International Journal of Engineering & Technology, 7 (4.35) (2018) 782-790

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Exploratory Experiment on Co-Authorship Network using Social Network Analysis Metrics and Measures

Pritheega Magalingam[1]*, Ganthan Narayana Samy[2], Nurazean Maarop[3], Wan Nazirul Hafeez Wan Safie[4], Muhammad Khairul Rijal[5], Lim Yee Fang[6], Abdullah Sakib[7], Muhammad Yassin[8]

[1]Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia Kuala Lumpur
[2]Ministry of Education Malaysia, Putrajaya, Malaysia
[3]Schlumberger Business Support Hub, Bandar Utama, Petaling Jaya, Malaysia
[4]Telekom Malaysia Berhad, Kuala Lumpur
**Corresponding author E-mail: mpritheega.kl@utm.my*

Abstract

This paper contributes in understanding and gaining meaningful insight about the relationship among the scientist in the co-authorship network using social network analysis. We argue that the relationship analysis is not always a straightforward process. In the past one single measure, for example, the egocentric or centrality measure was used to describe the scientific collaboration patterns separately. In this paper, various analysis such as centrality analysis, ego network, community detection, largest clique and word frequency have been used to examine and interpret the collaboration among the authors. This research is not dominated by known researchers but involves an overall exploration of the network. Our research is mainly guided by the creation of research issues, assessing the type of dataset and the objectives for presenting the co-authorship relationships. It is important to identify the motive of the selected measures in order to achieve the predefined objective. Specific methodology and procedures are designed to solve each research issue respectively. This study reveals that the network interpretation should not be solely based on one network measure, but an explorative analysis results need to be considered because it allows exploring the hidden information through the changes in the network structure, topology patterns and nodes' position.

Keywords: *Degree Centrality, Betweenness Centrality, Ego Network, Community Detection, Clique.*

1. Introduction

In a co-authorship network, the nodes represent the authors and edges denote collaboration of authors. It has long been realized that the co-authorship of articles in learned journals provides a window on patterns of collaboration within the academic community. Co-authorship of a paper can be thought of as documenting a collaboration between two or more authors, and these collaborations form a ‘‘co-authorship network,’’ in which the network nodes represent authors, and two authors are connected by a line if they have co-authored one or more papers. The structure of such networks can reveal insight into the relationship between the authors.

Relevant social network analysis (SNA) measures and the aim of this paper are discussed here. One of the ways to analyse a network relationship is to study the ego-network from the main network. Ego networks consist of a focal node ("ego") and the nodes to whom ego is directly connected to (these are called "alters") plus the ties, if any, among the alters. Egos can be any crucial node in the network that is chosen to be analysed or determined by influential nodes [18]**.** The influential node can be identified using network centrality measures. The formal concept of centrality namely degree, closeness and betweenness centrality were initially introduced in social network analysis by Freeman in 1979 [10]. Following the development of centrality concepts, eigenvector centrality was presented by Bonacich [6]. Centrality is a measure to assess a node's position critically. Determining a node's importance using centrality values has been commonly applied in social network analysis [2]. However, the idea of what forms a central node in different network application and the type of commodity flowing through a network affect the interpretation of an important node or influential node. Therefore, selection of a suitable centrality measure is essential by characterizing the network's structural properties. Further, the community structure of the co-authorship network denotes the collaborative research work among the authors and co-authors, and communities are formed on the basis of mutual interest. Therefore, analysing the structure of the community of a co-authorship network allows exploring the hidden information about its functional and structural properties. In this paper, we have addressed five issues from the co-authorship dataset as stated below:

1. What is the most appropriate centrality measure to identify the influential author in the co-authorship network of scientists?
2. Could the relationship of authors be measured in an isolated and balanced ego network?
3. What is the relationship among the scientist in the biggest community identified by different algorithms?
4. What is the core area of research and favourite publisher of authors in the co-authorship network?

This paper can also guide students or employees who are in their initial stage of exploring the social network analytics on deriving relevant questions and how to implement the techniques proposed in this paper for their academic research or industry projects.

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

2. Related Work

This part includes all the related work that involves different social network measures mainly focusing on centrality, egocentric network, clique and community analysis. Over years, researchers have devoted great efforts in understanding network characteristics, structures and topologies, and network evolution[21].

2.1 Node Centrality

Node centrality aids in discovering imperceptible features in network analysis. Different centrality metrics emphasize different features and aspects of the network structure. Hence, appropriate selection of centrality measures is important for a specific type of network analysis [27]. The commonly known centrality measures include degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality [2]. In the study on the robustness of node centrality against network manipulation, Niu et al. [21] suggested that the robustness of node centrality can be improved if a centrality measure can well separate the centrality score of nodes to identify an influential node. Delgado-Garcia et al. [8] adopted centrality metrics in identifying the influential authors in the co-authorship networks of Latin American Computer Science research groups. Their study found that densification occurred as the size of networks emerged from the co-authorships increased over time. Apart from centrality measures, Guércio et al. [13] described how the topology analysis can be applied with centrality values in identifying the influential node. Impact of the loss of an influential node is verified by node removals [13].

Basically, in social network analysis, high degree node is likely more influential in the network. However, it is not practical to apply this concept to all networks considering the specific network properties and its origin. In this case, betweenness centrality and closeness centrality would provide more precise justice to certain influence level [15]. Betweenness centrality is a measure of information transmission through a node where the sum of the fraction of all-pairs shortest paths that pass through the particular node. A node with high betweenness may have high influence and control over information passing between others and removal of that node from the network will disrupt communications between other nodes. While closeness centrality measures the mean distance from one node to another. This can be used to estimate the nearest node for the information transmission from a node; source to destination in the network [2, 21]. Eigenvector centrality value is computed based on the centrality of its neighbours. For example, n is proportional to the sum of the centralities of other nodes that connected with n [21].

