with the highest number of collaborations with other network members. As a result, the highest degree of centrality was for the author KS Aloufi, who had a value of 15.11%, as shown in Figure 22.



Author: M AlRodhaan, Years Range: 13, PapersNum: 111, Citations: 2298
Research Area:Networks, Affiliations: King Saud University

M AlRodhaan Similar Authors:
W Alsalih: 85.25 %
O Cheikhrouhou: 82.86 %
AS Almogren: 79.09 %
F Azzedin: 57.67 %

Figure 20. Interactive network's functionality.

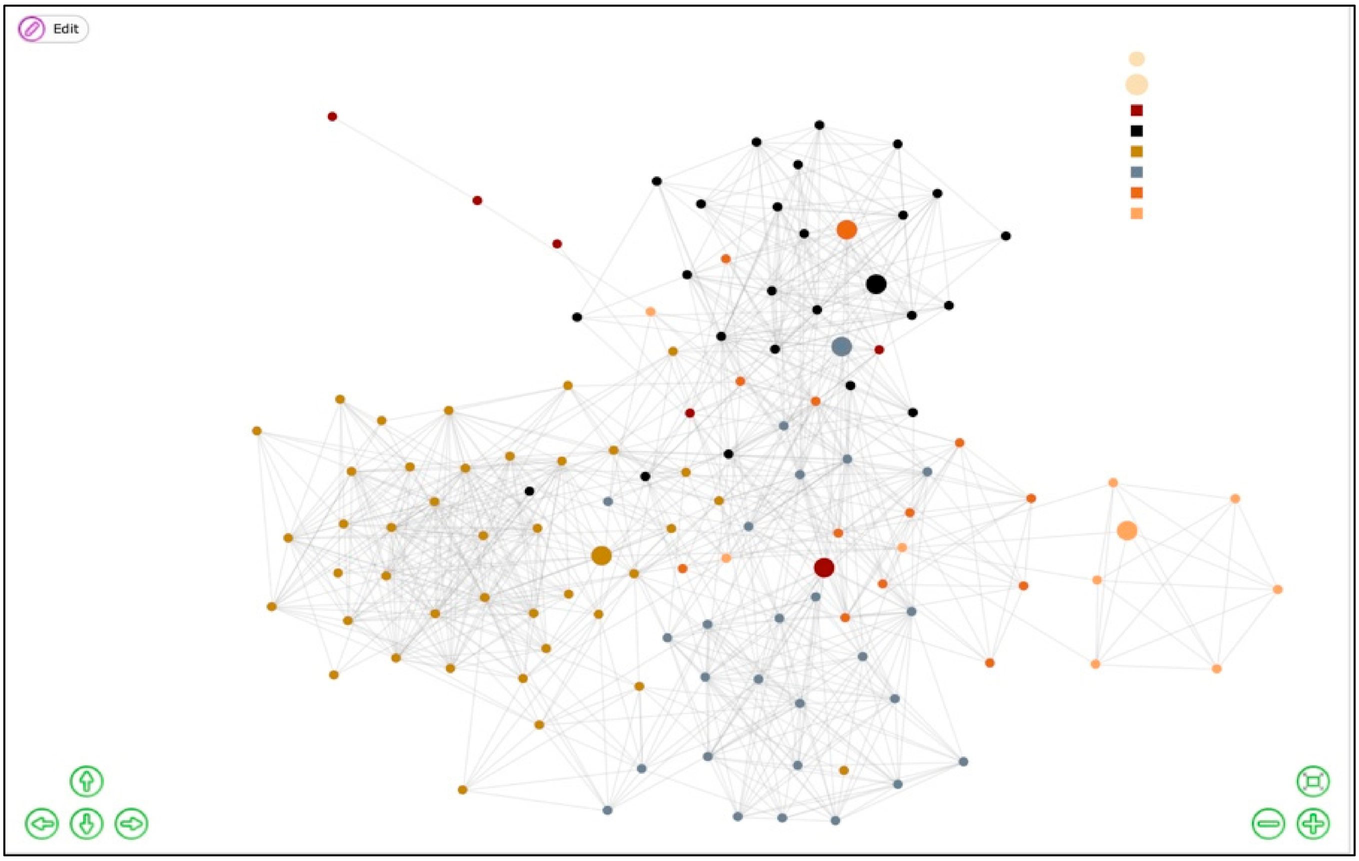
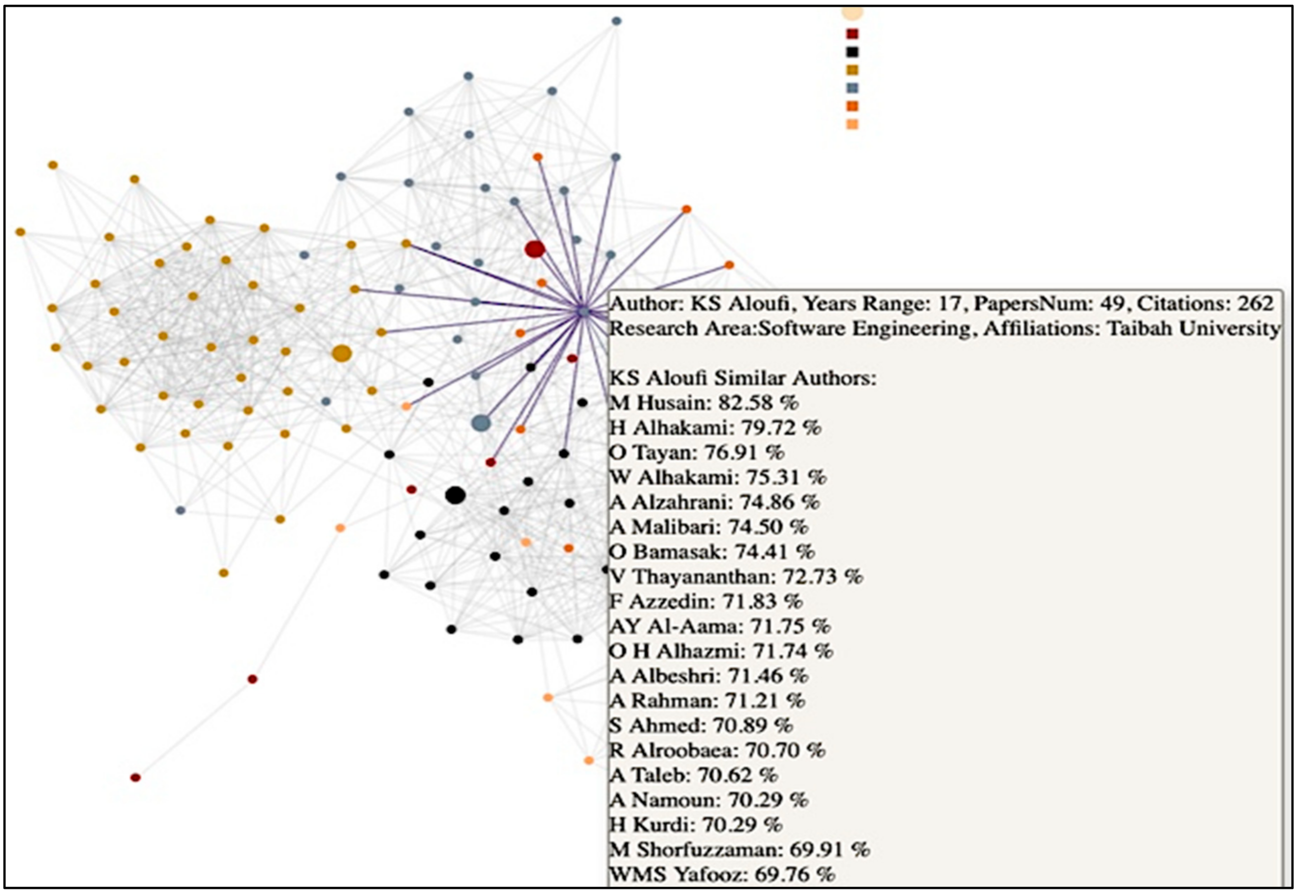


