Applying Centrality Measures for Impact Analysis in Coauthor ship Network

Adeel Ahmed[1], Riqza Shabbir[2], Atifa Afzal[3], Muhammad Akmal[4], Sahar Fatimah[5]
[1]Department of Information Technology, The University of Haripur, KPK, Pakistan
[2,3,4,5]Department of Computer Science, National University of Modern Languages, Islamabad, Pakistan
[1]aahmedqau@gmail.com, riqzashabbir@gmail.com[2], atifaafzal2@gmail.com[3],
akmalm929@gmail.com[4], saharfatimah3@gmail.com[5]

***Abstract*— Nowadays social networking is an essential part of everyone's life to communicate with different people around the globe. Due to improvement in expertise networks are growing rapidly and becoming more complex. Through social networking, we can identify different communities that help us to get information about different people and their work in different fields. In social networks, community detection is one of the hot areas. In this paper, we have analyzed a co-authorship network of political science and ranked the authors on the basis of common centrality measures. Finding reveals that these common centrality measures can be useful indicators for impact analysis.**

Keywords: community; undirected graph, clustering; scientific collaboration; social network; centrality measures

I. INTRODUCTION

Clustering is one of the data mining algorithms that partition the graph into clusters of the same group. Clustering is the most important unsupervised learning technique. It is used to find the groups in unlabeled data. There are many clustering algorithms available. Modularity is normally used as a measure of how good clustering is? In this paper, we performed an analysis on coauthorship network. These authors belong to the field of Political Science related to different research and academic institutes present in all over the world.

Figure 1 shows the coauthorship network of Political Science.

In this paper, we have applied common centrality measures such as Betweenness centrality, Closeness centrality, Eigenvector centrality, Degree centrality, and discuss the usability of centrality measures for the author's ranking and observed that these measures can be useful for impact analysis.

II. RELATED WORK

George et al. [1] proposed a method based on clustering. They assigned a new definition to a cluster using graph edit distance in the probabilistic graph. The method uses ground truth data and protein-protein interaction, it outputs probabilistic graph partition into groups. George et al. model probabilistic graph and computed an Edit distance between probabilistic and deterministic graph.

Mardala et al. [2] described Ant Colony based approach for detecting communities. This approach is based on model-based technique. Nodes are represented as ants, first the network is divided into bisection, partition vector x represented with 0 and 1, node is assigned 0 if it belongs to group 1 and assigned 1 if it belongs to group 2. Modularity Maximization is used for zero rows and column sum property and then from these the vertex values are checked and nodes assigned to group 1 or group 2 based on these values. Each ant performed local search within it, one harming neighbor, if they differ only in one component. Pheromone value is updated using iteration and best solution. The experiments performed on real social networks with friendship testbed network. The result shows that ACO provide a better solution than existing networks up to 50%. 1-opt-local search procedure also contributed for high quality solution.ACO based approach provides the best result and improvement in modularity values as compared to existing algorithms.

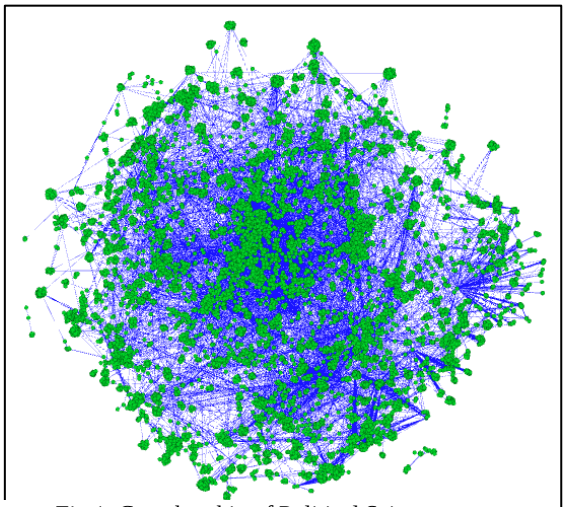

Fig 1. Coauthorship of Political Science

Sui et al. [3] come up with a new Genetic algorithm for Overlapping Community Detection (GaoCD) for finding communities based on link clustering. Algorithm finds link communities by using partition density objective function and then map these communities to node communities. Algorithm works by partitioning M links into different partitions P and create C subsets. Partition density is calculated and it is the average density D of all communities in each community. Two edges are considered as adjacent if they share one node, in this approach individual g of m genes represented as edges. Adjacent edges are identified if two edges are linked to same node. Then bridge link is created between two cliques. Overlapping communities are then identified by gathering node incidents to the edge in community. Experiments performed on typical overlapping structure, artificial network, and real network. Overlapping structure shows that it accurately reveals

communities for all networks and properly partition them. LFR is considered as the benchmark in artificial network, experiment tell that GaoCD always achieve best performance as compared to ABL and GA-Net+. Experiments for real network performed on a real networks, it shows that ABL finds small communities where GaoCD find denser communities in all sizes.

