A Survey on Scientific Collaboration Networks (Case Study: Morocco)

R. Haffadi, F.E. Senhaji, A. Benhiba

December 25, 2020

Abstract

Science is the key of human evolution and survival and it's leaving a golden era with a continuously growing number of scientific research in different fields and disciplines.For their mutual benefits researchers often tend to collaborate with each other.Studying the co-authorship network has been the subject of intense interest in recent years because analysing this type of network can help to determine the main areas of specialization of universities and research centres, as well as the type of internal and external collaborations of their researchers.In this paper we be investigating bibliometric data of papers published in journals through the use of the Scopus Database API Interface. Then analyse it using graph visualization software and statistical tools. After getting familiarize with the data the next step would be building the network and analysing it using measures of social network analysis in order to explore different behavioural patterns of co-authorship networks.

1 Introduction

Scientific collaborations are defined as "interactions taking place within a social context among two or more scientists that facilitate the sharing of meaning and completion of tasks with respect to a mutually shared, super-ordinated goal" [1], those collaborations frequently emerge from, and are perpetuated through, social networks. Since social networks may span disciplinary, institutional, and national boundaries, it can influence collaboration in multiple ways [1]. Social network analysis has produced many results concerning social influence, social groupings, inequality, disease propagation, communication of information, and indeed almost every topic that has interested 20th century sociology [2].

Before the big data era scientific collaborations between scholars was an important subject. After the emerging of new technologies like Big Data analysis the topic become more and more important with the increasing necessity for new breakthroughs in different disciplines like Physics, Biology and others.

By jointly publishing a paper, researchers show their knowledge-sharing activities, which are essential for knowledge creation and help to set fairness in the scientific community. For example skilled researchers from around the world who have no access to lager amount of resources can benefit a lot from experienced researchers and resources from big laboratories and organizations. Scientific collaboration has even been called a "springboard for economic prosperity and sustainable development" [3]. As most scientific output is a result of group work and most research projects are too large for an individual researcher to perform.

Social networks has been utilized a lot in different domains. They play a critical role in determining the way problems are solved, institutions are run, markets evolve, and the degree to which individuals succeed in achieving their goals [4]. Social networks have been analyzed to identify areas of strengths and weaknesses within and among research institutions, businesses, and nations as well as to direct scientific development and funding policies [5].

Not only the application of network statistics is useful in characterizing the nature of scientist networks, it also provides a powerful tool to study and predict scientific performance such as productivity or research impact. A study on the effects of co-authorship on the performance of scholars using regression model and social network analysis showed that researchers who have a strong connection to only one co-author among a group of connected co-authors perform better than those who have many connections to the same group. The study also suggests it is possible to use professional social network of researchers to predict future performance [6].

2 Literature Review

Last few decades scientists and researchers are trying to establish procedures and metrics for evaluating journal quality and the scientific output of researchers based on journal and author metrics ([7]; [8]; [9]; [10]).

Metrics used to evaluate the journal performance are often based on the journal impact factor, which is measured using Thomson's Journal Citation Reports (JCR), Scimago Journal Country Rank (SJR), etc.Evaluating authors is based on metrics such as the total number of citations [11], as well as the h-index [12], or variations of it [13], [14]

In recent years, Social Network Analysis (SNA) has become one of the go to technologies for studying social networks structures through the use of graph theory [15]. [6] presented a theoretical model that applied measures often used in SNA to explore collaboration (co-authorship) networks of scholars.[16] presented a review about the status of the research in SNA by analyzing a large number of papers in terms of institutional and individual contribution, citations, topic coverage, etc. of several American, British and Australian universities. In a recent paper, [17] analyzed the intensity of collaboration within a German university using SNA.

3 Data Specifications

Nowadays there is several scientific databases some of them are open access (e.g., Google Scholar), while others are accesses thanks to some subscriptions made by some public or private organisations which are the affiliations of the authors (e.g., Web of Science, Scopus [18]). [19] have analysed the characteristics of scientific databases, such as scientific databases (e.g: Scopus), search engines (e.g: Google Scholar) and social networks (e.g: academia.edu or ResearchGate).Certain studies have been devoted to investigate the differences, advantages and disadvantages of different types of scientific databases. For example [20].These problems are minimized in literature databases whose data sources are editorial, institutions, etc [19]. Here comes popular literature scientific databases such as Web Of Science and Scopus

In our case we will be working on Scopus database who is considered reliable due to the accuracy of it bibliometric information.In addition according to[21] Scopus has over 25,100 titles from more than 5,000 international publishers (including 5,500 full open access journals and Over 9.8 million conference papers) with different source types, serial publications (journals, trade journals,book series and conference materials) and Non-serial sources. Scopus also offers a specific coverage of metadata consisting of document types, Keywords and index terms added to improve search recall.Furthermore,Scopus includes cited references, affiliation data (more than 70,000 affiliation profiles), as well as Author profiles (16,000) which is particularly relevant for analyzing citation metrics for authors, as well as specific articles by an author. The data can also be used to find authors or reviewers to collaborate with or for hiring purposes.In fact The Scopus Author Identifier assigns each author in Scopus a unique number and groups together all of the documents written by that author. To determine which author names should be grouped together under a single identifier number, the Scopus Author Identifier uses an algorithm that matches author names based on their affiliation, address, subject area, source title, dates of publication citations and co-authors.One more relevant technique used overcome the author id ambiguity is the ORCID (Open Researcher and Contributor Identifier) a nonprofit organization dedicated to solving the name ambiguity problem in scholarly research by assigning a unique identifier to each author. From their Scopus Author Profile, authors can import their list of publications in Scopus and their Scopus Author Identifier into ORCID. Once an author connects their ORCID record with their Scopus profile, a link to their ORCID record will appear on their profile page. Scopus and ORCID share and sync their data on a monthly basis.