According to Sun and Rahwan [27], the degree centrality quantifies the importance of an author using local structural property. On contrary, betweenness centrality and PageRank index quantify higher order interactions using distance or random-walk-based approaches considering the network global properties. The study [27] indicated the collaborations with high betweenness authors need not be strong considering the weight, but it tends to be in a more central position with the network structure and the paths between nodes. High betweenness author more likely to be on the shortest path and plays important role in collaborations connecting different communities.

Ahmed et al. [2] deduced that although some co-authors have frequent publications with other co-authors if those authors are working in the same institute or collaborate with other institutes in the same city, that might cause the authors to have high centrality values only in a certain position of the network. An author with high publication might not be linked to many authors if he or she do not publish papers with sparsely located co-authors [2]. The differences in the centrality position cannot highlight the most important or influential author.

Thus, the first analysis of this paper intends to examine the most appropriate centrality measure to identify the most influential author in the co-authorship network by evaluating the impact of the removal of nodes based on different node centrality measures. The changes in the network structure, topology patterns and nodes' position are inspected closely with the corresponding characteristics and functions of the co-authorship network. The node (author) that gives significant impact to the network are quantified as the most influential node (author) in the co-authorship network.

2.2 Ego Network

Co-authorship networks represent the patterns of human collaborations in the production of scientific knowledge [4]. By studying the patterns in the network, it would provide the insight of relationship between authors such as the distance between authors, number of collaborators or clustering. One of the ways to understand the relationship between the authors in the co-author network is to form ego-network based on specific nodes [18]. According to Newman [19], co-authorship networks have high clustering co-efficient and small average distance between pairs of nodes, thus being "small world" networks. In addition to that, Abbasi [1] also used ego-networks to analyse co-authorship network.

Gasko [11] proposed to use clustering methods to identify groups within a network and measures the authors in relation to their collaboration network. Gasko also set weight to show the number of paper published together by the authors while degree shows the numbers of collaboration. The result of Gasko's study shows that maintaining a strong co-authorship relationship with one primary co-author within a group of linked co-authors is better than to maintain multiple relationships with the same group of linked co-authors. Gasko also found that the productive authors tend to collaborate and often cite colleagues with the same research interests while highly cited authors do not collaborate but cite each other.

2.3 Community in a Network

Community detection is a commonly used approach to explore collaborative patterns in social networks. Usually, communities are constructed on the basis of mutual interest which leads the groups to be connected among themselves through this same interest [28]. Within this community, authors and co-authors share their knowledge with each other and this association displays their common interest in the particular field [2]. Therefore, analysing the authors' community reveal the relationship pattern among the authors as well as its functional and structural properties. However, many researchers denote that determining community in a large network is a challenging task due to the lack of clear definition of the community [9, 16, 17, 29, 30].

Various computational approaches have been constructed to detect communities in the network. Some of these methods have worked efficiently in determining the community namely Walktrap[24], Infomap [25] and Louvain [5]. Walktrap makes use of a random-walk based similarity between vertices and between communities and uses modularity in a hierarchical agglomerative clustering scheme to derive an optimal vertex clustering structure [22]. Infomap also uses the concept of random walk and based on information theory and Louvain magnify the modularity function. All of these approaches are effective for synthetic and real-world benchmarks but when it deals with real data, their behaviour may differ, each one revealing different possible structures. In this study, all three methods were used to identify the authors' community and then by extracting the biggest community, we analysed the networks and compare the results in terms of the relationship of the nodes within the community.

2.4 Clique and Word Frequency

In engineering application, largest clique approach has been used in braiding application from textile engineering [12]. They devel-

oped new models that instruct the machine routing for collision avoiding between the thread-spools from their source to their destination. These are achievable by applying branch and bound algorithm for largest clique problem to compute the controls. Cliques also have been used in various integrated circuit (IC) wafer technology. IC partitioning derived from cluster-based partitioning that recursively collapsing small cliques in the graph shows 49.6% improvement in smaller cut sizes than conventional direct partitioning [7] and give efficient IC design for computing partially specified Boolean functions [23].

On the other hand, implementations of modelling using cliques have been used extensively in Bioinformatics to solve many problems. Food webs that comprised of predator-prey data use cliques to model the ecological network [3]. By searching for cliques in a protein-protein interaction network, it helps to find clusters of protein [26] and explain a bi-clustering problem data in which the clusters are required to be cliques. Using a similar technique, we will find the underlying reason behind the largest clique formed in the co-authorship network.

In brief, this paper consists of an exploratory study on co-authorship network using different social network analysis measures to identify the influential author, the relationship of authors in isolated and biggest community and the reason for forming largest clique among the authors.

3. Dataset and Tools

The dataset is obtained from Gephi sample datasets, a co-authorship network of scientists working on network theory and experiment, as compiled by M. Newman in May 2006 [20] that consist of 1589 vertices and 2742 edges. There are eight columns that comprise source, target, type, Id, label, timeset and weight. Source and target show the direction of the edges. The weight shows the strong or weak relationship between the source and target. In the vertex dataset, there are three columns; id, label and timeset. The Id field is referred to the source field in the edge dataset. The label contains the name of author for the nodes.

In this study, we use R language as a programming language to analyze the network. The tool used in the experiment is R Studio with Igraph package added into the R library. Igraph has a collection of network analysis tools and is free to use.

4. Methodology

The methods used for the explorative experiment are described based on each research question, where each question is transformed into "how" question particularly for this section in order to explain well the methods and for better presentation. Below show the methods discussed for each question.

How to identify the most appropriate centrality measure to detect the influential author in the co-authorship network?

