IJDACR
ISSN: 2319-4863

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 6, Issue 8, March 2018)

Analysis of Co-Authorship Network of Scientists Working on Topic of Network Theory

Manu Kohli
School of Informatics and Computing
Indiana University Bloomington, USA
kohlim@umail.iu.edu

Saurabh Jain
School of Informatics and Computing
Indiana University Bloomington, USA
jainsa@umail.iu.edu

Abstract— **The co-authorship of articles in peer review journals contributes to knowledge base of research community. Coauthorship can be termed as a collaboration between two or more authors if they publish a scientific work together. The coauthorship of papers can reveal interesting patterns that can be analysed and therefore strengthened by developing enhanced framework of collaboration. In this document we have analysed co-authorship network of scientists whose major work has been in the area of network theory. We have used principle of network sciences where the network nodes represent authors who are connected with edges if they have co-authored papers together. We have studied network topology and thereafter determined multiple statistical figures, important authors, prevalent community structure, measures of various centralities, such as closeness and betweenness, to identify connectedness of authors within a network. We have formed null hypothesis on assortivity degree and clique number using a similar graph of same degree sequence using configuration model. In addition we have measured robustness of the network and have summarized opportunities that can improve author's collaboration leading to inter-discipline collaboration and high quality scientific research.**

Keywords — **Node, Edges, Network science, Centrality, Betweenness, Eigenvector, community, robustness, assortivity, shortest path, average degree, collaborative.**

I. Introduction

The collaboration between authors usually leads to join research exercises that can immensely contribute to the scientific community. The collaboration between authors is usually dependent on mutually interesting research topics and competencies. Three different types of networks associated with scientific research, publications and bibliometric can be analyzed using network sciences [1, 2].

A. Collaboration networks

The studies on these type of network show the collaboration of authors or institutions in jointly performing scientific research. One of the most basic type of collaboration network is the co-author network, in which multiple nodes represent authors and edges represent joint publications on which authors have collaborated together.

B. Semantic Networks

Study on Semantic networks analyze the occurrence of designated set of keywords related to focus area of study in publications. Based on the search outcomes a network is formed and analysis is performed.

C. Citation Networks

Publication citation networks show the relationships among various scientific publications and authors based on their citation records.

The focus of this paper is to understand various statistical measures that are calculated by applying principle of network sciences on co-authorship networks. We will then evaluate how these measures can contribute to enhance the collaborative network between the authors. This can immensely help research communities and promote inter-disciplinary research where limited research budgets can find way to most deserving scientist or an institution evaluated through bibliometric analysis and peer review process. [2]

To analyze the trends of co-authorship of papers we intend to use dataset compiled by M. Newman in May 2006 [3, 4, 5]. The proposed data is compiled from the bibliographies of two review articles on network sciences with a few additional references added by hand. The data consists of 1589 scientists that can be represented as nodes in the network and 2784 edges, if the scientists have co-authored papers together.

The outcome of the study allows identification of central authors, so called experts in the scientific area of analysis. The study also identifies essential communities of authors in the network and at same time present various centrality measurements that can improve communication and

IJDACR International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 6, Issue 8, March 2018)

collaborative opportunities between communities and authors. The study of community structure and clustering network offers a potential opportunity to establish intertropical and interdisciplinary connections that are still not realized [6].

We have also studied on robustness of the network to identify the possible collaboration opportunities between authors and further strengthen the network.

II. Literature Review

We have performed literature review limited to application of network science principles on the networks related to publications and bibliometric. The collaborative study of coauthorship analysis by using publication data has been carried out extensively [7] and has produced significant outcomes useful to scientific community. The collaboration network displays property of small world in which randomly chosen researchers are separated by few intermediate potential collaborators [8]. The analysis of network therefore offers an opportunity of collaboration between authors who may be unrelated with each other but are separated only by a distance of one or two degrees.

Friendship paradox is a phenomenon that shows that your friends have on average more friends than you have. Friendship paradox can also be observed on the networks related to publications and coauthorship [9]. For example a co-author of an individual on an average is expected to have more citations than the individual.