Jaewon Yang et al [4] proposed an approach based on modularity.

Xuewu Zhang et al [5] addresses CNM algorithm proposed by newman is used. Algorithm finds non-overlapping and also overlapping communities and then validity function finds closeness between nodes and communities. Algorithm first collect data from online social network and finds non-overlapping community by using the module "Get non-overlapping communities". Second, after finding non-overlapping communities, algorithm will traverse all the nodes resulting from merging of each pair of communities.

Kamal Sutaria et al [7] proposed a community detection algorithm based on modularity. All vertices which are placed inside each community are identified in the last stage of modularity or partition is gained. At first dataset file is read after initializing, the file contains edges list with the origin and destination vertex. Then it checks for each edge in a file, if both vertex new then both vertices are placed in same community index and also add it to partition and visited vertex list. Then if any of the two vertex already visited the new join or split event is performed. Event performed on the basis of strongly connected property conditions. If vertex i already visited and vertex j is not, then using betweenness vertex j is moved to vertex i community, otherwise, vertex j will create its new community.

Chawla et al. [8] used a simple approach with topological clustering coefficient similarity (CSS), Common Neighboring Similarity (CNS) and Node Attribute Similarity (NAS) on weight of edges, node attributes are used to calculate better weight.

Jaewon Yang and Jure Leskovec [6] represented methodology which allows us to compare network communities quantitatively. Using a spectral clustering algorithm with heuristic parameter free community detection method that scales for more than hundred million nodes. 13 different communities are examined and divided into four classes. This methodology performed in three major steps: (1) in first step, ground-truth communities are defined from 230 large social and information networks. (2) In second step, 13 structured communities that are commonly used are evaluated quantitatively for robustness and sensitivity. (3) In third step, the local spectral clustering method is used that scales to hundreds of millions of nodes.

III. ABOUT DATASET

For applying four most widely used classic measures (closeness centrality, degree centrality, betweenness centrality and pageRank) to co-authorship network, Microsoft Academia Research dataset has selected in this research. Microsoft Academia Research is an experimental service developed by Microsoft that explore how authors, students, scholar and researchers find contents. It also shows relationships among subjects, author and contents. (71) Dataset from last 50 years contains 7 different files;

Authors.txt: This text file contains 114+ million records which has fields ID and Name of all Authors who have published research Papers.

Affiliations.txt: This text file contains 19843 records which has fields ID and Name of institutions/universities to which Authors is affiliated.

FiedOfStudy.txt: This text file contains Name of all Fields in which Authors related to and published papers such as Physics, Chemistry and Computer Science etc.

FiedOfStudyHierarchy.txt: This text file contains Parents and Child Field of Study up to 4 level such as Database, Computer Networks, Data mining, Data Warehouse and HCI etc.

Papers.txt: This text file contains 126+ million records which contains all Papers that different Authors Published in different time spans and in different conferences or journals.

PaperAuthorAffiliation.txt: This file contains records of Papers which are published by different Authors and also Affiliation name to which Authors and Papers are affiliated.

PaperKeywords.txt: This file contains Fields of Study papers that has been published in different years by different Authors.

IV. RANKING AUTHORS ON THE BASIS OF CENTRALITY MEASURES

We have ranked authors on the basis of four common centrality measures. These measures are betweenness centrality, closeness centrality, eigenvector centrality, and degree centrality.

A. *Ranking Authors on the basis of Betweenness Centrality*

Betweenness Centrality measures how often a node appears on the shortest paths between nodes in the network. The Betweenness Centrality of a node v is given by the expression:

$$G(v) = \sum_{s \neq v \neq t} \frac{\sigma st(v)}{\sigma st}$$

Here σ_{st} is the total number of shortest paths from node S to node v and $\sigma_{st}(v)$ is the number of those paths that passes through *v*. Figure 2 shows the highest betweenness centrality of node '5F59DCDC'.