The overall structure of the co-authorship network is sparse with many isolated clusters or nodes scattered over the network. Hence, we focus only on the relationships among the authors in the giant component instead of examining the whole network to determine the appropriate centrality measure for identifying the most influential author. First and foremost, the clustering function in R is applied to calculate the maximal connected components and 396 clusters are identified. The giant component constitutes 379 nodes which represent authors and their connections. The position of the giant component (blue color) in the co-authorship network is shown in Figure 1.

The centrality values by degree, betweenness, closeness, and eigenvector are identified and ranked. The ranking of each centrality values is shown in Table 2. Node removal is performed to see the impact of the loss of high centrality nodes. The network measures namely diameter, average path length, density, reciprocity, transitivity and assortativity are measured for the giant component before and after removing the specific nodes which had the maximum centrality value of the four centrality measures respectively.

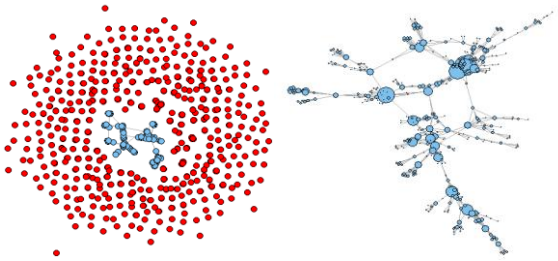


Fig. 1: Giant component (nodes are in blue colour) in the Co-authorship Network of Scientists and the giant component is zoomed and shown at the right side of the picture.

In order to better visualise the relationships of authors after removing the nodes with maximum centrality values individually in each scenario, the networks are filtered with different degree range. The relationship between the authors and network patterns are evaluated. The interpretations are validated by network measurement of the subgraphs of each node with maximum degree, betweenness, closeness, and eigenvector centrality respectively.

How to measure the relationship of authors in an isolated and balanced ego network?

An experiment has been done to answer the research question; Could the relationship of authors be measured in an isolated and balanced ego network? To answer the question, three methods have been used in the experiment. The methods are

I. Cluster the network graph
II. Create ego networks
III. Applying weight onto network edges

Clustering method has been used as it allows better analysis of the network. All nodes are put into a group which normally been defined based on the properties of the nodes or the distance among nodes. Each node will be assigned colour based on their group. As mention by Abbasi [1], centrality values can be used to measure the productivity of scientific knowledge in the network.

For this experiment, four centrality measures that have been used are degree, betweenness, closeness and eigenvector. Nodes with the highest degree, betweenness, closeness and eigenvector will be pulled out from the main network to create their own ego-network. All network edges have their weight. Weight represents the number of papers that being co-written by the authors. The edges' weight will be pulled from the dataset. Higher edge thickness shows there is more than one paper that has been written together.

How to identify the relationship among the scientists in the biggest community?

In order to find the relationship among the scientists in the biggest community, first, the isolated nodes need to be eliminated. Solo nodes inside a network considered as an individual community and in order to reduce this insignificant community that does not have any effect on the network, these nodes are better to be excluded. Omitting the nodes whose degree less than 1 returned a network without solitary vertex.

Next step is to apply community detection algorithms to identify the number of communities that exists in this network. Walktrap, Louvain and Infomap community detection algorithms are used to conduct this step. Walktrap finds 188 communities, Infomap algorithm detects 209 and Louvain finds 175 communities. The objective of this part of research is to find the relationship among the scientists in the biggest community. Hence out of hundreds of communities detected by three algorithms the biggest community was extracted. After extracting the biggest community, different centrality measures are applied to identify the highest centrality node in this network. Once the central nodes of the biggest communities are identified, the community is then plotted and analysed to find the relationship pattern of the scientists in this biggest community.

How to find the core area of research and top publisher of authors in the co-authorship network?

Dataset were loaded into R and transformed into graph data frame using igraph package in R. Vertices label also been loaded into the data frame for later step that is to identify the corresponding co-authors. Next, we used clique function in igraph library to discover the membership and size of the largest cliques in the graph.

In addition, we used Harzing's Publish & Perish software [14] as a data collection tool to scrap and query corresponding membership authors publications data and metadata from Google Scholar. This is a free tool used to look up for scholarly citations from the online sources according to the search metrics (eg: authors name etc.). After the largest clique have been identified, this process was carried out to bring additional attributes to the dataset followed by Word of Cloud techniques that focuses on the frequency of word occurrence in title & publisher of the publication to support our assumptions on the underlying reason of largest complete clique in co-authorship network. We faced some challenges in interpreting the author's name as their first name are ambiguous and some refer to other authors with the same short naming style (refer to Table 1).

Table 1: Authors' Name

No.	Authors (Dataset)	Authors (Finalize)
1	ROTHBERG, J	JM Rothberg
2	GIOT, L	L Giot
3	UETZ, P	P Uetz
4	CAGNEY, G	G Cagney
5	MANSFIELD, T	TA Mansfield
6	JUDSON, R	RS Judson
7	KNIGHT, J	JR Knight
8	LOCKSHON, D	D Lockshon
9	NARAYAN, V	V Narayan
10	SRINIVASAN, M	M Srinivasan
11	POCHART, P	P Pochart
12	QURESHIEMILI, A	AQ Emili
13	LI, Y	Y Li
14	GODWIN, B	B Godwin
15	CONOVER, D	D Conover
16	KALBFLEISCH, T	T Kalbfleisch
17	VIJAYADAMODAR, G	G Vijayadamodar
18	YANG, M	M Yang
19	JOHNSTON, M	M Johnston
20	FIELDS, S	S Fields

We used the following search keywords and Boolean to perform our query:

JR Knight OR "V Narayan" OR "JM Rothberg" OR "P Uetz" OR "L Giot" OR "G Cagney" OR "TA Mansfield" OR "RS Judson" OR "D Lockshon" OR "M Srinivasan" OR "P Pochart" OR "AQ Emili" OR "Y Li" OR "B Godwin" OR "D Conover" OR "T Kalbfleisch" OR "G Vijayadamodar"

We then explored the Word of Cloud techniques that focus on publication title & publisher attributes among authors to find most common number frequency of word occurrence.

