University of Sindh Journal of Information and Communication Technology (USJICT)
Volume 5, Issue 3, September 2021
ISSN-E: 2523-1235, ISSN-P: 2521-5582
Website: http://sujo.usindh.edu.pk/index.php/USJICT/

Network Analysis of co-authorship system of University of Sindh authors on Science Direct

Nazish Basir[1], Bisharat Rasool Memon[1], Abdul Waheed Mahesar[1], Yasir Arfat Malkani[2] , Sehrish Nizamani[3]

[1]Department of Information Technology, University of Sindh
[2]Institute of Mathematics and Computer Science, University of Sindh
[3]Department of Information Technology, University of Sindh , Mirpurkah Campus
nazish.basir@usindh.edu.pk, bisharat.memon@usindh.edu.pk, waheed.mahessar@usindh.edu.pk, yasir.malkani@usindh.edu.pk sehrish.basir@usindh.edu.pk

Abstract: Many complex systems have been modelled and analyzed as complex networks. These systems are huge and complex in terms of number of interconnecting components. In this research, we have analyzed the co-authorship network of Sindh university authors on science direct to understand the connectivity pattern of authors who published their articles over time. This research has found that the connectivity pattern of the authors is highly heterogonous due to the emergence of hubs in this system. Further, this network has shown highly clustered behavior with small world effect. These findings based on network analysis suggests that the co-authorship system is depending on few authors frequently publishing multiple papers in this network.

Keywords: social network analysis; complex networks; collaboration networks; two-mode networks

I. INTRODUCTION

Network science deals with the study of network representations of various real-life systems including physical, biological, and social systems with the aim of describing predictive models of these systems and the underlying phenomena. When a real-life system or phenomenon is modelled as a network, the distinct elements in the system are denoted by nodes/vertices, and the connection or interactions among them are denoted by links or edges. The field of network science lies at the intersection of mathematical graph theory, statistical mechanics, data mining and visualization, inferential modelling, and social structure. The goal of network science is to understand the properties and behavior of networked systems and ultimately to draw insights and conclusions regarding the real-life systems such networks represent [1, 2, 3].

When the pattern of connections or relationships among social entities is to be studied from a network-theoretic perspective it is aptly known as social network analysis [4, 5, 6]. These entities are often persons, but may also be groups, organizations, nation states, websites, and scholarly publications. A particular class of social networks, known as co-authorship networks, are especially important in social network analysis because these have been used extensively to determine the structure of scientific collaborations and the status of individual researchers [7, 8, 9, 10,11]. This paper explains the overall process of analysis of co-authorship network of all papers of University of Sindh which are published in Science Direct until now.

The rest of the paper is structured as follows: Section 2 discusses at length the relevant background and literature. Section 3 describes the methodology adopted for this study including data collection, pre-processing, cleaning and the main network analysis methods applied. Section 4 presents and discusses the results of this study. Finally, Section 5 summarizes the main findings and concludes the paper.

II. BACKGROUND AND RELATED WORK

Many researchers tried to observe co-authorship system from different perspectives of network. The authors [12], formalized the scientific collaboration network as two-mode network. They modeled authors and research articles as two-mode network. In this network two authors considered connected if they were co-authors on a given paper. A study of the co-authorship network of Canadian scientists for the period of 14 years since 1996 to 2010 undertaken by Ashkan Ebadi et al [13] resulted in formalization of the multiple regression models in order to estimate their impact on the underlying network structure. Similarly, the co-authorship system in diverse fields such as nano science, pharmacology, and statistics in Spain from 2006-2008 was studied by María Bordons et al [14]. By making use of the g-index as an intermediary measure and using the Poisson regression model they identified a relationship between type of research and performance of the authors in terms of their contributions. Micael S Couceiro et al [15] examined performance of researchers from different scientific fields both within the immediate network and within the larger network. Primarily they created a weighted adjacency matrix using the dataset of the published articles in publications having standardized

