

CASE STUDY

Available Online at www.jgrcs.info

CASE STUDY – CENTRALITY MEASURE ANALYSIS ON CO-AUTHORSHIP NETWORK

Dr. V. Umadevi¹

¹Department of Computer Science and Engineering, BMS College of Engineering, Bangalore, Karnataka, India
Email: umadevi.cse@bmsce.ac.in

Abstract: Study of social networks reveal communication patterns which are of interest to researchers. Co-authorship network is one type of a social network. These networks represent the publication work carried out by researchers. Co-authorship networks analysis is useful in understanding the structure of scientific collaborations and status of individual authors. Centrality measure calculation is one of the many tasks of social network analysis. Focus of this paper work is on centrality measure analysis carried out on the co-authorship network using Gephi, a social network analysis tool.

INTRODUCTION

Co-authorship networks are an important class of social networks. Analysis of these networks reveals features of academic communities which help in understanding collaborative scientific works and identifying the prominent researchers. Structured analyses of scientific publications and visualizations synthesizing the results can help all interested stakeholders in the scientific process to be more aware about content and connections and thus may serve as decision support [1]. Significance of analyzing co-authorship networks are:

- Co-authorship networks will be beneficial to researchers in maintaining a social relationship with their colleagues (the co-authors) or to explore papers published by their colleagues with some other co-author.
- Analysis of co-authorship networks reveals the contribution structures of a scientific community by disclosing the collaboration of authors in terms of co-authoring papers.

Focus of this work is analyses of structural properties in the co-authorship network using centrality measures. A co-authorship network of scientists working on network theory was considered as a case study. Gephi, an open source social network analysis tool was used for extraction of centrality measures from co-authorship network to rank the authors.

CO-AUTHORSHIP NETWORK

Co-authorship network is a network which is used to express the existence of co-authorship relation between authors of scientific papers. Co-authorship relations are relations representing whether an author has written a paper with another author in the past. A researcher's publication data often reflects his/her research interests and their social relations [2]. Two scientists were considered connected if they had authored a paper together. In Co-authorship

network, nodes represent authors of the paper and an edge exists between two nodes if authors have co-authored a paper together.

Representation of Co-authorship Network

We make use of a simple network model of co-author which is an undirected, binary graph G , in which each edge represents a relationship between co-authors. For example, consider a research Paper 1, the authors of which are A1, A2 and A3. Similarly consider a research Paper 2 which is co-authored by authors A1, A2 and A4. The Co-authorship networks of these authors are shown in Fig. 1.

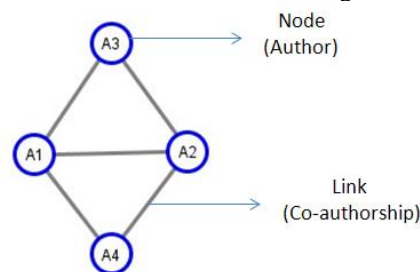


Figure 1. Sample of co-authorship network

GEPHI TOOL

Gephi is an interactive visualization and exploration platform for all kinds of networks and complex systems, dynamic and hierarchical graphs [3]. This tool aids to explore and understand graphs. The goal is to help data analysts to make hypothesis, intuitively discover patterns, and isolate structure singularities or faults during data sourcing. It is a complementary tool to traditional statistics, as visual thinking with interactive interfaces is now recognized to facilitate reasoning. This is software for exploratory data analysis, a paradigm which appeared in the field of research for visual analytics. This tool can handle networks up to 50,000 nodes and 1,000,000 edges. Gephi provides state-of-the-art layout algorithms, both for efficiency and quality. The statistics and metrics framework

offer the most common metrics for social network analysis and scale-free networks.

NETSCIENCE CO-AUTHORSHIP NETWORK

A co-authorship network of scientists working on network theory and experiment, as compiled by M. Newman in May 2006 is being used in this paper for analysis. The network was compiled from the bibliographies of two review articles on networks, M. E. J. Newman, SIAM Review 45, 167-256 (2003) and S. Boccaletti et al., Physics Reports 424, 175-308 (2006), with a few additional references added by hand. The version given here contains all components of the network, for a total of 1589 scientists (or nodes), and a total of 2742 links (or edges) [4].

CENTRALITY MEASURES

Social Network Analysis (SNA) has been increasingly used as a structured way to analyze the extent of informal relationship among people, teams, departments, or even organizations, within various formally defined groups. SNA makes visible these otherwise invisible patterns of interaction, to identify important groups in order to facilitate effective collaboration [5]. Various social network analysis metrics such as centrality measures have been applied to co-authorship network to identify the prominent scientists.