5. Result and Discussion

In this section, results are presented based on each analysis. The analysis is matched with the respective research question as discussed in Section 1.

5.1. Centrality Measure Analysis

Table 2 shows "BARABASI, A" ranked first in high degree and eigenvector centrality whereas "HOLME, P" ranked first in high betweenness and closeness centrality.

Table 2: Top 4 Authors Ranked by Degree, Betweenness, Closeness, and Eigenvector Centrality of the giant component of the Co-authorship Network of Scientists

Centrality	Ranking			
	1	2	3	4
Highest Degree	BARABASI, A	JEONG, H	NEWMAN, M	OLTVAI, Z
Highest Betweenness	HOLME, P	JEONG, H	NEWMAN, M	BOGUNA, M
Highest Closeness	HOLME, P	EDLING, C	LILJEROS, F	JEONG, H
Highest Eigenvector	BARABASI, A	JEONG, H	ALBERT, R	OLTVAI, Z

A deeper analysis on this, we found that some authors have high values in one or two centralities but not in other centralities are due to the collaborations with authors from same institution or location. This supports a study by Ahmed et al. [2]. Next, we compare the characteristic and network pattern due to the impact of the loss of high centrality nodes.

Table 3: Summary Statistics of Different Network Measures with and without Nodes with Max Centrality Value Nodes

Network	Removed Vertices	Diameter	Average Path Length	Density	Reciprocity	Transitivity	Assortativity
Giant Component as a whole	-	9.33	6.042	0.0128	1	0.431	-0.082
Giant Component Without Max Degree Node	BARABASI, A	9.33	6.066	0.0124	1	0.448	-0.091
Giant Component Without Max Betweenness Node	HOLME, P	9.75	6.378	0.0126	1	0.434	-0.085
Giant Component Without Max Closeness Node	HOLME, P	9.75	6.378	0.0126	1	0.434	-0.085
Giant Component Without Max Eigenvector Node	BARABASI, A	9.33	6.066	0.0124	1	0.448	-0.091

The diameter of the network remains the same with or without the removal of highest degree and eigenvector centrality node, "BARABASI, A" (refer to Table 3). However, the removal of highest betweenness and closeness centrality node, "HOLME, P" has resulted in an increase of the network diameter. The longer diameter indicates the connection of the network becomes weaker; after the node removal. The information transmission from one node becomes less efficient and it takes more time to reach the other nodes.

The average path length and transitivity of the network has increased for both networks without "BARABASI, A" and "HOLME, P". Compare to the network without "BARABASI, A", the difference in average path length is higher for the network without "HOLME, P". The impact is more significant when "HOLME, P" is removed from the network.

Transitivity or clustering coefficient measures the probability that the adjacent vertices of a vertex are connected. High transitivity in the network without "BARABASI, A" shows the existence of strong ties which lead to the formation of fully connected clusters. Evaluation from the network topology shows network without "HOLME, P" (Figure 3) displays less adjacent vertices connected to vertices compared to the network without "BARABASI, A" (Figure 2). In this network, "HOLME, P" plays important role in bridging different clusters and facilitates the inter-cluster collaborations. Network without "HOLME, P" displays lower transitivity compare to the network without "BARABASI, A".

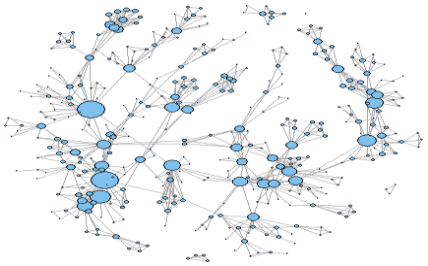


Fig. 2: Giant Component Without Max Degree/Eigenvector Node (without "BARABASI, A")

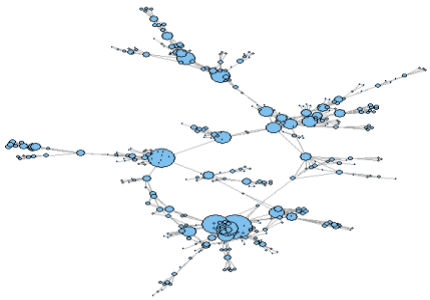


Fig. 3: Giant Component Without Max Betweenness/Closeness Node (without "HOLME, P")

Meanwhile, the reciprocity for all networks is 1. Density is a common measure to show how closely knit a network is. The networks become less dense after the node removal. However, the network without "BARABASI, A" is lower in density compared to the network without "HOLME, P". This proves that "BARABASI, A" position in the network increases the cohesion of his own clusters, whereas "HOLME, P" links to authors in other clusters and hence make the network sparser as illustrated earlier. On the other hand, assortavity is how likely similar nodes tend to connect to each other. High negative value of assortavity in the network without "BARABASI, A" shows that nodes are unlikely to connect to each other compared to the network without "HOLME, P" due to the inter-clusters collaborations as discussed above.

We filtered the networks after performing node removal to better visualise the structure and pattern of the networks. From the network patterns obtained, it is obvious that the impact of removing the node with highest betweenness and closeness centrality that is "HOLME, P" is more significant. The absence of "HOLME, P" cut off connections and collaborations between authors in different clusters that dispersed over the network. Although "BARABASI, A" has highest degree centrality, but he will not link to other high degree nodes in other clusters such as "NEWMAN, M" without "HOLME, P". Network pattern of "BARABASI, A" shows that he frequently collaborate in his own cluster. "BARABASI, A" might be the author with frequent publication, but he is not the most influential author in the co-authorship network. An influential author would be able to facilitate collaborations between authors from different clusters. The network structure without "HOLME, P" appeared to be sparse and less collaborative from the co-authorship network structure. "HOLME, P" with highest closeness centrality also has higher reachability to other nodes in the network. In summary, "HOLME, P" which has the highest betweenness and closeness centrality value is the most influential author in the co-authorship network.

