International Journal of Applied Science and Engineering

Investigating scientific collaboration networks in Iraq using cloud computing and data mining

Nagham A. Sultan *, Dhuha B. Abdullah

Department of Computer Science, University of Mosul, Mosul, Iraq

ABSTRACT

The academic performance of Iraqi authors and institutions is an important aspect that needs to be investigated in the field of research. Special scientific criteria establish a researcher's level. These include citations, published articles, and others. In this study, the aim is to analyze the academic performance patterns of Iraqi authors and institutions in Google Scholar. Collaboration patterns between Iraqi scientists and universities were explored using sophisticated data collection and analysis techniques. A crawler that worked in parallel with Google Scholar was created with SerpApi, thus collecting amounts of scientist profile data stored in MongoDB Atlas. The data were processed using Spark within the Amazon Web Services (AWS) Cloud. Complex network analysis and Cytoscape software were used to comprehensively investigate the co-authorship networks of Iraqi scientists. In addition, a set of complex network criteria was applied to reveal patterns of cooperation. After visualizing the network diagram and analyzing the relationships between the nodes, the virtual nodes and communities in the network were identified, revealing critical insights into the cooperation patterns of Iraqi scientists and universities. The results have important implications for research and development policies in Iraq and demonstrate the power of complex network analysis to reveal valuable insights into academic research collaboration.

Keywords: Bigdata, Parallel scraping, MongoDB, Apatch spark, Collaboration network.

Received: June 18, 2023
Revised: July 11, 2023
Accepted: July 21, 2023

Corresponding Author:
Nagham A. Sultan
naghamajeel@uomosul.edu.iq

Publisher:
Chaoyang University of Technology
ISSN: 1727-2394 (Print)
ISSN: 1727-7841 (Online)

1. INTRODUCTION

The growth of the bibliometrics discipline has always depended on the availability of large-scale metadata sources about scientific publications. These sources are ultimately the raw materials that bibliometrics use to carry out their analyses, so having access to them has been essential to developing the field (Delgado et al., 2019). The resurgence and increase in interest in the scale of the academic web can be attributed to the introduction of academic search engines such as Google Scholar and Microsoft Academic Search. These academic search engines aim to index the entirety of contemporary academic knowledge. Google Scholar is a search engine that provides access to scholarly literature across various disciplines. It indexes articles, theses, and conference papers and includes citation data for each item. Researchers can use Google Scholar to find articles related to their research, track citations to their work, and create profiles to showcase their publications. Google Scholar also offers an API for developers to build custom applications that integrate with its data (Orduña et al., 2015).

A scraper is a software program that automatically extracts data from websites. It can extract information like titles, authors, abstracts, and other metadata from Google Scholar pages (Azhar et al., 2019). Big data requires specialized tools and techniques to process and analyze it effectively. To work with big data, data scientists and engineers must have a solid understanding of distributed computing, data processing, and statistical analysis. In addition, they must be familiar with big data technologies like Hadoop, Spark, and NoSQL databases. Big data also raises ethical concerns around data privacy, security, and bias, which must be addressed in any data processing and analysis project; one of this solution is Apache Spark (Ramirez et al., 2018).

Apache Spark is an open-source distributed computing system that allows for the parallel processing of large data sets. At the same time, AWS provides cloud-based infrastructure for Spark, including managed Spark clusters and data warehousing. Using Spark in the AWS cloud allows researchers to scale their data processing and analysis capabilities on demand without needing expensive on-premises hardware. AWS also provides extensive data services, like Athena, Redshift, and Kinesis, that can be used with Spark to build a complete data processing and analysis pipeline (Buyya et al., 2016; Laghari et al., 2022). The Spark framework's ability to deliver low latency and high throughput for enormous data sets, in comparison to many other frameworks, is primarily responsible for the massive amount of popularity it has garnered in big data processing (Sivarajah et al., 2017; Ali et al., 2023). There are a number of things that set it apart from the competition and make it desirable from other data mining frameworks. Speed, powerful caching, deployment, real-time support, and multilingual support are some of the benefits of Spark. Using controlled partitioning, Spark speeds up to 100 times quicker than Hadoop Map/Reduce (Jaiswal et al., 2020). MongoDB is a popular NoSQL database that uses a document-oriented data model. It provides scalability, flexibility, and ease of storing and managing extensive data. MongoDB is commonly used in big data applications because it handles unstructured and semi-structured data, like documents, images, and videos. MongoDB is also highly scalable, allowing users to easily add new nodes to a cluster as their data grows. It is used in conjunction with other big data technologies, like Hadoop and Spark, to provide complete data processing and analysis (Osipov, 2019; Laghari et al., 2021).

Moreover, Amazon Elastic Compute Cloud (Amazon EC2) offers scalable computing power in the AWS Cloud. Utilizing Amazon EC2 allows for quicker application development and deployment because there is no longer a requirement to make an upfront hardware investment. Using Amazon EC2, virtual servers can be started, security and networking can be set up, and storage can be controlled. Estimating traffic is unnecessary because scaling up or down can accommodate fluctuations in demand or popularity (Laghari et al., 2018; Lula et al., 2020). In recent years, there has been an increase in interest in technologies that are concerned with indexing academic research and researchers, including many methodologies to deal with managing infrastructure, providing an academic environment, and making it a competitive statistical platform. Some concerns have been addressed in the literature in this context.

