Extraction, Visualisation and Analysis of Co-Authorship Based Academic Social Networks

Tasleem Arif

Department of Information Technology,
Baba Ghulam Shah Badshah University Rajouri,
Jammu & Kashmir, India

***Abstract*— In an online social network environment we establish relationships by sharing status, by way of likes, or tweets and retweets. However, these relationships are casual whereas the relationship established through co-authorship is much more formalized. Through this co-authorship relationship, researchers form academic social networks. In order to study these networks the co-authorship data has to obtained and used. Digital libraries like DBLP, Microsoft Academic Search, etc. provide a rich source of co-authorship information on the Internet. In addition to these digital libraries institutional websites also prove to be a rich source of co-authorship information of people working with that institution. Analysis of this co-authorship relationship provides a whole lot of information about authors and research activities carried out in an institution. In this paper we use social network analysis metrics to study these academic social networks obtained from the underlying co-authorship relationship. We obtained and analyzed social network both at institutional as well as individual author level to understand their research collaborations. It was observed that at the institutional level people have very few collaborations with people within their organization.**

***Keywords*— Extraction & Visualisation, Academic Social Network, Co-authorship, Digital Libraries.**

I. Introduction

The advent of Internet, particularly Web 2.0 has changed the way we live, communicate and maintain relationships. In fact researchers also seem to take advantage of the advances in Information and Communication Technology to enhance their research collaborations [1] and advances in web technologies have been deriving developments at the research front [2]. Online literature management services or digital libraries like DBLP, CiteSeer, Microsoft Academic Search, Google Scholar, etc. store a lot of data about authors and their associated publications. These digital libraries provide structured information about co-authorship relationships in addition to other publication attributes. This data can be obtained either as static snapshots or in a dynamic manner using Web mining techniques.

It is necessary to understand the research collaborations from social network analysis point of view because such an analysis answers various important questions related to collaboration patterns, flow of knowledge, etc. Although there are other form of academic collaborations but co-authorship has proved to be the most documented and tangible form of research collaboration [3]. Co-authorship relationship is a result of jointly writing a research publication and to answer the above questions and understand these collaborations one needs to focus on joint publications [4]. Thus we have a graph or more specifically a social network with authors of these publications as nodes and the co-authorship relationship between them as edges. In order to analyse and understand this form of collaboration one needs to have disambiguated publications data [5].

In this work we obtained publications data from institutional websites for understanding the research collaborations among people of a particular institution and publications data from DBLP and Google Scholar was obtained to understand the research collaborations of individual authors. From this data we obtained co-authorship relationship from the downloaded publications after these publications were disambiguated. There are a number of techniques that have been proposed in the literature for the purpose of author name disambiguation. The reader can refer to [6], [7] for a detailed survey and discussion about these name disambiguation techniques. We performed the publications disambiguation task using [8].

The rest of this paper is organized as follows: In section II we present the related work; Section III presents a brief idea about various social network analysis metrics; Section IV presents various aspects of social network extraction, visualization and analysis. In Section V we conclude the work and give some future directions.

II. Related Work

The use of Social Network Analysis to study research collaborations is not new phenomenon. In fact it has been used a number of times to study the collaborations and collaboration patterns between institutions at local, national and international level [1]. Studies conducted in this field [1] claim that research activities play an important role in creation as well as dissemination of new knowledge and collaborative research has proved to enrich scientific discoveries as well as new innovations and development of new industries [4].

The goal of social network analysis in research collaboration is to analyze the underlying structure of a social network formed by co-authorship relationship or any other academic relationship to infer knowledge about an individual researcher or a research group or an institution, etc. Some studies have tried to understand the structure and pattern of research collaboration through co-authorship relationship using social network analysis measures. Study of cooperation through co-authorship relationship using social network analysis measures has been addressed in various domains and through various case studies: in the field of Chinese humanities and social science [9], in DBLP listed conferences viz. KDD, VLDB, ICML and WWW [10], in biological medicine, physics, and computer science [11], in papers published in Scientometrics journal [12].

In addition to answering questions related to individuals social network analysis measure help understand institutional related questions, e.g. one can have an idea about the health of an institution, collaboration pattern within an institution, etc. [3]. Good health indicates that an institution is growing with the passage of time, producing more research papers, is alive having more collaborations, attracting more grants, etc. [3]. Such groups and institutions can be potential candidates for starting a new research project or for other research related activities.

