Analysis of Collaboration Network

Naitik Dodia (201501177) and Brijeshkumar Patel (201501238)
Dhirubhai Ambani Institute of Information & Communication Technology, Gandhinagar, Gujarat 382007, India
CS-454, Introduction to Complex Networks. Prof. Mukesh Tiwari

In this project, we study the GR-QC (General Relativity and Quantum Cosmology) collaboration network. The purpose of the project is to analyze the behaviour of this network by computing important statistical parameters, both at macroscopic and microscopic level. Further, the goal would be to reason the parameters in the scope of the network.

I. INTRODUCTION

A collaboration network consists of various authors who are geographically distributed and heterogeneous in terms of operating environment, culture, social capitals and goals, who come together by means of some technology to achieve common goals. The network is described by a set of vertices and a set of undirected edges where the vertices denote the authors and there is an edge between two authors if there is at least one research paper in which they have collaborated. The papers studied for creation of this network fall under the topic of General Relativity and Quantum Cosmology. The data covers papers in the period from January 1993 to April 2003 (124 months). Here, we are considering the papers related to single high-level topic because adding other high level topics would just add similar components to the network which finally result in same analysis.

II. ANALYSIS

The number of nodes in the network are 5242 and the number of edges is 14496. We will calculate some relevant centrality measures and study the degree distribution.

A. Degree distribution and Degree centrality

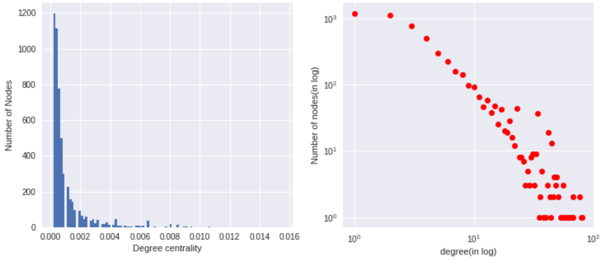

FIG. 1: a) (left) Degree centrality histogram b) (right) Degree distribution (in log-log scale)

From 1, we can see that a very large number of nodes have very less degree and very less number of nodes have very high degree. From 1 b) the log-log plot of degree distribution follows almost linear relationship, so we can say degree distribution follows the power law. Hence, the network is scale-free (number of nodes doesn't affect the trend of degree distribution).

B. Betweenness Centrality

If we consider each author related to some set of subtopics, then betweenness of a node will mean the intersection of the subtopics of its neighbours union-ed with its own subtopics. Betweenness centrality in this context means the power of the node to the flow of ideas from one subtopic to another subtopic irrespective of its degree. From 2, we can see that 4000 have their betweenness centrality as zero, so the number of degree 1 or degree zero nodes is 4000. This should result in high correlation between degree centrality and betweenness centrality but it is not the case. The correlation between these two centrality measures is 50.39%. This is because 591 nodes contribute to negative correlation because their betweenness centrality is higher than their corresponding degree centrality in their normalized form. This means that there are 591 authors whose degree is low (they have less intersection of their neighbours) but whatever the intersection of subtopics is, it is very essential in flow of subtopics.

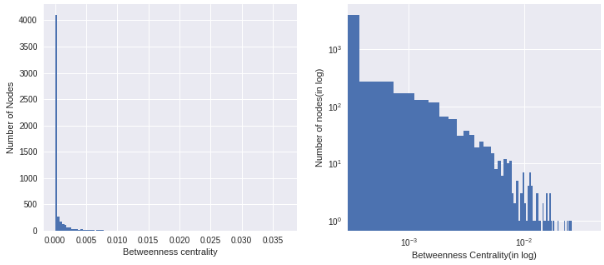

FIG. 2: a) (left) Betweenness centrality histogram b) (right) Betweenness centrality (in log-log scale)

C. PageRank

As our network is undirected, Pagerank is implemented as per [1]. As per this implementation, each undirected edge is converted into a pair of bidirectional edges and PageRank is applied with damping factor 0.85.