identifier ISBN, ISSN, and then using graph partition methodology the graph was divided into clusters, and computed the data using MATLAB script. In another study Xu Qianwen Ariel and Victor Chang [16] analyzed bibliographic data of 166 authors from three institutions to understand internal structure of co-authorship network in Shanghai, China. They found eigenvector centrality, betweenness centrality, authority and hub position, and efficiency were significant to g-index. In last they perform the Spearman correlation test to compare academic performance and SNA metrics. Hu, K., et al. [17] analyzed co-author and co-cited reference network of research done in the field of chlorophyll fluorescence. Firstly, active author communities were filtered by using the countries with high-citation-per-paper publications. Then the network-based methods were used to find out author groupings in these countries which they categorized to analyze author's focus area. In last, by using the co-cited reference networks the knowledge distribution timeline is presented. Another study by Zhao Yang, and Ning Wang [18] is focused on co-author network of domestic supply chain finance field. They used 15 year's data from 2005 to 2019 from CNKI. By using UCINET tool the co-author network was analyzed to determine subnet patterns of author, cohesive subgroups, centrality, network density, and structural holes. The study was used to identify the weaknesses in field of supply chain finance. A similar study by Qiqi Jiang [19] is the analysis of co-author network graph in the field of economics. The author analyzed SSCI academic literature in economics from 2004 to 2019 by using Python and Pajek. The author proposed the need of corporation among researchers. Xin Lu and Wentao Zhang [20] proposed a study in which they constructed the co-authorship network of Chinese sports culture field in by collecting the CNKI database's data from 2000 to 2019. The authors analyzed the structural characteristics and distribution of the network by using density, average distance, centrality, and clustering coefficient and presented their recommendations.

III. METHODOLOGY

The methodology adopted here includes several phases. Starting from data collection which involved searching for relevant publications within the Scopus database. The second stage was data pre-processing which involved cleaning and formatting of the initial result set, identifying any spurious results, and unifying any redundant results in order to achieve uniformity and consistency. Then followed the network analysis phase including the computation of several network and graph-theoretic measures.

A. Network Extraction

The article database was queried for authors who were affiliated with University of Sindh, with the initial search providing 302 results. The results were then exported into an appropriate data format. From the initial result set, the primary authors who worked on a paper were extracted and a corresponding entry was created for a network edge list, against which their co-authors were recorded. This represented the co-authorship collaboration network as an edge list. The result of that extraction was an excel sheet with 302 rows for the primary authors, and a variable number of columns per row for the collaborating co-authors. A snapshot of the extracted result data set is shown in Fig. 1.

283	Laghari, Abdul Hafeez	Memon, Shahabuddin	Nelofar, Aisha	Khan, Khalid Mohammed	Yasmin, Arfa	
284	Qureshi, Munawar Saeed	Mohd Yusoff, Abdull Rahim bin	Shah, Afzal	Sirajuddin		
285	Wadhwa, Sham Kumar	Kazi, Tasneem Gul				
286	Zia-Ul-Haq, Muhammad	Iqbal, Shahid	Ahmad, Shakeel	Imran, Muhammad	Niaz, Abdul	Bhanger, M.I.
287	Ibupoto, Z.H.					
288	Mari, R.H.	Aziz, M.	Jameel, Dler	Al Saqri, Noor		
289	Iqbal, Javed	Wattoo, Feroza Hamid	Wattoo, Muhammad Hamid Sarwar	Malik, Rukhsana	Tirmizi, Syed Ahmad	Imran, Muhammad
290	Derazshamshir, Ali	Shaikh, Huma				
291	Ikram-ul-Haq	Mukhtar, Zahid	Jaleel, Cheruth Abdul	Azooz, Mohamed Mahgoub		
292	Jamil, Waqas	Solangi, Sorath	Ali, Muhammad	Khan, Khalid Muhammad	Taha, Muhammad	Khuhawar, Muhamma
293	Imran, Syahrul	Taha, Muhammad	Ismail, Nor Hadiani	Kashif, Syed Muhammad	Rahim, Fazal	Jamil, Waqas
294	Bughio, Mansoor A.	Junejo, S.A.				
295	Khan, Khalid M.	Jamil, Waqas	Taha, Muhammad	Perveen, Shahnaz		
296	Hussain, Abdullah I.	Anwar, Farooq	Chatha, Shahzad A.S.	Latif, Sajid	Sherazi, Syed T.H.	Ahmad, Ashfaq
297	Taha, Muhammad	Kashif, Syed Muhammad	Shah, Syed Adnan Ali	Jamil, Waqas	Sidiqqui, Salman	Khan, Khalid Mohamn
298	Mujeeb-Kazi, A.	Kazi, Alvina Gul	Rasheed, Awais	Mahmood, Tariq	Bux, Hadi	Farrakh, Sumaira
299	Solangi, Sarfraz H.					
300	Taha, Muhammad	Ismail, Nor Hadiani	Jamil, Waqas	Kashif, Syed Muhammad	Ali, Muhammad	Rahim, Fazal
301	Brohi, Imdad A.	Solangi, Sarfraz H.	Lashari, Rafiq A.			
302	Solangi, Imam Bakhsh	Bhatti, Ashfaque Ali	Kamboh, Muhammahad Afzal	Memon, Shahabuddin	Bhanger, M.I.	
303						
304						

