Evolution of Scientific Collaboration Networks

Gaurav Madaan
Computer Science Department
Thapar University
Patiala, India
gauravmadaan95@gmail.com

Shivakumar Jolad
Department of Physics
IIT Gandhinagar
Ahmedabad, India

Abstract—We study the structure and evolution of scientific collaboration network by using collaboration network constructed from DBLP Computer Science Bibliographic database [1], from year 1936 to 2013 using social network analysis techniques. We have found many interesting features such as collaboration between scientists is increasing with time and few numbers of scholars publish a large number of papers while most of the authors publish a small number of papers, which is consistent with Lotka's law on frequency of publications [2]. The degrees of the vertices in the collaboration graph follow a "Power law" pattern i.e., the number of vertices of degree x is proportional to a negative power of x. The clustering coefficient of collaboration graph comes out to be very high which means that there are more chances for two authors to co-author a paper if they have a common collaborator. We also found that the collaboration graph follows various real graph properties like WPL (Weight power law), DPL (Densification power law) etc. We try to apply the Lorenz curve and Gini coefficient on the collaboration graph to study the variation in concentration of collaboration between researchers with time.

Keywords—bibiliometrics; collaboration networks; network structure; power laws; research collaboration; scale free networks;

I. Introduction

Social network analysis is an active research field in the social sciences where researchers try to understand social influence and groupings of a set of people or groups. A scientific collaboration graph is a graph modeling social network where the vertices represent authors of that network and where two authors are joined by an edge whenever they have co-authored a paper. Inspired by various attempts to apply social network analysis to db community, such as Newman's work on structure of scientific collaboration networks [3], Barabasi's work on study of dynamics of scientific collaboration networks [4], we analyze the dynamics of evolution in collaboration network, made by database researchers, using DBLP database consisting of bibliographic information about 2.4 million papers by around 1.3 million authors from 1936 to 2013 and see if we can find any interesting patterns underlying the db community and their publication behavior. We primarily use Python based package ***Networkx*** [5] for our research work.

II. Related literature

Collaboration networks are being studied extensively in the past few years. Collaboration between scientists can be used as a measure for the qualitative examination of literature. The intensity of collaboration varies significantly between different disciplines, for e.g, in the literature related to humanities and arts, the collaboration between researchers is very less; while in the fields like mathematics, physics, medicine, computer science and engineering, there occurs to be a large extent of collaboration between researchers. Network analysis techniques have been widely used to study collaboration networks, for e.g. Newman in 2002 [3], tried to explore the structure of large scale collaboration networks using network analysis tools. He studied the structure of collaboration networks for various disciplines like physics, biomedicine and mathematics etc. Barabasi (2002) [4], studied the dynamics of evolution of the collaboration networks in mathematics and neuroscience. Elmacioglu and Lee (2005) [6], have done extensive research on dynamics of collaboration network based on data extracted from DBLP dataset containing 38,773 publications written by 32,689 authors. Finally, Franceschet M. (2011) [7], divided the largest computer science collaboration graph ever studied (extracted from DBLP database) into sub-networks emerging from conference and journal co-authorship. He studied various bibiliometric properties like scientific productivity, collaboration level in papers, network resilience, network clustering and concentration of collaboration using Lorenz curve and Gini coefficient.

III. Setup

We used DBLP Computer Science bibliographic dataset for our analysis. DBLP has bibliographic information about different kinds of entities like articles, proceedings, in proceedings, books, thesis etc. We extracted the entries for articles and proceedings for our analysis. The dataset we analyzed contains bibliographic information about 2.4 million papers authored by around 1.3 million authors from 1936 to 2013. DBLP is very well known for the accuracy provided in the data. An identification key is assigned to each publication entity in DBLP. DBLP reduces the *name problem* (occurrence of a scholar with several names or several scholars with the same name) up to an extent by using full names of authors instead of initials. The DBLP data is freely available on the internet. The entire dataset can be downloaded for analysis purpose. We downloaded this file in 2014, parsed the file

978-1-4799-5666-1/14/$31.00 ©2014 IEEE

using Python parser and built the following types of networks using Python based package ***Networkx*** –

1. Author-Paper bipartite network- From the dataset available, we created a graph which has two types of nodes – authors and papers. If an author has authored a paper then there will be an edge between the author and the paper. This graph is the most complete representation of the collaboration.

