Study of Collaboration Network in Dept. of Computer Science & Engineering, IIT Kharagpur

THE FINAL REPORT
FOR THE TERM PROJECT
ON
COMPLEX NETWORK THEORY

by

Arnab Sinha

Mithun Dhali

Lalit Narayan Paswan

under the guidance of

Dr. Niloy Ganguly

Department of Computer Science and Engineering
Indian Institute of Technology
Kharagpur
April 2006

Contents

List of Figures

1 Introduction

Collaboration networks in the scientific communities are a well-studied subject for its inherent complexity and motivation to predict or analyze certain features among the persons involved. In various literatures [1, 2], we find the researchers investigated into certain parameters like *small-world*, *betweenness centrality*, *vertex centrality* etc to interprete the obtained data. In this term project, we would like to investigate the co-authorship collaboration network among the students, researchers, and primarily the faculty members of the *Department of Computer Science and Engineering, IIT Kharagpur*.

2 Motivation

Before discussing the various parameters, it is utmost important to know what we are interested in looking into. Primarily we are interested in investigating into the **health** of the research community. The next question is what do we mean by the health of the collaboration network in CSE, IIT-KGP. We try to answer this question in the following subsections.

2.1 What do we mean by Health of A Research Community?

We admit that this is a difficult question to answer. Broadly, health of anything stands for the current state-of-being of something. That might be an individual or an institution. By the term "*good health*", we understand the physical well-being for an individual, while for an institution, we mean that the institute is thriving, full of activity, prospering, inviting more students, researchers, collaborators etc. Well, this definition is not complete. In order to get the complete picture, we need to understand how different people interprete the *health* of an institute/department.

2.2 A Student's Perspective

Suppose, a new student wants to join the department in the postgraduate level. He has several questions in mind. We list them as follows.

- Does any faculty-member work in his field of choice in the department?
- If yes, how active is he? how many recent publications does he have in that field?
- Does he have any strong group? If yes, how many research scholars are there? How strong is that group?

- Does the faculty member has a variety of research interests enabling him with greater flexibility to choose some other field, in case he feels like doing so?
- Which group/faculty member is likely to *grow* in near future?

2.3 A Company's Perspective

Again suppose, a company wants to fund a new project in the department for the first time. The company is likely to be interested in the following questions.

- Does there exist some strong and active group who works in the same field?
- If yes, have they done any other project in recent past in the same direction?
- Suppose the company know some renowned senior faculty in the department, and also finds that another junior faculty collaborates with him regularly. It is likely, that they might be offering the project to the junior faculty member if they find the senior professor busy. So they are interested in knowing the other members of the hub led by the renowned senior professor.
- Suppose, that there is considerable proportion of faculty members who once were alumni of the department. So they might be interested to know whether the senior professor supervised the junior faculty in question.

2.4 An Administrator's Perspective

Sometimes the Ministry of Human Resources would like to enquire whether the department is at all productive in recent times. They are likely to enquire the following.

- How many students and faculty members are currently there in the department?
- How much research contribution are they making? (through journal and conference publications)
- Who all are elligible for the promotion? The number of publications, students under him, leadership qualities etc reflect the eligibility quite precisely.

2.5 Assimilation of Perspectives

We understand that different individuals/authorities have different views and queries regarding the department. Essentially, **the union** of answers to their views gives us the complete vision regarding the health of the department. So, after proper assimilation of the concept of health, we want to answer their questions

quantitatively, and hence we will be using the various tools/parameters which justify those intuitive and abstract answers.

3 Data Collection and Network Building

We will be collecting the data from the following sources.

- **Department Faculty Homepage**: The website is following, http://www.facweb.iitkgp.ernet.in
- **Annual Report**: The report is with the CSE office.
- **Personal Contacts**: By visiting different labs and talking to researchers there.

We will be collecting the following data on each paper.

1. **Name of authors**
2. **Title of Paper**
3. **Year of publication**
4. **Journal/Conference**

We will be building the network for a given span of years (say year-wise) to study the *evolution* undergoing in the network. The collaboration network is essentially a graph (G) where the vertices (V) represent authors and the edge between them represents the fact that they are related by the relation of *co-authorship*. Each edge has certain weight reflecting the number of papers written by a pair of members.

4 The Analysis of Collaboration Network

We will be dealing with the following aspects of the network, closely following the definitions in [1, 4].

4.1 How *Strong* are the Collaborative Ties?

The number of papers produced by a pair of collaborators indicates the strength of "*collaborative ties*" among them. Suppose, a paper k is written by n_k persons. Then it is natural to assume that a particular author is acquainted with $1/(n_k - 1)$ authors. May be this is a crude approximation since, it is natural that he might have spent much of the time with only a few persons among the co-authors. But for the time

being due to lack of data, we live with it. Also, suppose that $p_i^k = 1$ and $p_j^k = 1$ if author i and j co-authored paper k otherwise equal to zero. Then, w_{ij} represents the strength of the collaboration network given by,

$$w_{ij} = \sum_k (p_i^k \; p_j^k)/(n_k - 1)$$

Note that the equivalent vertex degree for our weighted network, i.e. the sum of the weights for each of an individual's collaborations-is now just equal to the number of papers they have co-authored with others [1].

$$\sum_{j(\neq i)} w_{ij} = \sum_k \sum_{j(\neq i)} (p_i^k \; p_j^k)/(n_k - 1) = \sum_k p_i^k.$$

Here we assumed that the distance between a pair of researchers is inversely proportional to the weight of their collaborative tie. Here we cannot use breadth-first search [1] since, shortest weighted path may not be the shortest path in terms of number of steps. We will be using Dijkstra's algorithm here.