Figure 1. First extraction.

B. *Data Cleaning*

In the first extraction it was observed that the names of certain authors were written in different naming styles. A number of naming styles were observed depending on the placement of first and last names, or the use of abbreviations for one or all of their given names. All such cases were identified, and the names of the authors were changed to a single standard naming format for consistency. An automated find-and-replace approach was adopted to correct all the names of authors, and the final results were verified manually. Table I shows the distribution of papers according to number of authors.

After the name corrections the final edge list was extracted from the excel sheet and the total number of edges were 3416 as shown in Fig. 2. As this is a collaboration network, it naturally gives rise to the presence of multiple edges from one author to another author. The graph was created by converting the excel file into CSV file.

TABLE I. DISTRIBUTION OF PAPERS ACCORDING TO NUMBER OF AUTHORS

Number of papers	Number of authors
3	11
4	10
14	9
29	8
24	7
30	6
33	5
45	4
53	3
42	2
25	1

Fig. 3 shows the network diagram which shows that there are many authors who are disconnected in small cluster and there is a huge cluster of authors connected together. The simplify function was used to remove all the multiple edges in the graph for further calculations.

3399	Kolachi, Nida Fatima	Shah, Faheem
3400	Jalbani, Nusrat	Shah, Abdul Qadir
3401	Sirajuddin	Sarfraz, Raja Adil
3402	Kolachi, Nida Fatima	Arain, Muhammad Bilal
3403	Jalbani, Nusrat	Memon, Ateeq-ur-Rehman
3404	Sirajuddin	Jamali, Muhammad Khan
3405	Kolachi, Nida Fatima	Jamali, Muhammad Khan
3406	Jalbani, Nusrat	Khandhro, Ghulam Abbass
3407	Sirajuddin	Syed, Nasreen
3408	Wadhwa, Sham Kumar	Arain, Muhammad Bilal
3409	Ansari, Rehana	Memon, Ateeq-ur-Rehman
3410	Shah, Abdul Qadir	Jamali, Muhammad Khan
3411	Wadhwa, Sham Kumar	Jamali, Muhammad Khan
3412	Ansari, Rehana	Khandhro, Ghulam Abbass
3413	Shah, Abdul Qadir	Syed, Nasreen
3414	Shah, Faheem	Jamali, Muhammad Khan
3415	Shah, Abdul Qadir	Khandhro, Ghulam Abbass
3416	Sarfraz, Raja Adil	Syed, Nasreen
3417		

Figure 2. The final edge-list.

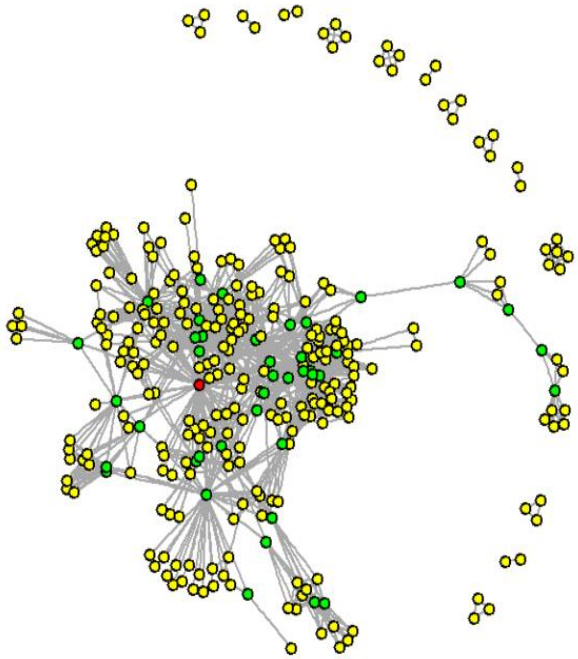

Figure 3. The co-authorship network.