5.2. RQ2: Ego Network Analysis

Figure 4 shows the result of clustering using cluster_walktrap algorithm. Cluster_walktrap algorithm searches a densely connected subgraph to create community via short random walks. There are many overlapping clusters and the largest connected component is located in the center of the network. There are two separate components disconnected from the largest component. Colour of the nodes is sparse all around the network. The total number of cluster formed is 429 and 396 number of (maximal) connected component in the network graph.

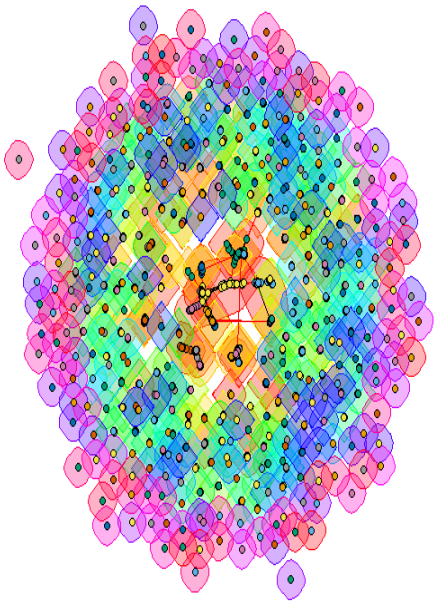


Fig. 4: Cluster by Walktrap algorithm

We picked two nodes that are 'Barabasi, A' (highest degree and eigenvector) and 'Holme, P' (highest betweenness and closeness) to create ego networks. From each of the ego network respectively, we extracted components of one, two and three neighbourhood. The changes in the connection of both nodes when the neighbourhood increases were identified. The result shows that when the neighbourhood increases, the connection between Barabasi, A and other nodes increases and formed a single cluster that is highly connected to each other. Meanwhile, when the neighbourhood of Holme, P increases, the connections grow within Holme, P cluster and other clusters also begin to appear. Next, the network measure of 'Barabasi, A' and 'Holme, P' Ego-Network are calculated separately. Table 4 shows comparative results of both ego network.

Table 4: Comparison Between 'Barabasi, A' and 'Holme, P' as an Ego-Network

Ego	Structure	Diameter	Density	Average Path Length	Reciprocity	Transitivity
'Barabasi, A'	Adjacent nodes to ego with 1 neighbour	2	0.1831933	1.816807	1	0.4001617
	Adjacent nodes to ego with 2 neighbours	3.73333	0.0785777	2.575505	1	0.364094
	Adjacent nodes to ego with 3 neighbours	4.5	0.04597701	3.490421	1	0.4117202
'Holme, P'	Adjacent nodes to ego with 1 neighbour	2	0.328095	1.67619	1	0.5396825
	Adjacent nodes to ego with 2 neighbours	2.69999	0.08095884	2.849842	1	0.4175754
	Adjacent nodes to ego with 3 neighbours	5.4	0.02987421	4.016903	1	0.3586751

5.2.1. Adjacent Nodes to Ego with 1 Neighbour

The result in Table 4 shows the ego network for 'Holme, P' with one neighbour is denser, smaller and higher transitivity than ego network for 'Barabasi, A'. Based on these facts, it can be concluded that 'Holme, P' has a stronger relationship (direct ties) with other nodes in his ego-network and with a strong node (high degree node) from another network compared to Barabasi, A's ego-network. Barabasi, A has the highest degree but needs his co-authors to exert influence in the network.

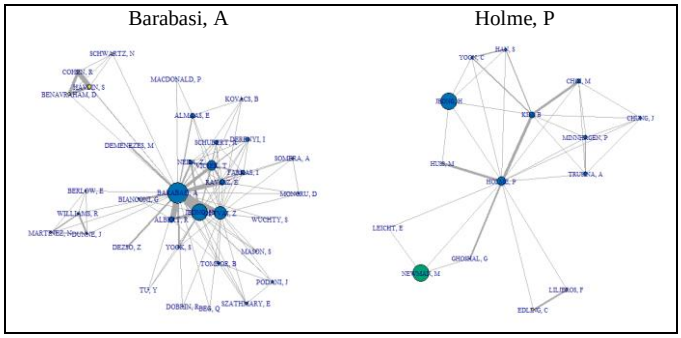


Fig. 5: Adjacent nodes to ego (Barabasi, A and Holme, P) with 1 neighbour

5.2.2. Adjacent Nodes to Ego with 2 Neighbours

The result of this experiment is Barabasi, A's ego-network has a shorter average path but longer diameter, less density and transitivity compared to 'Holme, P' ego-network. It can be concluded that with high betweenness centrality (Holme, P), the probability to have a connection with nodes from the different cluster is higher while the highest degree node (Barabasi, A) more likely to connect to many nodes from the same cluster. This also shows that in order to spread your influence, it is important to maintain links with other important nodes from different clusters in the network than to only have relationships within the same cluster.