4.2 Who has more *"Connections?"*

Faculty members with more connections are assets to any department, since they are the persons who attract students, funds and projects in the department. They are likely to be the future leaders in the department as well as involved in cutting-edge and relevant research activities. In order to answer the above query, we aim to find out the *geodesic path length* between each pair of vertices. This would also help in analyzing the *vertex centrality* of the vertices. The faculty with more connections is often the more preferred one over the other faculties. We would be calculating the shortest paths by using the modified breadth-first algorithm mentioned in [1]. In this context, we are more interested in finding out the number of steps between a pair of members. Thereby we would like to know whether the network is a ***small-world community***. Moreover, the average path length between the researchers in CSE, IIT-KGP can be of our interest (if we want to study the diameter of the network). This would be good measure to find who are the better connected people in the department.

4.3 Who are the *Hub Leaders*?

In order to identify the *leaders* in the network, the quantity of interest in many social network studies is the "*betweenness*" of an actor i, which intuitively hints that persons with high "betweenness" are indispensable to the department due to the information flow they assist in. In research, information flow is very important since one should know about the facts like the current areas of academic as well as commercial value. Formally, it is defined as the total number of shortest paths between pairs of actors that pass through i [5]. This quantity signifies the most influential people in the network. These vertices with high betweeness when removed typically result in increase in distances [6]. We will be following the algorithm given in

[1] with a complexity of $O(mn)$, m and n being the number of edges and number of vertices respectively. Often we find that researchers try to work with some senior/famous people in a certain field, which we denote as "*preferential attachment*". We call this effect as "*funneling*", i.e. working with just one or two famous people in a field one can easily establish relation with the other members of that field.

4.4 Whom does more people like to collaborate with?

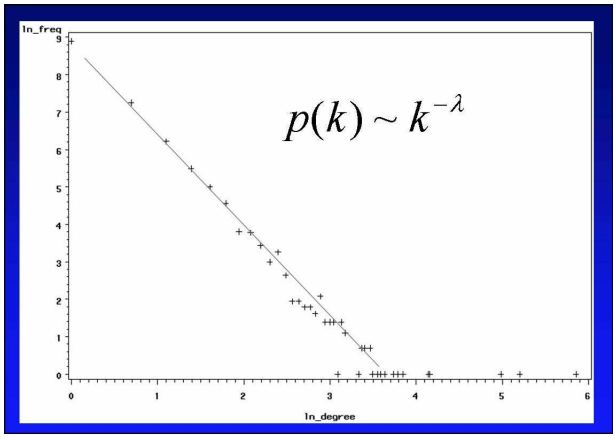

Figure 1: The degree-distribution of a typical scale-free network

Some people in the research community are found to be more fit for attracting new collaborators or new students. Hence, we would also like to investigate whether the network is scale-free (Barabasi & Albert, 1999). Scale-free networks appear when new nodes enter the network by attaching to already popular nodes [4]. Thus the degree-distribution appears to be pretty skew in nature (ref Fig 1).This also explains the phenomenon *preferential attachment* when the already popular nodes increase in connectivity with time. This kind of graph follows the *power-law*

$$p(k) = k^{-\lambda}$$

where, k is the degree and $p(k)$ represents the probability that any randomly chosen author would have a degree equal to k, and λ is a constant. We slightly modify the definition to find the *cumulative probability*.

$$P(k) = \sum_{i \geq k} p(i)$$

where, $P(k)$ represents the probability that a certain person chosen randomly has more than or equal to k publications.

4.5 Who is the *Boss*?

Suppose, a paper k is written by n_k authors(a_1, $a_2 \ldots a_{n_k}$). In Indian context it is important to study the order too. It is usual to find the name of the senior scientist(s) towards the end. We are interested in finding the seniority of the member by merit of his age/experience etc. The motivation behind computing this metric is the fact that many of the faculties earned their doctorate degree in this institute only. We propose a metric called **hierarchy rank($r(v)$)** for the vertex v. The author a_i (denoted by v, say) gets a weight i. In order to normalize, we define, the following,

$$r_k(v) = i/n_k.$$

The overall *hierarchy rank* for the node v would be following,

$$r(v) = \sum_{k \in Coll(v)} r_k(v)/d(v).$$

where, $d(v)$ is the sum of weights in the incident edges of vertex v is the normalizing factor, $Coll(v)$ being the papers co-authored by v. We can extend this parameter to determine who is likely to be the supervisor of a particular person. So in essence we want to capture the notion of *guru-shishya* relationship, which is primarily an Indian concept.

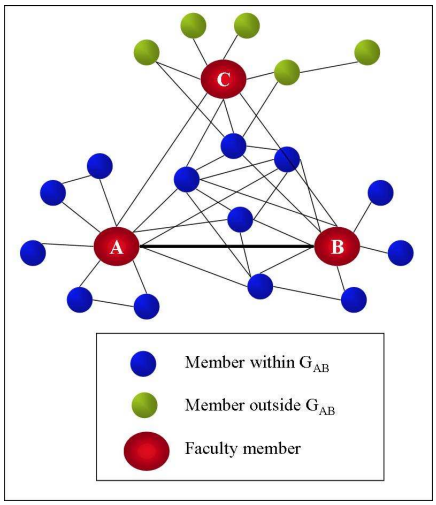

Figure 2: The Community Structure defining the hiearchy rank of A and B

In the department of CSE, it is often found that projects are investigated by more than one faculty member. For example, let say A and B(as shown in Fig 2) are two faculty members who are jointly involved in some project. In Indian context, as the project is same, A personally visits the lab where he

regularly meets and spends time with his own students as well as those of B too. Same goes with B also. Now we want to define the community structure with centered around A and B, let say G_{AB}. The vertex set $V(G_{AB})$ is defined as,

$$V(G_{AB}) \subseteq N_{AB} = \{v \mid v\ is\ adjacent\ to\ either\ A\ or\ B\ where\ A\ and\ B\ are\ adjacent\}$$

and G_{AB} is the subgraph induced by $V(G_{AB})$. Essentially, the vertices at a distance more than 2 from either A or B are less likely to be supervised by both, and hence that person should not be considered in the group led by both. The problem next lies in pruning the subgraph induced by N_{AB} as it comprises of all the neighbors of both A and B. We will be using the algorithm suggested in [7] to find the community structure by iteratively eliminating the edges with high betweenness (since edges connecting the communities have high betweenness). Now we are interested in finding out the *hierarchy rank* of both A and B within the network G_{AB}, we call it **hierarchy rank w.r.t** G_{AB}. We can deduce the following facts from their hierarchy ranks.