Usually network of co-authors display property of scale free graphs and preferential attachment [10]. In scale free network some nodes, authors in our case, form close dense structures referred as “hubs”. The nodes in these hubs have many more connections than others and the network follows a power-law distribution on the number of edges connected to a node.

An author who starts performing research is most likely to be acquainted and jointly perform research with established authors and scientist following law of preferential attachment. In preferential attachment established authors who are already connected with many other authors get connected to new authors in proportion to the connections they already have leading to power law distribution [11].

Both scale free and preferential attachment properties can be attributed largely to betweenness centrality [12]. Some authors show prominent position in the network and are able to channelize information flow in the network with the other authors based on shortest path [13].

The analysis of centrality measures on the co-authorship networks [14] provides extensive insights on importance of authors in the network. The Eigenvector centrality helps identifying authors who are connected to other important authors. This may stress the importance of getting connected to few important authors in the network rather than connected to large number of authors who may be not influential and well known.

III. Graph Analysis

In this section we will evaluate certain topological graph features that can provide additional insights about the graph and co-authorship patterns. We have used networkx package in python and Gephi tool to perform analysis. Gephi is an open source social network analysis tool used for graph visualization, community detection and perform various centrality measure calculations [15].

Usually the co-authorship networks display fat-tailed distribution, observed via average degree in the graph, in which a small proportion of scientists produce large number of papers [16]. Fat tail distribution has been observed in other coauthorship networks as well and we had observed similar results in our network as shown in Figure 1. Average degree in the graph displays fat tailed phenomenon that can substantially help propagation of information in the network. The weak ties observed as a result of fat tailed phenomenon may lead to professional relationships and co-authorship opportunities for new authors.

The below table (Table 1) shows some of the parameters that were observed for the graph using python networkx and Gephi tools. The average path length of the graph was observed to be 5.823 and Network diameter as 17. In co-authorship network average path length can be thought of as average number of steps taken from each author in the network along the shortest path to reach another author.

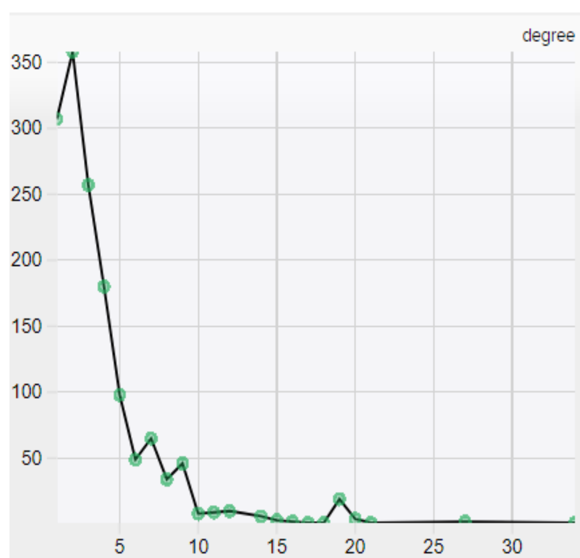

Figure 1: Average Degree- Fat Tail distribution
Source: http://networkrepository.com/ca-netscience.php

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 6, Issue 8, March 2018)

Table 1: Network Measure for Co-author Graph

Graph Parameters and remarks	Measure
Nodes - Number of Authors in network	1589
Edges - Number of paper co-authored	2743
Maximum Degree -Maximum number of authors in a paper	34
Minimum Degree -Single Authored papers	0
Assortativity - Possible association of similar authors	0.46
Number of triangles clusters of co-authors	11300
Fraction of closed triangles - fraction of pairs of authors with a common person who has co-authored with two	0.70
Average clustering coefficient measure of the degree to which nodes in a graph tend to cluster together	0.637

Scientists having maximum collaboration i.e. connected to most scientists in the network, observed via degree distribution metric, can be seen in figure 2.We then calculated all the paths of length less than or equal to 4 between Barabasi and Newman, two authors with highest degree that can be observed in Figure 3.