2. Collaboration network- In this type of network, there are only one type of vertices– authors. There exists an edge between two authors if they have co-authored a paper. The disadvantage of this type of network is that there is loss of information while formation of a collaboration network from the author-paper bipartite network, for e.g. if three authors are mutually linked in the collaboration graph then we cannot say, just by looking at the collaboration network, that there exists a single paper co-authored by all the three authors or three different papers. In spite of the loss of information while formation of collaboration graph, it is very helpful in the study of various collaboration related properties and their variation with time.

IV. Terminology

Knowledge of some fundamental terms is required to understand the work described in the next sections. Some of these terms related to social network analysis are explained below-

1. *Clustering coefficient*- It is a measure of the likelihood that two associates of a node are associates. A large clustering coefficient implies more number of cliques in the network.

Clustering coefficient of a node n in an undirected graph can be defined as-

$$C_n = 2e_n / (k_n (k_n\text{-}1)),$$

where k_n is the number of neighbors of n and e_n is the number of connected pairs between all neighbors of n.

The clustering coefficient for the whole graph will be the average of clustering coefficient of all nodes in the graph.

2. *Degree of a node*- In an undirected graph, it is the number of edges connected to a node. So in collaboration network degree represents the number of authors with which the author is connected.

3. *Weight of an edge*- In the collaboration network, the weight of an edge between two authors will represent the number of papers coauthored by them.

4. *Bipartite Graph*- A bipartite graph is a graph whose nodes can be divided into two disjoint sets U and V such that every edge in E connects a node in U to one in V. G = (U, V, E)

5. *Connected Component*- It is the sub graph in which all nodes are connected to each other by paths.

6. *Giant connected component*- It is the largest connected component in the network.

7. *Lotka*'s law- "The number of authors making n contributions is about $1/n^2$ of those making one; and the proportion of all contributors that make a single contribution is about 60 percent" [5].

8. *Equidistribution Line*- This is the Lorenz curve obtained when each author is having same collaboration i.e. every author has the same number of collaborators.

V. Analysis

To analyze and depict the structure and dynamics of collaboration network, we stored the whole collaboration network and year wise collaboration networks in 'gpickle' format using Python based package *Networkx.* We used Python package *Matplotlib* for visualization purposes.

5.1 *Some fundamental properties of the collaboration network* – Global properties of the collaboration network (1936-2014) of all authors in DBLP dataset are as follows-

TABLE I. FUNDAMENTAL PROPERTIES OF NETWORK

Property	***Value***
Total number of authors	13,69,347
Total number of papers	24,44,271
Mean papers per author	1.8
Mean authors per paper	2.78
Collaborators per author(avg.deg)	7.75
Size of giant component	11,60,078
As a percentage	84.72 %
Clustering coefficient	0.6

5.2 *Identifying trends in collaboration* – To understand how the trend of publication varies with time, we separately calculated the number of papers having single authors, 2 authors and 3 or more authors for each year from 1936 to 2013. The plots we obtained are shown in next figures.

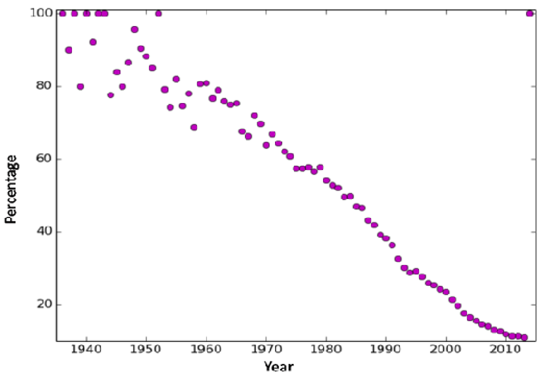

Fig 1: Percentage of single author papers per year

We found that the percentage of single author papers is almost consistently decreasing with time and the percentage of papers having 3 or more authors is increasing with time. This shows that as the computer science collaboration network evolves, new scholars come and collaborate with each other resulting in increase in extent of scientific collaboration.