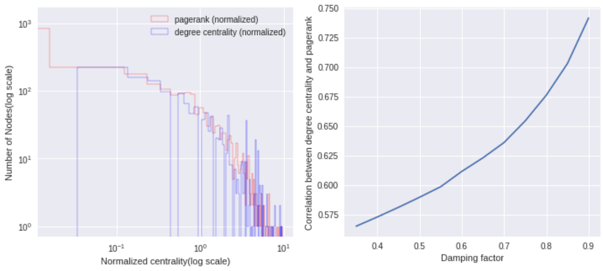

FIG. 3: a) (left) Normalized Pagerank (red) and Degree centrality(blue) b) (right) correlation vs alpha(damping factor)

From 3, a) we can see that PageRank behaves very similar to degree centrality. Statistically, the correlation coefficient between the two centralities is 70.3%. This is because of the fact that for each node in-degree is equal to out-degree, so now, degree plays the main role in assigning the rank.

$$PR(p) = \frac{1-d}{N} + d\sum_{i=1}^{k} \frac{PR(p_i)}{deg(p_i)} \tag{1}$$

where d = damping factor, N = number of nodes

Damping factor contributes by stating the probability of deciding the rank from a node's neighbours. So, as the value of damping factor increases more and more dependence of assigning the rank on its degree increases. So, at *d = 1* the pagerank is converted to degree centrality.

D. Clustering and Communities

Researchers working under a main topic share interests in many subtopics. This sharing of interests forms communities in the form of their interactions by publishing research papers. The modularity between communities will be increases if there is very less reference of a group of subtopics with other such group.

1. *Maximal Cliques*

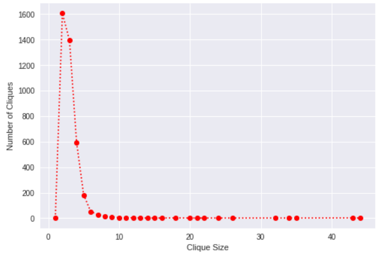

FIG. 4: Histogram of clique size vs number of cliques

The maximum number of maximal cliques corresponds to k = 2 and the maximum clique size is 44. And the whole distribution of number of k-cliques vs k is centered towards k [2,7] and is very sparse over larger values of k i.e. there are very less number of higher value k-cliques. This means that there are very less number of authors in a single research paper (which is obvious). Maximum clique size = 44 doesn't necessarily mean that there are 44 co-authors in a single research paper, rather, it means that this group of people have collaborated internally with very high frequency of research papers.

2. *k-core*

Clique is very strict definition for deciding communities. The definition of k-core[2] relaxes this definition by adding freedom of (n-k) degrees to every node. So, maximal cliques are extended by a new vertex which will be connected to k vertices of the existing community.

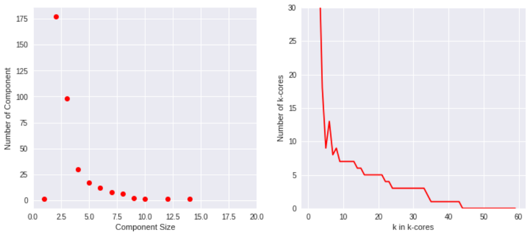

FIG. 5: a) (left) Dotted Histogram of component size b) (right) line histogram of k in k-core

The maximum value for 5 (a) is for components of size 2 (number of size 2 components = 177) i.e. there are many authors who have contributed with only one other author. Also there is only one component of more than 4000 nodes, the information about all other sizes of components is shown in the plot. This means that there is one large component and others are insignificant to that.

5 (b) shows number of k-cores vs k (as in definition of k-cores). As the maximum clique size is 44 we can see in 5 (b) that for k ¿ 44 the number of maximal k-core decreases to zero. Comparing 4 and 5 (b) we can see that there is relaxation with respect to number of k-cores rather than cliques, in the plot and this relaxation decreases with increase in value of k.

3. *Greedy modularity maximization*

As per [2] greedy modularity maximization works first selecting each node as its own cluster. Then we amalgamate the clusters which maximize the modularity and continue this process. Eventually, a single cluster will be formed and the algorithm will stop. Now, we look back at the steps that we carried and we consider the clustering at which the modularity score is maximum.

By using this algorithm we find that the number of clusters formed in our data is 426 with maximum cluster size = 1039. And the communities formed in the large component is 72. From this we can say that, though being large component there 72 subtopic in the network which have very less reference between one another.

III. CONCLUSION

From the above analysis we can say that, the large component formed plays an important role in deciding network structure. Though we are analyzing the network corresponding to collaboration under single topic there will be clusters formed due to subtopics or due different ideologies corresponding to the researchers.

[1] Discovering author impact: A PageRank perspective - Erjia Yang, Ying Ding

[2] Networks: An Introduction by M.E.J Newman