- whether A is senior/more experienced to B in this field.
- if $r(A) >> r(B)$, it is more likely that A supervised B.
- if $r(A) \approx r(B)$, both A and B are the leaders in the group.
- study of the evolution of community hierarchy ranks, can extract useful information regarding the dynamic aspects of a group, e.g. whether B is elligible to take the responsibility of the group in absence of A, etc.

5 Tool Architecture

The architecture of the tool that we developed is shown in the Fig 3. The data is first fed into the java program in a specific format. The format contains the name of the paper, the authors and lastly the other details. We parse the input to store it in our database in a suitable data-structure. We run the microsoft excel as well as the pajek [8] in the backend to get the results as shown in the Fig 3.

6 Results and Analysis

We have studied and tried to come up with meaningful interpretations from our study which are given as follows.

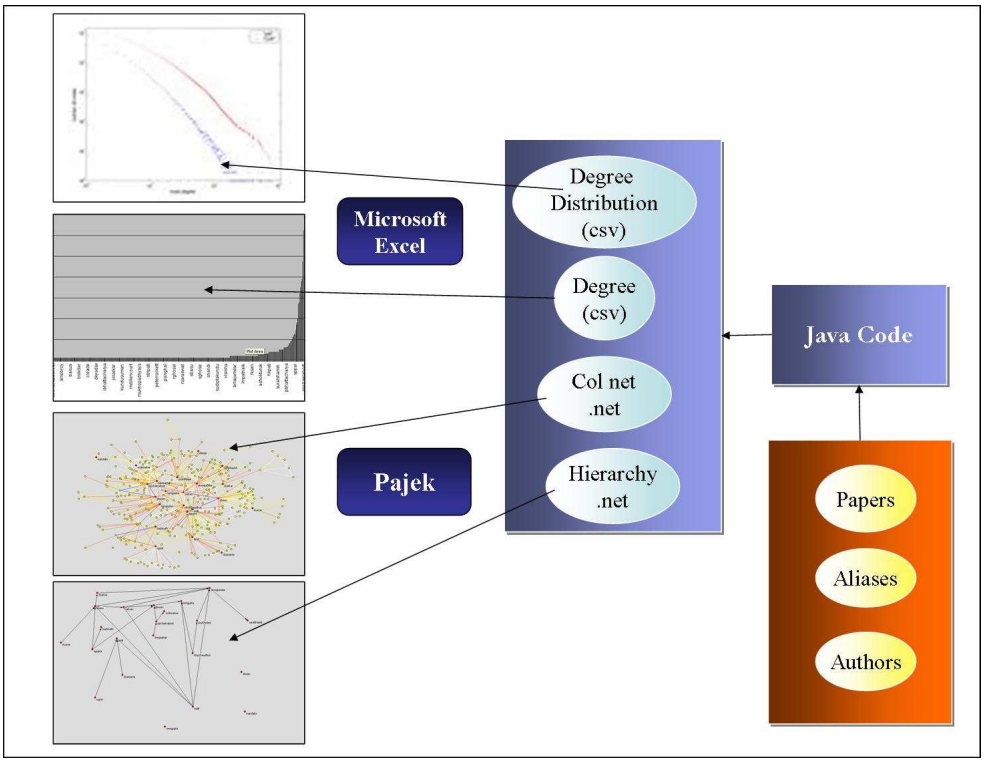

Figure 3: The Tool Architecture

6.1 Collaboration Network

In the Fig 4 we have shown the overall collaboration situation in the department. In the network we have labeled edges blue(with highest collaborative bonds), red(moderate collaborative strength), yellow(weak collaborative bond), white(poor strength of collaboration) respectively.

In this figure we can identify the following features.

1. The distinct research communities headed by Prof PPC, ISG, AKM, AB are identified.
2. The various degree of collaborative bonds are also identified. e.g. the bond between PPC and SG, SS and AB, AB and RM are among the very strong bonds in the dept.
3. Some of the young faculties like CRM, D.Samanta and experienced faculties like A.Pal, SPP are found less collaborative.

We have also studied the dynamic aspects of the collaboration network in the department. Fig 5 and fig 6 shows the collaborative scenarios in 1995-1999 and 2000-2003 respectively.