We also found that till 1995, the percentage of 2 author papers is steadily increasing, but after that, there is a sudden fall in the curve. One possible explanation of this type of curve can be that after 1995, the trend of publication of 2 author papers is decreasing but the trend of publication of papers having 3 or more authors is increasing consistently with time as can be seen in Fig 3.

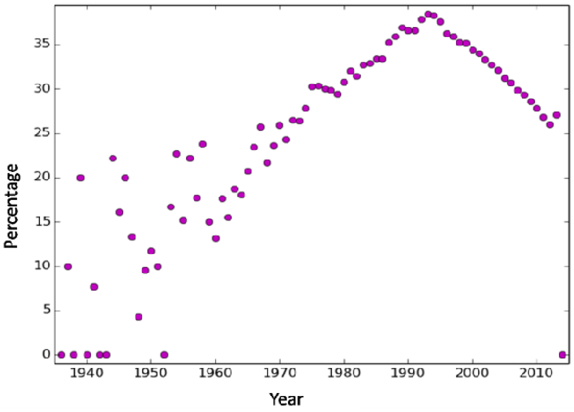

Fig 2: Percentage of papers having 2 authors

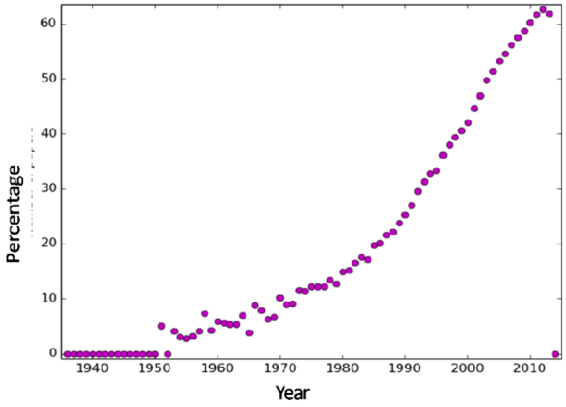

Fig 3: Percentage of papers having 3 or more authors

5.3 Study of dynamics- In this part, we studied the variation in various properties of collaboration network with time and try to find interesting patterns in collaboration.

Inspired by [6], we plotted the variation in number of new authors joining the network each year, number of active authors for each year (who publish one or more articles in that year), the distribution of number of papers per author for each year, number of authors per paper and number of collaborators per author each year etc. For the analysis, we extracted the papers and their authors for each year. We separately constructed the collaboration network for each year and stored them in *gpickle* format using *Networkx*. We used the following formulae to study the dynamics of collaboration network-

1. Number of Active authors each year- To calculate the number of active authors for each year, we calculated the number of nodes in the collaboration graph for each year.

Number of active authors in year y = Number of nodes in collaboration network of year y

2. Number of new authors joining the network each year- The new authors joining in year y are the authors who have not published anything before year y.

New authors = Number of authors who have published in year y – Number of authors of year y who have already published something before year y

3. Mean papers per author for an year- We added the number of papers published by each author and then divided the sum by number of active authors in that year. Numerically, it is equal to the number of edges in the author- paper bipartite graph of that year divided by number of active authors in that year.

i.e. *Mean papers per author = Number of edges in bipartite graph of the year/ Number of active authors in the year*

4. Mean authors per paper- We added the number of authors for each paper published in an year and divided it by the total number of papers published in that year. Numerically,

Mean authors per paper = (Number of edges in the bipartite graph of that year / Number of papers published in that year)