The data used in this work has been downloaded from Scopus dataset. Using the scopus engine to filtrate Moroccan publications in the period of early 2010 to April 2020.

4 Exploratory Data Analysis

After extracting the data, we tried to explore it in order to get acquainted with it.So, our data consist of 22.000 rows with 10 attributes (Affiliations,Author(s) ID, Authors, Authors with affiliations, Cited by, Document Type, Source, Source title, Title, Year). Affiliations contains the full specification of the affiliation; her name,university, city and country.Author(s) ID contains the list of IDs of the author(s),Authors with affiliations design the name of the author with the affiliation/ lab to which he belongs ,the attribute Cited by shows the number of times the paper was cited by others this index will be useful for metrics measurements ,Source title specifies if the paper is an article or a conference paper, Title formulate the full name of the paper.The following figure gives an overview:

Affiliations	Author(s) ID	Authors	Authors with affiliations	Cited by	Document Type	Source	Source title	Title	Year
Systems' Architecture Team, Ensem, Hassan II U...	36701835800;24725091300;	Moutaouakkil F., Medromi H.	Moutaouakkil, F., Systems' Architecture Team, ...	1.0	Conference Paper	Scopus	ICCTD 2010 - 2010 2nd International Conference...	Control robot by multiagent control architecture	2010
Laboratoire de Chimie Physique Générale, Facul...	36911796000;24823985800;6602908622;7004122780;...	Saoiabi S., El Asri S., Laghzizil A., Coradin ...	Saoiabi, S., Laboratoire de Chimie Physique Gé...	18.0	Article	Scopus	Materials Letters	Nanoporous surface of organofunctionalized hyd...	2010
University Caddy Ayyad, Bd. Prince My Abdellah...	6602701872;35760797400;35760820700;7006304181;	Mesquine F., Bakka O., El Bahja H., Vega P.	Mesquine, F., University Caddy Ayyad, Bd. Prin...	5.0	Conference Paper	Scopus	Proceedings of the 15th IEEE International Con...	Observer based regulator problem for WWTP with...	2010

Figure 1: data overview

5 Collaboration Network Analyses

5.1 Histogram of Number of Documents per Countries

For the first analytical aspect, we choose to explore the number of produced papers per number of collaborating countries. Figure 2 shows the more the number of contributing countries increases the more the number of documents drop by a significant amount, that means that big general studies are very little and hard to manage. In our case the average number of collaborating countries after dropping the outliers -very strange values that affect greatly the results, due to the noise in data- is 1.65 approximately 2 countries. So, we can conclude that the most popular type of collaboration is between two countries.

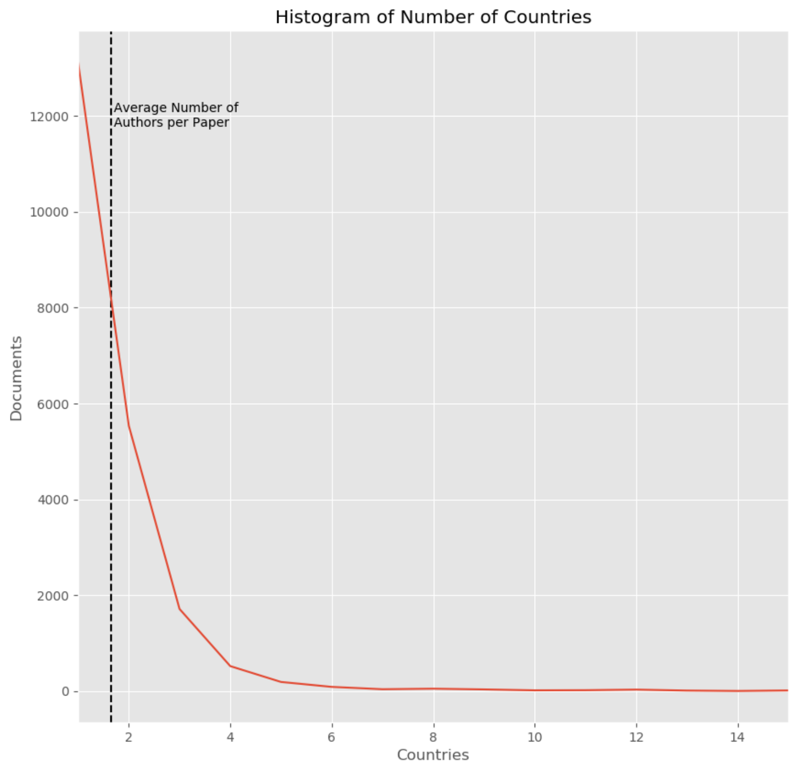


Figure 2: documents count in function of number of collaborating countries

5.2 Histogram of Number of Documents per authors

Figure 3 is representing the histogram of the number of authors in function of there publications, it shows that the distribution of documents per authors can be approximated by a normal distribution.