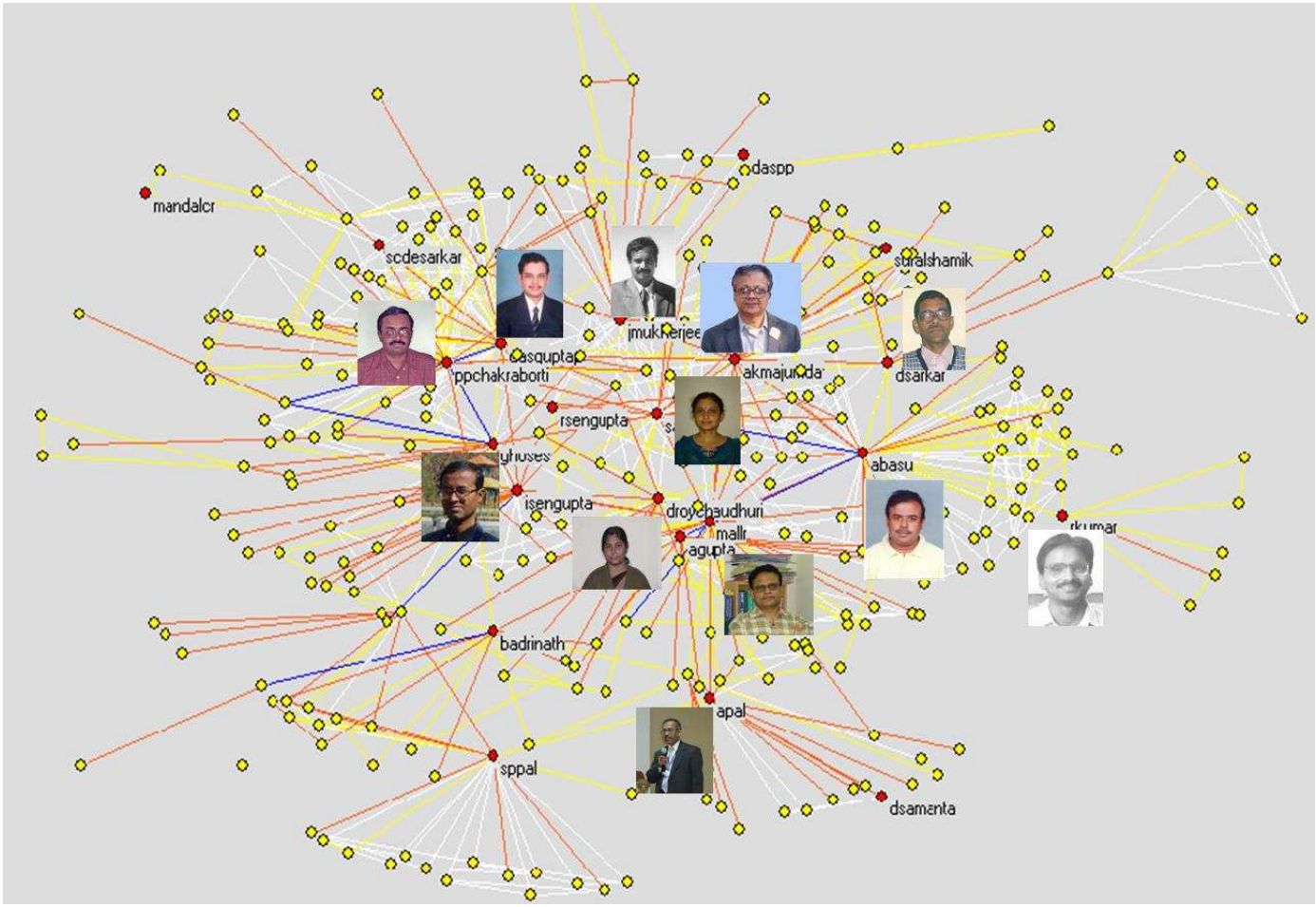

Figure 4: The Collaboration Network 1995-2003

We present our findings as follows.

1. We can easily find that the collaborations have hiked over the years.

2. The prominent hubs in 1995-1999 are those of PPC (along with SCD, SG and PDG), AKM, and JM respectively. While in 2000-2003, the network has grown a lot. New hubs under the leadership of ISG (along with DRC), AB (along with SS) have come up.

