Scientific Collaborations

Aaloke Mozumdar
aaloke19004@iiitd.ac.in
IIIT-Delhi
India

Gitansh Raj Satija
gitansh19241@iiitd.ac.in
IIIT-Delhi
India

Karan Abrol
karan19366@iiitd.ac.in
IIIT-Delhi
India

Karanjot Singh
karanjot19050@iiitd.ac.in
IIIT-Delhi
India

Rohan Jain
rohan19095@iiitd.ac.in
IIIT-Delhi
India

ABSTRACT

Collaborative network analysis is a field with growing interest. Many researchers have studied the collaborative networks of research publications across different domains, universities, and found some interesting insights as well. Currently, we judge an author's profile by their number of publications, citations and other such metrics. However, we believe a measure of a coauthorship network of a researcher will also provide valuable insight to their research endeavour. In this paper, we analyze the collaborative network of research work done by individuals associate with the Indraprashtha Institute of Information Technology, Delhi, and try to build a metric to evaluate the impact of a researcher on the collaborative network.

CCS CONCEPTS

• **Information Retrieval**; • **Collaborative Network Analysis**;

KEYWORDS

Collaborative Networks, Social Network Analysis

ACM Reference Format:
Aaloke Mozumdar, Gitansh Raj Satija, Karan Abrol, Karanjot Singh, and Rohan Jain. 2018. Scientific Collaborations. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION AND MOTIVATION

Collaboration in any domain of life tends to increase the productivity of the task. Different scientific fields have seen many new discoveries in the past century, all of which have been because of the contribution of numerous researchers building upon each other's research. When researchers collaborate, they bring different skills to the table and pave the way for new scientific developments.

Studies in the past have proven that collaboration between authors leaves a positive influence on Research. This brings about a need to connect researchers to increase collaboration. The relational tie between authors may help identify long-term collaborations, common research interests, preferred conferences, and research groups under formation. Furthermore, as social ties evolve, new research interests and new collaborations can be identified. This can also help in identification of possible hidden collaboration nets. [4]

Thousands of researchers have published millions of work, and most of the publications have multiple authors. There authors can be working individuals, researchers or professors, and even students. Authors who have been associated with IIIT-D as a professor or as a student have thousands of publications when combined. In our work, we try to do a collaborative network analysis of a subset of research work done by IIIT-D associated authors.

2 PROBLEM STATEMENT

Analyse the scientific collaborative network among researchers at IIIT Delhi, by studying publication data and using it to create a web platform to link researchers (both faculty and students). Further, define metrics to determine the impact of researchers on the overall collaborative network and use these to identify budding researchers and subsequently incentivize them.

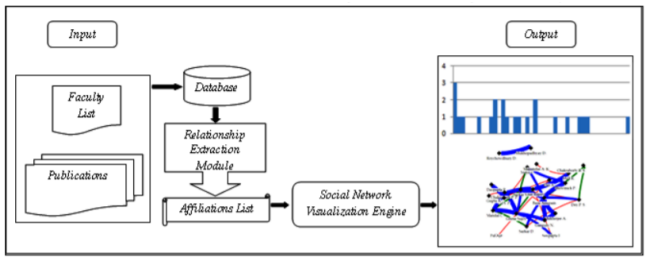

Figure 1: Social Network Extraction System Architecture [1]

3 LITERATURE REVIEW

Collaborative Network analysis has been a growing field amongst researchers. We look at various works which provide key insights and help us formulate the scope of our research.

In Cooperative Authorship Social Network [4] The authors propose an approach to build a co-authorship research social network that disseminates new publications and research connections to

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

individuals who subscribe to the service over a multi-layered architecture. The data is collected from the DBLP repository with around 677,345 authors.
First, the information about research groups and researchers is mined from the Web. Next, Their publications are organized in semi-structured data in a Digital Library to be processed by a DL interactive process that builds the network. The limitations of this study are the lack of an evolved and complete collaborative network and the lack of extensible and Web-scalable features.