IV. Graph Nature and Preferential Attachment

The coauthorship networks usually displays the growth pattern based on preferential attachment [17]. That means the established authors tend to collaborate more with each other and resulting graph are usually scale free in nature. The clustering coefficient helps evaluating existence of ties between author A and C if ties are observed between A and B and between authors B and C [18]. In our case the result of 0.63 suggest good collaboration opportunity within the co-authorship network. The clustering coefficients can be visualized through Figure 4 where majority of nodes with value of 1 shows high connectivity , however quite of few nodes of single author with no connectivity with others show clustering coefficient of zero.

A component in a network consist of nodes, authors in our case, where any author in the network can be reached from any other by traversing a suitable path of intermediate collaborators. We had evaluated largest component in the graph using maximum connected component functionality using python. For the largest component we observed that it contains number of nodes as 379, number of edges as 914 with average degree of 4.8. That means that only small hub within the network is dense and forms the largest component.

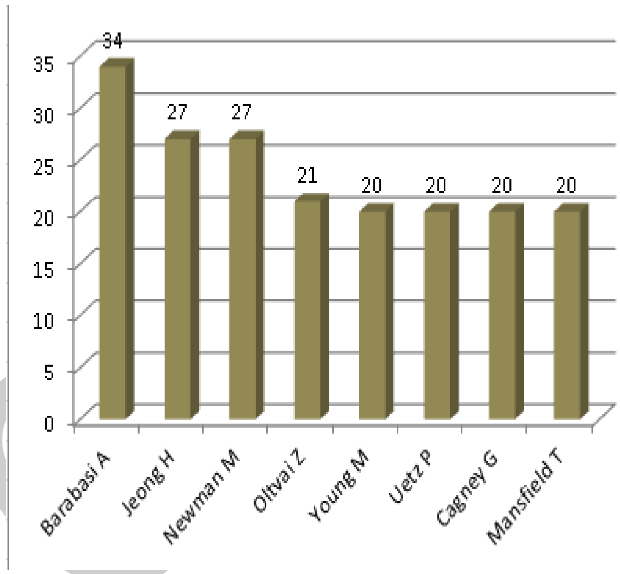

Figure 2: Most connected Authors - collaboration in network
Source: Author

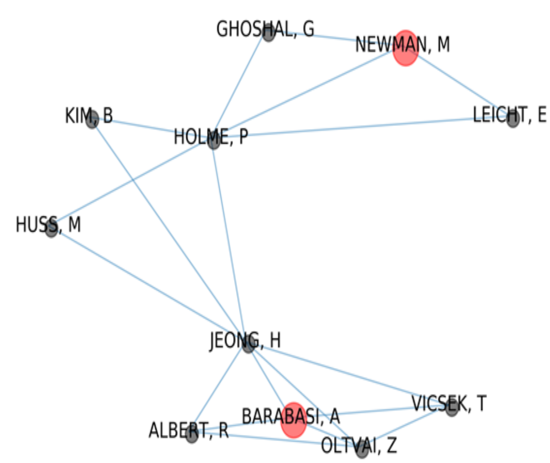

Figure 3: Shortest path of length 4 between Newman and Barabasi
Source: Author

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 6, Issue 8, March 2018)