The status of an author is usually expressed in terms of its centrality, i.e. a measure of how central the author is to the network graph. Central authors are well connected to other authors and metrics of centrality will therefore attempt to measure an author’s degree (number of in- and out-links), average distance to all other authors, or the degree to which geodesic paths (or shortest paths) between any pair of authors passes through the author [6]. There are four measures of centrality that are widely used in network analysis: Degree centrality, Betweenness, Closeness, and Eigenvector centrality.

Degree Centrality:

Degree centrality equals to the number of ties that a node has with other nodes in the network graph. The equation which is used to express the degree centrality is as follows:

$$C_D(n_i) = d(n_i)$$

where $d(n_i)$ is the degree of node n_i

Nodes with higher degree or more connections are more central to the structure and tend to have a greater capacity to influence others [7].

Betweenness Centrality:

Betweenness centrality is based on the number of shortest paths passing through a node. Nodes with a high betweenness play the role of connecting different groups. Geodesic or the shortest path is the path between a pair of nodes which involve a minimum number of nodes in between, which connect the two nodes. In the following formula, g_{jik} is all geodesics linking node j and node k which pass through node i ; g_{jk} is the geodesic distance between the nodes of j and k .

$$C_B(n_i) = \sum_{j,k \neq i} \frac{g_{jik}}{g_{jk}}$$

In social networks, nodes with high betweenness are the brokers and connectors who bring others together. Being between means that a node has the ability to control the flow of knowledge between most others. Individuals with high betweenness are the pivots in the network knowledge flowing. The nodes with highest betweenness also result in the largest increase in typical distance between others when they are removed.

Closeness Centrality:

A more sophisticated centrality measure is closeness which emphasizes the distance of a node to all others in the network by focusing on the geodesic distance from each node to all others. Closeness can be regarded as a measure of how long it will take for the information to spread from a given node to others in the network. Closeness centrality focuses on the extensity of influence over the entire network. In the following equation, $C_c(n_i)$ is the closeness centrality, and $d(n_i, n_j)$ is the distance between two nodes in the network.

$$C_c(n_i) = \sum_{i=n}^N \frac{1}{d(n_i, n_j)}$$

Eigenvector Centrality:

Eigenvector centrality is a measure of the influence has a node in a network. It assigns relative scores to all nodes in the network based on the well known principle that connections to high-scoring nodes contribute more to the score of the node in the question than equal connections to low-scoring nodes [8]. In general, connections to people who are themselves influential will lend a person more influence than connections to less influential people. If we denote the centrality of node i by x_i , then we can allow for this effect by making x_i proportional to the average of the centralities of i ’s network neighbors:

$$x_i = \frac{1}{\lambda} \sum_{j=1}^n A_{ij} x_j$$

where λ is a constant. Defining the vector of centralities $x = (x_1, x_2, \dots)$, we can rewrite this equation in matrix form as

$$\lambda x = Ax$$

and hence we see that x is an eigenvector of the adjacency matrix with eigenvalue λ . Assuming that we wish the centralities to be non-negative, it can be shown that λ must be the largest eigenvalue of the adjacency matrix and x the corresponding eigenvector.

The eigenvector centrality defined in this way accords each node a centrality that depends both on the number and the quality of its connections: having a large number of connections still counts for something, but a node with a smaller number of high-quality contacts may outrank one with a larger number of mediocre contacts. Eigenvector centrality turns out to be a revealing measure in many situations [9].

EXPERIMENT SETUP AND RESULTS

The objective of this experiment is to perform centrality measure analysis on the netscience co-authorship network described in earlier section. Fig. 2 shows a block diagram of the experiment setup. Initially the co-authorship network data is loaded into the Gephi tool. Then on, the calculation of centrality measures is carried out. Further using the functions of Gephi tool the centrality measurements are calculated to rank the authors in the last step of the process.

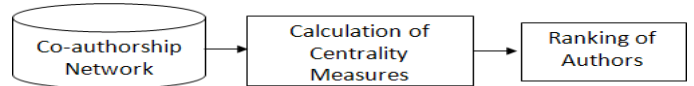


Figure 2: Block diagram of experimental setup

Visualization of Co-authorship network based on Degree centrality measure by Gephi tool is shown in Fig. 3. In this figure, the expanded view shows author, BARABASI A, with a bigger size of the node than the other nodes in the network. This bigger size of the node is the result of higher degree centrality value as compared to all other authors in the network.