In the research paper [3], the authors propose an approach to analyse the structure of the co-authorship network among researchers at the Italian Institute of Technology. The paper uses multiple metrics like centrality measures, density, degree distribution, etc. to measure an author's productivity and impact. The authors concluded that researchers who played a bridging role between researchers published better quality papers. They also found that researchers collaborated within the institute more than external collaboration.
For future research, they emphasized the importance of human capital of the institute to calculate its growth. The most important limitation of the study was that they used only the citation and publication counts to evaluate research performance, other bibliometric indicators could have been used.

In Scientific Co-authorship Social Networks [1], the authors study the collaborative networks of the computer science departments of 4 major IITs - Delhi, Kanpur, Kharagpur and Madras to investigate the hubs, connections and strength of collaboration ties using metrics - Betweenness Centrality, Degree Centrality, Clustering Coefficient and Average Degree respectively. It was observed that high degree centrality doesn't directly imply high betweenness centrality and vice versa. Further, concerning the strength of collaboration ties, it was observed that low clustering coefficients were associated with a lack of connectivity and very few nodes being arranged into cliques. A major limitation of this study is the effects of this study on overall research throughput have not been observed.

In 2005, the Ministry of Health, Brazil, launched a program to study certain 'neglected' diseases that are prevalent in the poor and marginalized regions [5]. The program aimed to foster technological innovation in the research that it was funding. In this study, the authors used SNA to develop new approaches to analyse the productivity of research being conducted in a region. The study attempted to map co-authorships between authors using the authorship data from publications of seven diseases having at least 1 Brazilian author. The authors have majorly focused on the component analysis of the overall network structure to reveal valuable insights into possible collaborations. They also studied the cut-points of the network which helped identify nodes responsible for connecting several other institutes in different regions. While these two metrics give essential insight on the collaborations made during the program, they fail to bring to light various other aspects of the collaboration network which might be essential to analyze the network that can be achieved by introducing other metrics.

Researchers from the University of Southern California [2] recently conducted a study for analyzing interactions between researchers and institutions. They were able to establish a superlinear relation between the number of active researchers and institute size. They also made an interesting observation that establishment of new institutions can 'trigger' even more potential institutions. IIIT-Delhi being a relatively new institution can be foundation for one such 'trigger'. This work doesn't provide any metric to evaluate an author's collaboration level however, this research motivates us to lay the foundation of a collaborative network analysis for the research done by our institute.

The above works motivated us to analyze the trends in co-authorship and research in our Institute and see if any fruitful and insightful results can be gained.

4 DATA

4.1 Data Sources and Collection

IIIT-Delhi has around 100 active professors who have together contributed in more than 5000 publications in their research endeavours. We also have an active community of Undergraduate and Postgraduate students who are actively involved in the research domain.
We retrived Google Scholar IDs of IIIT-D faculty by scaping the Irins portal which contains their research profiles. For the faculty, whose data could not be retrieved in this manner, we manually collected their google scholar IDs which will help us extract their publication data in a structured manner. There were some faculty who did not have a scholar ID. For the current scope of the project, we decided to exclude these professors from our dataset.

As mentioned in the previous sections, we planned to also incorporate research data of students who are persuing/ have pursued undergraduation, graduation or postgraduation degree from the IIIT-D institute. We identified various PhD Students from the IIIT-D website, and manually extracted their google scholar IDs.
We used the SerpApi [6] Tool to collect publication data of all the

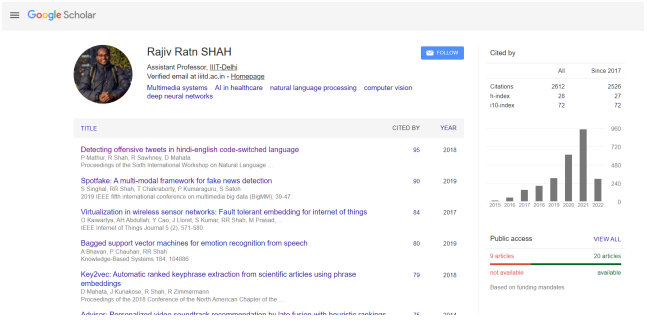

Figure 2: Sample Author Profile in the Google Scholar Platform. SerpApi tool helps extract all the information present in the profile

authors in our database from the Google Scholar platform. The tool provided us with meta data of the author (name, email, affiliations,

interests) along with their publication and coauthor list. We used these coauthor list to find more student researchers of IIIT-D by doing a regex based search on their verified email. The added authors who had a 'Verified email at iiitd.ac.in' to our database, after which we had 145 unique researchers with more than 6000 unique publications.

