Graph Analysis of Citation and Co-authorship Networks of Egyptian Authors

Mariam Ayman*, Sohaila Kandil*, Youssef El-harty*, Alaa Moheb*, Ahmed Abdullah*, Omar Wassim*
*Department of Computer Engineering, Egypt-Japan University of Science and Technology, Alexandria, Egypt.

***Abstract*—The study conducts a comprehensive analysis of citation networks focusing on publications by authors affiliated with Egyptian institutions. Leveraging the Semantic Scholar platform and its API, a citation network ana a co-authoship network graphs are constructed to visualize the interconnections among these publications and their authors mainly using the python package for graph analysis (Networkx). The primary objective is to identify influential Egyptian publications and assess the centrality of nodes within the citation network. Through meticulous data collection including web scraping techniques, we obtained a cleaned dataset comprising publications by authored affiliated with Egyptian institutions. The analysis addresses challenges related to data quality, technical intricacies, and time constraints, resulting in a reliable and robust dataset. The findings provide valuable information on the impact of Egyptian publications, offering insights into the scholarly influence of authors associated with Egyptian institutions. This research equips researchers and academics interested in evaluating the impact of Egyptian publications with valuable data for future studies, collaborations, and policy decisions.**

***Index Terms*—graph analysis , citation network , co-authorship network , authorship pattern , network metrics , Networkx**

I. Introduction

Citation networks play a crucial role in assessing the impact and influence of scholarly publications within specific domains. Our study focuses on conducting a comprehensive analysis of citation networks for authors affiliated with Egyptian institutions. The primary objective is to identify influential publications and evaluate the centrality of nodes within these networks.

Egyptian publications have made significant contributions to advancing knowledge across various fields, making it essential to understand their impact. By analyzing the citation networks of Egyptian authors, we aim to uncover influential publications and explore the connections that exist among them. This analysis will enable us to identify articles that have garnered a substantial number of citations, indicating their widespread influence within the scholarly community.

To achieve our goals, we utilized the Semantic Scholar platform and its API to gather data on publications authored by individuals affiliated with Egyptian institutions. We constructed a citation network graph to visually represent the interconnections among these publications. Each article or paper is represented as a node, and citation relationships are depicted as edges.

The findings of this research provide valuable insights into the scholarly influence of authors associated with Egyptian institutions. By identifying influential publications and assessing the centrality of nodes within the citation networks, we gain a deeper understanding of their impact. Additionally, the analysis sheds light on authorship patterns and co-authorship relationships. Moreover, it reveals certain behaviors adopted by authors to enhance their scholarly reputation.

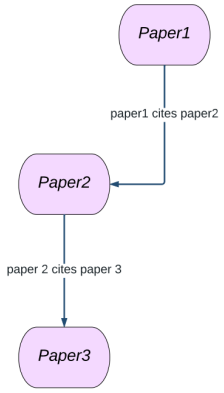

Fig. 1. Sample of citation network

II. Citation Network

A citation network is a valuable tool for studying the relationships and interconnections among academic papers. In this network, each paper or article is represented as a node, and a directed edge exists between two nodes if one paper cites another. This allows us to visualize the flow of information and references within the scholarly community.

By analyzing the citation network, we can gain insights into the influence and impact of specific papers. Nodes that receive a significant number of incoming citations are considered influential, as they indicate the widespread recognition and relevance of the research. These influential nodes play a crucial role in shaping the academic discourse and guiding future studies.

In Figure 1, we provide a sample visualization of a citation network. Each node represents a paper, and the directed edges depict the citation relationships between them. This network provides a visual representation of how papers are interconnected through citations, forming a web of knowledge.

III. Co-authorship Network

A co-authorship network is a type of social network that focuses on the collaboration relationships between authors. In this network, each author is represented as a node, and an edge