Using the above techniques, we observe the following patterns-

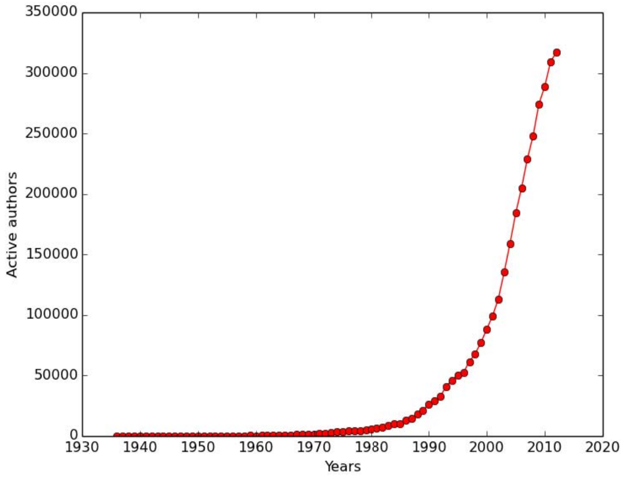

Fig 4: Number of active authors each year

We can see from Fig 4 that the number of active authors is increasing consistently from the beginning due to increasing demand to publish more. The numbers of new authors joining the network each year are shown in Fig 5. From Fig 6(a), we can see that number of papers per author is also increasing with time and reaches to about 0.65 at the end. The reason for this is the increase in collaboration between the scientists. The number of authors per paper (Fig 6(b)) and the number of papers per author both are increasing with time.

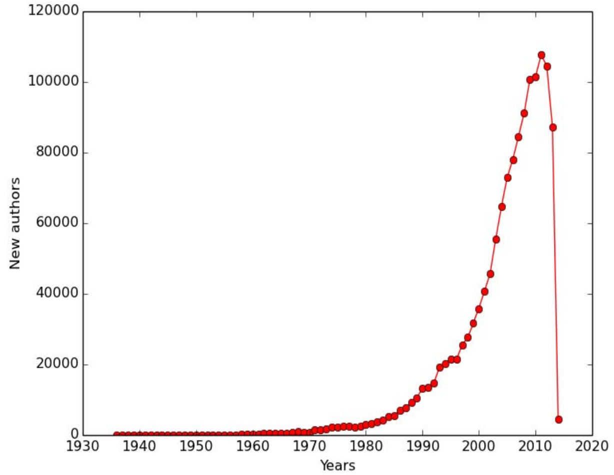

Fig 5: Number of new authors joining each year

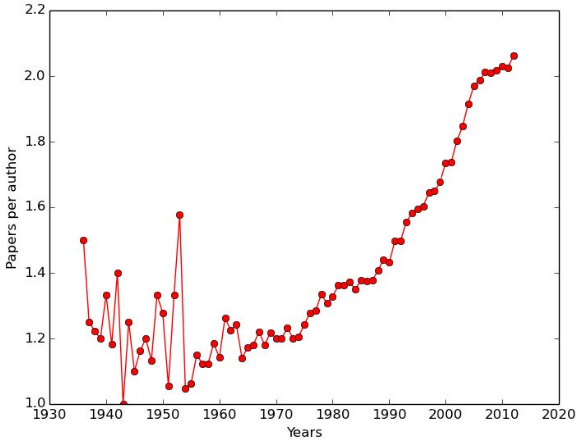

Fig 6(a): Number of papers per author each year

Next we plotted the total number of authors and number of papers published by them to see the variation in one with respect to change in the other. The plot we obtained (Fig 7) is following Lotka's law for frequency of publications [2].

Next we studied one of the most important characteristics of real world graphs i.e. degree distribution. In the collaboration network, as described earlier, the nodes are the authors and the degree of a node will be the number of authors with which the author has collaborated at least once. *Degree distribution* is the plot in which values of degrees are on x-axis and the fraction of nodes having a specific degree are plotted on y-axis corresponding to that degree on x-axis. The degree distribution is shown in Fig 8.