4.2 Publication Data Preparation

For all the 6,527 research publications in our corpus we had the list of authors that contributed to the project. However, SerpApi tool did not link the author names to the scholar IDs of the authors. Moreover, their was no standardization of names. For instance, we looked at several publications of Dr. Rajiv Ratn Shah, and found the following variations of his name.

- R Shah
- RR Shah
- Rajiv Ratn Shah
- R Ratn Shah

We went through several profiles and built a suitable regex to map these variations to the original author name. Since we followed a rule based approach, we looked for any exceptions/misclassified samples in our data. We identified around 100 samples for which the name variation belonged to multiple different authors. We corrected these samples manually by going over the publication title and indicating the original author. This way, we were able to map all the publications in our database with all the authors that contributed to it.

We created an Author Class of storing information about all the authors present in our corpus. The structure of the Class looks like the following:

- **Name:** Name of the Author
- **Type:** Faculty or Student
- **Category:** IIIT-D or Non IIIT-D
- **Pen Names:** List of all possible variations of the name that can exist in the corpus
- **articles:** List of articles of the author

4.3 Data Analysis

We have a look at the data statistics of the corpus we generated for publications by individuals associated with IIIT-D. The table1 shows the division of total unique publications along with there author category. Evidently, we see that number of publications of

Author Type	Number of Authors	Number of Unique Publications
Faculty	75	5789
Students	70	1259
Combined	145	6527

Table 1: Century Wise Distribution

the faculty is significantly higher (due to longer time in the research domain) than the publications by student, despite taking similar number of authors.

Before creating a network graph, we have a look at a visualization of how our graph will look like. This is shown in image 3

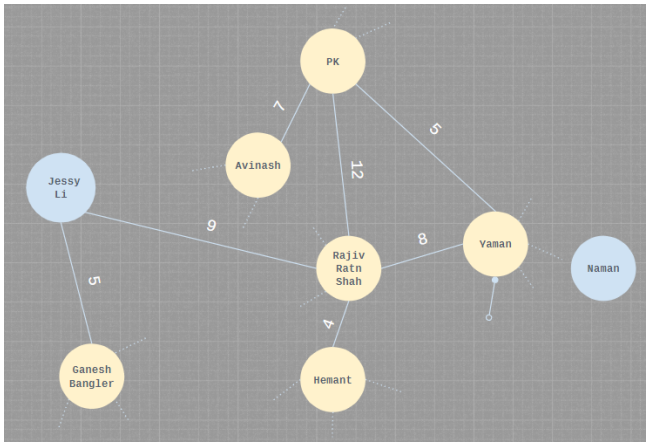

Figure 3: We see that the researcher nodes are connected with each other through edges of different weights, which represents the collaborations two authors have had. The yellow color nodes represent researchers from IIIT-Delhi, and the blue color nodes are external collaborators. Note: The values in the graph are for representational use only and may be different from actual values.

5 COLLABORATIVE NETWORK ANALYSIS

In this section we share details about the Collaborative network graph formed using our dataset and see the results of some baseline metrics on the network graph.