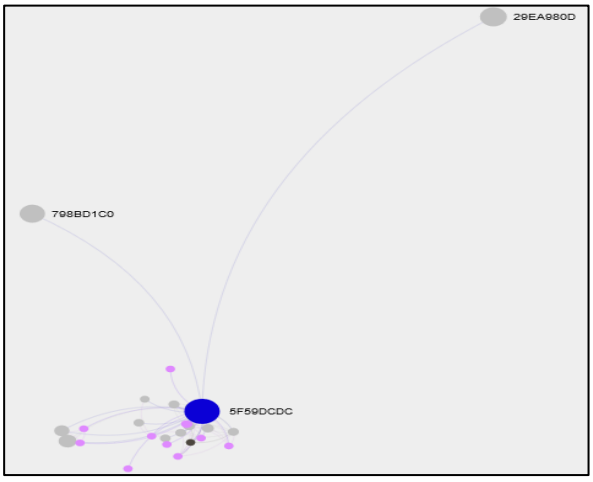

Fig. 2: 5F59DCDC with highest Betweenness Centrality

Table 1: Statistics related to Dataset

Nodes	98866
Edges	101589
Volume	101589

Table 2: Authors Ranking on the basis of centrality Betweenness and Closeness Centrality

ID	Betweenness Centrality	ID	Closeness Centrality
5F59DCDC	217253.858043	000E1FBA	1.0
OBC96107	210189.232145	000F0038	1.0
7535A3F5	199176.356706	2BFC6D41	1.0
215ABA27	172252.159141	273723	1.0
14ABE527	164023.475612	7ABBEE08	1.0
2A7D2E02	161336.695569	004208DF	1.0
77602149	156279.759736	287961ED	1.0
5E06E669	149189.033545	00448F4E	1.0
16C3726B	146837.482299	0046A55D	1.0
1F48549C	146005.221937	005CCB20	1.0
7D5D270A	145772.735692	005F784E	1.0

Table 2 shows the nodes with highest betweenness centrality. In this table, the first highest betweenness centrality of the node whose ID is 5F59DCDC and betweenness centrality is 217253.858043. The node '5F59DCDC' shows that this author belongs to 'John F. Kennedy School of Government, Harvard University' and also linked with 'World Bank'. Second highest betweenness of the node OBC96107 is 210189.232145. This node belongs to 'Harvey Mudd College Middle East Technical University'.

B. *Ranking Authors on the basis of Closeness Centrality*
Closeness is the reciprocal of the farness, that is:

$$C(x) = \frac{1}{\sum_y d(y,x)}$$

where d(y, x) is the distance between vertices x and y. Table 2 shows the nodes with the highest closeness centrality that is 1.0. The node named' 7964578A' having closeness that is 0.314607 and the author belongs to 'Australian Research Centre for Population Oral Health, School of Dentistry, Faculty of Health Sciences, The University of Adelaide, South Australia'. From top ten authors, whose closeness centrality is 1.0, belongs to 'Yale University, Political Science, Economics' and 'Columbia University', respectively. In figure 3, the graph shows the nodes in blue color have highest closeness centrality.

C. Ranking Authors on the basis of Eigenvector Centrality
Eigenvector centrality is computed as,

$$xv = \frac{1}{\lambda} \sum_{t \epsilon M(v)} xt = \frac{1}{\lambda} \sum_{t \epsilon G} av, txt$$