Figure 21. Authors' interactive co-collaboration network.

Table 13. Descriptions of authors' network entities.

Type	Number	Type	Number
Node Type	Author	Nodes Number	120
Edge Type	Similarity	Edges Number	892
Cluster Number	6	Expert Nodes	6



Author: KS Aloufi, Years Range: 17, PapersNum: 49, Citations: 262
Research Area:Software Engineering, Affiliations: Taibah University

KS Aloufi Similar Authors:
M Husain: 82.58 %
H Alhakami: 79.72 %
O Tayan: 76.91 %
W Alhakami: 75.31 %
A Alzahrani: 74.86 %
A Malibari: 74.50 %
O Bamasak: 74.41 %
V Thayananthan: 72.73 %
F Azzedin: 71.83 %
AY Al-Aama: 71.75 %
O H Alhazmi: 71.74 %
A Albeshri: 71.46 %
A Rahman: 71.21 %
S Ahmed: 70.89 %
R Alroobaea: 70.70 %
A Taleb: 70.62 %
A Namoun: 70.29 %
H Kurdi: 70.29 %
M Shorfuzzaman: 69.91 %
WMS Yafooz: 69.76 %

Figure 22. The author with the highest degree of centrality.

4.5.2. Universities' Interactive Network

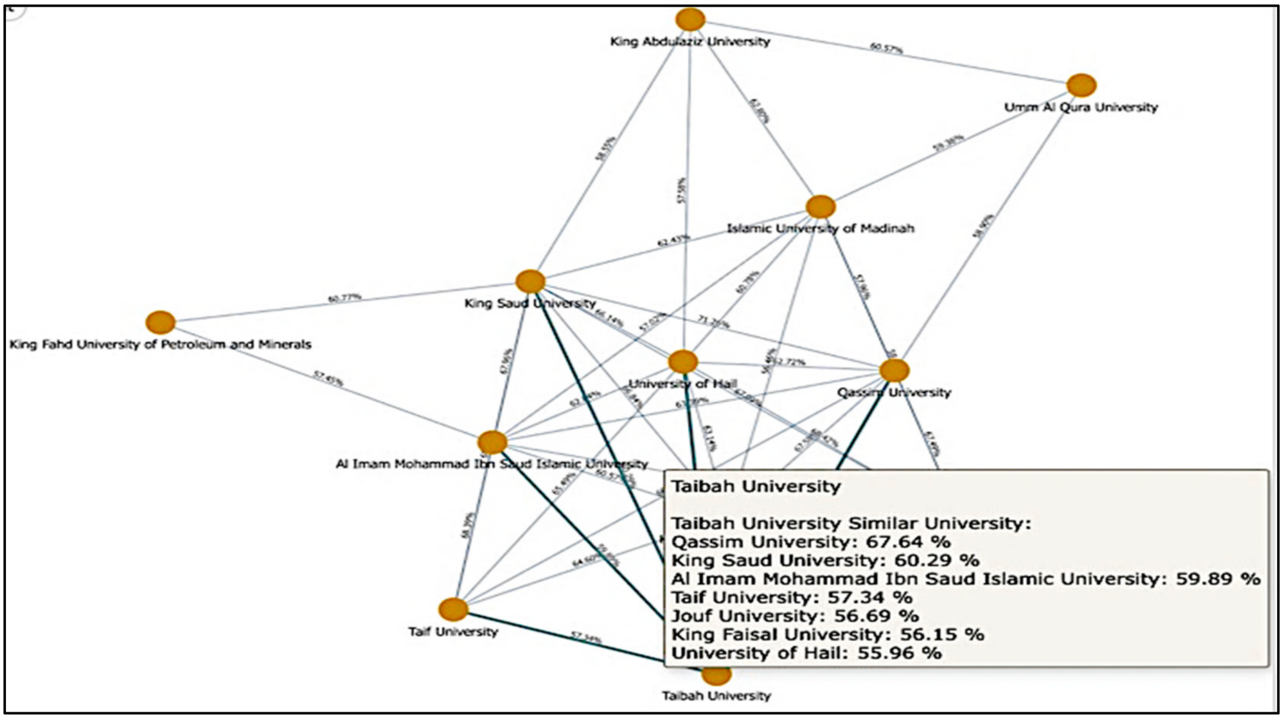
Collaboration across universities is sought after in order to share knowledge globally between different cultures and from different experts. However, this study focused on visualizing the network relationships between universities in the Kingdom of Saudi Arabia in the form of a case study. Nevertheless, this research may go on to inform the development of future trends in research areas of computer science. The similarities between the institutions examined were computed by utilizing the keywords used in their publications and by applying cosine similarity. Table 14 presents the universities' similarities in a matrix format.