The average number of co-authoring is 4.79 and we can notice that when the number of authors increase the number of publications in the other way decrease by a lot, our explication to that is because when the number of authors is significantly high the difficulty of management increase dramatically. The largest number of publications generally have 2 to 3 authors which is higher than the number of documents per a single author.this shows that researchers often opt for creating teams of two to three members often one of them is supervisor and the others are junior researchers.

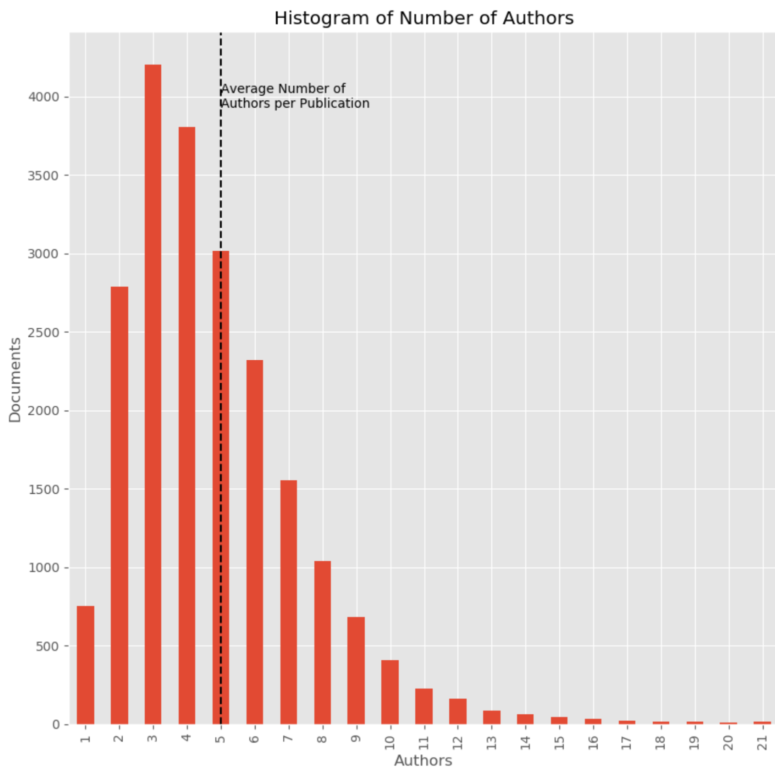


Figure 3: Histogram of the number of authors and their publication

5.3 Histogram of Number of Documents per affiliations

To describe a collaboration network we elaborate on two main concepts. The first one is inter-institutional collaboration -collaboration between institutions- which gives us information about how the institutions of different countries collaborate. The second is international collaboration -collaboration between institutions from different countries- which give information about how international projects are organized.

Figure 4 is a pie chart describing the percentage of the number of total documents in function of the number of collaborating affiliations. As we can see over 59% of the scientific papers are published by 1 affiliation, whilst 25.2% are published by the collaboration of two affiliation, 7.8% are due to the collaboration of 3 affiliations.The rest of the percentages are considered negligible.

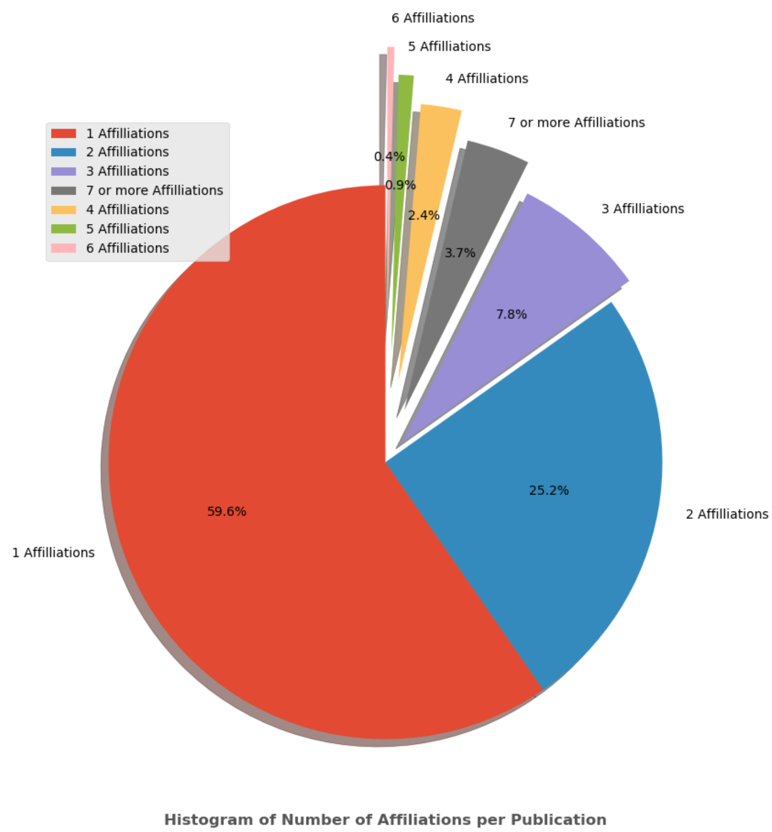


Figure 4: Histogram of the number of collaborating affiliations and their publication

5.4 International Collaboration Between Morocco and Other Countries

One way of describing international collaboration is by calculating the number of publications with one or more collaborative countries -with Morocco- and see how is changing over time.