Al Husaeni et al. (2022) presented a comprehensive guide to using bibliometric analysis to study digital learning articles published before and after the COVID-19 pandemic. This research aims to provide a step-by-step guide for using bibliometric analysis tools, Vosviewer and Publish or Perish, to analyze research publications related to digital learning before and after the pandemic. The study uses Google Scholar data to identify and analyze relevant publications and to visualize co-citation and bibliographic coupling networks. The authors present a step-by-step guide to using the software tools, including data collection, cleaning, and analysis. They also provide practical examples of applying the means to analyzing digital learning articles and examining patterns of authorship, citation networks, and co-citation networks. The study results demonstrate the utility of Vosviewer and Publish or Perish as powerful tools for bibliometric analysis, providing users with a range of analytical features to explore patterns in research publications and citations. Fujita et al. (2021) proposed a new method to identify promising researchers based on the analysis of co-authorship networks from academic literature. In the study, the authors used network centrality measures to evaluate the prominence of individual researchers in a co-authorship network. They identified the most central researchers in the network based on various measures of centrality, including degree centrality, betweenness centrality, and eigenvector centrality. The study found that network centrality measures could effectively identify promising researchers who traditional citation-based methods might have overlooked. The authors concluded that co-authorship network analysis using network centrality measures could be a helpful tool for identifying good researchers and facilitating collaboration in academic communities. Martín-Martín et al. (2018) proposed a novel method for depicting academic disciplines using Google Scholar Citations data. They argue that traditional bibliometric analysis approaches, such as journal impact factors, fail to capture the multidisciplinary nature of research accurately and can overlook essential connections between different fields. To address this issue, the researchers propose a new method for analyzing citation data from Google Scholar Citations that considers the interdisciplinary nature of research. They developed a visualization tool called the Academic Disciplines Graph (ADG), which uses network analysis to identify clusters of related research areas based on citation patterns. The researchers demonstrate the utility of their approach by applying it to the field of bibliometrics, analyzing the citation patterns of articles published in the Journal of Informatics. They found that their system identifies a broader range of research areas than traditional bibliometric methods and provides a more accurate representation of the interdisciplinary nature of research. This paper has some contributions, as follows:

1. There is no standard approach for scraping data from Google Scholar, and No study was conducted on the authors of the Iraqi universities. Therefore, A scraper tool that can collect data from Google Scholar using cloud computing technologies has been proposed.
2. This paper shows how to effectively store and process data with MongoDB Atlas and Apache Spark on AWS. Uses the cloud to achieve scalability, flexibility, and availability for managing massive datasets.

3. Explores the co-authorship networks of Iraqi scientists and universities using Complex network analysis techniques and network metrics to find patterns of institutional collaboration in Iraq.
4. In this study, the network of Iraqi authors is built and compared to networks found in previously published works of literature. demonstrates the location of Iraq's university system on a map and how geographic distances affect university collaboration.

The rest of this paper is organized as follows: Section 2 describes the approach followed in performing this research. Section 3 presents the obtained results in terms of the generated network and discusses the results obtained. Finally, this research is concluded in Section 4.

2. RESEARCH METHODOLOGY

In this research, we focus on studying patterns of collaboration between Iraqi scholars and universities using Google Scholar data and complex network analysis. By using innovative data collection, processing, and analysis techniques, the aim is to shed light on the research landscape in Iraq and identify potential areas for collaboration and improvement through the following methodology.

2.1 Parallel Scraping, Storage and Processing

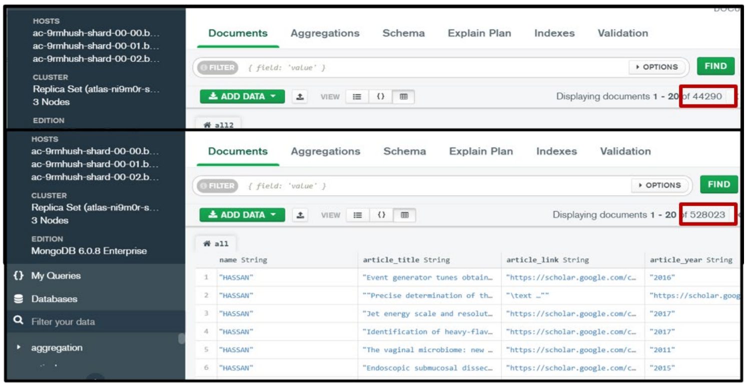
To collect the data, multiple computers are used to implement parallel web scraping and store the results in a MongoDB Atlas cloud database. VirtualBox is used to create three virtual instances of operating systems on each computer. Then the necessary dependencies, such as Python and any relevant libraries, are installed on each VM, and the scraping script is run on each VM independently. First designed three Python scripts that worked in parallel to crawl into Google Scholar using the SerpApi. The scripts were designed to extract author data from their profiles, including author names, affiliations, publication titles, h-index, i10 index, and citations. To ensure that the collected data was of high quality and accuracy, several parameters were set up to control the crawling process. For example, the number of requests per minute is limited to avoid overloading the Google Scholar server and to implement a delay between each request. After collecting the data, it is stored in the Cloud MongoDB Atlas, a cloud-based database management system. MongoDB Atlas is chosen because it provides high scalability, flexibility, and availability, which are crucial for managing large amounts of data. The data of the authors collected from Google Scholar using scrapers for Iraqi universities amounted to 44,290 authors, and the data of the published papers of these authors amounted to 528,023 research papers, as shown in Fig. 1. The proposed algorithm to scrape data from Google Scholar is showed in algorithm 1.