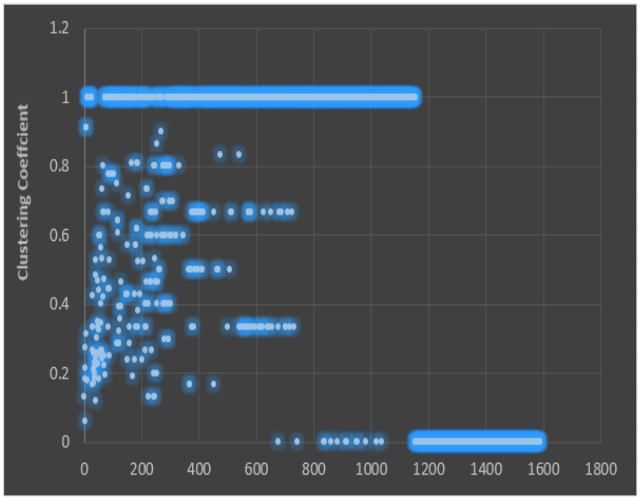

Figure 4: Clustering Coefficient
Source: Author

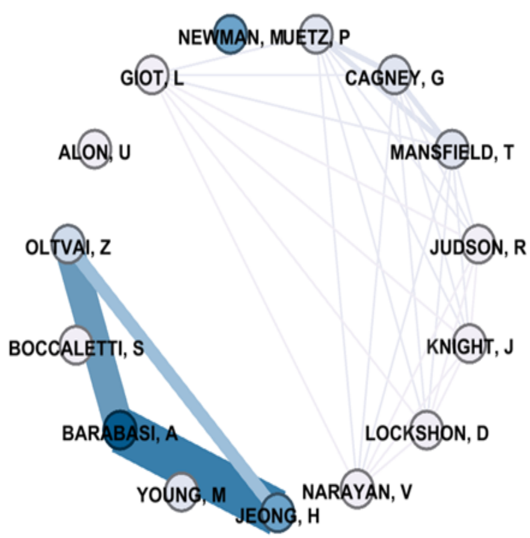

Figure 5: Weighted Edges between top 20 scientists
Source: Author

A. Centrality Measures

The centrality measures allows us to find important node, in our case most important author, in the network. The analysis could help in information flow, knowledge distribution, protecting network from breaking.

Betweenness centrality in case of our graph measures how many authors will pass through a specific author to reach another author through minimum hops. The author through which maximum traversals happen has the highest betweenness centrality score. In the case of our co-authorship network the figure 6 shows the number of possible traverses and percentage of path through 15 authors having maximum betweenness centrality score. Figure 7 shows Sub graph of top 20 scientists having highest betweenness centrality:

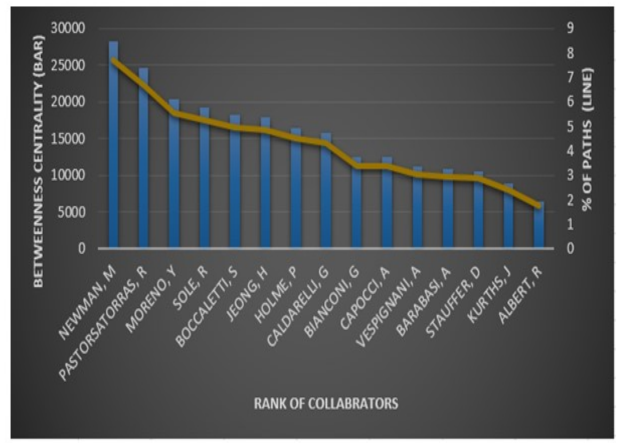

Figure 6: Betweenness Centrality
Source: Author

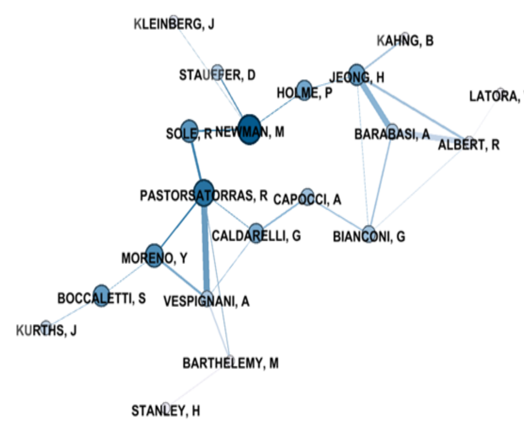

Figure 7: Betweenness Centrality of top 20 Scientist
Source: Author

Closeness measure is based on the length of the average shortest path between an author and all the other authors in the graph through shortest route. Eigenvector states that central node should be one connected to powerful nodes (Quality vs Quantity). The table 2 below shows top 15 authors in terms of closeness and eigenvector centrality measures.

IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 6, Issue 8, March 2018)

Table 2: Closeness and Eigenvector centrality Measures Top 15

Author Name	Eigenvector Measure
UETZ, P	1
CAGNEY, G	1
MANSFIELD, T	1
GIOT, L	0.991754
JUDSON, R	0.991754
KNIGHT, J	0.991754
LOCKSHON, D	0.991754
NARAYAN, V	0.991754
SRINIVASAN, M	0.991754
POCHART, P	0.991754
QURESHIEMILI, A	0.991754
LI, Y	0.991754
GODWIN, B	0.991754
CONOVER, D	0.991754
KALBFLEISCH, T	0.991754

Author Name	Closeness Measure
ACEBRON, J	1
BONILLA, L	1
PEREZVICENTE, C	1
RITORT, F	1
SPIGLER, R	1
AERTSEN, A	1
ALBERTS, B	1
BRAY, D	1
LEWIS, J	1
RAFF, M	1
ROBERTS, K	1
WATSON, J	1
ULANOWICZ, R	1
BARJOSEPH, Z	1

B. Community Detection

The modular structure in the graph can be observed using optimization algorithms that can subdivide graph into logical community structures. However these algorithms are challenged with respect to the limit where they should stop subdividing graph to further communities [19, 20].

Gephi, an open source network analysis tool is used to perform calculation of centrality measures and community detection. The Gephi tool was applied to the network with filter criteria applied to degree of nodes ranging from 5 to 35. We initially used 1 as a resolution without use weights option to see 64 communities in picture. With resolution of 5 we observed 44 communities with following observations:

- Majority of communities observed were disconnected from each other
- Few of the communities had lot of interaction.
- With application of closeness centrality to above filtered graph, we observed following results:
- Disconnected communities have much lower closeness centrality
- Nodes which are forming connectivity with other communities have higher closeness centrality.
- These are the scientists who are closer to other scientists and thus will be more probable to perform cross discipline research.

The output of Gephi tool shows four major interlinked communities (Figure 8) connected via nodes of high betweenness centrality.

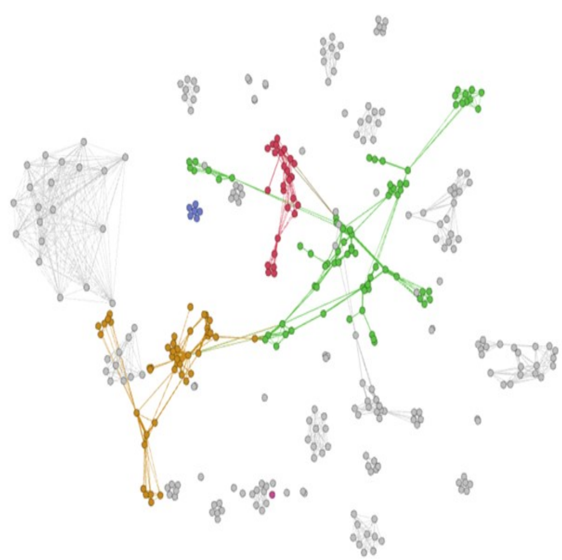

Figure 8: Community visualization
Source: Author

IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 6, Issue 8, March 2018)

C. Network Robustness

Network robustness can be measured as a response of the network on removal of nodes [21]. For networks related with co-authorships it provides an opportunity for the network to grow and expand. The robustness of the network can be measured with respect to random attacks or targeted attacks. Random attacks can be executed by removing any node in the network based on the probability of selection of that node. The effect of removal of that node shows the robustness of the network. However for targeted attacks specific node is selected from the network which when selected that lead to disintegration of the scale free properties of the network.

We performed critical point calculation for both random and targeted removal and observed on the graph representing network of co-authors that usually display scale free properties and observed following results. The results can also be visualized with the graph shown below.

- Critical point is reached much earlier when we do targeted attacks. This observance is in accordance with expected result.
- Critical point in case of random attack for 10 percent level is observed after removing 50 percent of nodes.
- Critical point in case of targeted attack for 10 percent level is observed after removing 5 percent of nodes.

Calculating critical point using random removal and targeted removal as shown in Figure 11.