IV. RESULTS AND DISCUSSION

The Network had on large connect component along with other small components as always observed in other co-author graphs. The smaller components of authors indicate groups of authors who publish together within that group, but do not have any publications in collaboration with authors outside of that group.

A. *Comparison with Random Network*

A random graph R was generated with `random.graph.game()` with 315 nodes and probability of 0.3 the similar number of edges were generated for comparison with co-author graph. Table II shows different matrices applied on the co-author graph and the random graph.

TABLE II. COMPARISON OF DIFFERENT VALUES BETWEEN THE CO-AUTHORSHIP GRAPH AND THE RANDOM GRAPH

	Co-authorship graph	**Random graph**
Average path length	3.2	2.8
Clustering co-efficient	0.3	0.03
Betweenness	Min: 0 Max: 12426.97	Min: 0 Max: 934.2999
Eccentricity	Min: 1 Max: 10	Min: 0 Max: 5
Closeness	Min: 1.01×10^{-05} Max: 7.79×10^{-05}	Min: 1.01×10^{-05} Max: 9.12×10^{-04}
Maximum degree	76	19

The number of differences were observed during comparison of this co-authorship systems with an equivalent random graph: The average path length "2.8" of random graph was less than the average path length "3.2" of co-author graph. The clustering coefficient "0.03" of random graph was less than the clustering coefficient "0.3" of co-author graph. This shows that the Network is Small World Network. In case of betweenness, co-authorship graph sows higher values due to the presence of hubs in this network. Finally, the maximum degree shows the difference of coauthors in both these networks.

B. Existence of Hubs

Another property of small world graph is the presence of hubs in the network. These are made up of nodes that have an extremely high degree (co-authors) compared to other nodes in the network.

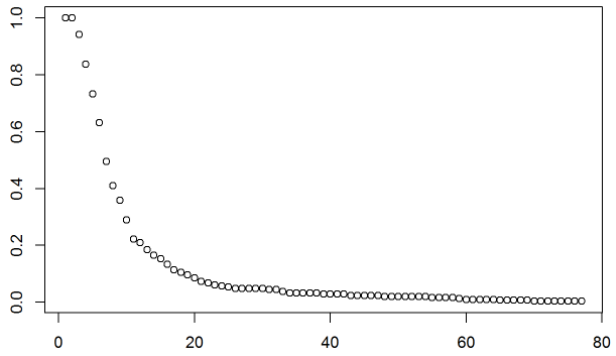

Figure 4. The degree plot.

The degree distribution plot in Fig. 4 indicates the existence of hubs, showing that a small number of authors have a very large degree while the majority of the authors have smaller degree. The largest clique that was found was of 12 nodes all connected together. These findings clearly show that the few authors have published many papers and many have few. This leads to inhomogeneous distribution power-law.

C. Community Detection

Another property that is often observed in networks is the existence of community structure in which nodes are interconnected within tightly knit groups known as communities [21, 22, 23], while at the same time the communities are loosely connected among themselves.

The fast greedy algorithm was used to detect communities in the graph [24] and the Fig. 5 illustrates the graph which was generated after the community separation. The modularity of the generated graph is 0.59.

The walk trap algorithm was used to detect communities in the graph [25] and the Fig. 6 illustrates the graph which was generated after the community separation. The modularity of the generated graph is 0.6.

The Louvain method [26, 27] was used to detect communities in the graph and the Fig. 7 illustrates the graph which was generated after the community separation. The modularity of the generated graph is 0.63.

It is seen that all three community detection methods gave nearly same results with a modularity of approximately 60%.