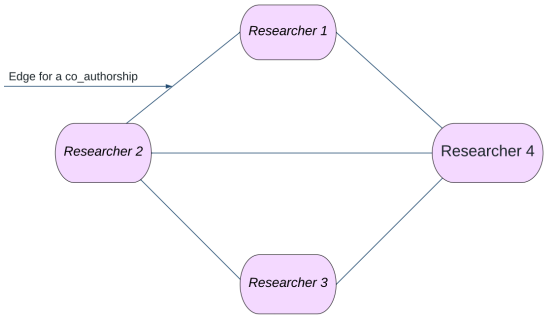

Fig. 2. Sample of co-authorship network

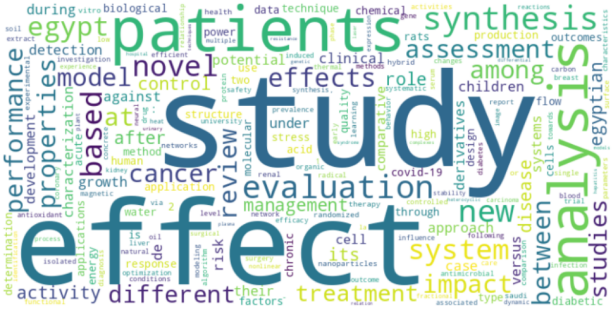

Fig. 3. A word cloud for common words in the papers' titles

connects two authors if they have collaborated on one or more publications together.

The co-authorship network allows us to understand the patterns of collaboration within the scholarly community. By examining the connections between authors, we can identify clusters or communities of researchers who frequently work together. These collaborations often indicate shared research interests, expertise, or institutional affiliations.

In Figure 2, we present a visualization of a co-authorship network. Each node represents an author, and the edges represent co-authorship relationships between authors who have collaborated on publications. The weight of an edge can be interpreted as the number of times two authors have co-authored publications together, providing a measure of the strength of their collaboration.

Studying the co-authorship network helps us identify influential researchers and research groups who have established extensive collaborative networks. It also provides insights into the dynamics of knowledge production, as authors with strong collaborative ties may have a higher likelihood of producing impactful research.

Both the citation network and co-authorship network offer valuable perspectives on the scholarly landscape associated with Egyptian institutions. By analyzing these networks and assessing various network metrics using tools like Networkx, we can gain a deeper understanding of the impact, influence, and collaboration patterns within the academic community. These insights are essential for researchers, academics, and policymakers interested in evaluating the scholarly contributions of Egyptian publications and fostering future collaborations and policy decisions.

IV. Literature Review

V. Umadevi [1] explores the Co-authorship networks of papers, which depict collaborations among researchers involved in scholarly publications. In this context, the Co-authorship network is established by representing researchers as nodes, while connections between nodes (edges) symbolize collaborative efforts between researchers. To facilitate the visualization and analysis of these networks, the study employs the Gephi tool, a widely utilized software specifically designed for network visualization and exploration.

A paper puplished by Azimjonov and Alikhanov [2] present a methodology for collecting metadata from papers, including titles, abstracts, keywords, body text, conclusion, and references. They utilized targeted keyword searches and manual extraction processes to extract the required metadata. To enhance the extraction process, they utilized Java programming language and Optical Character Recognition (OCR) libraries.

Our study improves upon this approach by employing web scraping techniques on the Google Scholar website to collect data of Egyptian authors. Additionally, we utilize the Semantic Scholar API to retrieve a comprehensive dataset of papers associated with each researcher identified through web scraping.

V. Methodology

A. Data Collection

During the data collection phase, we acquired our dataset from the Semantic Scholar website, where so many papers are puplished. Subsequently, we meticulously stored the gathered data in a CSV (Comma-Separated Values) file.

In our study, we obtained a dataset consisting of 31,508 research papers. Each paper in the dataset includes at least one researcher who is either from Egypt or affiliated with an Egyptian university. In our dataset, each paper includes important information such as its title, unique paper ID in semantic scholar, publication year, and a list of authors with their respective names and IDs in semantic scholar. Furthermore, we collected the number of references cited in each paper and added a list of all the references, including their titles and corresponding IDs for each paper in the dataset.

To obtain our dataset, we utilized the Semantic Scholar website database by employing API requests with our private Semantic Scholar API key. By following the documented API requests provided by Semantic Scholar [3], we accessed their database, and retrieved information such as paper titles, IDs, references, and authors' details. However, not all papers available on the Semantic Scholar website included author affiliations. Therefore, when requesting paper affiliations from Semantic Scholar, a substantial number of papers did not provide any affiliation information. Considering that the percentage of Egyptian papers among those published on Semantic Scholar is very small, it became inefficient to retrieve

all papers' data from Semantic Scholar and then filter out the non-Egyptian papers. This approach would not be optimal, particularly when aiming to gather data from approximately 30,000 papers. To address this challenge, we devised a two-stage process for collecting the dataset. The first stage involved gathering information about Egyptian researchers and those affiliated with Egyptian universities. Subsequently, we utilized the Semantic Scholar API to retrieve the papers published by these researchers.