5.1 Proposed Approach

We convert our dataset to a networkx graph which helps in performing network analysis. We make use of the following Group Metric to analyse the graph:

- **Density** - It is a measure of the number of edges in the graph compared to the total number of possible edges. The more dense a graph, the more collaborative is its nature

The above metric tell us about the nature of our collaborative network graph. But to analyze the effect of the nodes (authors) on the network individually, we use the following single node metrics to evaluate our network graph:

- **Degree Centrality** - It is the number of researchers that a particular researcher is collaborating with.
- **Closeness Centrality** - It signifies the fact that a central node is also the one who has direct collaborations with many other researchers and thus can reach them through a very short path in the graph.
- **Betweenness Centrality** - It highlights the ability of a researcher to play a mediating role between other researchers thus playing a central role.
- **Clustering Coefficient** - It is a measure of the transitive nature of a researcher.

Further, note that collaboration can be of two types

- **Internal Collaboration** - Collaboration only amongst individuals associated with IIIT-D i.e. both the authors should belong to IIIT-D. It is represented by a homogenous graph.
- **External Collaboration** - It signifies collaboration between an author associated with IIIT-D collaborates with an author who is not associated with IIIT-D. It is represented by a heterogenous graph. The graphs only includes authors who have at least 1 publication.

All the above metrics are performed for both the homogenous graph and the heterogenous graph. Homogenous graph gives the values of these metrics over internal collaboration, whereas heterogenous values signify the metrics over external collaboration.

5.2 Results

On performing the above experiments we get a good insight to the collaborative network graph of research work done by individuals associated with IIIT-D.

5.2.1 ***Results on Group Metrics.*** We report several group metrics over both the heterogenous and the homogenous graphs.
We can see from table 2 that nodes in heterogenous graphs is

Metric	Heterogeneous Graph	Homogeneous Graph
No. of Nodes	5855	143
No. of Edges	22651	571
Density	0.0562	0.0013

Table 2: Group Metrics on the Graph

significantly higher than nodes which represents IIIT-D authors (143). This signifies that individuals associated with IIIT-D have collaborated with a wide network authors who are not from IIIT-D. Moreover, the heterogenous graph is around 40 times more dense than the homogenous graph representing that external collaboration is more prevelant than internal collaboration.

5.2.2 ***Results For Single Node Metrics.*** For each metric, we report 10 authors having the highest score for each of the metric.

Name	Homogenous Degree Centrality
anubha gupta	48
shivam sharma	37
amarjeet singh	31
anuradha sharma	29
vibhor kumar	28
gaurav gupta	27
angshul majumdar	27
tanmoy chakraborty	26
rajiv ratn shah	26
shikha singh	24

Table 3: Authors having top 10 Homogenous degree centrality

Name	Heterogenous Degree Centrality
gaurav gupta	456
ponnurangam kumaraguru	342
anubha gupta	334
kuldeep yadav	320
mukesh mohania	295
rajiv ratn shah	287
tanmoy chakraborty	264
amarjeet singh	244
vibhor kumar	225
gajendra ps raghava	207

Table 4: Top 10 Heterogenous degree centrality

Name	Homogenous Closeness Centrality
anubha gupta	0.56
shivam sharma	0.52
amarjeet singh	0.5
anuradha sharma	0.49
gaurav gupta	0.49
vibhor kumar	0.48
tanmoy chakraborty	0.48
shikha singh	0.48
angshul majumdar	0.47
rajiv ratn shah	0.47

Table 5: Top 10 Homogenous closeness centrality

Name	Heterogenous Closeness Centrality
anubha gupta	0.4
amarjeet singh	0.39
shivam sharma	0.39
gaurav gupta	0.38
vibhor kumar	0.38
anuradha sharma	0.38
ponnurangam kumaraguru	0.37
shikha singh	0.37
tanmoy chakraborty	0.37
richa gupta	0.37

Table 6: Top 10 Heterogenous closeness centrality

From the tables, we see that Dr. Anubha Gupta has a very high score for most of the metrics like Degree Centrality, Closeness Centrality and Betweenness centrality. It means that this author has a very active role in the network by directly collaborating with many different authors, and also being more likely than others in playing a mediating role between two different authors. We also see some authors like Mr. Shivam Sharma having high scores for homogenous degree centrality as compared to heterogenous degree centrality which indicates that a large fraction of their collaboration