Table 14. Similarity matrix between universities based on keywords.

Universities	University 1	University 2	University 3	University 4	University 5
University 1	1.000000	0.497273	0.367437	0.574453	0.418921
University 2	0.497273	1.000000	0.407681	0.683875	0.538110
University 3	0.367437	0.407681	1.000000	0.470206	0.593579

The collaboration network of universities was mapped and analyzed using a threshold. In this case, the nodes represent the universities, while the edges represent their similarities.

Additionally, the interactive functionality from the authors' network was included, as shown in Figure 23.


King Abdulaziz University
Umm Al Qura University
Islamic University of Madinah
King Saud University
King Fahd University of Petroleum and Minerals
University of Hail
Qassim University
Al Imam Mohammad Ibn Saud Islamic University
Taif University
Taibah University

Taibah University

Taibah University Similar University:
Qassim University: 67.64 %
King Saud University: 60.29 %
Al Imam Mohammad Ibn Saud Islamic University: 59.89 %
Taif University: 57.34 %
Jouf University: 56.69 %
King Faisal University: 56.15 %
University of Hail: 55.96 %

Figure 23. Universities' interactive co-collaboration network.

4.6. Results

This research adopted different evaluation metrics to assess the quality of the inferred topics and the proposed matrix. This section describes the two different evaluation methodologies: ground-truth evaluation and quantitative analysis. Hence, the authors have been clustered based on their topics, and the evaluation of the clustering performance was completed in terms of precision, recall, F1-measure, accuracy, and Cohen's kappa. The cluster labels of the topics were obtained from the model along with their original labels from the ground-truth table using a binary representation. Table 15 shows the overall assessment statistics, while Figure 24 is a heatmap representation of the results.

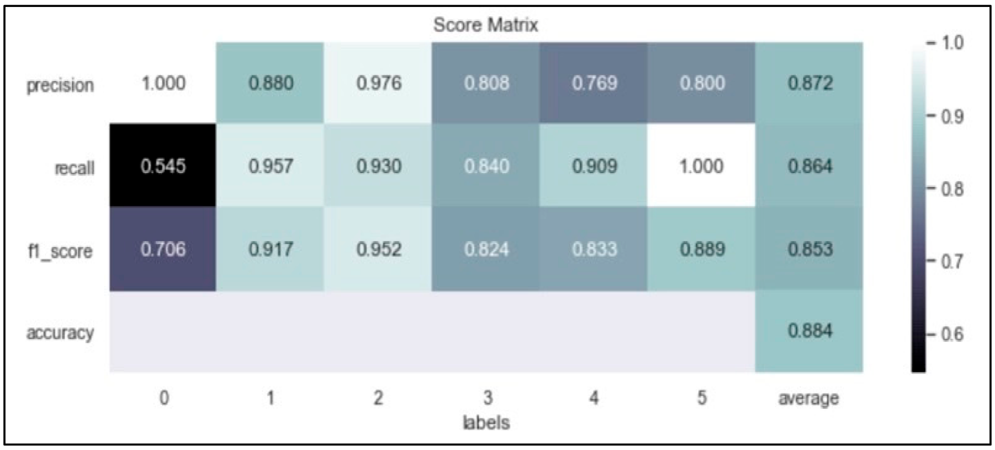


Figure 24. Confusion-matrix evaluation results.

Table 15. Evaluation results.

Metric Name	Score
Precision	87.2
Recall	86.4
F1-Measure	85.3
Accuracy	88.4
Cohen's kappa	85.1

Based on these results, the proposed model delivers significantly higher levels of accuracy when predicting the actual topic and therefore correctly assigns the authors' topics that are related to each other in the ground-truth table. Additionally, quantitative analysis was conducted, which equated to the evaluation process of the proposed advanced similarity measure and the robustness of the results; the aim of this was to discover the influence of the selected factors on the similarity results' accuracy, which was conducted by selecting different cases from the obtained data and applying the proposed approach. Three factors will be assessed in order to evaluate the validity of the claim made in this research: authors' keywords, publication statistics, and topic distribution. As a result, this qualitative analysis covers all applicable cases, which include the following instances:

- Case 1: Slightly differ in terms of keywords, similar with regard to publication statistics, and highly similar in relation to topic distribution.
- Case 2: Slightly differ in terms of keywords, similar with regard to publication statistics (theoretically highly different), and highly similar in relation to topic distribution.
- Case 3: Differ in terms of keywords, similar with regard to publication statistics, and highly different with regard to topic distribution.
- Case 4: Highly different in terms of keywords, similar with regard to publication statistics, and highly different in relation to topic distribution.