We performed web scraping on the Google Scholar website to obtain a dataset comprising 13,027 entries that includes author names and affiliations. To accomplish this, we utilized the Selenium library in Python. To maintain adherence to the scraping guidelines and prevent exceeding the data scraping rate imposed by Google Scholar, we implemented a structured approach. Specifically, we collected data for each university in a single session, ensuring sufficient time intervals between each scraping instance. The data collection process focused on researchers from prominent Egyptian universities such as Ain Shams University, Cairo University, EJUST university, Alexandria University, Banha University, Assiut University, and Zewail University.

To collect the dataset, we utilized the Semantic Scholar API to request the data of each author identified in the initial stage. These requests enabled us to collect a dataset containing the titles, publication IDs, and publication times of their respective papers. Additionally, we used further API requests with each paper's ID to retrieve information regarding its references, number of citations, and the names of all authors associated with the publication. We gave our collected datset a unique name "AlGoNet" to be referenced afterwards.

B. Data Cleaning

Data cleaning was performed in two stages. Initially, during the data collection process, we implemented a code segment to check for duplicate entries before adding them to the CSV file. Subsequently, we employed the Pandas library in Python to eliminate duplicates and empty fields from the dataset, ensuring its integrity and consistency.

C. Analysis of Citation Network

1) Temporal Analysis of Publications: Analysing our dataset to get insight about the publication trend, we found the oldest paper was published in 1903 while the most recent is in 2023. As shown from the trend in figure 3, most of the publications were recently published within the years from 2000 to 2023.

2) Degree Distribution Analysis: The degree distribution shows how the number of nodes in the network is distributed across different degrees where the degree of a node represents the number of citations it received from other nodes in the network. As shown in figure 4, the majority of nodes in the network have a low degree (less than 50) which indicates that most publications by Egyptian authors have a relatively low number of citations. However, the long tail in the distribution represents the small number of highly cited publications that have a much higher degree. The degree distribution also appears to follow a power law distribution, with a few highly connected nodes (publications) and many nodes with few connections. This pattern is often observed in citation networks and is known as the "rich get richer" phenomenon, where highly cited publications tend to receive more citations over time, leading to a skewed distribution of citations. Overall, the degree distribution analysis suggests that there are a few highly influential publications by Egyptian authors that have received a large number of citations, while the majority of publications have received relatively few citations.

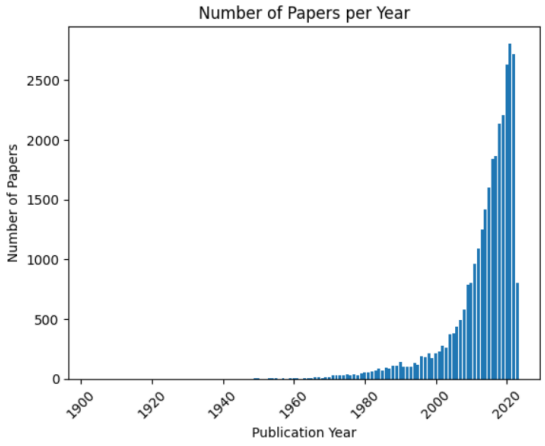

Fig. 4. Temporal Analysis

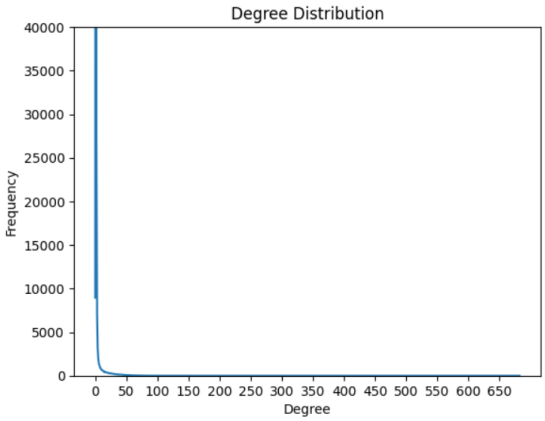

Fig. 5. Degree Distribution Analysis

3) Authorship Pattern Analysis: It is evident from figure 5 that out of all the included papers (30905 papers), four-authored papers (6115 papers) are little ahead than three-authored papers (5640 papers) followed by two-authored papers (3964 papers), while single-authored papers (2856 articles) are at the back foot. The remaining papers having more than 4 authors (12330 papers) are the majority keeping in mind that the limit of the API used in collecting the dataset was 500 authors per paper. Therefore, although the maximum author count is 500 authors for a total of 46 papers, those papers may include more than 500 authors. Hence, it is inferred that,