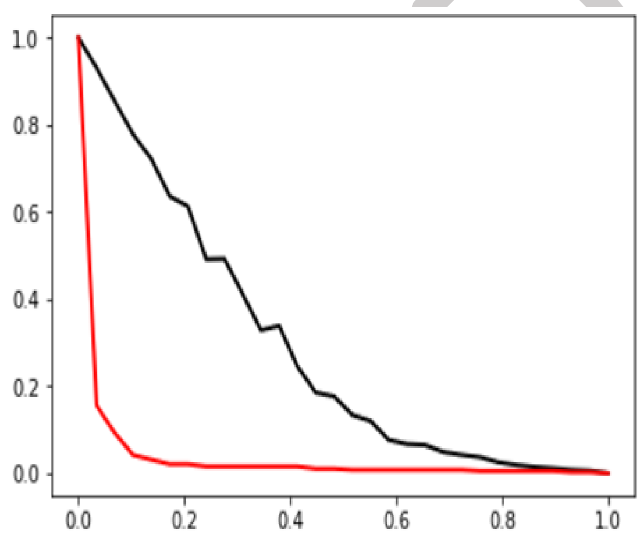

Figure 9: Random and targeted attack on the Co-author network
Source: Author

We carried out similar exercise of finding critical point in case of random graph of similar dimensions. In case of random graph we see that the critical point is reached much later when we compare with our co-authorship graph with both random and targeted attacks.

- Critical point in case of random attack for 10 percent level is observed after removing 65 percent of nodes.
- Critical point in case of targeted attack for 10 percent level is observed after removing 30 percent of nodes.

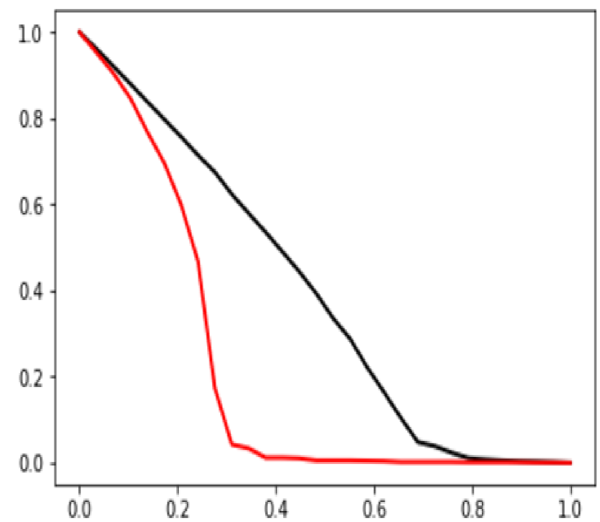

Figure 10: Random and targeted attack on the random graph
Source: Author

The following are the comparison results of attacking real graph representing a co-authorship network that is scale free in nature with respect to a random graph.

- In case of real life graph (network of co-authors) we see that the critical point is reached very early in case of targeted and random attacks.
- In targeted attack removing 5 percent of the nodes decreases the largest connected community size below 1 percent.
- In case of real life scale free graph it is expected that the few nodes of graph with be part of hub and will have high degree. Removing these nodes from network is expected to disintegrate the graph fairly quickly.

V. HYPOTHESIS FORMULATION WITH RESPECT TO CONFIGURATION MODEL

Null model forms a basis of hypothesis [22] specifically in the area of network science where the actual graph is compared to a similar graph on some feature, such as community structure, assortivity and so on. Our hypothesis is formed on that basis that the graph with which real graph is compared is constructed using configuration model with the same degree sequence as that of real graph [23].

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 6, Issue 8, March 2018)

In the following section we will evaluate two hypothesis formed using configuration model.

A. Hypothesis Formulation

1) Hypothesis 1

Usually scientist tend to collaborate with each other in pursuit of scientific research. The question we will like to ask is does scientist who have published large number of paper usually do with other scientist who have also published equal or large number of papers know as rich man club [24]. Does popular scientist tend to collaborate with other popular scientist, a phenomenon that can be measured using degree of assortivity [25].

We have used networkx as a package in python language to measure assortivity for our graph found it to be 0.46. The first finding on degree assortivity suggest that network is fairly assortative though the score yielded is not very high but does suggests somewhat that our hypothesis is true. We them compared it to a null model in order to confirm with our hypothesis.

We performed calculation of zscore to compare our graph with the graph that was generated using configuration model. The zscore of the model when compared with real graph yielded a score of 19.45. We have used a histogram function to plot the distribution of assortativity coefficients that suggest that assortativity in real graph is very far away from the average assortativity in our configuration models showing a behavior that is far away from similar random graphs.