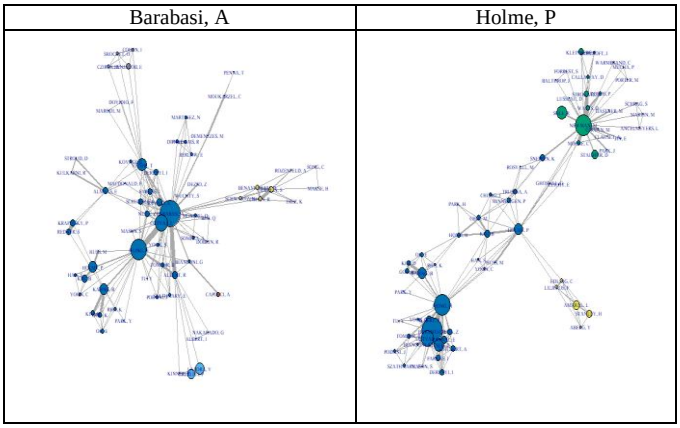


Fig. 6: Adjacent nodes to ego (Barabasi, A and Holme, P) with 2 neighbours

5.2.3. Adjacent Nodes to Ego with 3 Neighbours

There is a significant increase in the number of nodes when 3 neighbours of Barabasi, A is extracted from the Barabasi, A's ego-network. The edge thickness of the connection between nodes is also high indicates the weight where more than one paper has been written together. Meanwhile, Holme, P's 3 neighbourhood network extracts more clusters. As a result of more cluster in the network, 'Holme, P' ego-network has a longer diameter, longer average path, lower density and lower transitivity when compared to Barabasi, P's ego network. It should be noted that in Holme, P's ego network, a connection between clusters is more likely to happen. To summarize this, highest betweenness node is important to synergize with other clusters and more likely able to reach many places in the network while the highest degree node is a productive author with his work.

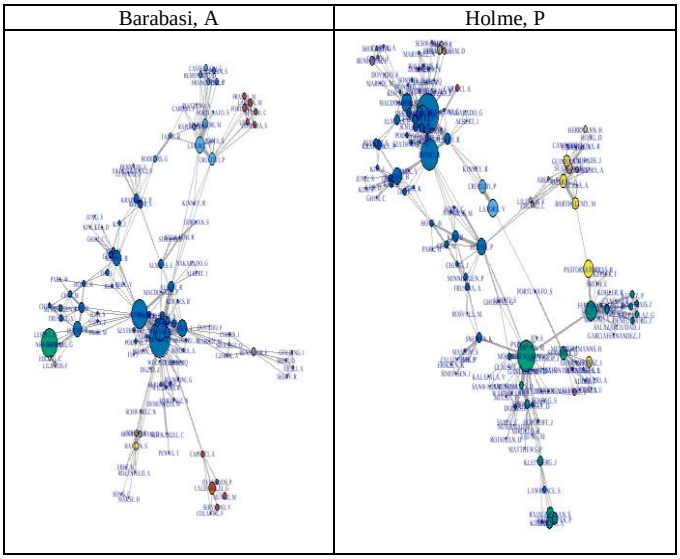


Fig. 7: Adjacent nodes to ego (Barabasi, A and Holme, P) with 3 neighbours

5.3. RQ3: Community Detection Analysis

Three biggest community are identified using three different algorithms. These results are shown in Table 5.

Table 5: Centrality measures in the extracted biggest community identified by three algorithms

Algorithm	Biggest community's size (by node)	Highest Degree Centrality	Highest Betweenness Centrality	Highest Closeness Centrality	Highest Eigenvector Centrality
Walktrap	85	Barabasi, A	Jeong, H	Jeong, H	Barabasi, A
Louvain	88	Barabasi, A	Jeong, H	Jeong, H	Barabasi, A
Infomap	31	Barabasi, A	Barabasi, A	Barabasi, A	Barabasi, A

5.3.1. Largest Community Identified by Walktrap and Louvain Algorithm

After applying this edge betweenness algorithm the biggest community have been extracted. This community consists of 85 vertices. Different centrality measures have been applied in this community to identify the most central node and the relationship with other nodes. It is found that "Barabasi, A" has the highest degree and highest eigenvector centrality on the other hand "Jeong, H" has the highest betweenness and highest closeness centrality. Following the similar procedure, it is found that the biggest community found using Louvain Algorithm has a total of 88 nodes. The same nodes, "Barabasi, A" and "Jeong, H" stood up as the most central nodes.

5.3.2. Largest Community Identified by Infomap Algorithm

Unlike the previous two communities, this community is comparatively very small which has only 31 nodes in it. Moreover, after applying the centrality measure, it is seen that "Barabasi, A" alone acquire the most central node in this community.

After analyzing the above three communities, it is found that in the first two communities "Barabasi, A" has the highest degree and highest eigenvector centrality. However, highest betweenness and closeness centrality goes to "Jeong, H". Even after plotting the community it is visible that "Jeong, H" is connected with the nodes across the community whereas "Barabasi, A" is not connected with the entire community rather only densely connected with a certain group of nodes. On the other hand, in the third community, "Barabasi, A" is well connected with the majority of the nodes, therefore, here "Barabasi, A" acquire the highest value for all centrality measures.

From the first two communities, we can conclude that even though "Barabasi, A" has co-authored with a maximum number of researchers but he is connected to the same group. Which may denote that, "Barabasi, A" has co-authored with researchers who belong to the same institute or in the same research field. Whereas "Jeong, H" has co-authored with multiple groups of linked co-authors who may come from different institutes or from different research field which also supports the findings by [11, 28].

In the small community found using Infomap algorithm, which consists of only 31 nodes, authors did not create any subgroup except few cliques. The community was plotted and it is visible that most of the researchers have co-authored with "Barabasi, A" and "Oltvai, A". Surprisingly, "Jeong, H" who had the highest betweenness and closeness centrality in previous two community has sparse co-authorship with the researchers in this community. By analyzing the third community it can be presumed that all of these researchers are from the same institute or from same research field where "Barabasi, A" and "Oltvai, A" are very prominent researchers. Whereas, "Jeong, H" may not from the same institute or he is not expert in this research field, therefore "Jeong, H" is not as a central node in this community as in the previous two communities.