3. Also the hub under the leadership of AKM has diminished in the order of their scientific contributions.

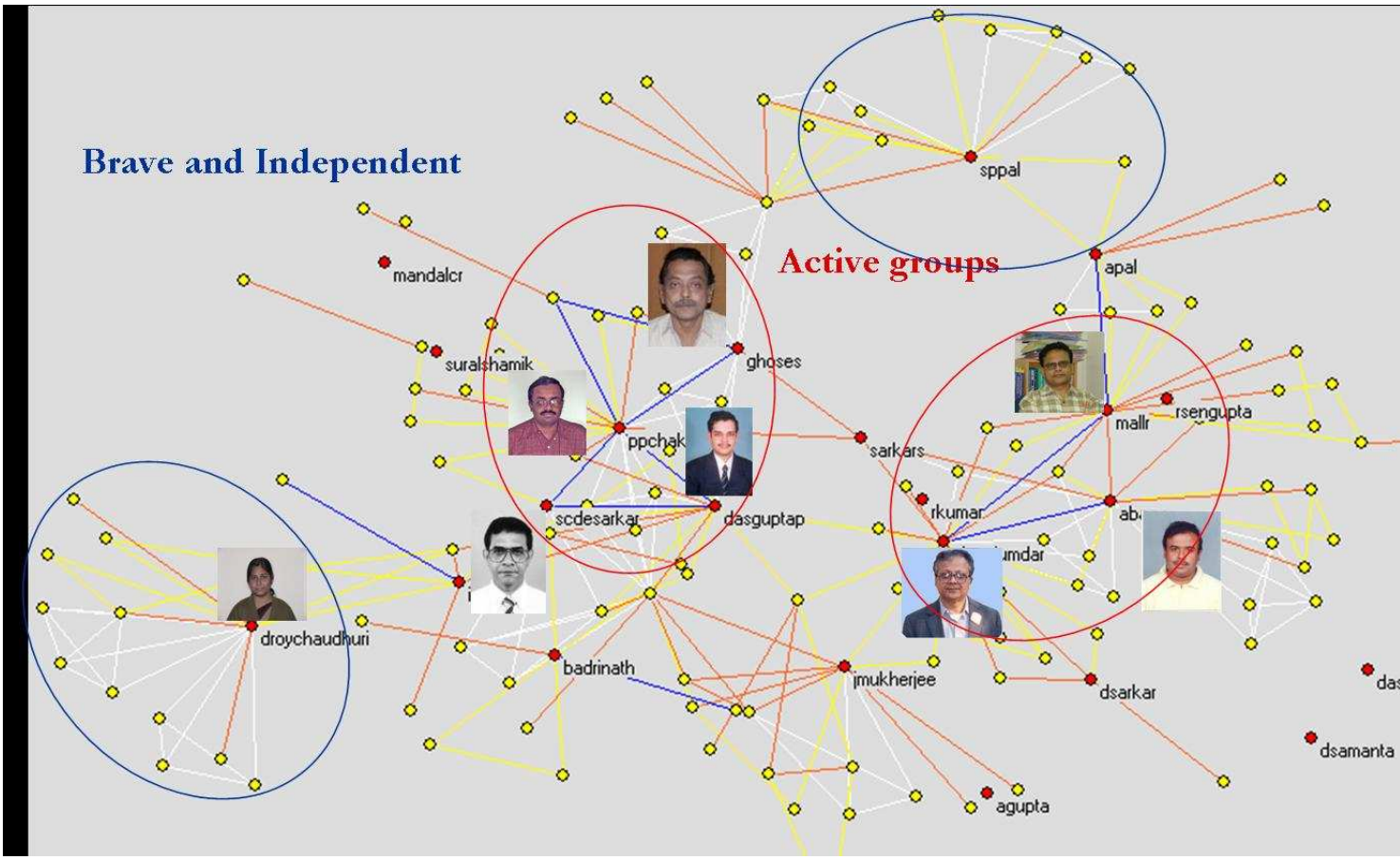

Figure 5: The Collaboration Network 1995-1999

4. Senior faculty members like SG and SCD became busy with administrative work-loads and hence we find them less productive, while faculty member like RM, AB remained to be equally productive.

5. New collaborative ties between AG and AB, ISG and RM have emerged.

6. We can also predict that groups led by AB, PPC (along with PDG) will thrive in near future in a greater way.

6.2 Hierarchy Rank

In this subsection we studied the novel parameter - hierarchy rank. This network is essentially a *poset* where $arc(i, j)$ represents that i is junior to j. The edge weights quantifies the seniority and contributions

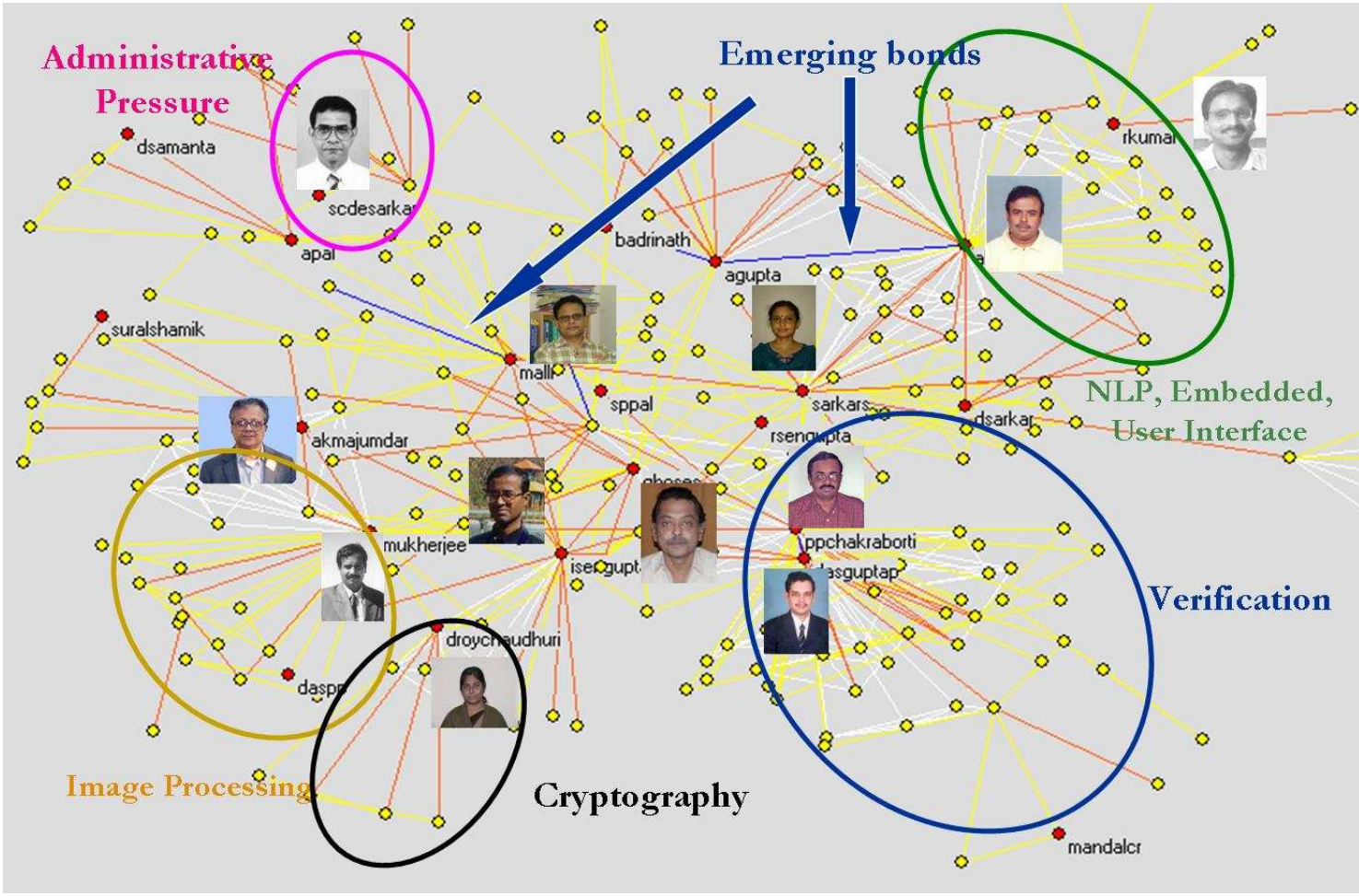

Figure 6: The Collaboration Network 2000-2003

in the work done together by i and j.