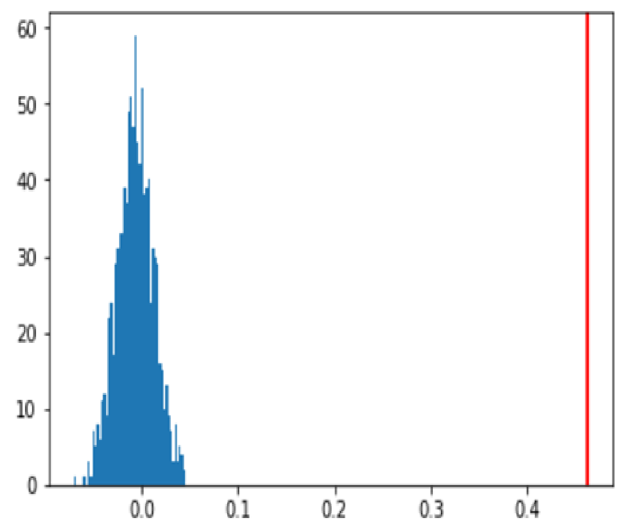

Figure 11: Null Hypothesis- Assortivity degree with real graph and configuration model
Source: Author

2) Hypothesis 2

The second hypothesis we assumed was that we will not encounter large clique in the network as we do not expect large group of scientists collaborating with each other. We used network x function to calculate clique number (size of the largest clique) for our graph that yielded the result of 20 that means that at least one group of 20 scientists is collaborating with each other. This is against our second hypothesis where we assumed that we will not discover large clique in the graph analysis. We thereafter used configuration model to calculate that yielded clique of 3. The zscore calculation was performed to compare clique of real graph with graph generated from configuration model. The result suggest and plotted in Figure 2 shows that groups/clique in real graph is very far away from the average closed groups/clique in the configuration models.

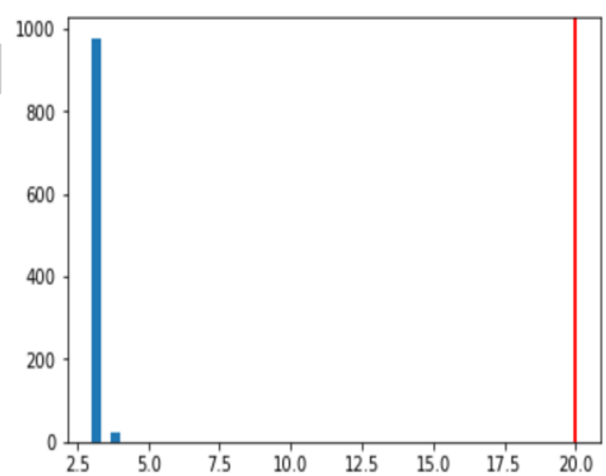

Figure 11: Null Hypothesis- Clique number of real graph and configuration model
Source: Author

Since we see a sufficiently high value of real closed groups/clique when compared to that of configuration model that suggest that our hypothesis is incorrect. Therefore null hypothesis based on development of configuration model designed on degree sequence is alone not sufficient compare real graph and the graph generated from configuration model.

IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 6, Issue 8, March 2018)

VI. CONCLUSION

In this paper, we have discussed network of co-authors who have been involved in research in area of network science. The network in analysis showed 'fat tailed' distribution and scale free network properties. The network displayed community structures with closely knitted 4 communities that were well integrated. The network displayed typical properties of scale free network and showed considerable robustness against random attacks, however with targeted attacks the network disintegrated. We also performed null hypothesis assumption and compared the graph under analysis with a similar graph with same degree sequence on the measures of assortivity and clique number. We observed that both the hypothesis did not received expected scores for hypothesis acceptance. This shows that there may be another factor for example high degree of nodes with no co-author criteria to affect our hypothesis results.

REFERENCES

[1] E. Yan and Y. Ding, "Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other," Journal of the Association for Information Science and Technology, vol. 63, no. 7, pp. 1313–1326, 2012.