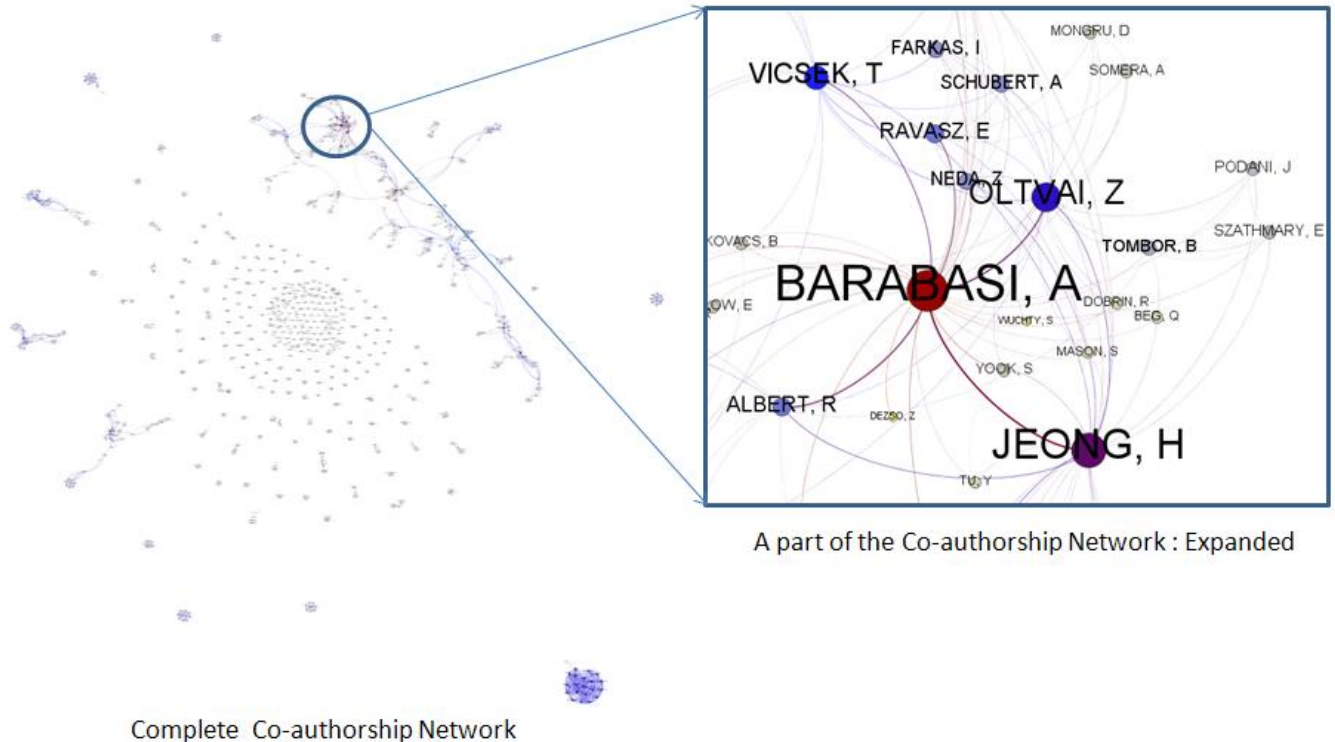


Figure 3. Structure of co-authorship network visualized using Gephi tool

Fig. 4 shows the degree distribution of all authors in the network. From the graph we see that the number of authors with degree two is high, which shows that the number of papers with two authors are highest. Maximum number of contacts an author has in his network is 34. Number of authors with zero degree are 128 which shows there are 128 papers with single authorship.

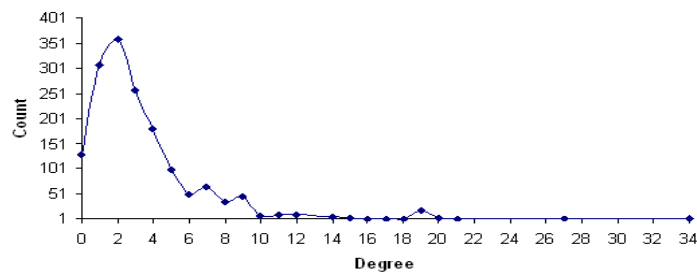


Figure 4. Degree distribution of authors

Table I shows the top ten authors ranked on Degree, Betweenness, Closeness and Eigenvector centrality measures. Degree centrality measures the author’s collaboration scope, closeness centrality measures the author’s position and virtual distance with others in the field, and betweenness centrality measures the author’s importance to other authors’ virtual communication. Betweenness Centrality is a measure of the influence an author has over the spread of information anywhere in the network, and indeed some high-ranking authors play the important role of a link in connecting different groups.