The following are our observations.

1. The hierarchy rank shows that AKM, ISG, SG, AB are among the most senior faculties in the department.

2. It is interesting to note that the arc between SG and SS has the highest weight (0.91) and she did her PhD under SG only. We find similar behavior among D. Samanta and A.Pal also. So we can say that if the hierarchy rank value is more than 0.75 it is more likely that the two authors are linked through supervisor-student(*guru-shishya*) relationship.

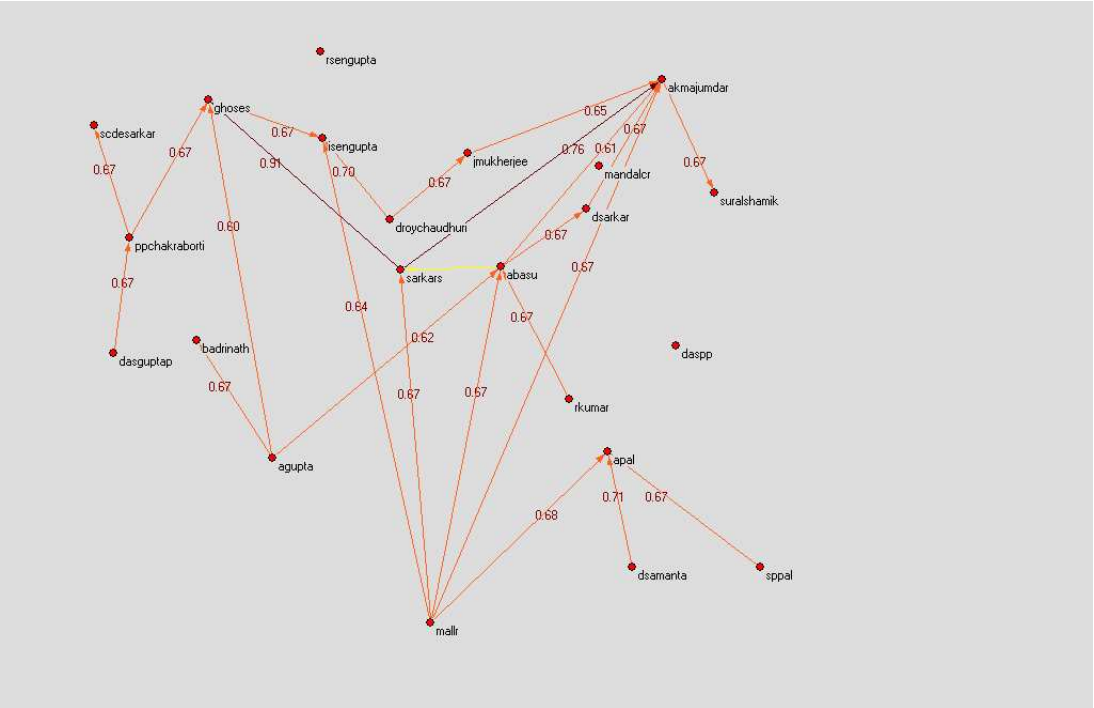

Figure 7: The Hierarchy Network 1995-2003

We also study the evolution in the hierarchy network.

The evolution of the hierarchy network reveals interesting features.

1. Fig 8 shows that arc between PDG and PPC had a very high coefficient of 0.84 but in the fig 9 that has decreased to 0.62 only. This indicates that PDG has carried independent research after gathering adequate experience working under PPC.

2. Also AKM has mentored quite a lot of faculty members in the department in 1995-2003.

3. Moreover, we find that there can be error also. We find Shamik Sural senior to AKM. The reason behind this error is as follows. They have co-authored very few papers where the contribution of AKM was greater. Hence, this error cropped up. Nevertheless, when the number of co-authored papers are greater this kind of errors vanish.