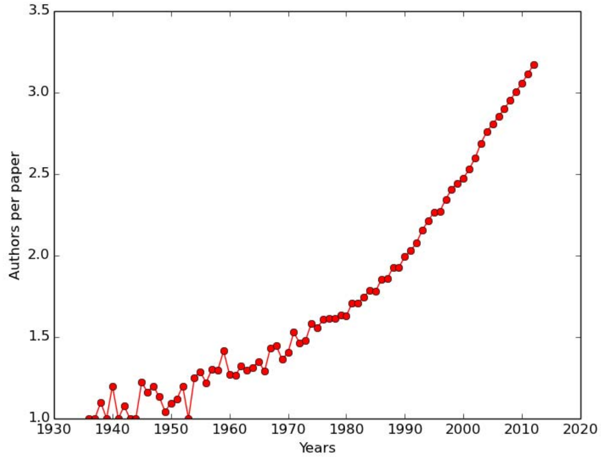

Fig 6(b): Number of authors per paper

From this figure, it is very difficult to analyze that the degree distribution is following a power law ($p(x) = Cx^{-a}$, where a=exponent of power law, see [8] for details) or not. The basic strategy, to check if a curve follows a power-law or not, is to draw the histogram on logarithmic scale and see if it is a straight line or not. But it is poor way to check a power-law because of the noise we get in the right side of the curve (Fig 9) known as 'noisy tail'. There are various procedures to reduce the noise in the curve, one of them being logarithmic binning [8]. We applied logarithmic binning to our data and we found that the degree distribution is following power-law with an exponent of around -1.5.

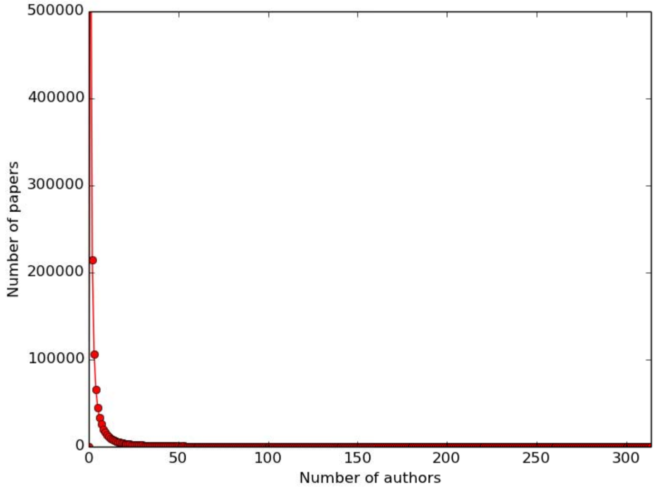

Fig 7: Lotka's law

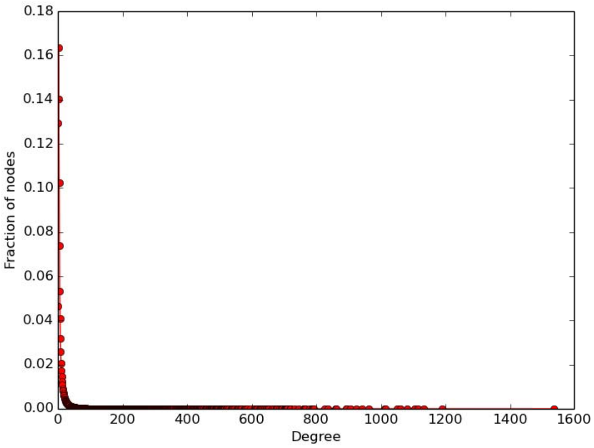

Fig 8: Degree distribution of collaboration graph

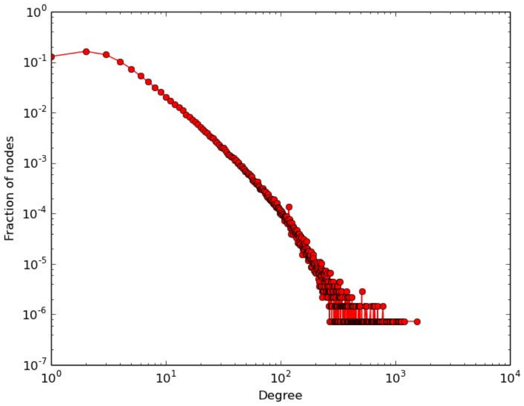

Fig 9: Degree distribution on log-log scale (with noise)

Giant Component- Next we analyzed the properties of the giant connected component and their variation with time. First we calculated the percentage of number of nodes contained by the giant component for each year, taking into account, the cumulative data up to the year (Fig 10). The fluctuations from 1936 to 1960 are because the network is in the development stage i.e. the growth of giant component has not yet started. After 1960, the giant component starts to grow and then reaches up to ~84% of the whole network at the end of 2013.