Figure 5 represent the sum of Moroccan international scientific papers grouped by years. We can observe a kind of stability over the years.The number oscillates generally between 700 and 900 with a minimal values 731 in 2014 and the biggest number was 988 in 2015.

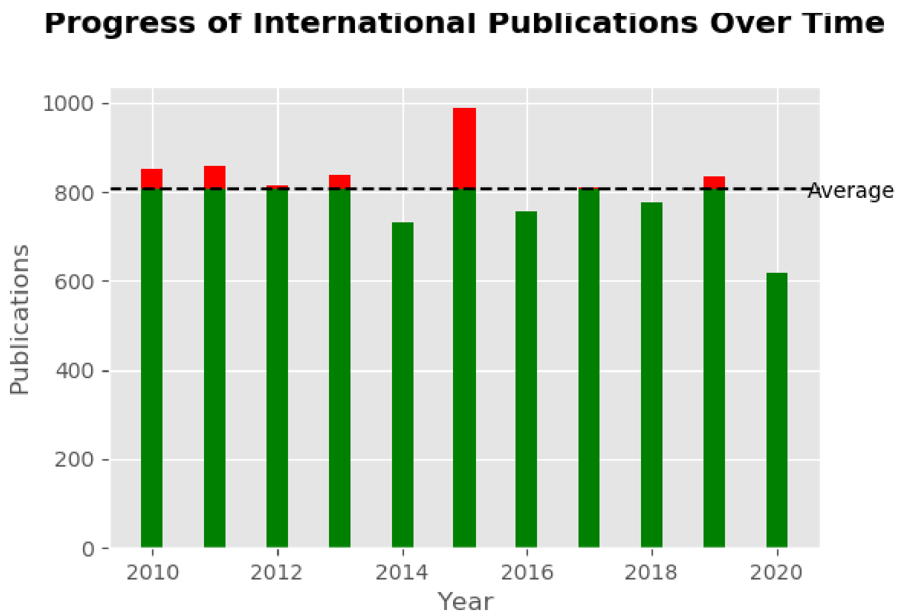


Figure 5: tabulation of the sums of international collaboration of Morocco by year

Another analytical aspect that we were interested in is the top most countries collaborating with morocco.The majority of countries were European countries in addition to the united states and china. France was the first position with remarkable gap as we can see Morocco and France has a total of 4568 document which reflect the strongest of scientific collaboration boundary.

Figure 7 shows the top collaborative authors in this field of study.The first column represent the Author-ID, the second attribute represent the sum of the published papers of each author, and the last attribute contain the name of the authors.