In the MongoDB Atlas cloud, A cluster was constructed of three nodes, one master and two workers, to iterate data across these nodes to ensure it is always available even if one node fails, as shown in Fig. 2. In contrast, storing data in a local file is not scalable, secure, or flexible in terms of data organization. MongoDB is highly scalable, able to handle large amounts of data and high traffic.

```
Algorithm1: scraping scholarly data (author and article information)
Input: CSV file containing author IDs.
Output: Author and article information in Cloud MongoDB Atlas.
1. Start
2. Import the essential libraries, such as Pymongo, SerpApi, and Pandas.
3. Using the connection string that is provided, connect to the MongoDB database.
4. Examine the author IDs in the CSV file.
5. Remove any null or unnecessary characters from the CSV file.
6. Create an empty DataFrame df to hold the information about the retrieved article.
7. To store the retrieved author data, create an empty DataFrame df2.
8. Go through the appropriate amount of author IDs in iterations.
    a Using the current author ID, set the Google Scholar author search's engine, author_id, number, and Api_key parameters.
    b Carry out the search utilizing the SerpApi library and obtain the search outcomes.
    c Take the necessary data, including author and article details, out of the search results.
    d Add the DataFrame df2 with the author information that was extracted.
    e Add the information from the retrieved article to the DataFrame df.
9. Save the DataFrame df to the "articles.csv" CSV file with UTF-8 encoding.
10. UTF-8 encoding the DataFrame df2 into the CSV file "author.csv"
11. Open DataFrames df and df2 and read the CSV files "author.csv" and "articles.csv" into.
12. Create dictionaries from the DataFrames df and df2.
13. In the MongoDB database, add the dictionaries to the appropriate collections.
14. End.
```

Another benefit of storing data in MongoDB is its enhanced security features, such as authentication, authorization, encryption, and auditing, which protect data from unauthorized access or manipulation. Therefore, using the MongoDB cloud and cluster to store data provided benefits that significantly improved the efficiency and security of data management.

The Spark cluster was placed on the AWS EC2 platform to maximize the value of the acquired data, as shown in Fig. 3. The cluster, comprising a Node Master and two worker nodes, was constructed as shown in Fig. 4 to process the acquired data in a distributed and parallel manner. The processing phase included cleaning, filtering, transforming, and preparing the data. MongoDB's seamless interaction with Spark makes a smooth and uninterrupted workflow possible, maximizing data transmission potential. By utilizing the features of AWS EC2 servers, we were able to benefit from the inherent scalability and reliability of a cloud platform. This enables cluster resources to be quickly scaled based on data volume and computing needs, ensuring top performance and affordability. EC2's robust security protocols and automated backup systems further ensure the safety and integrity of priceless research data.

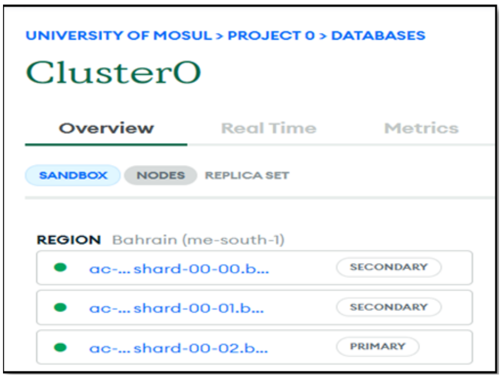


Displaying documents 1 - 20 of 44290

Displaying documents 1 - 20 of 528023

	name String	article_title String	article_link String	article_year String
1	"HASSAN"	"Event generator tunes obtain…	"https://scholar.google.com/c…	"2016"
2	"HASSAN"	""Precise determination of th…	"\text …""	"https://scholar.goog
3	"HASSAN"	"Jet energy scale and resolut…	"https://scholar.google.com/c…	"2017"
4	"HASSAN"	"Identification of heavy-flav…	"https://scholar.google.com/c…	"2017"
5	"HASSAN"	"The vaginal microbiome: new …	"https://scholar.google.com/c…	"2011"
6	"HASSAN"	"Endoscopic submucosal dissec…	"https://scholar.google.com/c…	"2015"