Moreover, expertise extraction analysis was employed, which is the evaluation process of the proposed metric to extract the expertise evidence in order to employ expertise identifications, allowing us to assign the experts. We completed this by selecting randomly different cases from the obtained data and applying the proposed approach. First, we converted the weighted topics' distribution for each paper into vector format by considering only the highest topic weight as the main topic for the paper and replacing it with 1 and 0 for all of the other topics' weights. We then added the corresponding citation counts for each paper to their identified topic. For each author in each topic, we counted the number of identified papers and added them to their corresponding citations; in addition, the averaged topic weight was included, which had previously been extracted. Table 16 shows the selected factors for each author where for each topic there are three factors to be computed for instance. Papers 0 represents the number of papers in Topic 0; Citations 0 is their corresponding citation count for papers in topic 0; Topic Weight 0 is the averaged weight for topic 0 publications distribution; and Expertise 0 is the average weight for topic 0, which denotes the number relating to the author's expertise with regard to Topic 0 based on his publications.

After ranking the resulting expertise for each topic, the author with the highest weight was considered to be the expert in that area. Table 17 highlights the top-three academics with the most experience in each topic.

Table 16. Process of extracting the authors' expertise.

Authors	Factors	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Author 1	Papers	0	1	5	14	21	1
	Citations	0	3	17	95	76	8
	Topic Weight	0.08907328	0.09769872	0.6399102	1.12391406	2.61450467	0.25997465
	Expertise	0.053444	0.858619	4.783946	22.474348	20.968703	1.955985
Author 2	Papers	19	8	17	3	14	0
	Citations	345	31	412	22	418	0
	Topic Weight	2.69323321	1.48435841	1.92540924	0.57017606	4.03028581	0.33310417
	Expertise	74.4159399	8.69061505	86.9552455	5.34210563	88.2181715	0.1998625

Table 17. Ranking authors' expertise in topics.

Topic	Authors	Papers	Citations	Topic Weight	Expertise
Topic 0	Author 1	7	517	0.89587796	105.337527
	Author 2	5	507	1.53028574	103.318171
	Author 3	19	345	2.69323321	74.4159399
Topic 1	Author 1	29	1009	4.03656339	210.021938
	Author 2	45	927	5.84844025	197.909064
	Author 3	33	657	4.73728603	140.842372

5. Discussion of the Results

The use of academic search engines, such as Google Scholar, is increasing. It is therefore becoming more important for authors to identify well-ranked papers as well as authors with similar research interests. Additionally, the use of network representation has proven its remarkable functionality by facilitating and accelerating the identification of comparable peers. The main aim of the network in this research was to assess and extract an author's expertise in order to direct them towards a highly recommended collaboration. Furthermore, to identify similar authors in the most efficient manner using academic search engines, this research concentrated on identifying similar authors using three key factors: the author-selected keywords, publications statistics, and the topics of the publications. The extracted main topics were evaluated with an accuracy of 88.4%.

However, when evaluating the resulting similarity with selected cases in the previous section, it was indicated that the greatest priority out of these factors is the publication topic distribution, which dominated the accuracy of the results. As shown in Table 18, Case 1 had slightly different keywords, but had a high statistical similarity with other publications, even though they are theoretically slightly different. Therefore, the topic similarity has finalized the result with accurate similarity, while neglecting the statistical publications' dissimilarity. Additionally, since they have a high level of topics' distribution similarity, they also share the same main research area, which was extracted from the topic distribution. This confirms that the topic extraction was very accurate. On the other hand, Case 2 has a high rate of variance in relation to other publications' statistics in spite of being recognized as a similar as topic distribution, which has standardized the results again into meaningful similarity. Case 3 differs in terms of keywords and theoretically differs slightly with regard to its similarities with other publications, as the topic distribution here was very different, which indicates that they are dissimilar authors in relation to their publication areas. Consequently, they also differ in terms of their main research area. Meanwhile, Case 4 contains an important observation regarding the topic distribution, which represents the weight of each topic according to the authors' publication: scholars may have knowledge of a topic but might not be an expert in it. Therefore, the topic distribution gives an accurate assessment of the results. In this case, even when the authors are dissimilar in two respects but are similar in some topics, they may still collaborate and should not be considered totally dissimilar.

Table 18. Summary of the evaluation cases.

Cases	Keywords	Publications' Statistics	Topics' Distribution	Similarity Result	Main Topic
Case 1	Slightly Differ	Slightly Differ	Highly Similar	Highly Similar	Same Topic
Case 2	Slightly Differ	Highly Differ	Highly Similar	Highly Similar	Same Topic
Case 3	Different	Slightly Differ	Highly Differ	Highly Differ	Different Topic
Case 4	Highly Different	Highly Differ	Slightly Similar	Slightly Similar	Different Topic

This research considered the degree of the topics' distribution over the authors' publications, and the findings showed that authors suitable for collaboration are not only limited to those linked to the same topic. Two authors may work on the same topic with the same level of knowledge, yet it may not be their field of expertise; however, they can still collaborate in that research area. On the contrary, finding an expert in a topic neglects the other topic distribution and only concentrates on the weight of this topic distribution, including the author's publication numbers and citation numbers for that area. The influence of the selected factors in relation to the resulting expertise levels were analyzed. In contrast to the similarity analysis, the topic distribution in the expertise identification was not the most dominating factor, whereas the citations were noted as the most influential element. In academia, the influence of authority means that a paper with a higher citation number has a greater impact because it indicates that the content of the paper is worthy and valuable. Moreover, the kind of influence in relation to the topic weight can be increased through the utilization of the abstract rather than through keywords alone. Therefore, the metric used was able to accurately identify the expert on a particular topic in instances where the topic was also their main area of interest. Overall, the proposed approach obtained robust results when predicting the most influential co-collaborations and identifying the experts while also transforming this tedious process into an interactive network visualization.