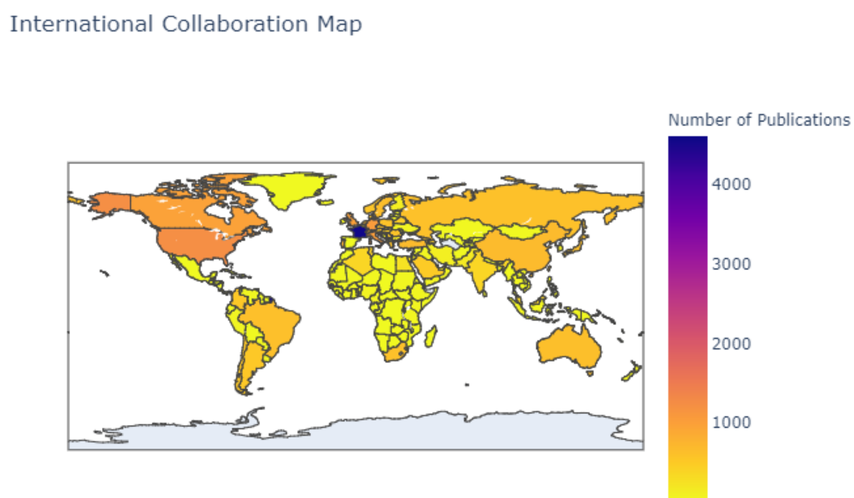


Figure 6: top 10 countries collaborating with Morocco

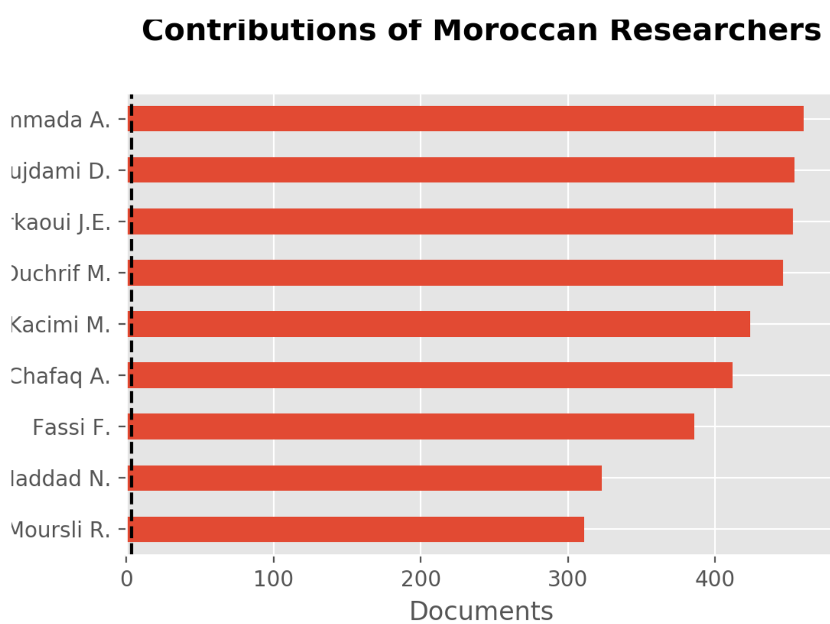


Figure 7: top 10 collaborative Moroccan authors

6 Graph analyses

Social network analysis is the study of social entities (actors) and their interactions and relationships. The interaction and relationships are represented as a graph, where each node represents an actor (user), and the edge between two nodes represents their relationship. In our work, we employ social network analysis metrics such as centrality, betweenness... in order to identify the "influential" actors in a social network, in terms of their position in the graph and their connections/interactions with other users.In this context, actors collaborate with the actors they trust and are influenced by their opinions. Moreover, trust and influence are reinforced for certain actors in the circle of trust and decrease for others.

Community detection One of the most relevant features of graphs representing real systems is community structure, or clustering, i. e. the organization of vertices in clusters, with many edges joining vertices of the same cluster and comparatively few edges joining vertices of different clusters. Such clusters, or communities, can be considered as fairly independent compartments of a graph[22]. Detecting communities

is of great importance in sociology, biology and computer science, disciplines where systems are often represented as graphs. This problem is very hard and not yet satisfactorily solved, despite the huge effort of a large interdisciplinary community of scientists working on it over the past few years.The aim of community detection in graphs is to identify the modules and, possibly, their hierarchical organization, by only using the information encoded in the graph topology.

6.1 Social Network Analysis Metrics

Centrality The three centrality metrics, namely degree, closeness, and betweenness centrality, identify "key" users of the graph, in terms of information dissemination. Let n denote the size of the graph (i.e. the number of actors/users).

DC σ_D represents the simplest CM and determines the number of direct contacts as an indicator of the quality of a network member's inter-connectedness [23]. Using the adjacency matrix $A = (a_{ij})$ it can be formalized as follows:

$$\sigma_D(x) = \sum_{i=1}^{n} a_{ix}$$

As a consequence, the centrality score $\sigma_D(x)$ for a node x is higher, the more contacts a node x has.the major disadvantage of DC is that indirect contacts are not considered at all. Therefore, a reduction of the distance from one actor x to another actor y resulting from an additional relationship in most cases does not increase the value of the CM.The intensification of a connection of shortest length between x and y does also not increase the value of this CM, since DC only considers direct contacts. However, in an undirected, unweighted graph a direct connection between the actors x and y can exist only once.Through a new relationship both actors involved win one additional direct contact. So the DC of both members equally increases by 1 and the ranking of the actors thus always remains the same.

Closeness Centrality CC σ_C is based on the idea that nodes with a short distance to other nodes can spread information very productively through the network[24]. In order to calculate the CC $\sigma_C(x)$ of a node x, the distances between the node x and all other nodes of the network are summed up by using the reciprocal value we achieve that the CC value increases when the distance to another node is reduced, i.e. when the integration into the network is improved. Formally, this means

$$\sigma_C(x) = \frac{1}{\sum_{i=1}^{n} d_G(x, i)}$$

For CC the reduction of the distance to at least one other actor when adding another relationship leads to a smaller value of the denominator. Consequently, in this case the CM value of the considered actor increases However,according to the formula only the distances between the different actors are taken into account. Therefore, a larger number of paths with shortest length between two actors does not positively affect the value of this CM.

Betweenness Centrality In case of BC σ_B a network member is considered to be well connected if he is located on as many of the shortest paths as possible between pairs of other nodes. The underlying assumption of this CM is that the interaction between two non-directly connected nodes x and y depends on the nodes between x and y. According to Freeman[25] the BC $\sigma_B(x)$ for a node x is therefore calculated as:

$$\sigma_B(x) = \sum_{i=1, i \neq x}^{n} \sum_{j=1, i<j, j \neq x}^{n} \frac{g_{ij}(x)}{g_{ij}}$$

with g_{ij} representing the number of shortest paths from node i to node j, and $g_{ij}(x)$ denoting the number of these paths which pass through the node x.

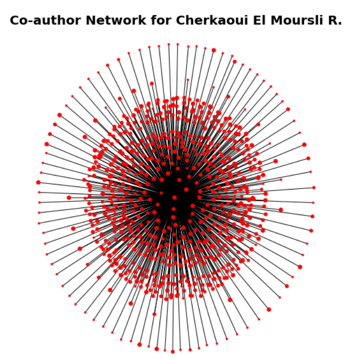
Centralization The general procedure involved in any measure of graph centralization is to look at the differences between the centrality scores of the most central point and those of all other points. Centralization, then, is the ratio of the actual sum of differences to the maximum possible sum of differences. The three different ways of operationalizing this general measure which Freeman discusses follow from the use of one or other of the three concepts of point centrality.