Fig. 1. Scrape data and store it in Mongodb cloud



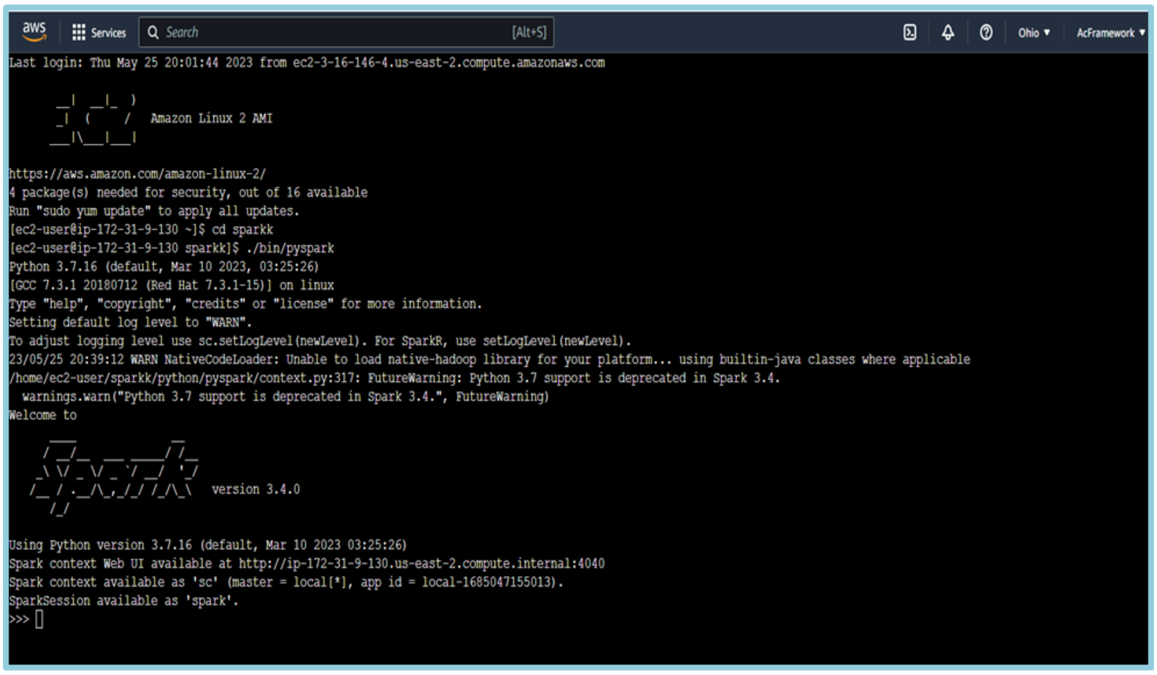
REGION Bahrain (me-south-1)

ac-… shard-00-00.b… SECONDARY

ac-… shard-00-01.b… SECONDARY

ac-… shard-00-02.b… PRIMARY

Fig. 2. MongoDB Atlas cluster in AWS



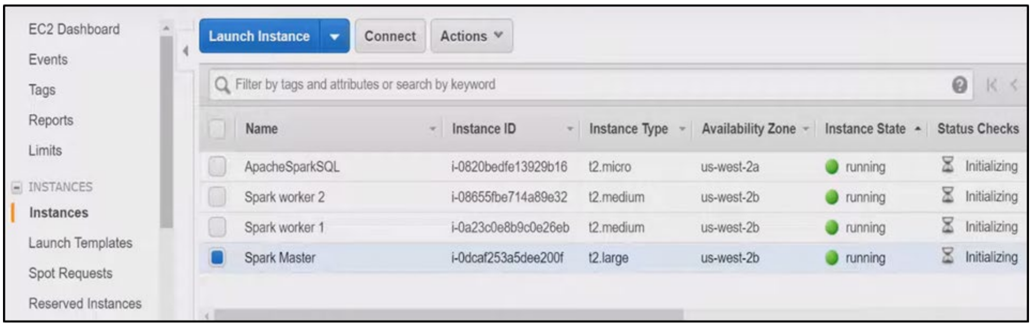
```
Last login: Thu May 25 20:01:44 2023 from ec2-3-16-146-4.us-east-2.compute.amazonaws.com

       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/
4 package(s) needed for security, out of 16 available
Run "sudo yum update" to apply all updates.
[ec2-user@ip-172-31-9-130 ~]$ cd sparkk
[ec2-user@ip-172-31-9-130 sparkk]$ ./bin/pyspark
Python 3.7.16 (default, Mar 10 2023, 03:25:26)
[GCC 7.3.1 20180712 (Red Hat 7.3.1-15)] on linux
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/05/25 20:39:12 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
/home/ec2-user/sparkk/python/pyspark/context.py:317: FutureWarning: Python 3.7 support is deprecated in Spark 3.4.
  warnings.warn("Python 3.7 support is deprecated in Spark 3.4.", FutureWarning)
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.4.0
      /_/

Using Python version 3.7.16 (default, Mar 10 2023 03:25:26)
Spark context Web UI available at http://ip-172-31-9-130.us-east-2.compute.internal:4040
Spark context available as 'sc' (master = local[*], app id = local-1685047155013).
SparkSession available as 'spark'.
>>> 
```

Fig. 3. Run Spark in AWS console



	Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks
□	ApacheSparkSQL	i-0820bedfe13929b16	t2.micro	us-west-2a	● running	Initializing
□	Spark worker 2	i-08655fbe714a89e32	t2.medium	us-west-2b	● running	Initializing
□	Spark worker 1	i-0a23c0e8b9c0e26eb	t2.medium	us-west-2b	● running	Initializing
■	Spark Master	i-0dcaf253a5dee200f	t2.large	us-west-2b	● running	Initializing

Fig. 4. Spark cluster in EC2

The carefully planned research methodology, which includes the Google Scholar scraper, MongoDB storage, and Spark cluster, allows us to extract essential insights from enormous amounts of research data. With the help of this all-encompassing strategy, it can perform complex analyses, detect obscure trends, and provide new research results.