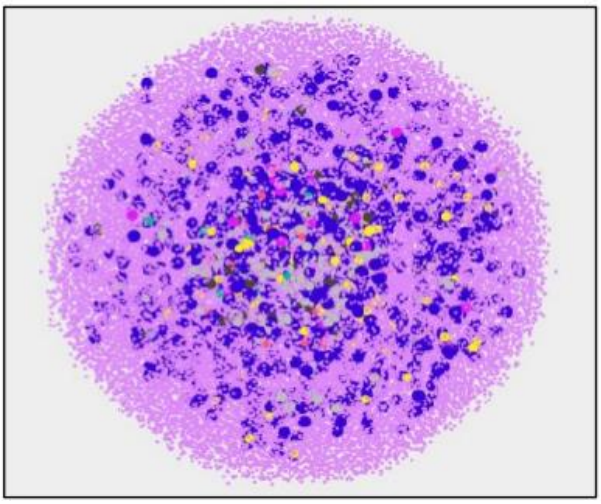

Fig 3: Political Sc. Network with highest Closeness Centrality

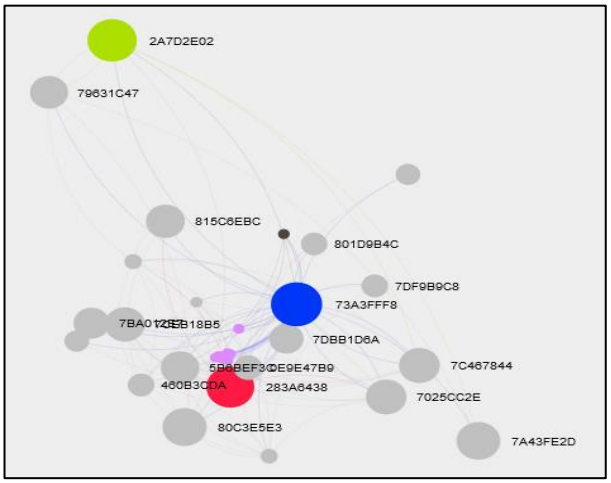

Fig 4: Graph with highest Eigenvector Centrality

Table 3 shows the highest Eigenvector centrality. In the table, the node with the highest Eigenvector Centrality is 73A3FFF8 and has value 1.0.

Table 3 Authors Ranking on the basis of Eigenvector Centrality and Degree Centrality

ID	Degree Centrality	ID	Eigenvector Centrality
12F4FdCC	68	73A3FFF8	1.0
2A7D2E02	67	2A7D2E02	0.953535
16C3726B	56	283A6438	0.91285
050D9082	54	163726B	0.874171
14ABE527	52	80C3E5E3	0.818782
2AF7FE2A	51	7A43FE2D	0.811711
755A55D7	49	7C467844	0.7358523
0B1D8191	43	79DCC5E7	0.734681
25FEB091	43	7025CC2E	0.730326
084238FB	42	726C9401	0.721169
766E0394	42	7CEB1885	0.714344

Second highest Eigen centrality of 2A7D2E02 is 0.953535. Third highest Eigenvector centrality of 283A6438 is 0.91285.The figure 4 shows the community of highest Eigenvector centrality.

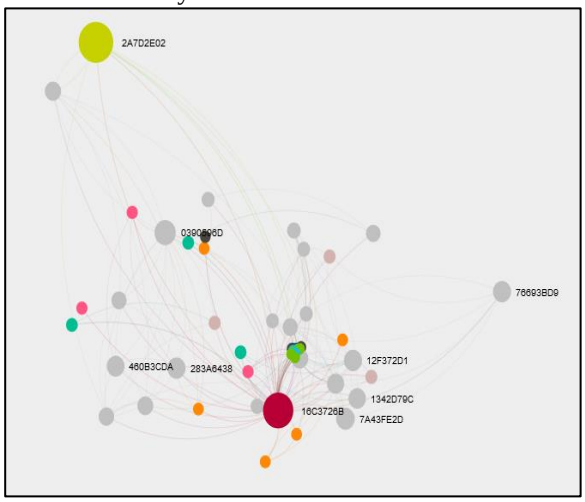

Fig 5: Nodes with highest Eigenvector Centrality

D. Ranking Authors on the basis of Degree Centrality

Degree is a simple centrality measure that counts how many neighbors a node has, and is measured as

$$Cp = d(ni) = Xi+= \sum_{j} Xij$$

In our dataset the nodes with the degree value 1 are 12056. The node named with'12F4FdCC' shows that the '12F4FdCC' belongs to 'Tropical Agricultural Research and Higher Education Center (CATIE), Turrialba, 7170, Costa Rica' with the Degree Centrality 68. Author named '2A7D2E02' belongs to , 'School of Public Health and Family Medicine, University of Cape Town, Anzio Road, Observatory, Cape Town 7925, South Africa' and 'Infectious Disease Epidemiology Unit, School of Public Health and Family Medicine University of Cape Town, Observatory, Cape Town, South Africa'. Figure 6 shows the community of the node with the highest degree centrality.