4. We can logically predict that ISG might be replacing AKM as mentor for young faculty members in near proximity.

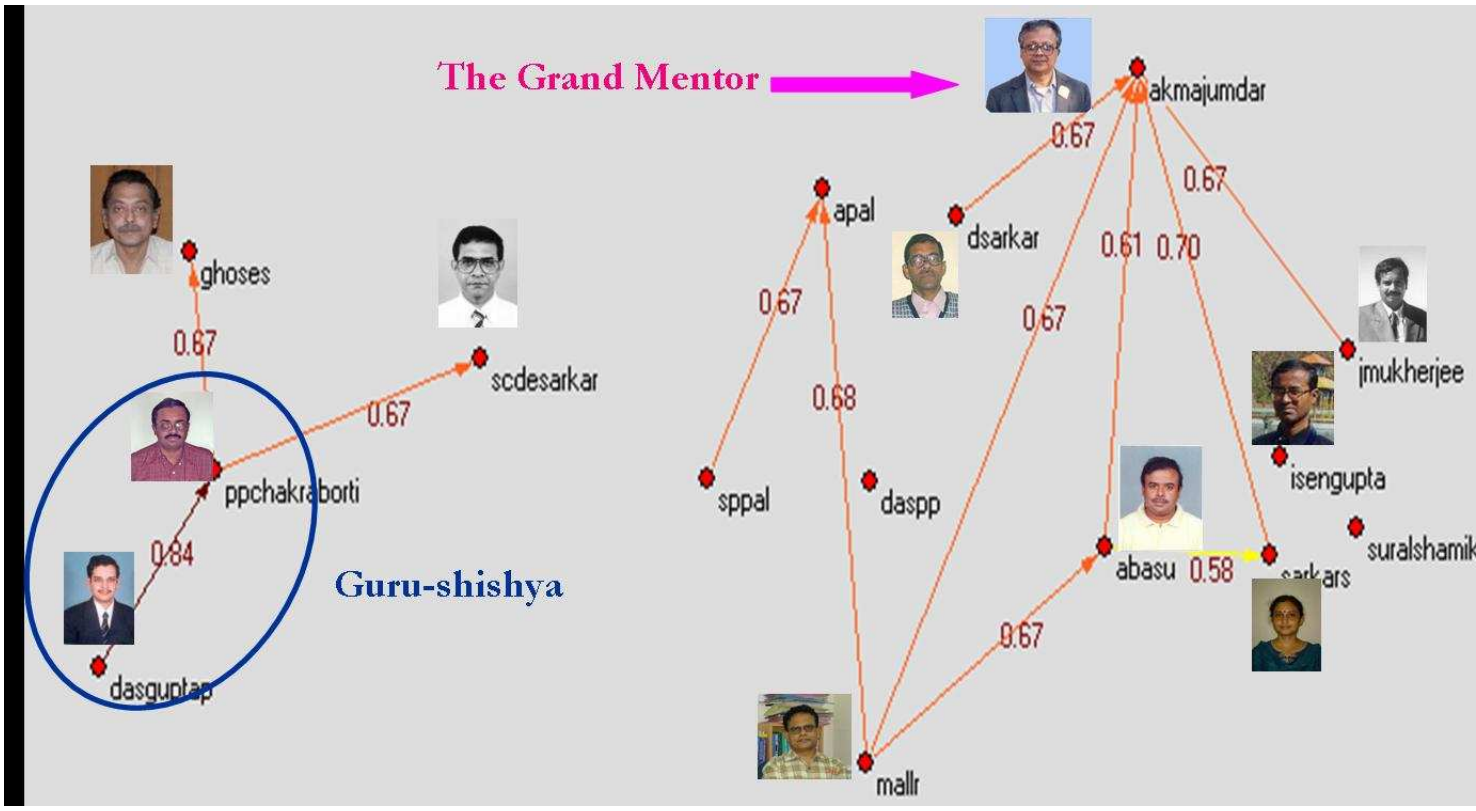

Figure 8: The Hierarchy Network 1995-1999

6.3 Degree Distribution

We have plotted the degree distribution as well as cumulative degree distribution for the two different spans of our investigation i.e. 1995-1999 and 2000-2003. We have found that the number of publication has increased during the second span. Also, it is interesting to note that the curve has got higher inclination in the latter mentioned period. This shows that the department is healthy and active. It has grown over the years. In fact the average number of publication for a person has also increased during the period. In the Fig 12 we find that it is following the powerlaw. We have shown the cumulative degree distribution for both the spans of investigation.