III. Social Network Analysis Metrics

There are different levels of social network analysis: actor; dyadic; triadic; subset; or network. Centrality and prestige analyse the social network at actor level, whereas distance and reachability at dyadic level, balance and transitivity at triadic level, and cliques, cohesive subgroups, components at subset level. At network level metrics like connectedness, diameter, centralization, density, prestige, etc. are used for analysis purposes. Social network metrics such as degree centrality, betweenness centrality, closeness centrality and network centrality; average degree; clustering co-efficient; density; and characteristic path length play a very important role in the analysis academic or research collaboration from a network perspective.

Several network analysis measures as proposed in [13] can be used to indentify influential nodes and discover community structures of the extracted social networks. We are interested in capturing the internal connectivity as well as attributes of key nodes in the network. Centrality measures like Degree Centrality, Closeness Centrality, Betweenness Centrality, Eigenvector Centrality, Katz Centrality and Alpha Centrality play an important role in graph theory and network analysis to measure the importance or prestige of actors or nodes in a network[1].

Degree Centrality of a node in the network is the number of links incident on it and is used to identify the nodes that have highest number of connections in the network. It a direct measure and does not takes into account the importance of the incident nodes. However eigenvector centrality takes care of the importance of incident nodes as well. Having connections with other important nodes in the network gives a node a higher value of eigenvector centrality. Betweenness centrality measures the fraction of all shortest paths that pass through a given node. Nodes with high betweenness centrality are considered central and indispensable to the network due to their role in the flow of

[1] Centrality: www.wikipedia.com/centrality

information in the network. Nodes with the high betweenness act as gate keeper. Analytical results obtained [11] testify that in an academic social network actors (scientists in this case) having high value of betweenness centrality in a network play a positive role in advancing scientific cooperation.

Clustering coefficient is a measure of the connectedness of a node's neighbourhood and is directly proportional to the degree of connectedness of its neighbours. The clustering co-efficient of a network as expressed in [14] is equal to the average of the clustering co-efficient of all the nodes in the network. It indicates the degree to which nodes in a network tend to cluster together and it is therefore considered to be a good measure if a network demonstrates "small world" behaviour [14]. Stanley Milgram's [15] theory of the "*6 Degree of Separation*" utilises the average path length metric. A graph is considered small world if its average clustering coefficient is significantly higher than a random graph constructed from the same set of vertices.

The average degree of all the nodes in the network is a measure of how collaborative the authors are. The Density of a graph quantifies the number of connections between various actors in the network. The graph is considered dense if the number of edges in the graph approaches the maximal number of edges which one can have in that graph and sparse otherwise.

IV. Social Network Extraction & Analysis

The social networks that we are interested to analyse in this work are co-authorship based academic social networks. Two types of social networks viz. institutional academic social networks and egocentric academic social networks were extracted and analysed. The social network graphs presented in this section provide a candid picture of the cooperation through co-authorship relationship between different authors. The institutional network graphs have been obtained from the publications data obtained from the Websites of the institution under consideration whereas egocentric social networks that have been presented in this section have been obtained from the disambiguated publications data obtained from DBLP and Google Scholar Author Search facility using Web mining techniques.

The institutional or departmental academic social networks presented in this section provide an overview of internal collaborations. For the purpose of such a visualization and analysis from institutional perspective we extracted co-authorship social networks from publications data of four Indian Institutes of Technology (IITs) under consideration. Publications data of people working for Computer Science Departments over a period of seven years from 2005 to 2011 was obtained from Websites of these institutions. A total of 1017 co-authored publications of 107 authors and 2375 co-authorship relationships formed the basis of this analysis. Analysis of these networks and the values they returned for various metrics under consideration presented glaring picture about collaboration between people within these department.