6. Conclusions

In academia, the publication of research papers and books is common to every science and research field, where hundreds of scholarly articles are being published online on a daily basis. In relation to this, one of the most important academic activities is searching for the right collaborator who will help to improve a scholar's research quality at the same time as accelerating the research process. Consequently, recommending academic collaborators based on scholarly big data is becoming increasingly relevant. It has also been observed that research interests play a significant role in the selection of academic collaborators and extracting expertise. When authors often write across multiple domains of interest, it may provoke a topic drift in general recommendation systems. As a result, creating a personalized collaborator recommendation system is a viable alternative.

This research aimed to propose a model for improving the extraction of the recommended author to collaborate with by relying on their explicit and implicit topics of interest as well as ascertaining levels of author expertise in a specific area. Our approach focused on addressing the gaps found in the literature as few academics in this field have considered scholars' multi-factors as an enhanced method to find potential collaborators or to identify the expert among them. Additionally, there was a lack of visualizations with regard to an interconnected network of authors that incorporate experts for each area, depict the co-collaboration of authors' universities and the availability of benchmark datasets that include all of the scholars' publications.

The proposed model was undertaken in four phases: data acquisition; data pre-processing; extracting, linking, and mapping; and interactive network visualization. First, the data acquisition phase involved collecting the relevant data from publicly available academic authors' publications in the field of computer science from twelve Saudi Arabian universities. In the second phase, the data was pre-processed due to the different languages that were found; subsequently, non-English publications and keywords that added no meaning were discarded. Meanwhile, the third phase consisted of three main parts: topic-distribution extraction; identifying experts in the selected topics; and computing authors' similarity-based keywords, publication statistics (range of active years, number of papers, and number of citations), and the topics' distribution. The last phase was designed to construct an interactive network that facilitates the representation of the proposed measure while also clustering the authors based on their research area. Additionally, a network of the authors' universities was constructed based on the papers' keywords, which convey universities' research areas perfectly. In conclusion, from the experimental results, the most influential factor for accurately recommending a collaborator was the topics' distribution, as it had an accuracy rate of 88.4%. Meanwhile, in relation to the country that was examined, the dominant factor in the process of expertise identification was the citation count due to its indication of valuable impact within academia. This research has helped to develop our understanding of different factors that may affect the authors' similarity quality. Furthermore, various academic disciplines might utilize such a predictive model to identify potentially beneficial collaborations.

On the other hand, there are certain limitations to the findings presented in this study that should be discussed further. This research relies on using the author-selected keywords for extracting the topics and finding the similarity; however, some of the publications from the collected dataset did not have keywords, whether it was a book chapter, thesis, or an old publication. Additionally, this research includes all the authors' publications, yet some of the collected papers were not publicly available and could not be accessed. In the process of extracting the topics of the papers, some of them could not be recognized or fit in the selected topics due to their differing in subject. Additionally, the lack of a benchmark dataset is a major limitation in this field of study.

Therefore, based on the factors identified in this research, further research on the similarity can be conducted in the future where instead of neglecting the publications that do not include keywords, replacing the missing keywords with the titles of the papers. Additionally, for more accurate results for the topics distributions, the publication time ranges for each topic could be included and a heterogeneous co-collaboration network that includes both the authors with their affiliations and computes their similarities could be built. The recommendations would be improved if potential and real collaborations were combined in a single network.

Author Contributions: Conceptualization, A.A.A., W.M.S.Y. and A.A.; methodology, A.A.A. and W.M.S.Y.; software, A.A.A.; validation, A.A.A. and W.M.S.Y.; formal analysis, A.A.A.; investigation, A.A.A. and W.M.S.Y.; resources, A.A.A. and W.M.S.Y.; data curation, A.A.A.; writing—original draft preparation, A.A.A.; writing—review and editing, A.A.A., W.M.S.Y. and A.A.; visualization, A.A.A.; supervision, W.M.S.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The authors confirm that the data supporting the findings of this study are publicly available in https://github.com/abrar-ali-m/Data-Authors-Similarities/blob/main/data-final.csv (accessed on 13 January 2022).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Algorithm A1 Authors' Publication Statistics

```
Input: $authors = \{u_1, u_2, u_3, \ldots, u_n\}, papers = \{p_1, p_2, \ldots, p_m\},$
       $year(p) = publication\ year\ of\ p,\ citation(p) = citation\ count\ of\ p$
Output: Number of paper, citation, years range
Begin
      $p_j = 0$
      for $\forall\ author \in authors$ do
        if $paper_i$ was written by $author_j$ then
    $p_j = p_j + 1$
    $PNum(author) \leftarrow p_j$
          end if
        end for
  for $\forall\ author \in authors$ do
    $Cite(author) \leftarrow \sum_{i=1}^{m} citation(p_i),\ author(p_i) == author$
          end for
        for $\forall\ author \in authors$ do
    $years \leftarrow \{year(p)\ \forall\ p \in papers,\ author(p) == author\}$
    $F_p \leftarrow x \in years,\ 0 < x < y,\ \forall\ y \in years,\ x \neq y$
    $L_p \leftarrow x \in years,\ 0 < x > y,\ \forall\ y \in years,\ x \neq y$
    $Years(author) \leftarrow L_p - F_p$
            end for
End
```