Modularity Modularity is one measure of the structure of networks or graphs [26]. It was designed to measure the strength of division of a network into modules (also called groups, clusters or communities). Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules. Modularity is often used in optimization methods for detecting community structure in networks. However, it has been shown that modularity suffers a resolution limit and, therefore, it is unable to detect small communities.

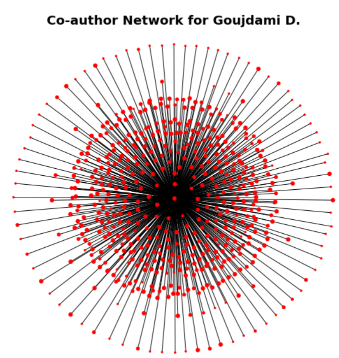
7 Results and Analysis

7.1 Examples of Co-authoring Networks

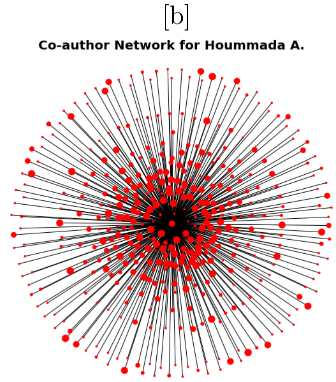
In this section we be presenting some results accompanied with there corresponding analyses.Let's start with co-authoring networks examples:



[a]



[b]



[c]

Figure 8: Network of Collaborators for Three Different Moroccan Researchers

To begin with figure 8 [a],[b] represent the network of co-authoring for Cherkaoui El Moursli R.and Goujdami D. respectively, we can say that this two authors have basically three circles of co-authorship reflecting the differ-

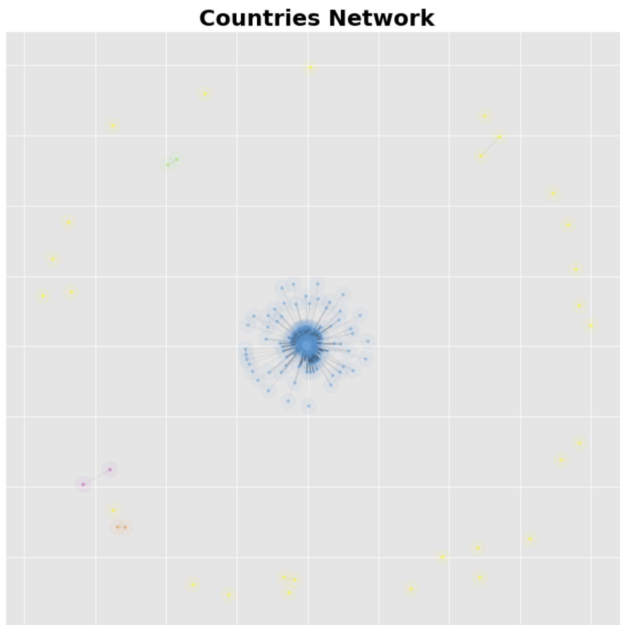


Figure 9: Countries Network

ent co-authoring frequency. The last author is Hoummada A. and we can see from the illustration the nodes are of a perceptible size and are mostly centred.This facts can means that the author is more likely to have a stable team of collaborators sharing information within there close entourage.

7.2 Countries Network

One regards the countries collaboration network we have found 30 community and we can notice that the centrality is high proving the fact that there is some central nodes (important nodes) that holds up the whole network and assure the maintenance of the connection and manage the information flow. When it comes to modularity factor the value is very small(2.37e-05) which makes sense because communities attempt to communicate and search for new collaboration opportunities.Finally we can tell that there is some countries that are interested in scientific collaboration research and are investing in it by providing to there searchers there required support and therefor opening the world before them.