Table I

Values Of Various Local Co-Authorship Graph Metrics

IIT	**Delhi**	**Kanpur**	**Kharagpur**	**Madras**
Vertices	31	26	27	22
Total Edges	4	6	39	13
Connected Components	28	20	6	10
Maximum Vertices in a Connected Component	3	4	21	13
Maximum Edges in a Connected Component	3	3	38	13
Average Geodesic Distance	0.61539	0.96552	2.40899	2.52071
Graph Density	0.0086022	0.0184616	0.1111112	0.0562771
Average Degree	0.258	0.462	2.889	1.182
Average Betweenness Centrality	0.000	0.154	12.037	6.136
Average Closeness Centrality	0.113	0.158	0.090	0.019
Average Eigenvector Centrality	0.032	0.038	0.037	0.045
Median PageRank	0.161	0.346	0.790	0.537
Average Clustering Coefficient	0.097	0.000	0.349	0.027

Table-I lists the values of various important network metrics obtained for the departmental co-authorship networks of the four IITs under consideration. The departmental co-authorship network graphs of IIT-Delhi, IIT-Kanpur, IIT-Kharagpur and IIT-Madras have been presented in Figure 1, 2, 3, and 4, respectively.

Table-I presents some interesting facts about the research collaborations of people working in the departments under consideration of these institutions. Graph Density specifies the degree of connection between people of that Department. From the Graph Density values it can be observed that the people working for IIT-Kharagpur collaborate and publish with other people in their department much more often than people of other three IITs followed by IIT Mardas, IIT Kanpur and IIT Delhi respectively. This is because of the fact that out of 27 vertices only 4 are isolated vertices. It means that more than eighty five percent of the people of this Department have joint publications with each other whereas this percentage decreases for other IITs and is least in case of IIT Delhi. The value of Average Clustering Coefficient for IIT-Kharagpur testifies to the fact that its local affiliation graph also exhibits Small World behaviour.

Figure-1 presents the co-authorship network formed by the collaborations of people under consideration of IIT Delhi and their co-authors. From the visual analysis of Figure 1 it can be observed that a few nodes in the graph are connected and majority of the nodes in the graph are isolated. It presents a grim picture of the status and frequency of collaboration within the department. This can be verified in social network analysis terms from the low values of various network metrics like Average Degree and high values for network metrics like Single Vertex Connected Components (isolated vertices). The departmental co-authorship networks of IIT-Kanpur, Kharagpur and Madras are shown in Figure-2, 3 and 4 respectively.

If we analyse the co-authorship graphs shown in various figures above in combination with the values of various network metrics listed in Table-I it can be observed that the IIT Madras has highest number of vertices as well as highest number of edges of all the IITs under consideration. However IIT-Kharagpur has the highest Average Degree because the ratio of number of edges to the number of vertices in the co-authorship graph of IIT-Khargapur is higher than other IITs.

Analysis of the graphs shown in Figure-1, 2, 3 and 4 coupled with the values of various network metrics listed in Table-I help us answering questions related to the internal collaboration status of that particular IIT. IIT-Kharagpur has least percentage of isolated vertices (14.81%) followed by IIT-Madras (40.91%) and IIT-Kanpur (65.38%). However IIT-Delhi has the highest percentage of isolated vertices (83.87%). This implies that for the period under consideration the people in IIT-Delhi has almost negligible intradepartmental research collaborations. Because of having least percentage of isolated vertices and a good number of edges in the local affiliation graph of IIT-Kharagpur strength of intradepartmental collaboration ties is strong (Average Degree=2.889, Graph Density=0.111112, and Average Clustering Coefficient=0.349). This situation is reverse in case of IIT-Delhi (Average Degree=0.258, Graph Density=0.00086022, and Average Clustering Coefficient=0.097).

The social network of an individual author concentrating only the immediate neighbours of an individual is called as egocentric network [16]. Such a network provides an insight into the collaborations and co-authorship relationships of a particular author with his immediate co-authors. There are two types of graphs that we obtained from the disambiguated publications data that we obtained from DBLP and Google Scholar. The width and the colour used for representation of an edge in the graphs shown in Figure-5 and 6 represents the degree or strength of co-authorship relationship with a co-author. Green edges represent that these two authors publish more often with each other whereas red edges represent co-authoring a single publication only. Figure-5 shows the collaboration graph of the author under consideration i.e. ‘Rashid Ali’ and Figure-6 shows the academic social network of the author under consideration.

The values of various important social network metrics for the egocentric academic social network of 'Rashid Ali'2 are presented in Table-II.