Algorithm A2 Authors' Similarities

```
Input: $authors = \{u_1, u_2, u_3, \ldots, u_n\},\ keywords,\ PNum,\ Cite,\ Years$
Output: Authors Similarity Matrix
Begin
      Compute similarity of keywords
    $Word_{sim} \leftarrow cosine(keywords_{author_1}, keywords_{author_2})$
      Compute author $PCYnum$ features
    $f_{author_1} \leftarrow [PNum(author_1),\ Cite(author_1),\ Years(author_1)]$
    $f_{author_2} \leftarrow [PNum(author_2),\ Cite(author_2),\ Years(author_2)]$
    $PCY_{sim} \leftarrow cosine\left(f_{author_1}, f_{author_2}\right)$
      Calculate similarity of topics
      Get authors $topics \leftarrow \{t_{p1}, t_{p2}, \ldots, t_{pm}\}$
    $T_{author_1} \leftarrow mean(t_p\ \forall\ p \in papers,\ author(p) == author_1)$
    $T_{author_2} \leftarrow mean(t_p\ \forall\ p \in papers,\ author(p) == author_2)$
    $Topic_{sim} \leftarrow cosine(T_{author_1}, T_{author_2})$
      Compute authors similarities
    $\alpha \leftarrow set$ weight
    $Author_{sim} \leftarrow \alpha * (Topic_{sim}) + (1 - \alpha) * (Word_{sim} + PCY_{sim})$
      return $Author_{sim}$
End
```

Algorithm A3 Extracting Authors' Expertise

```
Input: $authors = \{u_1, u_2, u_3, \ldots, u_n\}, \quad T_i = \{t_0, t_1, \ldots, t_k\},$
      $TWeight_i = \{tw_0, tw_1, \ldots., tw_m\}, Cite$
Output: Expert Authors
Begin
            for $\forall$ $author_i \in authors$ do
                    for $\forall$ $p \in papers$ do
                        for $\forall$ $t_i \in T_i$ do
       $P_{weight} \leftarrow t_i \forall\ p \in papers,\ author(p) == author_i$
                        end for
                    end for
                    Get the paper $main\ topic$
                    for $\forall$ $p \in papers$do
                      if $t_i = max\left(P_{weight}\right)$ then
     $t_i \leftarrow 1$
                      else
     $t_i \leftarrow 0$
                      end if
                  end for
        end for
        for $\forall$ $author_i \in authors$ do
    $PNum_{t_i} \leftarrow \sum_{i=1}^{k} paper(t_i)$
    $Cite_{t_i} \leftarrow \sum_{i=1}^{k} citation(t_i)$
        end for
          Get the weighted average for each $author$
  $\alpha \leftarrow set\ weight$
        for $\forall$ $author_i \in authors$ do
    $Avg_{Weight_i} \leftarrow \alpha * TWeight_i + (1 - \alpha) * (PNum_{t_i} + Cite_{t_i})$
        end for
        for $\forall$ $t_i \in, T_i$ do
    $Experts_{t_i} = max\left(Avg_{Weight_i}\right)$
        end for
        return $Experts_{t_i}$
End
```

Algorithm A4 Generating an Authors’ Graph

```
Input: $S_{u,\,u} = n \times n$ authors similarity matrix, $T_u = n \times m$ authors
topic distribution, $Experts_{t_i}$ = author with highest weighted average $t_i$
Output: Authors Co-collaboration Network
Begin
  $G \leftarrow Graph\ \langle V,\ E\rangle$
        Extracting main research area
        for each author in authors do
          for each $t_i$ in $T_u$ do
    $author_{topic} \leftarrow max(t_i \,\forall\, i\ \in\ T_u,\ author_{t_i} == author)$
            end for
    end for
    for each author in authors do
          Create node $V_{author}$ for author
        if $author \in Experts$ then
              Increase size of node $V_{author}$
        end if
            Clustering based on research areas
            Color $V_{author}$ corresponding to $author_{topic}$
            Add node $V_{author}$ to graph $G$
      end for
      for all $author_1,\ author_2,\ \in\ authors$ where $author_1 \neq authors_2$ do
          Set edge weight as similarity value between the authors:
    $weight \leftarrow S_{author1,\ author2}$
        If $weight > threshold$ then
                Create edge $E_{\,(V_{author1}\ ,\ \ V_{author2})}$ between nodes $V_{author1},\ \ V_{author2}$
                Set weight of edge to round ($weight$)
                Add edge $E_{(V_{author1},\, V_{author2})}$ to graph $G$
        end if
    end for
    return $G$
End
```

Algorithm A5 Similar Authors List

```
Input: Authors Co-collaboration Network
Output: Authors Ranked List
Begin
    for each node $V$ in graph $G$ do
   $N = list\ of\ all\ neighbors\ of\ node\ V\ in\ graph\ G$
            $W = list\ of\ all\ corresponding\ weights\ of\ E_{(v,\acute{v})}\ for\ each\ \acute{V}\ in\ N$
            Get sorted indexes of weights in descending order:
    $indexes\ \leftarrow argsort(W)$
            Sort and by $indexes$ to get ranked neighbors and weights:
    $N_{ranked} \leftarrow sort(N,\ key = indexes)$
    $W_{ranked} \leftarrow sort(W,\ key = indexes)$
            Update description text of node in with $N_{ranked}$ and $W_{ranked}$
    end for
    return $G$
End
```

References

1. Meishar-Tal, H.; Pieterse, E. Why Do Academics Use Academic Social Networking Sites? *Int. Rev. Res. Open Distrib. Learn.* **2017**, *18*, 1–22. Available online: http://www.irrodl.org/index.php/irrodl/article/view/2643/4044 (accessed on 13 November 2020). [CrossRef]
2. Xia, F.; Wang, W.; Bekele, T.M.; Liu, H. Big Scholarly Data: A Survey. *IEEE Trans. Big Data* **2017**, *3*, 18–35. [CrossRef]
3. Kong, X.; Shi, Y.; Yu, S.; Liu, J.; Xia, F. Academic social networks: Modeling, analysis, mining and applications. *J. Netw. Comput. Appl.* **2019**, *132*, 86–103. [CrossRef]