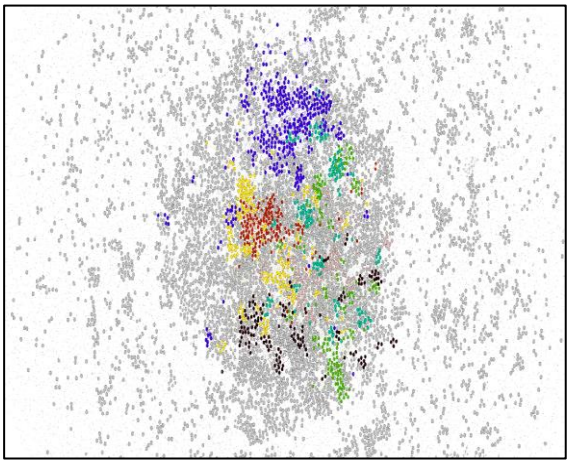

Fig 6: 2A7D2E02 having 2nd highest Degree Centrality

D. Ranking Authors on the basis of Modularity

The Equation of the Modularity is

$$Q = \frac{1}{2}\sum_{vw}[Avw - \frac{kvkw}{2m}]\frac{svsw+1}{2}$$

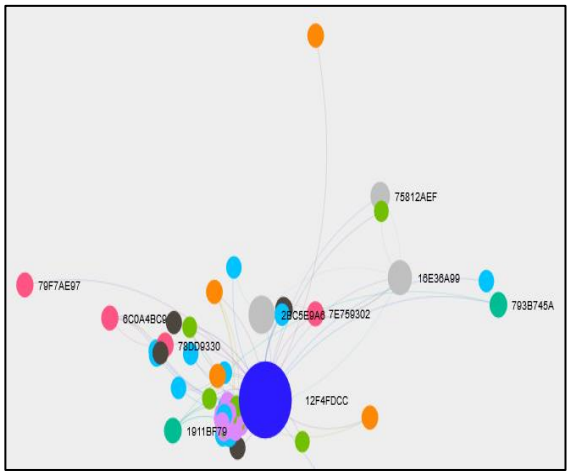

Fig 7: Modularity based Partitioning

In figure 7, we have found 14 modularity classes and number of nodes in each modularity class are varies. In our dataset, the number of communities found is 62 and modularity value is 0.838. In figure 7, the nodes in blue color have 696 modularity class which contain 346 elements. This is the largest community in our dataset. The nodes in yellow color have 365 Modularity class which contains 203 this is the second largest community in our dataset. The nodes with Green color have 264 Modularity class and contains 169 numbers of nodes this is the third largest community in our dataset. The nodes in red color have 198 modularity class and it contains 153 nodes.

V. CONCLUSION

Analyzing any social networks is the demand of new era. In this paper, we have analyzed a coauthorship network of Political Science and ranked the top ten authors of this field. Our purpose is to find the highly collaborative groups of people and the productive institute for this field. Different co-authors are collaborating with different communities some of them are overlapping in more than one community, and few co-authors have frequent collaboration in different fields. We have found that these centrality measures are good indicator for impact analysis. In future, we will find overlapping communities in Political Science.

REFERENCES

[1] Z. Halim. "Clustering large probabiistic graphs using multi-population evolutionary algorithm" , Information Sciences, vol. 317, pp: 78–95, 2015.

[2] S. R. Mandala, "Clustering social networks using ant colony optimization",Operational Research, vol. 131), pp:47-65,

[3] Y. Cai, "A Novel Genetic Algorithm for Overlapping Community,"Advanced Data Mining and Applications, 7th international Conference, ADMA, China, 97-108,2011.

[4] M.E.J. Newman, “Modularity and Community Structure in Networks”, Proceesings of National Academy of Sciences, vol. 103(23), pp:8577-8582, 2006.

[5] X. Xhang, “Overlapping Community Identification Approach in Online Social Networks”, Physica A: Statistical Mechanics and its Applications, Vol. 421, pp: 233-248, 2015.

[6] J. Yang and J. Leskovec, “Defining and Evaluating Network Communities based on Ground Trusth”, Knowledge and Information Systems, Vol: 42(1),181-213,2015.

[7] D. S. Bassett,”Robust Detection of Dynamic Community Structure in Networks”, Chaos, 23(1): 013142.

[8] V. Chawla and K. Steinhaeuser, “Identifying and evaluating community structure in complex networks”,2009.