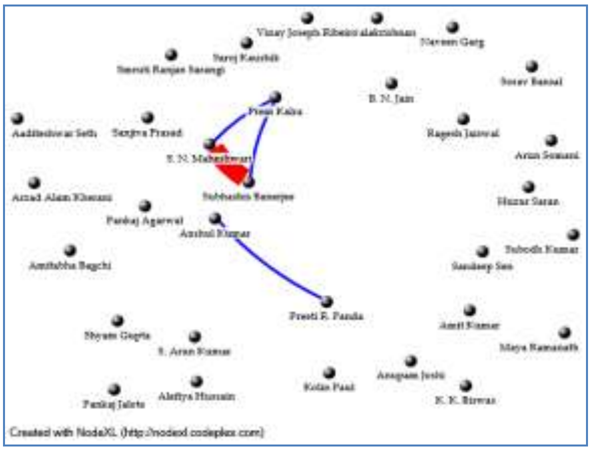

Figure 1: Academic Social Network of IIT Delhi

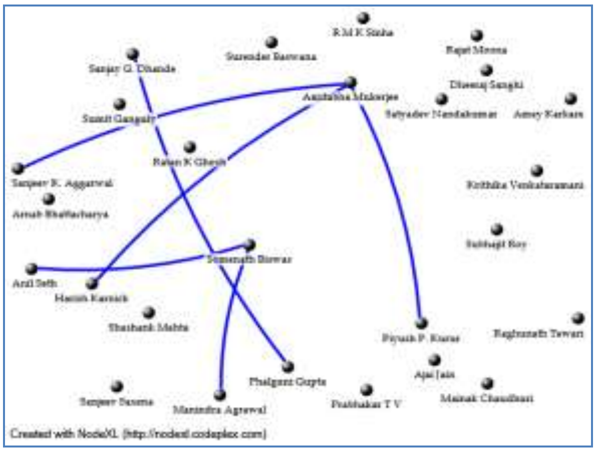

Figure 2: Academic Social Network of IIT Kanpur

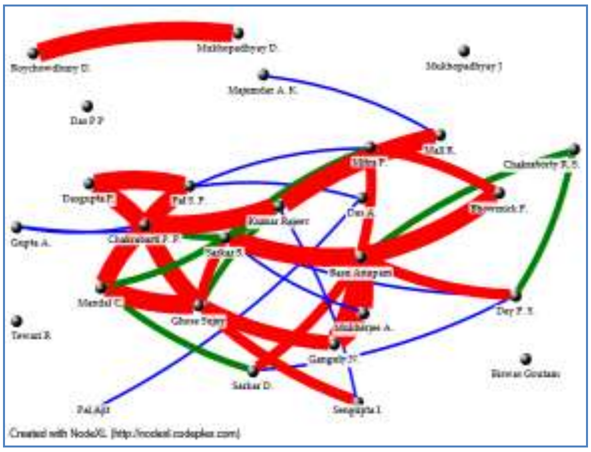

Figure 3: Academic Social Network of IIT Kharagpur

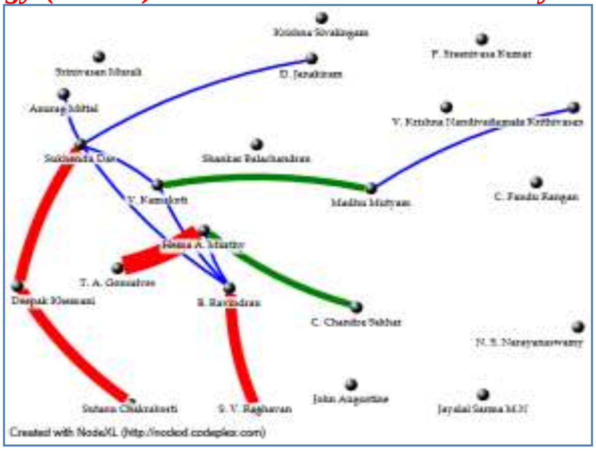

Figure 4: Academic Social Network of IIT Madras

TABLE-II
VALUES OF VARIOUS EGOCENTRIC CO-AUTHORSHIP GRAPH METRICS

Social Network Metrics	**Value**
Vertices	13
Total Edges	17
Connected Components	1
Maximum Vertices in a Connected Component	13
Maximum Edges in a Connected Component	17
Average Geodesic Distance	1.64497
Graph Density	0.27948718
Average Degree	2.615
Average Betweenness Centrality	4.692
Average Closeness Centrality	0.048
Average Eigenvector Centrality	0.077
Median PageRank	0.769
Average Clustering Coefficient	0.570