4. Caley, M.J.; O'Leary, R.A.; Fisher, R.; Choy, S.L.; Johnson, S.; Mengersen, K. What is an expert? A systems perspective on expertise. *Ecol. Evol.* **2014**, *4*, 231–242. [CrossRef] [PubMed]
5. Almuhanna, A.A.; Yafooz, W.M.S. Expert Finding In Scholarly Data: An Overview. In Proceedings of the IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), Toronto, ON, Canada, 21–24 April 2021; pp. 1–7. [CrossRef]
6. Shen, Q. Topic Discovery and Future Trend Prediction in Scholarly Networks. 2016. Available online: https://www.cs.sjtu.edu.cn/~{}wang-xb/wireless_new/material/Final2018/IEEE/%E6%B2%88%E7%BB%AE%E6%96%87-report.pdf (accessed on 13 January 2022).
7. Bai, X.; Liu, H.; Zhang, F.; Ning, Z.; Kong, X.; Lee, I.; Xia, F. An overview on evaluating and predicting scholarly article impact. *Information* **2017**, *8*, 73. [CrossRef]
8. Ibrahim, R.K.; Zeebaree, S.R.M.; Jacksi, K. Survey on semantic similarity based on document clustering. *Adv. Sci. Technol. Eng. Syst. J.* **2019**, *4*, 115–122. [CrossRef]
9. Alshareef, A.M.; Alhamid, M.F.; El Saddik, A. Recommending scientific collaboration based on topical, authors and venues similarities. In Proceedings of the 2018 IEEE 19th International Conference on Information Reuse and Integration for Data Science IRI, Salt Lake City, UT, USA, 6–9 July 2018; pp. 55–61. [CrossRef]
10. Leung, X.Y.; Sun, J.; Bai, B. Bibliometrics of social media research: A co-citation and co-word analysis. *Int. J. Hosp. Manag.* **2017**, *66*, 35–45. [CrossRef]
11. Kong, X.; Jiang, H.; Wang, W.; Bekele, T.M.; Xu, Z.; Wang, M. Exploring dynamic research interest and academic influence for scientific collaborator recommendation. *Scientometrics* **2017**, *113*, 369–385. [CrossRef]
12. Sun, N.; Lu, Y.; Cao, Y. Career age-aware scientific collaborator recommendation in scholarly big data. *IEEE Access* **2019**, *7*, 136036–136045. [CrossRef]
13. Czaputowicz, J.; Wojciuk, A. *International Relations in Poland: 25 Years after the Transition to Democracy*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 1–173. [CrossRef]
14. Kyvik, S.; Reymert, I. Research collaboration in groups and networks: Differences across academic fields. *Scientometrics* **2017**, *113*, 951–967. [CrossRef]
15. Kwiek, M. International Research Collaboration and International Research Orientation: Comparative Findings about European Academics. *J. Stud. Int. Educ.* **2018**, *22*, 136–160. [CrossRef]
16. Zhou, X.; Liang, W.; Wang, K.I.-K.; Huang, R.; Jin, Q. Academic Influence Aware and Multidimensional Network Analysis for Research Collaboration Navigation Based on Scholarly Big Data. *IEEE Trans. Emerg. Top. Comput.* **2018**, *9*, 246–257. [CrossRef]
17. Batagelj, V.; Maltseva, D. Temporal bibliographic networks. *J. Inf.* **2020**, *14*, 101006. [CrossRef]
18. Kumar, V.; Sendhilkumar, S.; Mahalakshmi, G.S. Author similarity identification using citation context and proximity. In Proceedings of the 2017 2nd International Conference on Recent Trends and Challenges in Computational Models, ICRTCCM 2017, Tindivanam, India, 3–4 February 2017; pp. 217–221. [CrossRef]
19. Al-Sultany, G.A. Exploiting Academic Factors for Improving Collaboration Recommendation System. *Int. J. Pure Appl. Math.* **2018**, *120*, 781–791.
20. Hernandez-Gress, N.; Ceballos, H.G.; Galeano, N. Research collaboration recommendation for universities based on data science. In Proceedings of the 2018 International Conference on Computational Science and Computational Intelligence, CSCI 2018, Las Vegas, NV, USA, 12–14 December 2018; pp. 1129–1132. [CrossRef]
21. Huang, S.-L.; Lin, S.-C.; Hsieh, R.-J. Locating experts using social media, based on social capital and expertise similarity. *J. Organ. Comput. Electron. Commer.* **2016**, *26*, 224–243. [CrossRef]
22. Köseoglu, M.A.; Okumus, F.; Putra, E.D.; Yildiz, M.; Dogan, I.C. Authorship Trends, Collaboration Patterns, and Co-Authorship Networks in Lodging Studies (1990–2016). *J. Hosp. Mark. Manag.* **2018**, *27*, 561–582. [CrossRef]
23. Liu, J.; Xia, F.; Wang, L.; Xu, B.; Kong, X.; Tong, H.; King, I. Shifu2: A network representation learning based model for advisor-advisee relationship mining. *IEEE Trans. Knowl. Data Eng.* **2019**, *33*, 1763–1777. [CrossRef]
24. Rathore, M.M.U.; Gul, M.J.J.; Paul, A.; Khan, A.A.; Ahmad, R.W.; Rodrigues, J.J.P.C.; Bakiras, S. Multilevel Graph-Based Decision Making in Big Scholarly Data: An Approach to Identify Expert Reviewer, Finding Quality Impact Factor, Ranking Journals and Researchers. *IEEE Trans. Emerg. Top. Comput.* **2021**, *9*, 280–292. [CrossRef]
25. Gao, S.; Li, X.; Yu, Z.; Qin, Y.; Zhang, Y. Combining paper cooperative network and topic model for expert topic analysis and extraction. *Neurocomputing* **2017**, *257*, 136–143. [CrossRef]
26. Berger, M.; Zavrel, J.; Groth, P. Effective Distributed Representations for Academic Expert Search. *arXiv* **2020**, arXiv:2010.08269.
27. Rampisela, T.V.; Yulianti, E. Academic Expert Finding in Indonesia using Word Embedding and Document Embedding: A Case Study of Fasilkom UI. In Proceedings of the 2020 8th International Conference on Information and Communication Technology, ICoICT 2020, Yogyakarta, Indonesia, 24–26 June 2020. [CrossRef]
28. Li, X.; Verginer, L.; Riccaboni, M.; Panzarasa, P. A network approach to expertise retrieval based on path similarity and credit allocation. *arXiv* **2020**, arXiv:2009.13958. [CrossRef]
29. Javadi, S.; Safa, R.; Azizi, M.; Mirroshandel, S.A. A Recommendation System for Finding Experts in Online Scientific Communities. *J. AI Data Min.* **2020**, *8*, 573–584. [CrossRef]
30. Isenberg, P.; Isenberg, T.; Sedlmair, M.; Chen, J.; Moller, T. Visualization as Seen through its Research Paper Keywords. *IEEE Trans. Vis. Comput. Graph.* **2016**, *23*, 771–780. [CrossRef]