2.2 Network Generation

After collecting the Iraqi authors data from GS, A dataset was created from two files: a node and an edge. The node file includes information about the collected universities and their attributes. The edge file consists of the relationships among the universities. The strategy of creating edges among universities is based on co-authoring papers between the universities. An edge is made between two universities if their authors co-authored or collaborated on a paper (two authors from different universities have a paper in common). The weight of an edge between two universities is increased according to the number of papers associated with both. This strategy for generating academic networks (including nodes and edges) is followed in the literature, such as the studies of (Abdullah, 2020; Mohammed et al., 2020; Sultan et al., 2020). The proposed algorithm for Network Generation as follow:

```
Algorithm2: Network generation for researcher author and co-authors data
Input:  Researchers names & coauthors data
Output: CSV File with co-author network generation
1. Start
2. Import libraries: Pandas, SerpApi (Google Search)
3. Open-source CSV files contain author and coauthor names as rows for each paper and are saved in a DataFrame (df).
4. Open another CSV file that contains author and coauthor names with an ID for each one, and save it in another DataFrame (df3).
5. Create a new empty DataFrame (df2) with the column's 'index', 'source', 'target', and weight.
6. loop through each row and column of df:
    a. Check if the value in the current column exists in the 'coauthor names' column of df2.
        i.  If yes, get the index of that row and increment the 'weight' value by 1.
        ii. If no, add a new row to df2.
    b. Get the index of the row in df3 where the 'names' column matches the value in column 'name' for the current row in df.
    c. Set the 'source' column and 'target' column of the new row to the 'index' value of that row in (df3), with the weight' column of the new row set to 1.
7. Write (df2) to the output CSV file in UTF-8 encoding.
8. End.
```

3. RESULTS AND DISCUSSION

The first step in generating the network starts with feeding the Cytoscape software with the dataset (nodes and edges). Cytoscape software is used to visualize networks and perform network calculations. The main characteristics of the Iraqi Co-Authorship Network are presented in Table 1. The table shows that the number of Iraqi universities (collected) is 176, including public and private universities. The number of edges is 322, representing the relationships among the Iraqi universities. The diameter, the farthest distance between two nodes in the network, is 17, considered high compared to the network size. It should be mentioned that the distance of 17 is the number of steps to get between the farthest two universities in Iraq, which is significant. This also means there is a lack of collaboration among the Iraqi universities.

The density of the edges in the network is too low compared to the size of the network, which also reflects a common tendency for collaboration between Iraqi universities. On the other hand, the average clustering coefficient (Average CC) reflects the tendency of nodes to cluster together. This means high levels of Average CC represent a high level of collaboration. The value of the Average CC should be between 0 and 1. According to Table 1, the Average CC is too low, confirming the previous findings. Also, the average path length of 3.275 ensures the results. Fig. 5 shows the visualization of the network. The coordinates of each university are projected on the Iraqi map to show their exact locations and provide a comprehensive view of the Iraqi universities in terms of scientific collaboration. In this figure, each university has a different color, and the size of the nodes represents the number of papers at the university (larger nodes reflect a large number of papers). The edges represent the relations between universities; thicker edges mean a high level of collaboration in authoring papers between two universities.

And it also showed an important observation: cooperation between universities depends on their geographical distances; it turns out the universities in the northern areas of Iraq have more collaboration between them. Similarly, the universities in the middle regions of Iraq have more collaborations than other sites in Iraq. This phenomenon can be interpreted as an indicator of a problem in the collaboration pattern between the Iraqi universities related to the geographical area. The Ministry of Higher Education and Scientific Research in Iraq should adopt a strategy to address this issue and offer additional support for collaboration between two universities that are farther apart.

In addition to the previous findings, the modularity of the communities of Iraqi universities is also investigated. The modularity of a network reflects the strength of the communities in its structure (Tomasini and Menezes, 2015). Fig. 6 demonstrates the level of modularity of all the communities of the authors in the Iraqi universities. It was found that there are a total of 50 university communities in Iraq with a maximum modularity level of 0.256, which reflects weak modularity. Moreover, the degree distribution of authors represents the frequency of co-authoring papers.

Fig. 7 depicts a power-law distribution of degrees, which means there are a few authors in Iraq who have a high frequency of co-authoring papers, with a majority of authors who have few collaborations. This finding tells us there is a gap between the senior authors and other authors, which should be addressed in the community of Iraqi authors.