7.3 National Networks

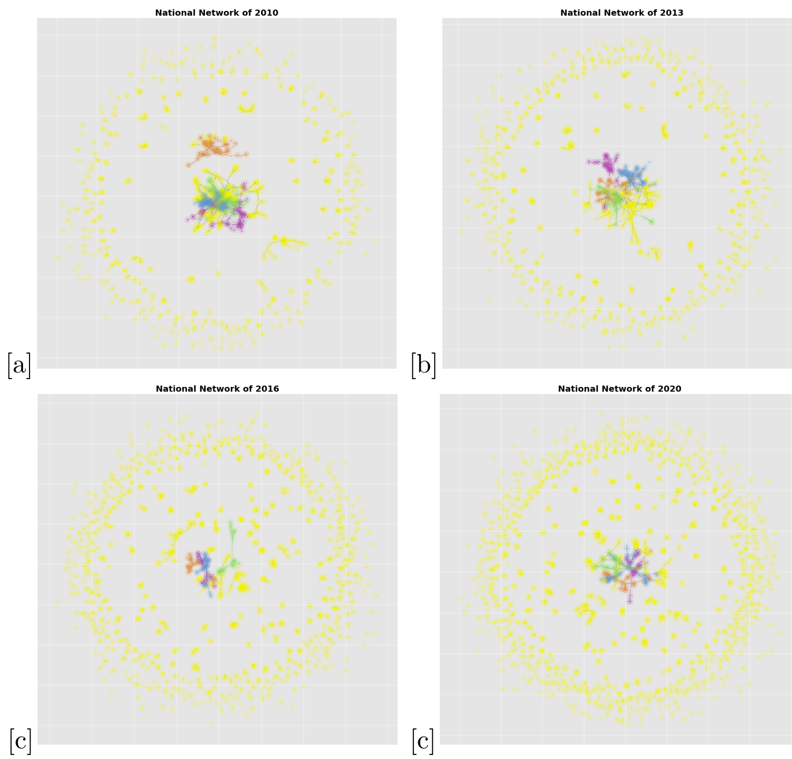


Figure 10: Network of Collaborators for The Moroccan Researchers Community

	Year	Authors	Edges	Density	Communities	Centralization	Modularity
0	2010.0	3110.0	11101.0	0.002296	285.0	0.023765	0.880119
1	2011.0	2958.0	9625.0	0.002201	321.0	0.021141	0.920913
2	2012.0	3083.0	9079.0	0.001911	346.0	0.024703	0.951247
3	2013.0	2968.0	8442.0	0.001917	369.0	0.026403	0.939488
4	2014.0	3695.0	12068.0	0.001768	346.0	0.019623	0.915539
5	2015.0	2893.0	7679.0	0.001836	417.0	0.011654	0.958989
6	2016.0	3365.0	8205.0	0.001450	503.0	0.015499	0.982610
7	2017.0	3315.0	8126.0	0.001479	514.0	0.012405	0.983467
8	2018.0	3443.0	8122.0	0.001371	538.0	0.010544	0.979421
9	2019.0	3419.0	8039.0	0.001376	569.0	0.010623	0.981763
10	2020.0	4016.0	9550.0	0.001185	633.0	0.007535	0.981969

Figure 11: Network Characteristics

As we can see in 11 although the number of authors wobbles a little bit over the years, the density is rather stationary. The number of detected communities varied by years with unique characteristics of Centralization Modularity.But in general we can notice that for example modularity is varying between 0.8 and 0.9 which can be explained by the strongness of the connections inside the communities while connections between different communities are rather soft and delicate. For the centralization factor it's low value can be interpreted by openness of the communities on each other on the national state.

7.4 International Networks

Comparing 13 with 11 We can note that the number of communities is almost the same over the years that can reflect the major affiliations internationally collaborating with other affiliations are probably the same and remain in the same community. One regarding the centralization factor we can see that it rather low which means that the information is distributed equally within the community.

8 Conclusion

Most scientometric studies are usually based on the use of journal metrics and author metrics, whereas an analysis of how large numbers of researchers collaborate is lacking.The characterization and understanding of collaboration benefits is a key precondition for a wide adoption of the collaborative networks paradigm in its various manifestation forms.In this paper we have used the Scopus Database to explore and analyse different aspect of scientific collaborations specifically Morocco case.This includes collaborations between authors,institutions, and countries over the last ten years. However more metrics

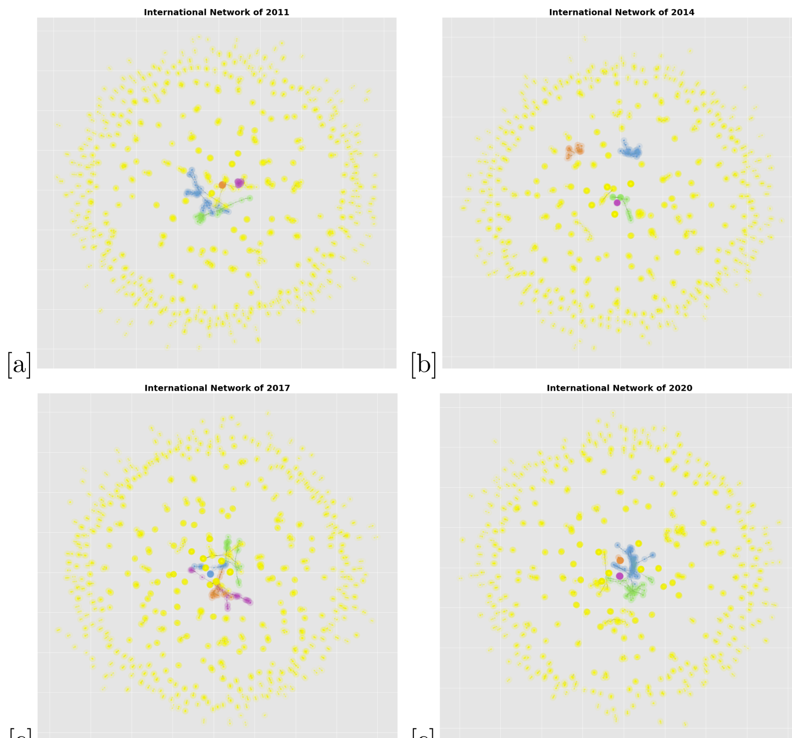


Figure 12: Network of Collaborators (International Only)

	Year	Authors	Edges	Density	Communities	Centralization
0	2010.0	3468.0	20051.0	0.003335	419.0	0.028689
1	2011.0	3610.0	16609.0	0.002550	411.0	0.018514
2	2012.0	3324.0	14807.0	0.002681	379.0	0.025012
3	2013.0	3850.0	18556.0	0.002504	416.0	0.028420
4	2014.0	3446.0	22651.0	0.003816	361.0	0.020863
5	2015.0	4614.0	35483.0	0.003334	429.0	0.019215
6	2016.0	3494.0	18052.0	0.002958	403.0	0.015941
7	2017.0	4329.0	38366.0	0.004095	409.0	0.021556
8	2018.0	4124.0	28987.0	0.003410	404.0	0.021577
9	2019.0	4796.0	46594.0	0.004052	429.0	0.017224
10	2020.0	3914.0	35183.0	0.004594	336.0	0.019689

Figure 13: International Network Characteristics

and analytical aspects could be developed to extract new patters and deeper insights.