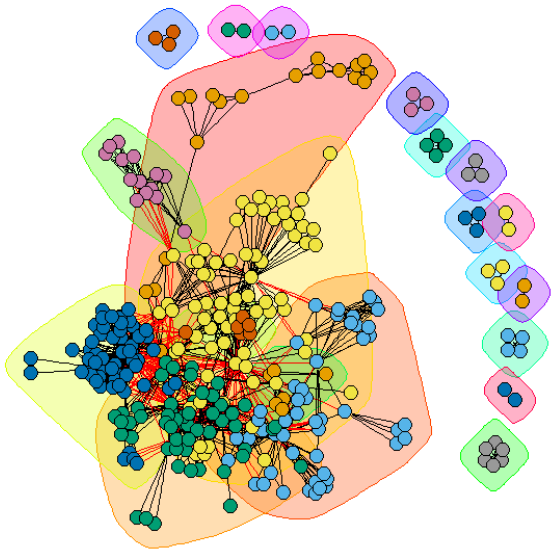

Figure 5. Collaboration network plot with fast-greedy algorithm.

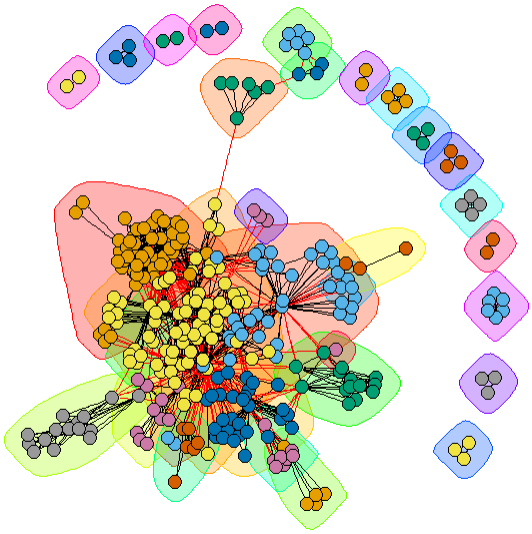

Figure 6. Collaboration network plot with walk-trap algorithm.

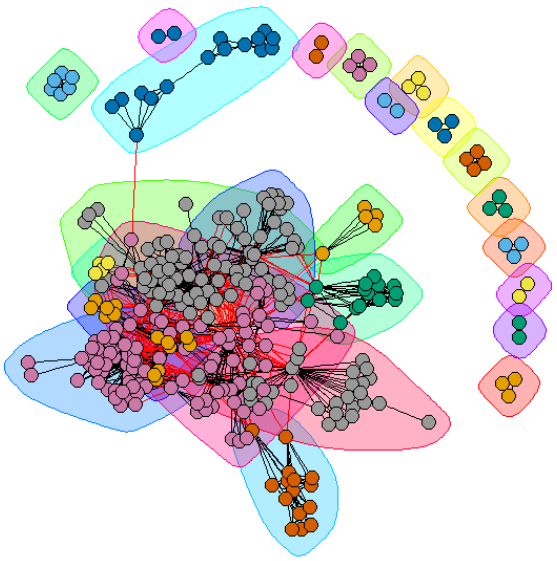

Figure 7. Collaboration network plot with Louvain method.

V. CONCLUSION

Many networked systems have been modelled and analyzed to understand their topological structures and overall behavior. The obtained co-author network was observed with overall similar properties as any other co-author network would have. We have found that few authors in this system have published many articles with many other collaborators and majority of authors have published few articles.

This is the reason that this network shows inhomogeneous pattern of connectivity and applied metrics shows variation in the obtained results of network analysis metrics. The presence of hubs clearly shows that few authors are actively engaged in their research and are publishing articles in this system quite frequently.

In this research, we have done static analysis of the network. On the other hand, this dataset is dynamic based on the issues published quarterly. The dynamic network analysis can be done to understand its longitudinal behavior with link prediction. Also weighted clustering coefficient can further endorse the results of small world effect of this network.

REFERENCES

[1] National Research Council. Network Science. The National Academies Press, Washington, DC, 2005.

[2] A.-L. Barabasi. Network Science. http://barabasi.com/networksciencebook/, 2015.

[3] T S Evans. Complex networks. Contemporary Physics, 45(6):455–474, #nov# 2004.