Clustering Coefficient- Next we calculated the clustering coefficient of giant component for each year and observed its variation with time. The plot can be seen in Fig 11. In 1960 the clustering coefficient is very large because of a few papers having very large number of authors. From year 1960 to 1975, the clustering coefficient for the graph is decreasing and then there is a sudden increase in its value and it reaches to 0.6 for the whole collaboration network. It implies that the chances for 2 authors to collaborate, if they have a common collaborator, are increasing with time.

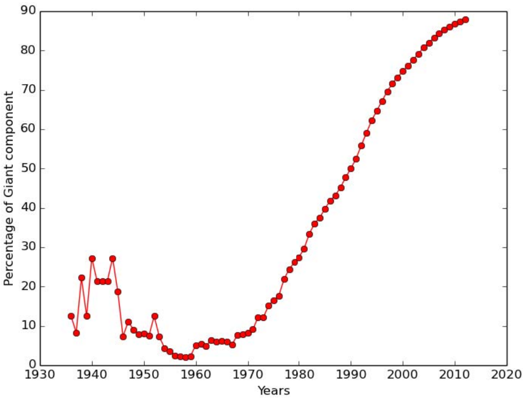

Fig 10: Size of giant component each year

We also tried to check whether our collaboration network follows some real world graph properties or not. Inspired by [9], we tried to observe the variation of size of 2nd and 3rd largest connected component with time. We found that the results were same as explained in [9] i.e. the 2nd and 3rd largest connected components have sizes oscillating about a constant value after the spike (gelling point, at which the giant component starts to grow) as we can see from Fig 12 and Fig 10.

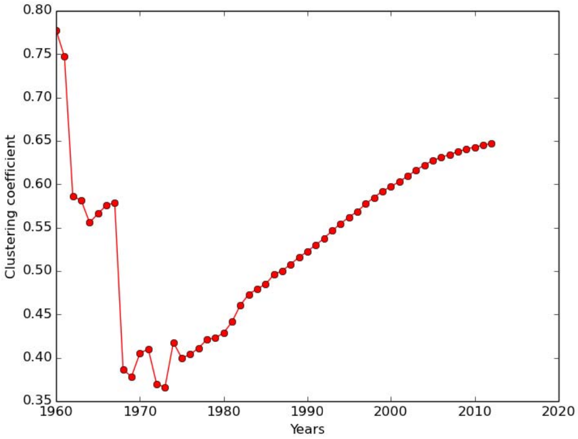

Fig 11: Clustering coefficient of giant component

The evolution of collaboration graph also follows the Weight Power Law ($W = kE^a$, where W is the weight and E is number of edges in network, k=constant, a=exponent, see [9] for details). It is a plot between the variations in weight of the network with change in the number of edges in the network with time. It also follows a power law with exponent of about 1.29. The log-log plot for weight power law is shown in Fig 13. The collaboration network also follows the Densification power law ($E=kN^a$, where E=Number of edges, N=Number of nodes), which is derived from the plot between the number of edges and the number of nodes that are being added to the graph with time (Fig 14), with an exponent of about 1.31.