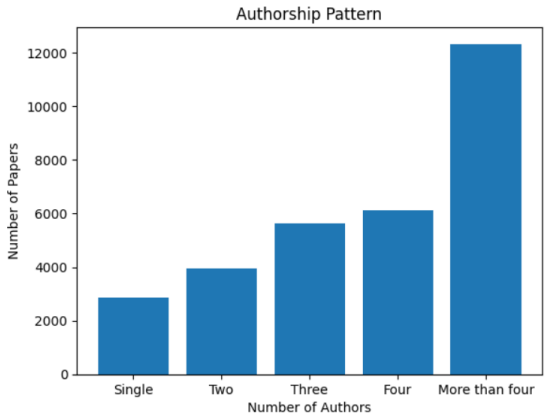

Fig. 6. Authorship Pattern Analysis

the trend of collaborative research has taken place among the Egyptian papers of this collected dataset.

4) Clustering Coefficients: The network has a low average clustering coefficient (= 0.019107480538823315) which suggests that the nodes in this citation network have relatively few connections to their neighbors, and there is a lack of tightly connected communities or clusters within the network. This could indicate that the citations between papers in the network are more random or sporadic, rather than forming cohesive groups or neighborhoods of interconnected papers. Another potential implication is that the papers in the network may have diverse topics or have different research areas, resulting in less clustering of citations.

5) Density Analysis: The density value of the graph is also low (= 3.5501576199356724e-06) which suggests that the number of actual connections (citations) in the network is relatively small compared to the maximum possible number of connections. This implies that there are many potential connections that are not present in the network, resulting in a sparse or sparsely connected graph. It may indicate that the papers in your citation network have relatively few citations or that the citation relationships are not densely interconnected. This could be due to various reasons, such as:

- A large number of papers with limited or no citations.

- A diverse range of topics within the network leading to fewer direct connections.

- The network size itself can affect the density as larger networks tend to have lower density values. This is because the number of possible connections increases exponentially compared to the actual connections.

6) Centrality Measures: Computing the degree centrality measure for all the included publications, we got an average degree of 2.8 with the highest degree (=682) for the paper (volumes 357â€“377 (1995)). On the other hand, approximately two-thirds of the publications have a degree of approximately zero, indicating very low connectivity or lack of connections with other papers in the network. For further analysis, we computed degree and eigenvector centrality measures for recent publications published after 2015.

*a) **Degree Centrality for Publications (>2015)**:* Degree centrality measures the number of direct connections a node has, i.e., the number of citations a paper has. It provides a measure of how influential a paper is based on the number of direct references it receives. The higest degree obtained (=0.0015) was for (managing gsh elevation and hypoxia to overcome resistance of cancer therapies using functionalized nanocarriers) paper published in 2021.

*b) **Eigenvector Centrality for Publications (>2015)**:* Eigenvector centrality considers both the number of connections a node has and the centrality of its neighbors. It assigns higher importance to nodes that are connected to other highly central nodes so it takes into account the global structure of the network, whereas degree centrality focuses solely on the immediate connections of each node. Simply, it captures the notion that being connected to influential nodes contributes to the overall centrality of a node, which may differ from the simple count of connections provided by degree centrality. The highest measure obtained from the analysis (=0.314) was for (efficacy and biological correlates of response in a phase ii study of venetoclax monotherapy in patients with acute myelogenous leukemia. Eigenvector Centrality) published in 2016.

7) Strongly Connected Components: As known from the study of strongly connected components, every node (representing a paper) is in precisely one strongly connected component since the equivalence classes partition the set of nodes. In our dataset, there are those individual papers that may be without any citations, or may having no references, or even may have references for other papers that are not included in the dataset. These papers are included in the total number of strongly connected components (=320885).