[2] C. Belter. (2012, Jun.) Visualizing networks of scientific research @ONLINE. [Online]. Available: http://www.infotoday.com/online/may12/Belter-Visualizing-Networks-of-Scientific-Research.shtml

[3] M. E. Newman, "The structure and function of complex networks," SIAM review, vol. 45, no. 2, pp. 167–256, 2003.

[4] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: Structure and dynamics," Physics reports, vol. 424, no. 4, pp. 175–308, 2006.

[5] M. E. Newman, "Finding community structure in networks using the eigenvectors of matrices," Physical review E, vol. 74, no. 3, p. 036104, 2006.

[6] "Who is the best connected scientist? a study of scientific coauthorship networks," in Complex networks. Springer, 2004, pp. 337–370.

[7] "The structure of scientific collaboration networks,"Proceedings of the National Academy of Sciences, vol. 98, no. 2, pp. 404–409, 2001.

[8] J. Davidsen, H. Ebel, and S. Bornholdt, "Emergence of a small world from local interactions: Modeling acquaintance networks," Physical Review Letters, vol. 88, no. 12,p. 128701, 2002.

[9] Y.-H. Eom and H.-H. Jo, "Generalized friendship paradox in complex networks: The case of scientific collaboration," Scientific reports, vol. 4, p. srep04603, 2014.

[10] K. B¨orner, J. T. Maru, and R. L. Goldstone, "The simultaneous evolution of author and paper networks," Proceedings of the National Academy of Sciences, vol. 101, no. suppl 1, pp. 5266–5273, 2004.

[11] [11] M. E. Newman, "Clustering and preferential attachment in growing networks," Physical review E, vol. 64, no. 2, p. 025102, 2001.

[12] A. Abbasi, L. Hossain, and L. Leydesdorff, "Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks," Journal of Informetrics, vol. 6, no. 3, pp. 403–412, 2012.

[13] F. Zare-Farashbandi, E. Geraei, and S. Siamaki, "Study of co-authorship network of papers in the journal of research in medical sciences using social network analysis," Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences, vol. 19, no. 1, p. 41, 2014.s

[14] 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.

[15] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian, "Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software," PloS one, vol. 9, no. 6, p. e98679, 2014.

[16] M. E. Newman, "Coauthorship networks and patterns of scientific collaboration," Proceedings of the national academy of sciences, vol. 101, no. suppl 1, pp. 5200–5205, 2004.

[17] C. S. Wagner and L. Leydesdorff, "Network structure, self-organization, and the growth of international collaboration in science," Research policy, vol. 34, no. 10, pp. 1608–1618, 2005.

[18] M. E. Newman, "Who is the best connected scientist? A study of scientific coauthorship networks," Phys. Rev. E, vol. 64, no. 016131, 2001.

[19] D. S. Bassett, M. A. Porter, N. F. Wymbs, S. T. Grafton, J. M. Carlson, and P. J. Mucha, "Robust detection of dynamic community structure in networks," Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 23, no. 1, p. 013142, 2013.

[20] M. Newman, "Community detection in networks: Modularity optimization and maximum likelihood are equivalent," arXiv preprint arXiv:1606.02319, 2016.

[21] X. F. Wang and G. Chen, "Synchronization in scalefree dynamical networks: robustness and fragility," IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications, vol. 49, no. 1, pp. 54–62, 2002.

[22] J.-G. Young, G. Petri, F. Vaccarino, and A. Patania, "Construction of and efficient sampling from the simplicial configuration model," arXiv preprint arXiv:1705.10298, 2017.

[23] F. Chung and L. Lu, "Connected components in random graphs with given expected degree sequences," Annals of combinatorics, vol. 6, no. 2, pp. 125–145, 2002.

[24] V. Colizza, A. Flammini, M. Serrano, and A. Vespignani, "Detecting rich-club ordering in complex networks," arXiv preprint physics/0602134, 2006.

[25] F. Buccafurri, G. Lax, and A. Nocera, "A new form of assortativity in online social networks," International Journal of Human-Computer Studies, vol. 80, pp. 56–65, 2015.