[2] Rashid Ali is an Associate Professor in Department of Computer Engineering Aligrah Muslim University Aligrah, Uttar Pradesh, India. He can be reached at: http://www.amu.ac.in/dshowfacultydata.jsp?did=30&eid=3011

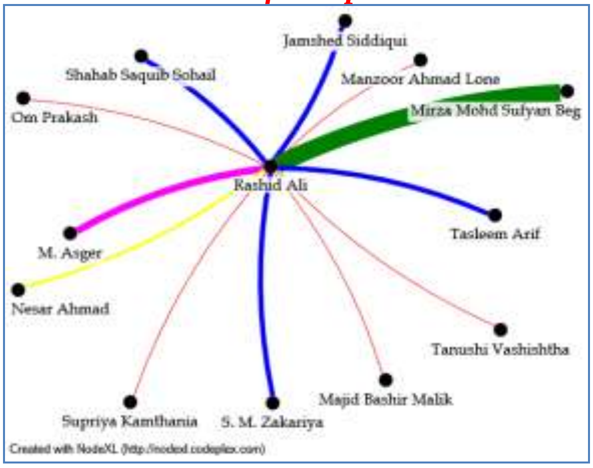

Figure 5: Collaboration Graph of 'Rashid Ali'

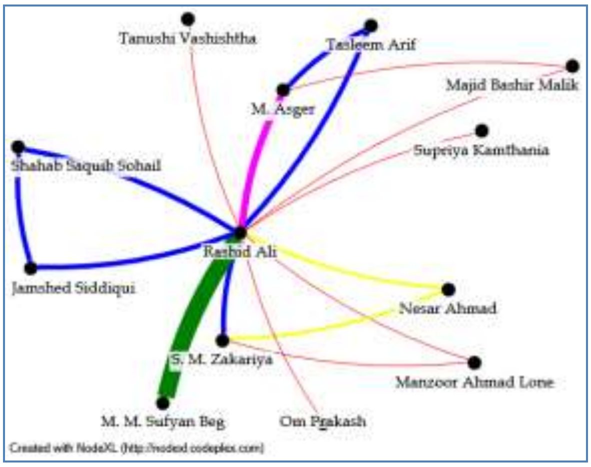

Figure 6: Academic Social Network of 'Rashid Ali'

From the analysis of Figure-5 it can be observed that Rashid Ali has most number of publications with Mirza Mohd Sufyan Beg, whereas he has least number of publications with five co-authors viz. Om Prakash, Supriya Kamthania, Majid Bashir Malik, Tanushi Vashishtha and Manzoor Ahmed Lone. The degree of co-authorship with other four co-authors viz. M. Asger, Nesar Ahmad, S. M. Zakariya, Tasleem Arif, Jamshed Siddiqui and Shahab Saquib Sohail lies in between the above two cases.

If we analyze Figure 6 it can be observed that out of the 12 co-authors of 'Rashid Ali' four authors viz. M. M. Sufyan Beg, Om Prakash, Supriya Kamthania, and Tanushi Vashishtha are not connected with any of the other co-authors. Whereas, rest of the eight co-authors are connected with other co-authors also. From the value of Graph Density in Table-II it can be observed that its value is very low considering the small size of the graph (number of vertices). This can be attributed to the fact that not all the co-authors of Rashid Ali are co-authoring with each other. But on the analysis of the publications of the author under consideration it was observed that with each new publication the density of the graph showed an upward trend. Considering the space requirements it was not possible for us to discuss the above mentioned growth pattern in this work.

Again, the value of Average Geodesic Distance listed in Table-II would have been near to one if almost all his co-authors would have been collaborating with each other because the value of Average Geodesic Distance is inversely proportional to the value of Graph Density.