"I think our work serves as a reminder that basic science preceded and influenced these translational breakthroughs through collaboration," the authors explain. "Public funding of basic research has many translational benefits; the inherently collaborative nature of scientific discovery leads to breakthroughs." [27].

References

[1] Diane H Sonnenwald. Scientific collaboration. *Annual review of information science and technology*, 41(1):643–681, 2007.

[2] Mark EJ Newman. Scientific collaboration networks. i. network construction and fundamental results. *Physical review E*, 64(1):016131, 2001.

[3]

[4] Alireza Abbasi and Jorn Altmann. On the correlation between research performance and social network analysis measures applied to research collaboration networks. In *2011 44th Hawaii International Conference on System Sciences*, pages 1–10. IEEE, 2011.

[5] Jason Owen-Smith, Massimo Riccaboni, Fabio Pammolli, and Walter W Powell. A comparison of us and european university-industry relations in the life sciences. *Management science*, 48(1):24–43, 2002.

[6] Alireza Abbasi, Jörn Altmann, and Liaquat Hossain. Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures. *Journal of Informetrics*, 5(4):594–607, 2011.

[7] Massimo Franceschet. Journal influence factors. *Journal of Informetrics*, 4(3):239–248, 2010.

[8] Peter Ingwersen. Scientific datasets: informetric characteristics and social utility metrics for biodiversity data sources. In *Library and Information Sciences*, pages 107–117. Springer, Berlin, Heidelberg, 2014.

[9] Sujay S Kaushal and Jonathan M Jeschke. Collegiality versus competition: how metrics shape scientific communities. *BioScience*, 63(3):155–156, 2013.

[10] Xuanyu Cao, Yan Chen, and KJ Ray Liu. A data analytic approach to quantifying scientific impact. *Journal of Informetrics*, 10(2):471–484, 2016.

[11] Boleslaw K Szymanski, Josep Lluis de la Rosa, and Mukkai Krishnamoorthy. An internet measure of the value of citations. *Information Sciences*, 185(1):18–31, 2012.

[12] Jorge E Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences*, 102(46):16569–16572, 2005.

[13] Lutz Bornmann, Rüdiger Mutz, and Hans-Dieter Daniel. Are there better indices for evaluation purposes than the h index? a comparison of nine different variants of the h index using data from biomedicine. *Journal of the American Society for Information Science and technology*, 59(5):830–837, 2008.

[14] Sozon Papavlasopoulos, Marios Poulos, Nikolaos Korfiatis, and George Bokos. A non-linear index to evaluate a journal's scientific impact. *Information Sciences*, 180(11):2156–2175, 2010.

[15] Bruce Hoppe and Claire Reinelt. Social network analysis and the evaluation of leadership networks. *The Leadership Quarterly*, 21(4):600–619, 2010.

[16] Xian Zheng, Yun Le, Albert PC Chan, Yi Hu, and Yongkui Li. Review of the application of social network analysis (sna) in construction project management research. *International journal of project management*, 34(7):1214–1225, 2016.

[17] Stefan Schlattmann. Capturing the collaboration intensity of research institutions using social network analysis. *Procedia Computer Science*, 106:25–31, 2017.

[18] Jeroen Baas, Michiel Schotten, Andrew Plume, Grégoire Côté, and Reza Karimi. Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies*, 1(1):377–386, 2020.

[19] Shahla Asadi, Halina Mohamed Dahlan, et al. Organizational research in the field of green it: A systematic literature review from 2007 to 2016. *Telematics and Informatics*, 34(7):1191–1249, 2017.

[20] Matthew E Falagas, Eleni I Pitsouni, George A Malietzis, and Georgios Pappas. Comparison of pubmed, scopus, web of science, and google scholar: strengths and weaknesses. *The FASEB journal*, 22(2):338–342, 2008.

[21] Scopus content coverage guide.

[22] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.

[23] Andrea Landherr, Bettina Friedl, and Julia Heidemann. A critical review of centrality measures in social networks. *Business & Information Systems Engineering*, 2(6):371–385, 2010.

[24] Murray A Beauchamp. An improved index of centrality. *Behavioral science*, 10(2):161–163, 1965.

[25] Linton C Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1978.

[26] Wikipedia encyclopedia https://en.wikipedia.org/wiki/modularity$_($networks$)$.

[27] Samet Keserci, Eric Livingston, Lingtian Wan, Alexander R Pico, and George Chacko. Research synergy and drug development: Bright stars in neighboring constellations. *Heliyon*, 3(11):e00442, 2017.