Eigenvector centrality is a measure of the importance of a node in a network. Here, an author is considered important if he/she is connected to other important authors. In the analysis, an author with a small number of influential contacts may outrank one with a larger number of mediocre contacts.

Analyzing the degree centrality of the netscience co-authorship network, the researcher who is most well-connected in the network science community is Barabasi A. Similarly based on Betweenness centrality and Closeness

centrality top ranked authors are Newman M and Baiesi M. Three of the authors Uetz P, Cagney G and Mansfield T are found to have high score of Eigenvector centrality.

Table: 1 Top ten ranked authors based on centrality measures

Rank.	Author Name	Degree	Name	Betweenness Centrality	Name	Closeness Centrality	Name	Eigenvector Centrality
1	Barabasi A	34	Newman M	28300.56	Baiesi M	10.79	Uetz P	1.00
2	Jeong H	27	Pastorsatorras R	24592.77	Corral A	9.80	Cagney G	1.00
3	Newman M	27	Moreno Y	20379.79	Paczuski M	9.79	Mansfield T	1.00
4	Oltvai Z	21	Sole R	19249.90	Gregoire G	9.07	Giot L	0.99
5	Young M	20	Boccaletti S	18200.00	Ye N	9.07	Judson R	0.99
6	Uetz P	20	Jeong H	17858.00	Liu Z	9.06	Knight J	0.99
7	Cagney G	20	Holme P	16506.04	Hari R	9.02	Lockshon D	0.99
8	Mansfield T	20	Caldarelli G	15786.01	Ilmoniemi R	9.02	Narayan V	0.99
9	Alon U	19	Bianconi G	12460.58	Knuutila J	9.02	Srinivasan M	0.99
10	Boccaletti S	19	Capocci A	12446.91	Lounasmaa O	9.02	Pochart P	0.99

CONCLUSION

In this paper centrality measure analysis carried out on netscience co-authorship network was deliberated. Analysis results assisted in identifying invisible patterns in the co-authorship network i.e., relationship between authors shown by visualization and top ranking authors.

Analyzing co-authorship information on a larger database of scientific publications will assist in identifying groups of people who work closely together. Focusing future research work on categorization and ranking of authors based on their area of research work will assist other authors to identify main stakeholders of their interested research domains. This will aid in strengthening and improving scientific collaboration work. Semantic analysis on larger co-authorship network for categorization and ranking of authors by machine learning algorithms will be fertile.

REFERENCES

[1]. Wolfgang Reinhardt, Christian Meier, Hendrik Drachsler and Peter Sloep, "Analyzing 5 years of EC-TEL proceedings," Proceedings of the 6th European conference on Technology enhanced learning: towards ubiquitous learning, Springer-Verlag Berlin, pp. 531-536, 2011.

[2]. Yang , Au Yeung, Ching Man, Weal, Mark J. and Davis, Hugh, "The Researcher Social Network: A Social Network Based on Metadata of Scientific Publications," Proceedings of WebSci'09: Society On-Line, Athens, 18 - 20 Mar 2009.

[3]. Gephi tool, <https://gephi.org/>

[4]. Datasets - UCINET Software, <https://sites.google.com/site/ucinetsoftware/datasets>

[5]. F. Cheong and B. J. Corbitt, "A Social Network Analysis of the Co-Authorship Network of the Pacific Asia Conference on Information Systems from 1993 to 2008," Proceedings of PACIS, pp.23-23, 2009.

[6]. Xiaoming Liu, Johan Bollen, Michael L. Nelson, Herbert Van de Sompel, "Co-authorship networks in the digital library research community," Journal of Information processing and Management, vol. 41, pp. 1462-1480, June, 2005.

[7]. Erjia Yan, Ying Ding, "Applying centrality measures to impact analysis: A coauthorship network analysis," Journal of the American Society for Information Science and Technology, Vol. 60, No. 10, pp. 2107-2118, 2009.

[8]. A. Noori, "On the relation between centrality measures and consensus algorithms," Proceedings of International Conference on High Performance Computing and Simulation, pp. 225 – 232, July 2011.

[9]. M. E. J. Newman, "Mathematics of networks," in The New Palgrave Encyclopedia of Economics, 2nd edition, L. E. Blume and S. N. Durlauf (eds.), Palgrave Macmillan, Basingstoke (2008).

Short Bio Data for the Author



Dr. V. Umadevi is an Associate professor in the Department of Computer Science and Engineering, BMS College of Engineering, Bangalore, India, having seven years of teaching and five years of research experience.