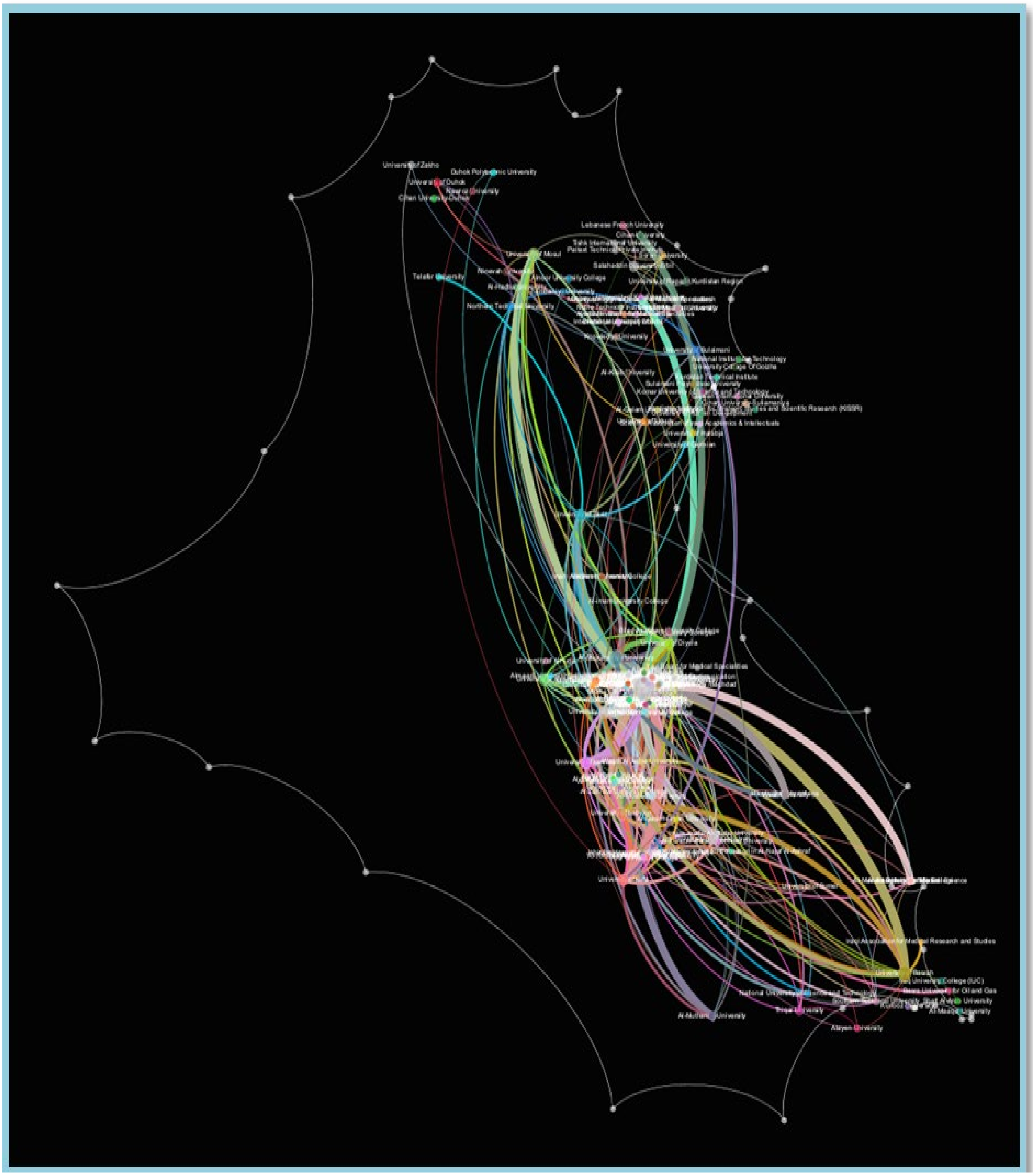

University of Zakho
Duhok Polytechnic University
University of Duhok
Nawroz University
Cihan University-Duhok
Lebanese French University
Cihan University
Tishk International University
Polad Technical Private Institute
Salahaddin University-Erbil
University of Mosul
Ninevah University
Telafer University
Alnoor University College
Al-Hadba University College
University of Raparin Kurdistan Region
Northern Technical University
Koya University
University of Sulaimani
National Institute of Technology
University College Of Goizha
Al-Kitab University
Sulaimani Polytechnic University
Komar University of Science and Technology
American University of Iraq Sulaimani
Kurdistan Institution for Strategic Studies and Scientific Research (KISSR)
Academics & Intellectuals
University of Halabja
University of Garmian
University of Diyala
Al-Mustansiriyah University
University of Kerbala
Al-Mustaqbal University College
University of Kufa
Al-Muthanna University
Iraqi Association for Medical Research and Studies
University of Basrah
Iraq University College (IUC)
Basra University for Oil and Gas
Southern Technical University
Shatt Al-Arab University
Al Maaqal University
National University of Science and Technology
Thiqar University
Alayen University
University of Sumer

Fig. 5. Visualization of the network

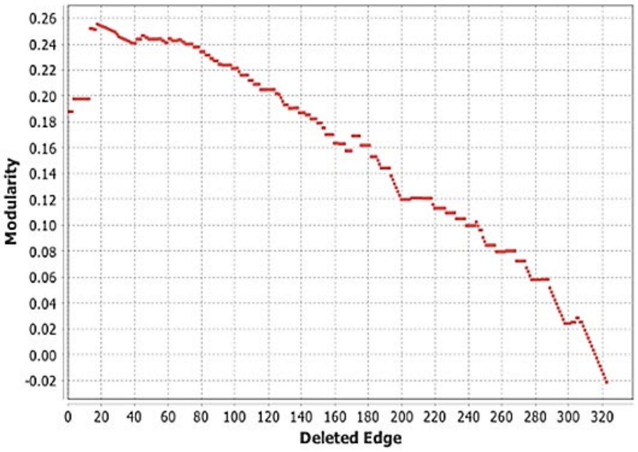


Fig. 6. Modularity of all the communities of the authors in the Iraqi universities