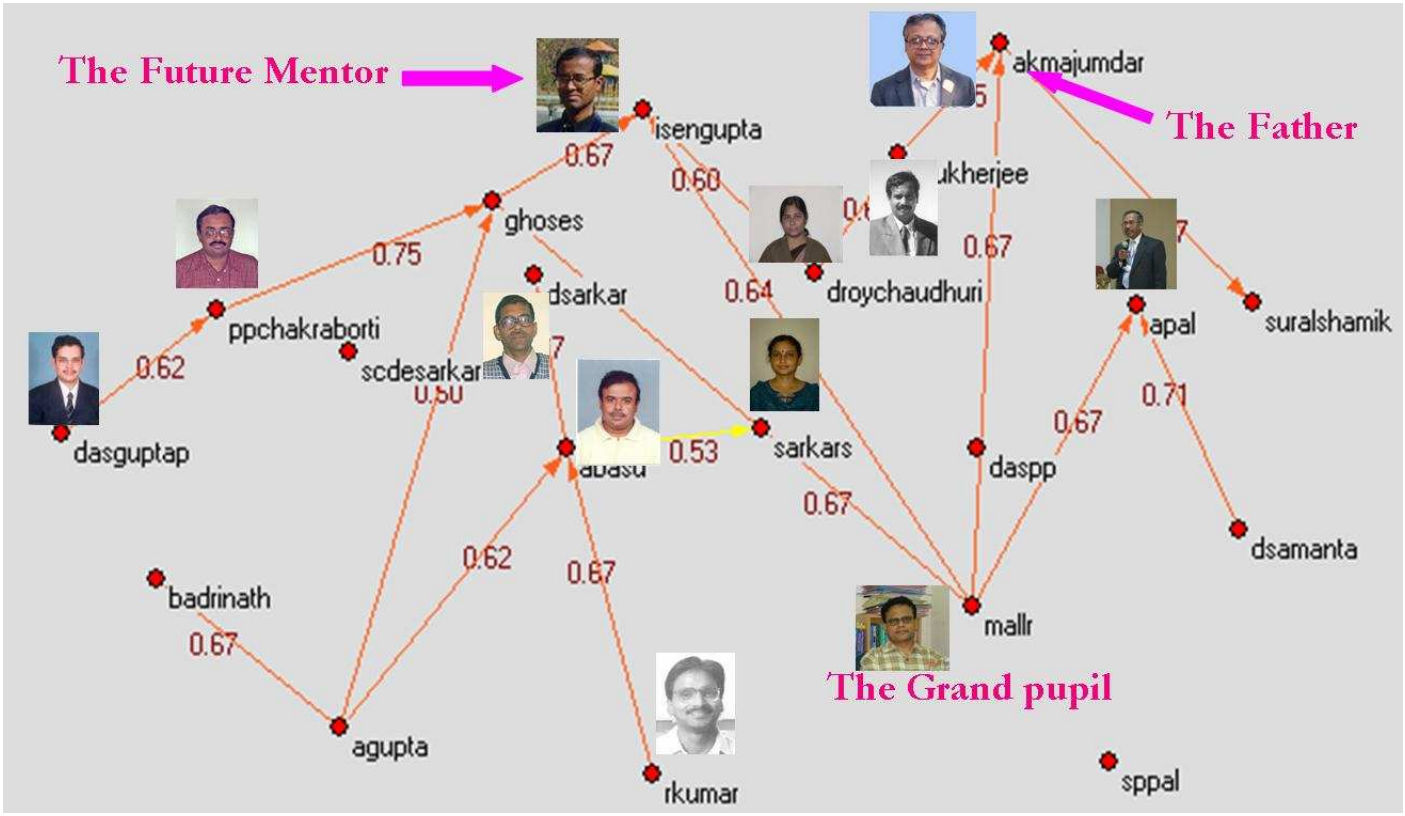

Figure 9: The Hierarchy Network 2000-2003

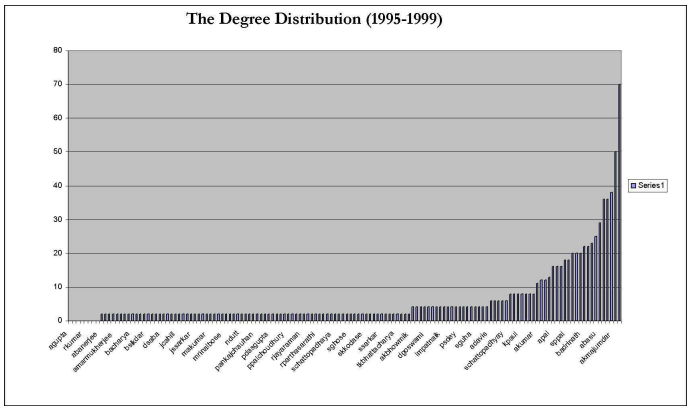

Figure 10: The Degree Distribution 1995-1999

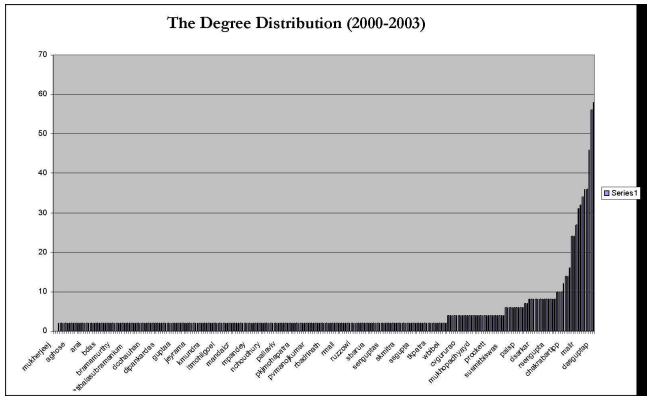

Figure 11: The Degree Distribution 2000-2003

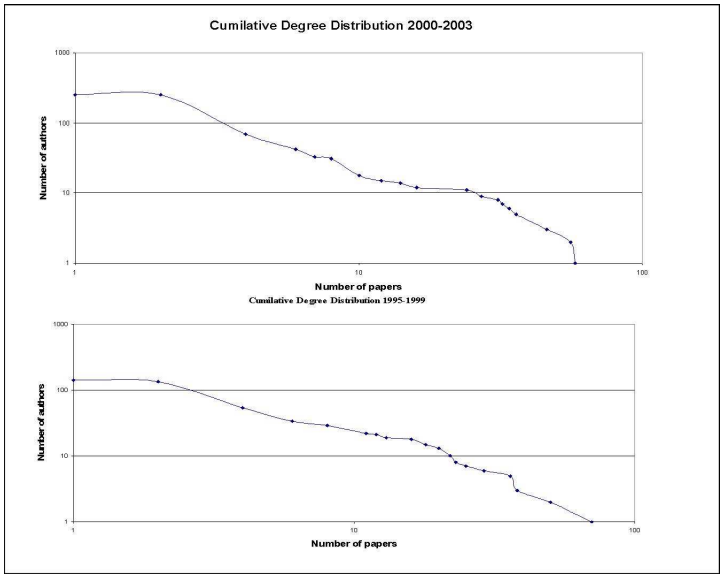

Figure 12: The Cumulative Degree Distribution

7 Conclusion

Through answering the above questions we can cover almost all the queries which are answerable studying the co-authorship network of a research community. We have revealed interesting facts about the department which are true in most of the cases (known to us from our internal sources of information). In order to find the changes in the network with time, we have also studied the mentioned parameters against time, i.e. we are interested in the **evolution of the network over time** and hence meaningful predictions regard-

ing the future of the scientific productivity of Department of Computer Science & Engineering, Indian Institute of Technology(IIT), Kharagpur.

References

[1] M. E. J. Newman, ”Scientific collaboration networks.II. Shortest paths, weighted networks, and centrality”, Physical Review E, Volume 64,016132

[2] M. E. J. Newman, ”The structure of scientific collaboration networks”, PNAS, January 16, 2001, vol 98, no.2, Page 404-409

[3] Carlos Cotta and Juan-Julian Merelo, ”The Complex Network of Evolutionary Computation Authors: an Initial Study”.

[4] James Moody, ”The Structure of a Social Science Collaboration Netowrk: Disciplinary Cohesion from 1963 to 1999”
www.sociology.ohio-state.edu/cjrc/bios/moody.html

[5] L.C.Freeman, Sociometry **40**, 35 (1977).

[6] S. Wasserman and K. Faust, *Social Network Analysis* (Cambridge University Press, Cambridge, 1994).

[7] Michelle Girvan and M.E.J.Newman, ”Community structure and biological networks”.

[8] http://vlado.fmf.uni-lj.si/pub/networks/pajek/