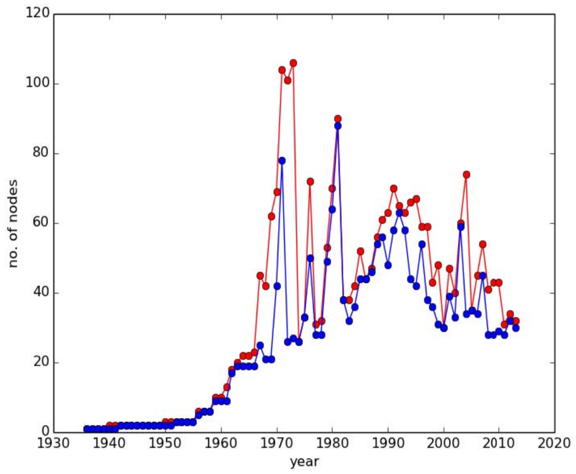

Fig 12: Sizes of 2nd and 3rd largest giant components

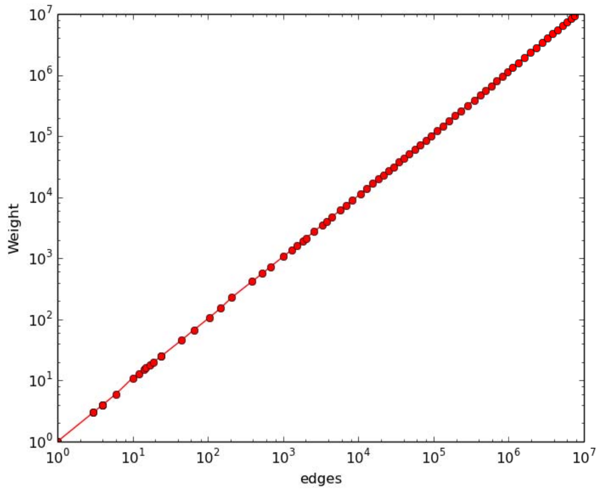

Fig 13: Weight power law

Inspired by [7], we applied Lorenz curve to our collaboration network (Fig 15) to observe the concentration of collaboration between scientists and calculated the Gini coefficient to estimate the amount of inequality in distribution of collaboration concentration (*Gini coefficient* is the ratio between the area contained between the Lorenz curve and the equidistribution line (Blue curve in Fig 15) and the area representing maximum concentration. In Fig 15, the ratio G=A/(A+B) represents Gini coefficient. Gini coefficient of 1 represents maximum inequality and G=0 represents perfect equality in the distribution). The Gini coefficient for the collaboration network comes out to be 0.61, indicating a large inequality in distribution. We also tried to observe how the Lorenz curve and the Gini coefficient are changing with time. The plot of variation of Gini coefficient with time is shown in Fig 16. We can see that the inequality (Gini coefficient) in the distribution is increasing consistently after 1975.