Name	Homogenous Betweeness Centrality
anubha gupta	0.15
shivam sharma	0.09
amarjeet singh	0.07
angshul majumdar	0.06
tanmoy chakraborty	0.05
anuradha sharma	0.05
vibhor kumar	0.05
ponnurangam kumaraguru	0.05
rajiv ratn shah	0.05
shikha singh	0.04

Table 7: Top 10 Homogenous Betweeness centrality

Name	Heterogenous Betweeness Centrality
gaurav gupta	0.1
anubha gupta	0.09
ponnurangam kumaraguru	0.07
kuldeep yadav	0.06
amarjeet singh	0.06
mukesh mohania	0.05
vibhor kumar	0.05
tanmoy chakraborty	0.05
n. arul murugan	0.04
rajiv ratn shah	0.04

Table 8: Top 10 Heterogenous Betweeness centrality

Name	Homogenous Clustering Coefficient
sarthak bhagat	1
sneihil gopal	1
pandarasamy arjunan	1
anupriya tuli	1
hitkul	1
mohd hamza naim shaikh	1
divya sitani	1
shagun kapur	1
ankita likhyani	1
dhriti khanna	1

Table 9: Top 10 Homogenous clustering coefficient

are of internal type.

We see that many authors have a very high (in some cases completely 1) score for clustering coefficient. Another interesting insight is that all these authors are students. This is because, students have a limited research network as compared to faculty and lesser number of publications,and they tend to work with individuals with whom they have already worked like their professors or batchmates. This leads to a high transitive nature of students authors

Name	Heterogenous Clustering Coefficient
bushra ansari	1
sneihil gopal	1
megha gaur	0.83
payel mukherjee	0.53
aditya chetan	0.52
harshit singh chhabra	0.5
neeraj pandey	0.48
akanksha farswan	0.46
shagun kapur	0.46
hitkul	0.46

Table 10: Top 10 Heterogenous clustering coefficient

especially in internal collaborating as their coauthors are also associated with each other. Clustering Coefficient for professors is low because of there wide network and association with many different individuals.

6 PLAN OF WORK

These are the 4 major milestones of our Project:

- Collecting Data: Publications indexed by the particular researcher's belonging to IIIT-Delhi. involved. (Completed)
- Co-Authorship network Assembly using that indexed Data (Completed)
- Analyze the Co-authorship network using traditional and important metrics (Completed)
- Analyze the sub Co-authorship network based on author type (faculty or student) and year (Ongoing)
- Identifying a novel scoring system for co-authorship of the researchers (Ongoing)
- Creating a Web Platform to link researchers and help visualize the collaborations that have happened over the years in IIIT-Delhi.(Future)

REFERENCES

[1] Tasleem Arif, Rashid Ali, and M. Asger. 2012. Scientific Co-authorship Social Networks: A Case Study of Computer Science Scenario in India. *International Journal of Computer Applications* 52 (08 2012), 38–45. https://doi.org/10.5120/8257-1790

[2] Keith A. Burghardt, Zihao He, Allon G. Percus, and Kristina Lerman. 2021. The Emergence of Heterogeneous Scaling in Research Institutions. arXiv:2001.08734 [physics.soc-ph]

[3] Enrico di Bella, Luca Gandullia, and Sara Preti. 2021. Analysis of scientific collaboration network of Italian Institute of Technology. *Scientometrics* 126, 10 (October 2021), 8517–8539. https://doi.org/10.1007/s11192-021-04120-

[4] Giseli Lopes, Mirella Moro, Leandro Wives, and José Palazzo Moreira de Oliveira. 2010. Cooperative Authorship Social Network. *CEUR Workshop Proceedings* 619.

[5] Carlos Medicis Morel, Suzanne J Serruya, Gerson de Oliveira Penna, and Reinaldo Guimarães. 2009. Co-authorship Network Analysis: A Powerful Tool for Strategic Planning of Research, Development and Capacity Building Programs on Neglected Diseases. *PLoS Neglected Tropical Diseases* 3 (2009).

[6] SerpApi. 2019. SerpApi. https://serpapi.com/.