V. Conclusions & Future Directions

Social Network Analysis plays an important role in explaining various important facts which otherwise would not have been possible. In this paper we obtained and analysed co-authorship social networks both at institutional and individual level. From the analysis of the four institutional co-authorship social networks it was observed that a very small number of people actually collaborate with other people in their organization. It can be considered as a very alarming situation and at some point of time may also be attributed to professional rivalries in the department. Considering that the institutions under consideration in this work have high academic prestige in the country and sometimes also serve as role model for other institutions of higher learning in the country, these trends may have adverse impact on other institutions.

On the other hand, if we analyse the ego-centric networks it can be observed that not all the people collaborate with other co-authors of the central author. In fact only few of the co-authors have direct co-authorship relationship with each other. This implies that they cannot be treated as a cohesive group.

In future we intend to study co-authorship relationship among people of leading academic and research institutions in the world and compare their collaboration patterns with the collaborations in

India. In addition we also intend to extract and analyse co-authorship social networks among various leading institutions in India.

REFERENCES

[1] H.W. Chang, and M.H. Huang, *Cohesive subgroups in the international collaboration network in astronomy and astrophysics*. Scientometrics, Vol. 101, No.3, pp. 1587-1607, 2014.

[2] D. Zhao and A. Strotmann, *The knowledge base and research front of information science 2006–2010: An author cocitation and bibliographic coupling analysis*. Journal of the Association for Information Science and Technology, Vol. 65, No. 5, pp. 995–1006, 2014.

[3] T. Arif, R. Ali, and M. Asger, *Scientific co-authorship social networks: A case study of computer science scenario in India*. International Journal of Computer Applications, Vol. 52, No. 12, pp. 38-45, 2012.

[4] G. Vidican, W. L. Woon, and S. Madnick, *Measuring innovation using bibliometric techniques: The case of solar photovoltaic industry*. Working Paper CISL# 2009-05, Massachusetts Institute of Technology, Cambridge, MA 02142, 2009.

[5] V.I. Torvik, M. Weeber, D.R. Swanson, and N.R. Smalheiser, *A probabilistic similarity metric for Medline records: A model for author name disambiguation: Research articles*. Journal of the American Society for Information Science and Technology, Vol. 56, No. 2, pp. 140–158, 2005.

[6] N.R. Smalheiser and V.I. Torvik, *Author name disambiguation*. Annual Review of Information Science and Technology,Vol. 43, No. 1, pp. 1–43, 2009.

[7] A.A. Ferreira, G.A. Gonçalves, and H.F.A. Laender, *A brief survey of automatic methods for author name disambiguation*. ACM SIGMOD Record, Vol. 41, No. 2, pp. 15-26, 2012.

[8] T. Arif, R. Ali, and M. Asger, *Author name disambiguation using vector space model and hybrid similarity measures*. In Proceedings of 7th International Conference on Contemporary Computing-IC3'2014, Noida, India: IEEE. pp. 135-140, 2014.

[9] F. Ma, Y. Li, and B. Chen, *Study of the collaboration in the field of the Chinese humanities and social sciences*. Scientometrics, May 2014.

[10] M. Coscia, F. Giannotti, and R. Pensa, Social Network Analysis as Knowledge Discovery process: A case study on Digital Bibliography. In Proceedings of 2009 Advances in Social Network Analysis and Mining, pp. 279-283, 2009.

[11] M. E. Newman, *The structure of scientific collaboration networks*. Proceedings of the National Academy of Sciences, Vol. 98, No. 2, pp. 404–409, 2001.

[12] H. Hou, H. Kretschmer and Z. Liu, *The structure of scientific collaboration networks in scientometrics*. Scientometrics, Vol. 75, No. 2, pp. 189–202, 2008.

[13] C. Chelmis, and V.K. Prasanna, *Social networking analysis: A state of the art and the effect of semantics*. In Proceedings of 3rd IEEE Conference on Social Computing (SocialCom), Boston, MA, 2011, pp 531-536, 2011.

[14] D.J. Watts and S.H. Strogatz, *Collective dynamics of 'small-world' networks.* Nature, Vol. 393, No. 6684, pp. 440–442, 1998.

[15] S. Milgram, *The Small World Problem*. Psychology Today, Vol. 2, No. 1, pp. 60-67, 1967.

[16] D. Fisher, *Using egocentric networks to understand communication*. IEEE Internet Computing, Vol. 9, No.5, pp. 20-28, 2005.