31. Zhou, Z.; Shi, C.; Hu, M.; Liu, Y. Visual ranking of academic influence via paper citation. *J. Vis. Lang. Comput.* **2018**, *48*, 134–143. [CrossRef]
32. Mokhtari, H.; Barkhan, S.; Haseli, D.; Saberi, M.K. A bibliometric analysis and visualization of the Journal of Documentation: 1945–2018. *J. Doc.* **2020**, *77*, 69–92. [CrossRef]
33. Liu, J.; Ren, J.; Zheng, W.; Chi, L.; Lee, I.; Xia, F. Web of Scholars: A Scholar Knowledge Graph. In Proceedings of the SIGIR 2020—43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Xi'an, China, 25–30 July 2020; pp. 2153–2156. [CrossRef]
34. Vahed, N.; Gavgani, V.Z.; Jafarzadeh, R.; Tusi, Z.; Erfanmanesh, M. Visualization of the Scholarly Output on evidence based Librarianship: A social network analysis. *Evid. Based Libr. Inf. Pract.* **2018**, *13*, 50–69. [CrossRef]
35. Harzing, A.W. Publish or Perish. 2007. Available online: https://harzing.com/resources/publish-or-perish (accessed on 21 January 2021).
36. QS University. QS University Rankings for Arab 2021 | Top Universities. 2021. Available online: https://www.topuniversities.com/university-rankings/arab-region-university-rankings/2021 (accessed on 3 February 2021).
37. Abdolahi, M.; Zahedi, M. A new method for sentence vector normalization using word2vec. *Int. J. Nonlinear Anal. Appl.* **2019**, *10*, 87–96. [CrossRef]
38. Mullen, L.A.; Benoit, K.; Keyes, O.; Selivanov, D.; Arnold, J. Fast, Consistent Tokenization of Natural Language Text. *J. Open Source Softw.* **2018**, *3*, 655. [CrossRef]
39. Kaur, J.; Kaur Buttar, P. A Systematic Review on Stopword Removal Algorithms. *Int. J. Future Revolut. Comput. Sci. Commun. Eng.* **2018**, *4*, 207–210. Available online: http://www.ijfrcsce.org (accessed on 11 February 2021).
40. Straka, M.; Straková, J. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, BC, Canada, 3–4 August 2017; pp. 88–99. [CrossRef]
41. Alshalan, R.; Al-Khalifa, H.; Alsaeed, D.; Al-Baity, H.; Alshalan, S. Detection of hate speech in COVID-19-related tweets in the Arab Region: Deep learning and topic modeling approach. *J. Med. Internet Res.* **2020**, *22*, e22609. [CrossRef]
42. Qaiser, S.; Ali, R.; Utara, U.; Sintok, M.; Kedah, M.; Ramsha, A.; Analytics, T. Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents Text Mining. *Artic. Int. J. Comput. Appl.* **2018**, *181*, 975–8887. [CrossRef]
43. Meng, G.; Xu, J.; Zhao, J.; Fu, L.; Long, H.; Gan, X.; Wang, X. Maximum Value Matters: Finding Hot Topics in Scholarly Fields. *arXiv* **2017**, arXiv:1710.06637. [CrossRef]
44. Yoon, S.J.; Yoon, D.Y.; Lee, H.J.; Baek, S.; Lim, K.J.; Seo, Y.L.; Yun, E.J. Distribution of citations received by scientific papers published in the imaging literature from 2001 to 2010: Decreasing inequality and polarization. *Am. J. Roentgenol.* **2017**, *209*, 248–254. [CrossRef] [PubMed]
45. Yang, S.; Xing, X.; Wolfram, D. Difference in the impact of open-access papers published by China and the USA. *Scientometrics* **2018**, *115*, 1017–1037. [CrossRef]
46. Lozano, S.; Calzada-Infante, L.; Adenso-Díaz, B.; García, S. Complex network analysis of keywords co-occurrence in the recent efficiency analysis literature. *Scientometrics* **2019**, *120*, 609–629. [CrossRef]