On the other hand, it is shown that we have 62 strongly connected components of size ¿ 1, which is unreasonable as there must be only one directed edge between any two papers (from the paper to be published to the paper already published). However, it is revealed after analyzing the output that there are 2 cases for that situation:

Case 1: There are some academic journals that are published in a set of supplement articles. These supplements may contain special articles, conference proceedings, abstracts, or other supplementary material that is related to the main journal's focus. Each article can cite any one of the others or all of them since they are considered separate and a supplement number is what differentiate them as for example:

In the citation "Eur Spine J. 2018 Sep;27(Suppl 6)," , here "Suppl 6" indicates that it is supplement number 6. From component number (8462) in the code output, we can see such case in the existing 3 papers that are part of 6 supplements of the same journal: 1- Title: the global spine care initiative: world spine care executive summary on reducing spine-related disability in low- and middle-income communities. 2- Title: the global spine care initiative: methodology, contributors, and disclosures. 3- Title: the global spine care initiative: model of care and implementation

Case 2: There is an adopted behaviour by some authors

that may seem illegal where authors cite papers, may be done manually, that are not published yet. Those papers are almost having the same publication year and month and some common authors. The target of doing so may be the desire of increasing the number of citations for both papers and consequently increasing the authors' ranks.

D. Analysis of Co-authorship Network

From the analysis, we got insight about the pairs of authors that are frequently cited together (edges in the network) where the weights of the edges in represent the number of times a pair of authors have collaborated or co-authored publications together. These insights include:

-Most Frequently Collaborating Pair: The pair of authors (A. Hussein and K. Guru) has the highest collaboration frequency, indicated by the maximum weight of 155. This suggests that they have collaborated on a significant number of publications, possibly indicating a strong research partnership or shared research interests.

-Least Frequently Collaborating Pairs: Within the top 50 pairs, there are three pairs with a minimum weight of 39. These pairs are (A. Soliman and S. Di Maio), (H. Elmansy and A. Kotb), and (M. Nagappan and A. Hassan). Their relatively lower collaboration frequencies compared to other highly collaborating pairs, indicating fewer instances of joint publications or research collaboration between them.

1) Degree Centrality Measures: The analysis of the co-authorship network revealed the top 50 authors with the highest degree centrality values, including: D. Nepogodiev, J. Glasbey, A. Bhangu, T. Drake, and M. Saad. These authors have extensive collaborations and strong connections within the network, indicating their prominent role in the research community. Their high degree centrality values suggest their influence in disseminating knowledge and their potential for fostering collaborations. They provide valuable collaboration opportunities and contribute to the overall connectivity and impact of the network. Their work is likely to have a significant influence on the field.

2) Connected Components Analysis: It revealed a total of 286 connected components. These components represent groups of authors who have collaborated with each other, forming cohesive subnetworks within the larger co-authorship network. Each connected component represents a distinct community of authors with strong collaborative ties. The presence of multiple connected components indicates the presence of various research communities or subfields within the overall research domain. By identifying those connected components, we gain insights into the structure and organization of collaborations as well as the clusters of authors who frequently collaborate and potentially uncovering specialized research areas within the broader field.

VI. Conclusion

In this paper, the analysis of the citation network provided insights into the temporal trends of publications, degree distribution, authorship patterns, clustering coefficients, and density. The majority of publications were recently published, indicating a growing research output. The degree distribution followed a power law distribution, with a few highly cited papers and a majority with fewer citations. Collaborative research was observed, with multi-authored papers being the most common. In addition, the analysis of the co-authorship network revealed pairs of authors who frequently collaborate, indicating strong research partnerships. Degree centrality measures identified the top 50 authors with the highest collaboration frequencies, highlighting their influential roles in the research community. Connected components analysis uncovered 286 distinct communities of authors with strong collaborative ties, indicating the presence of various research communities within the broader field.

Overall, our analysis provides valuable insights into the publication and collaboration patterns of Egyptian researchers. It highlights the impact of influential publications, the significance of collaborative research, and the presence of diverse research communities within the Egyptian research landscape. These findings can contribute to a better understanding of the research dynamics in Egypt and facilitate future collaborations and knowledge dissemination within the scientific community.

Acknowledgements

We would like to express our deepest gratitude to Eng. Zeyad Shokry for his invaluable guidance and assistance with the collection of our dataset (AlGoNet).

References

[1] V. Umadevi, "Case study–centrality measure analysis on co-authorship network," *Journal of Global Research in Computer Science*, vol. 4, no. 1, pp. 67–70, 2013.

[2] J. Azimjonov and J. Alikhanov, "Rule based metadata extraction framework from academic articles," *CoRR*, vol. abs/1807.09009, 2018. [Online]. Available: http://arxiv.org/abs/1807.09009

[3] Semantic Scholar, "Semantic scholar documentation," Online, Accessed 2023. [Online]. Available: https://api.semanticscholar.org/corpus/