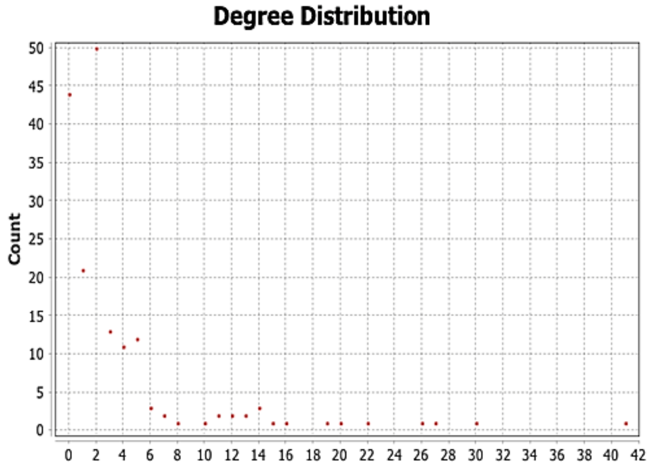


Fig. 7. Power-law distribution of degrees

This work proposes a scraper to retrieve the bibliographic information of academic authors from Google Scholar. Although the literature includes many methods to scrape authors information from online academic sources, the proposed method is considered unique because it can be performed completely in the cloud. This means researchers are able to automate the whole process and eventually obtain a dataset that is ready to be visualized, which is an important feature to be considered for the proposed scraper. Moreover, since Google Scholar is updated on a daily basis, the proposed scraper can be adjusted for any institution to automatically retrieve their bibliography data periodically and investigate the patterns in the data, which makes the process easier.

The generated network of Iraqi authors was benchmarked with two similar networks in the literature. Table 2 shows that the performance of the Iraqi network of authors has close values in terms of the average clustering coefficient and the average shortest path length. In fact, the reason behind using these metrics is that they can be used to compare two networks as a whole. The average clustering coefficient is used to measure the tendency of authors to collaborate with each other, as shown in the studies (Fronczak et al., 2002; Dabdawb and Mahmood, 2021; Hammadi et al., 2021). It shows a low tendency for the authors to collaborate with each other. The other metric is Average Shortest Path Length, which reflects the shortest path between authors to collaborate within the network. The value of 3.275 is considered acceptable compared to the ACM network, which means there are authors who play the role of a bridge between the communities of authors, and this is clear when observing the Average Clustering Coefficient. This metric is proven to be efficient in measuring the performance of a network, as shown in (Mahmood and Menezes, 2013). Furthermore, other network measurements play a significant role in evaluating a network. These metrics can also be used to analyze the spectral features of a network aiming for a more in-depth investigation, such as eigen-centrality, betweenness, and eccentricity. However, these kinds of metrics were left for future investigations of the network.

4. CONCLUSION

In conclusion, this paper shows the effectiveness of cutting-edge data collection and processing techniques in addressing complex research subjects. Through integrating Python, Spark, complex network analysis, and Cytoscape technologies, access to a large amount of data and revealing hidden collaboration patterns within the Iraqi academic's co-authorship network were achieved. Examining the Iraqi Co-Authorship Network reveals critical findings on the collaboration patterns among Iraqi institutions. The network, which consists of 176 universities, has a lack of coordination and sparse connectivity. The network's low edge density and high diameter value of 17 indicate limited institutional collaboration.

Additionally, the average clustering coefficient shows that there isn't much of a trend for university clusters to work together. The network's representation on the Iraqi map demonstrates how proximity to one another influences collaboration, with colleges in the country's north and center exhibiting better links with one another. Examining community modularity reveals weak community structures among Iraqi universities, displaying a fragmented network. The distribution of power-law degrees also points to a disparity between senior authors who collaborate frequently and those who collaborate less regularly. In conclusion, the analysis highlights the importance of addressing the collaboration patterns among Iraqi universities. Scholarly collaboration could be improved and promoted to promote a more unified research community in Iraq by encouraging more significant links, both geographically and across different levels of authorship.

ACKNOWLEDGEMENT

The authors are grateful to the University of Mosul, College of Computer Science and Mathematics, and Computer Science Department for helping us perform this research.

REFERENCE

Abdullah, D.B., 2020. Network-based bibliometric method for analyzing collaboration and publishing tendencies. In 6th International Engineering Conference Sustainable Technology and Development (IEC), 174–178.

Al Husaeni, D.F., Nandiyanto, A.B.D. 2022. Bibliometric using vosviewer with publish or perish (using Google Scholar data): From step-by-step processing for users to the practical examples in the analysis of digital learning articles in pre and post Covid-19 pandemic. ASEAN Journal of Science and Engineering, 2, 19–46.

Ali, M., Jung, L.T., Sodhro, A.H., Laghari, A.A., Belhaouari, S.B., Gillani, Z. 2023. A Confidentiality-based data Classification-as-a-Service (C2aaS) for cloud security. Alexandria Engineering Journal, 64, 749–760.