5.4. RQ4: Largest Clique and Word Frequency Analysis

The result shows that there were 741 maximal cliques found in the entire co-authorship network with the largest or maximum clique of a complete subgraph have membership size of 20 authors. For better visualization, we extracted the largest clique's subgraph from the main graph as shown in Figure 8. The red vertices' size represents the degree centrality relative to the main graph.

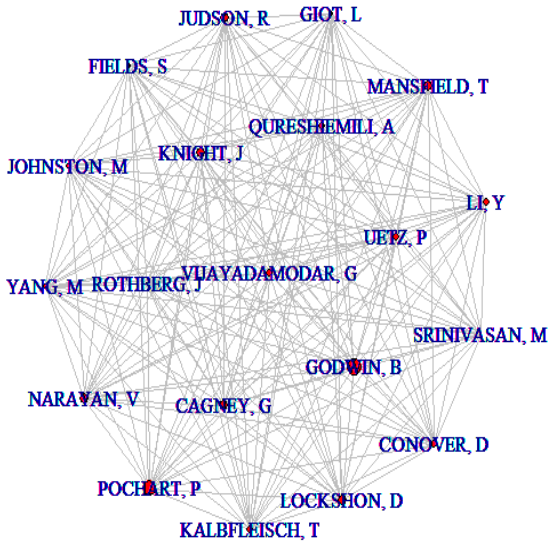


Fig. 8: Largest Cliques Membership

Next, we used these author memberships to find their respective publications title & publisher from Google Scholars search engine using Harzing's Publish & Perish software. Meanwhile, Table 6 shows part of publication Title and Publisher query results.

Table 6: Part of Publication Title & Publisher Query Results

Cites	Authors	Title	Year	Source	Publisher
256	K Dandekar, BI Raju…	3-D finite-element models of human and monkey fingertips to investigate the mechanics of tactile sense	2003	Journal of …	asmedigitalcollection.asme.org
50	MA Srinivasan	What is haptics?	1995	Laboratory for Human and Machine Haptics: The …	184.72.102.18
356	…, K Bailey, Y Balagurunathan, JM Rothberg…	Acidity generated by the tumor microenvironment drives local invasion	2013	Cancer research	AACR

Table 7: Total number of publications of each author in the largest clique membership

Authors	Total Publication
JM Rothberg	129
L Giot	17
P Uetz	163
C Cagney	60
TA Mansfield	29
RS Judson	130
JR Knight	45
D Lockshon	24
V Narayan	74
M Srinivasan	34
P Pochart	11
AQ Emili	4
Y Li	*NA*
B Godwin	23
D Conover	6
T Kalbfleisch	35
G Vijayadamodar	21
M Yang	59
M Johnston	149
S Fields	107
	1120

The results in Table 7 shows the total number of publications of each author in the largest clique membership. We had also removed duplicated publication from the total results. Note that author Y Li was marked as *NA* or not available as the query returned ambiguous result and it is impossible to validate which Y Li it referred to.

Based on the frequent word occurrence for publication titles (refer to Figure 9, left), within the authors' papers, diabetes keyword has the highest occurrence (62 out 3572) followed by protein, human, and acid. The first 10 keywords occurrence that is shown in Figure 9, left are usually used in bioinformatics fields; hence we can conclude that the co-authorship memberships in the largest cliques comprised of authors from bioinformatics fields.

Findings from the analysis (refer to Figure 9, right) shows that majority of authors publish their publication in Elsevier with 190 publications followed by library, online, wiley, springer, and nature.com. We analysed the authors' favourite publisher and their field of interest. Elsevier is information and analytics publishing company originated from Netherland with the main contribution in informatics in scientific, technical, and medical fields. The word library, online, and wiley are highly correlated with each other showing that it's actually a phrase of Wiley Online Library which is a well-known journal with a vast collection of research papers covering health, life and physical science as well social science. The same goes for Springer. Next publisher is Nature that is the home of high impact medical and scientific information across life, physical, chemical and applied sciences.



Fig. 9: Frequency of words using word cloud. Publication Titles (Left) and Publishers (Right).

6. Conclusion and Future Work

Social network analysis able to provide researchers with deep insight on how each node or vertex interact with others. We applied network topology analysis in evaluating the centrality values to determine the influential node. Overall, betweenness centrality is the most appropriate measure, followed by the degree centrality to apply in this co-authorship network in order to find the most influential author. However, interpretation of centrality measure could be different with regards to the specific property and function of a network.

Our network neighbourhood analysis also shows that the relationship can be measured by using ego network extraction. When the neighbourhood of an ego (node with highest degree centrality) increases, the connections between that ego and other nodes increases and forms a single cluster that is highly connected to each other. Meanwhile, when the neighbourhood of an ego (node with highest betweenness centrality) increases, the connections grow within that ego's cluster and other clusters also begin to attach. By using ego networks, insight can be found on a different pattern of relationships between the highest betweenness node or highest degree node with other nodes.

After analysing the extracted biggest community using three different algorithms, it can be concluded that authors and co-authors who are directly connected belong to the same cluster. They tend to form several subgroups inside the community either because they are from the same institute or have common research area. The experiment shows that the highest degree centrality author is connected to the most number of co-authors in a community due to the same reason. Whereas, the highest betweenness centrality author may work on many research areas hence authors from different research area are linked with that researcher. On the other hand, when the community size is relatively small then there is no noticeable subgroup is formed. That also implies that either all of the authors are devoted to the same research area or they are from the same institution.

Social network analysis involved complex studies to analyse the behaviour of authors and collaboration between authors in the fields that they work together. This in turn gave us the underlying reason for how the largest cliques in the co-authored publication are formed and what is their research area and favourite publishers. In this research, the limitation of the dataset hinders the capability to study the relationship based on homophily. The centrality measure did influence the relationship between the nodes and more attributes are needed to study whether homophily can state the relationship in the network and this will be our future work.