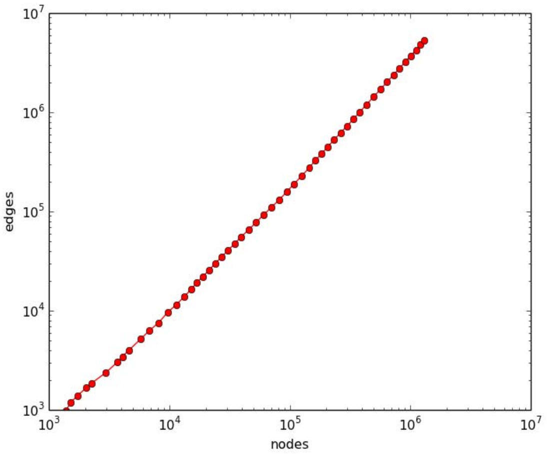

Fig 14: Densification power law

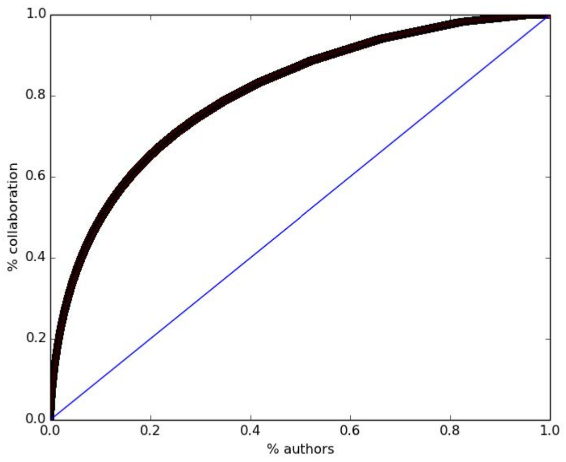

Fig 15: Lorenz curve for collaboration network

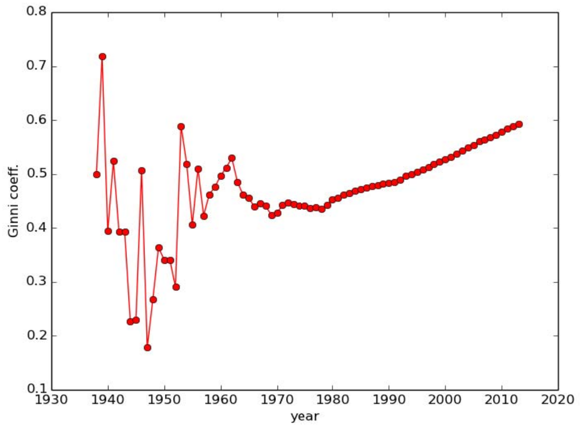

Fig 16: Gini coefficient for cumulative data up to the year indicated

Conclusions

We analyzed the collaboration network of scientists by studying various properties of the network. We also tried to understand how the network structure has changed over time. We studied the distribution of degree and the number of collaborators per paper over time. Both the distributions are well fit by power law forms. Frequency of publications also follows Lotka's law. We also found that clustering coefficient of graph is large and increases with time which means that two scientists are much more likely to be collaborated if they have a common collaborator. By applying the concepts of Lorenz curve and Gini coefficient, we also found that the concentration of collaboration in the network is not equally distributed and this inequality in the distribution is increasing with time.

Future work

We are also working to divide the whole collaboration network constructed from the DBLP database into different collaboration networks where each network will correspond to authors publishing in a specific conference and the evolution of these networks can be studied and compared to analyze how the evolution varies in different types of conferences. For example we can compare the evolution of collaboration networks constructed from conferences related to data mining and discrete mathematics and find some interesting patterns.

Acknowledgment

G. Madaan thanks Professor Anirban Dasgupta (IIT Gandhinagar), Mr. Priodyuti Pradhan and Professor Karamjit Cheema (Thapar University) for helping him throughout the work and believing in him.

References

[1] DBLP Computer Science Bibliography. [On-line]. Available: http://dblp.uni-trier.de/xml/ [Feb. 2, 2014]

[2] A.J. Lotka. "The frequency distribution of scientific productivity,". *Journal of the Washington Academy of Sciences*, vol. 16, pp. 317-323, 1926.

[3] M. E. J. Newman. "The structure of scientific collaboration networks," in *Proceedings of the National Academy of Sciences* 98(2), 2001, pp. 404-409.

[4] A. L. Barabâsi, Albert-Laszlo, H. Jeong, Z. Néda, E. Ravasz, A. Schubert and T. Vicsek. "Evolution of the social network of scientific collaborations," *Physica A: Statistical mechanics and its applications* 311(3), pp. 590-614, 2002.

[5] Networkx. [On-line]. Available: http://networkx.github.io/ [July 20, 2014]

[6] E. Elmacioglu and D. Lee. "On six degrees of separation in DBLP-DB and more," *ACM SIGMOD Record* 34(2), pp. 33-40, 2005.

[7] M. Franceschet. "Collaboration in computer science: A network science approach," *Journal of the American Society for Information Science and Technology* 62(10), 2011.

[8] M. E. J. Newman. "Power laws, Pareto distributions and Zipf's law," *Contemporary physics* 46(5), pp. 323-351, 2005.

[9] M. McGlohon, L. Akoglu, and C. Faloutsos. "Statistical properties of social networks," In *Social Network Data Analytics*. Springer US, 2011. pp. 17-42.

[10] J. Leskovec, J. Kleinberg, and C. Faloutsos. "Graphs over time: densification laws, shrinking diameters and possible explanations," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 2005, pp. 177-187.