Azhar, R.J.K., Nurhakim, L., Putra, R.E. 2019. Implementasi web scraping untuk menampilkan informasi tayangan film di bioskop: Book my show. Universitas Siliwangi, 1, 1–7.

Buyya, R., Calheiros, R.N., Dastjerdi, A.V. (Eds.). 2016. Big data: Principles and paradigms. Morgan Kaufmann, USA.

Dabdawb, M., Mahmood, B. 2021. On the relations among object-oriented software metrics: A network-based approach. International Journal of Computing and Digital Systems, 1, 901–915.

Delgado López-Cózar, E., Orduña-Malea, E., Martín-Martín, A. 2019. Google Scholar as a data source for research assessment. In: Glänzel, W., Moed, H.F., Schmoch, U., Thelwall, M. (eds) Springer Handbook of Science and Technology Indicators, 95–127.

Divakarmurthy, P., Menezes, R. 2013. The effect of citations to collaboration networks. Complex Networks, 424, 177–185.

Fronczak, A., Hołyst, J.A., Jedynak, M., Sienkiewicz, J. 2002. Higher order clustering coefficients in Barabási–Albert networks. Physica A: Statistical Mechanics and Its Applications, 316, 688–694.

Fujita, M., Inoue, H., Terano, T. 2021. Analyzing promising researchers using network centralities of co-authorship networks from academic literature. New Generation Computing, 39, 181–197.

Hammadi, D.S., Mahmood, B., Dabdawb, M.M. 2021. Approaches on modelling genes interactions: A review. Technium BioChemMed, 2, 38–52.

Jaiswal, A., Dwivedi, V.K., Yadav, O.P. 2020. Big data and its analyzing tools: A perspective. In 6th International Conference on Advanced Computing and Communication Systems (ICACCS 2020), 560–565.

Laghari, A.A., He, H., Khan, A., Kumar, N., Kharel, R. 2018. Quality of experience framework for cloud computing (QoC). IEEE Access, 6, 64876–64890.

Laghari, A.A., He, H., Khan, A., Laghari, R.A., Yin, S., Wang, J. 2022. Crowdsourcing platform for QoE evaluation for cloud multimedia services. Computer Science and Information Systems, 19, 1305–1328.

Laghari, A.A., Jumani, A.K., Laghari, R.A. 2021. Review and state of art of fog computing. Archives of Computational Methods in Engineering, 28, 1–13.

Lula, P., Dospinescu, O., Homocianu, D., Sireteanu, N.A. 2020. An advanced analysis of cloud computing concepts based on the computer science ontology. Computers, Materials and Continua, 66, 2425–2443.

Mahmood, B., Menezes, R. 2013. United states congress relations according to liberal and conservative newspapers. In 2013 IEEE 2nd Network Science Workshop (NSW), 98–101.

Martín-Martín, A., Orduna-Malea, E., Delgado López-Cózar, E. 2018. A novel method for depicting academic disciplines through Google Scholar Citations: The case of bibliometrics. Scientometrics, 114, 1251–1273.

Mohammed, A.J., Hasan, T.M., Mahmood, B. 2020. Citation networks Iraqi universities case study. In 2020 3rd International Conference on Engineering Technology and its Applications (IICETA), 41–46.

Orduna-Malea, E., Ayllón, J.M., Martín-Martín, A., Delgado López-Cózar, E. 2015. Methods for estimating the size of Google Scholar. Scientometrics, 104, 931–949.

Osipov, D. 2019. Development of a MongoDB-connected VR application. [Bachelor's thesis, TURKU University of Applied Sciences].

Ramirez-Gallego, S., Mourino-Talin, H., Martinez-Rego, D., Bolon-Canedo, V., Benitez, J.M., Alonso-Betanzos, A., Herrera, F. 2018. An information theory-based feature selection framework for big data under Apache spark. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 48, 1441–1453.

Shaikh, Z.A., Khan, A.A., Teng, L., Wagan, A.A., Laghari, A.A. 2022. BIoMT modular infrastructure: The recent challenges, issues, and limitations in blockchain hyperledger-enabled e-healthcare application. Wireless Communications and Mobile Computing, 2022, 1–14.

Sivarajah, U., Kamal, M.M., Irani, Z., Weerakkody, V. 2017. Critical analysis of big data challenges and analytical methods. Journal of Business Research, 70, 263–286.

Sultan, N.A., Mahmood, B., Thanoon, K.H., Khadhim, D.S. 2020. Network centralities-based approach for evaluating interdisciplinary collaboration. In 6th International Engineering Conference "Sustainable Technology and Development"(IEC), 216–221.

Sun, Y., Yin, S., Li, H., Teng, L., Karim, S. 2019. GPOGC: Gaussian pigeon-oriented graph clustering algorithm for social networks cluster. IEEE Access, 7, 99254–99262.

Tomasini, M., Menezes, R. 2015. Estimating memory requirements in wireless sensor networks using social tie strengths. In IEEE 40th Local Computer Networks Conference Workshops, 695–698.

Yu, J., Li, H., Liu, D. 2020. Modified immune evolutionary algorithm for medical data clustering and feature extraction under cloud computing environment. Journal of Healthcare Engineering, 2020, 1–12.