Acknowledgement

The authors are grateful to the UTM AIS Social Network Analytics subject instructor in 2018 and also to UTM KL for providing access to the online database library for our research findings.

References

[1] A. Abbasi, K. S. K. Chung, and L. Hossain, "Egocentric analysis of co-authorship network structure, position and performance," *Information Processing & Management*, vol. 48, no. 4, pp. 671-679, 2012.

[2] T. Ahmed, A. Ahmed, M. Ali, and M. Kamran, "Analysis of co-authorship in computer networks using centrality measures," in *Communication, Computing and Digital Systems (C-CODE), International Conference on*, 2017, pp. 54-57: IEEE.

[3] S. Allesina, A. Bodini, and C. Bondavalli, "Ecological subsystems via graph theory: the role of strongly connected components," *Oikos*, vol. 110, no. 1, pp. 164-176, 2005.

[4] V. Arnaboldi, R. I. Dunbar, A. Passarella, and M. Conti, "Analysis of co-authorship ego networks," in *International Conference and School on Network Science*, 2016, pp. 82-96: Springer.

[5] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.
[6] P. Bonacich, "Power and centrality: A family of measures," *American journal of sociology*, vol. 92, no. 5, pp. 1170-1182, 1987.
[7] J. Cong and M. L. Smith, "A parallel bottom-up clustering algorithm with applications to circuit partitioning in VLSI design," in *Proceedings of the 30th international Design Automation Conference*, 1993, pp. 755-760: ACM.
[8] J. F. Delgado-Garcia, A. H. Laender, and W. Meira, "Analyzing the Coauthorship Networks of Latin American Computer Science Research Groups," in *Web Congress (LA-WEB), 2014 9th Latin American*, 2014, pp. 77-81: IEEE.
[9] S. Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3-5, pp. 75-174, 2010.
[10] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social networks*, vol. 1, no. 3, pp. 215-239, 1978.
[11] N. Gaskó, R. I. Lung, and M. A. Suciu, "A new network model for the study of scientific collaborations: Romanian computer science and mathematics co-authorship networks," *Scientometrics*, vol. 108, no. 2, pp. 613-632, 2016.
[12] T. Grünert, S. Irnich, H.-J. Zimmermann, M. Schneider, and B. Wulfhorst, "Finding all k-cliques in k-partite graphs, an application in textile engineering," *Computers & operations research*, vol. 29, no. 1, pp. 13-31, 2002.
[13] H. Guércio, V. Ströele, J. M. N. David, R. Braga, and F. Campos, "Topological analysis in scientific social networks to identify influential researchers," in *Computer Supported Cooperative Work in Design (CSCWD), 2017 IEEE 21st International Conference on*, 2017, pp. 287-292: IEEE.
[14] A.-W. Harzing, "Publish or perish," 2007.
[15] M. U. Ilyas and H. Radha, "Identifying influential nodes in online social networks using principal component centrality," in *Communications (ICC), 2011 IEEE International Conference on*, 2011, pp. 1-5: IEEE.
[16] E. Y. Li, C. H. Liao, and H. R. Yen, "Co-authorship networks and research impact: A social capital perspective," *Research Policy*, vol. 42, no. 9, pp. 1515-1530, 2013.
[17] P. Magalingam, S. Davis, and A. Rao, "Using shortest path to discover criminal community," *Digital Investigation*, vol. 15, pp. 1-17, 2015.
[18] P. Magalingam, A. Rao, and S. Davis, "Identifying a criminal's network of trust," in *Signal-Image Technology and Internet-Based Systems (SITIS), 2014 Tenth International Conference on*, 2014, pp. 309-316: IEEE.
[19] M. E. Newman, "Modularity and community structure in networks," *Proceedings of the national academy of sciences*, vol. 103, no. 23, pp. 8577-8582, 2006.
[20] M. E. J. Newman, *Phys. Rev. E 74, 036104*, 2006.
[21] Q. Niu, A. Zeng, Y. Fan, and Z. Di, "Robustness of centrality measures against network manipulation," *Physica A: Statistical Mechanics and its Applications*, vol. 438, pp. 124-131, 2015.
[22] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, and P. Spyridonos, "Community detection in social media," *Data Mining and Knowledge Discovery*, vol. 24, no. 3, pp. 515-554, 2012.
[23] M. C. Paull and S. H. Unger, "Minimizing the number of states in incompletely specified sequential switching functions," *IRE Transactions on Electronic Computers*, no. 3, pp. 356-367, 1959.
[24] P. Pons and M. Latapy, "Computing communities in large networks using random walks," in *International symposium on computer and information sciences*, 2005, pp. 284-293: Springer.
[25] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proceedings of the National Academy of Sciences*, vol. 105, no. 4, pp. 1118-1123, 2008.
[26] V. Spirin and L. A. Mirny, "Protein complexes and functional modules in molecular networks," *Proceedings of the National Academy of Sciences*, vol. 100, no. 21, pp. 12123-12128, 2003.
[27] L. Sun and I. Rahwan, "Coauthorship network in transportation research," *Transportation Research Part A: Policy and Practice*, vol. 100, pp. 135-151, 2017.
[28] K. Sutaria, D. Joshi, C. Bhensdadia, and K. Khalpada, "An adaptive approximation algorithm for community detection in social network," in *Computational Intelligence & Communication Technology (CICT), 2015 IEEE International Conference on*, 2015, pp. 785-788: IEEE.
[29] X. Zhang *et al.*, "Overlapping community identification approach in online social networks," *Physica A: Statistical Mechanics and its Applications*, vol. 421, pp. 233-248, 2015.
[30] Z. Zhao, S. Feng, Q. Wang, J. Z. Huang, G. J. Williams, and J. Fan, "Topic oriented community detection through social objects and link analysis in social networks," *Knowledge-Based Systems*, vol. 26, pp. 164-173, 2012.