[4] Evelien Otte and Ronald Rousseau. Social network analysis: a powerful strategy, also for the information sciences. Journal of Information Science, 28(6):441–453, 2002.

[5] Alan Mislove. Online Social Networks: Measurement, Analysis, and Applications to Distributed Information Systems. PhD thesis, Rice University, Department of Computer Science, May 2009.

[6] John Scott. Social Network Analysis: A Handbook. SAGE Publications, 2000.

[7] Albert-Laszl Barabasi, H Jeong, Z N´eda, E Ravasz, A Schubert, and T Vicsek. Evolution of the Social Network of Scientific Collaborations. Physica A: Statistical Mechanics and its Applications, 311(3–4):590–614, 2002.

[8] M. E. J. Newman. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Physical Review E, 64(1):1–7, #jun# 2001.

[9] M. E. J. Newman. The Structure of Scientific Collaboration Networks. Proceedings of the National Academy of Sciences of the United States of America, 98(2):pp.404–409, 2001.

[10] L. Sun, "Coauthorship network in transportation research Coauthorship network in transportation research," no. April, 2017.

[11] E. Sarigöl, R. Pfitzner, I. Scholtes, A. Garas, and F. Schweitzer, "Predicting scientific success based on coauthorship networks," pp. 1–16, 2014.

[12] M. E. J. Newman, "The structure of scientific collaboration networks," Natl. Acad. Sci., vol. 98, no. 2, pp. 404–409, 2001.

[13] A. Ebadi and A. Schiffauerova, "How to become an important player in scientific collaboration networks?," J. Informetr., vol. 9, no. 4, pp. 809–825, 2015.

[14] M. S. Couceiro, F. M. Clemente, and F. M. L. Martins, "Towards the Evaluation of Research Groups based on Scientific Co-authorship Networks: The RoboCorp Case Study," Arab Gulf J. Sci. Res., vol. 31, no. 1, pp. 36–52, 2013.

[15] S. Uddin and A. Khan, "The impact of author-selected keywords on citation counts," J. Informetr., vol. 10, no. 4, pp. 1166–1177, 2016.

[16] Q. Ariel Xu, and V. Chang, "Co-authorship network and the correlation with academic performance", Internet of Things, vol. 12, 100307, 2020.

[17] K. Hu et al., "Co-author and co-cited reference network analysis for chlorophyll fluorescence research from 1991 to 2018", Photosynthetica, vol. 58, no 1, pp. 110–124, 2020.

[18] Y. Zhao, and N. Wang, "Research on Authors' Co-authorship Network in Supply Chain Finance in China Based on Social Network Analysis", WHICEB 2021 Proceedings, 2021.

[19] Q. Jiang, "Analysis of Productive Authors' Co-authoring Relationship in Economics", in 5th Internation Conference on Information in Education, Management and Business (IEMB 2021), 2021.

[20] X. Lu, W. Zhang, "Research on Co-authorship Network of Sports Culture in China Based on Social Network Analysis (SNA)", Academic Journal of Humanities & Social Sciences, vol. 4, no. 3, 2021.

[21] M. Girvan and M E J Newman. Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America, 99(12):7821–6, #jun# 2002.

[22] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E, 74:036104, September 2006.

[23] M. E. J. Newman and Michelle Girvan. Finding and evaluating community structure in networks. Physical Review E, 69(2):026113, #feb# 2004.

[24] M. E. J. Newman. Fast algorithm for detecting community structure in networks. Physical review E, 69(6):066133, 2004.

[25] G. K. Orman, Vincent Labatut, and Hocine Cherifi. Com- parative evaluation of community detection algorithms: a topological approach. Journal of Statistical Mechanics: Theory and Experiment, 2012(08):P08001, 2012.

[26] X. Que, F. Checconi, F. Petrini, and J. A Gunnels. Scalable community detection with the louvain algorithm. In 2015 IEEE Interna- tional Parallel and Distributed Processing Symposium, pages 28–37. IEEE, 2015.

[27] VD Blondel, JL Guillaume, R Lambiotte, and E Lefebvre. Fast unfolding of communities in large networks. arxiv. org. Physics and Society, 2008..