

# **A Multi-Level Analysis of Collaborations in Computer Science**

by  
Pramod Divakarmurthy

A dissertation submitted to  
Florida Institute of Technology  
in partial fulfillment of the  
requirements for the degree of

Doctor of Philosophy  
in  
Computer Science

Melbourne, Florida

May 2015

UMI Number: 3662784

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.

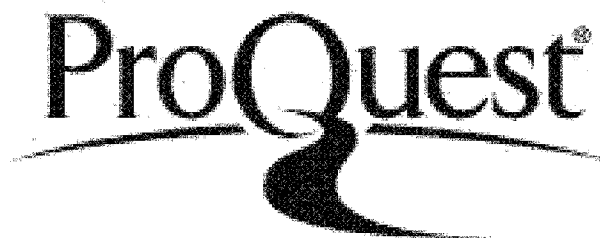


UMI 3662784

Published by ProQuest LLC 2015. Copyright in the Dissertation held by the Author.

Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against unauthorized copying under Title 17, United States Code.



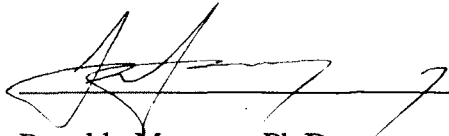
ProQuest LLC  
789 East Eisenhower Parkway  
P.O. Box 1346  
Ann Arbor, MI 48106-1346

We the undersigned committee hereby approve the attached dissertation

**A Multi-Level Analysis of Collaborations in Computer Science**

by


Pramod Divakarmurthy



---

Ronaldo Menezes, Ph.D.

Associate Professor, Computer Science  
Committee Chair, Dissertation Advisor



---

Eraldo Ribeiro, Ph.D.


Associate Professor, Computer Science  
Dissertation Committee



---

Gerald Marin, Ph.D.

Professor, Computer Science  
Dissertation Committee



---

Subasi Munevver, Ph.D.

Assistant Professor, Mathematical Sciences  
Dissertation Committee



---

Richard Newman, Ph.D.

Professor, Computer Science  
Department Head

# Abstract

Title: A Multi-Level Analysis of Collaborations in Computer Science

Author: Pramod Divakarmurthy

Committee Chair: Ronaldo Menezes, Ph.D.

Working in collaboration is common in today's highly connected scientific community. By collaborating, researchers can solve challenging multi-disciplinary problems, increase knowledge dissemination as well as productivity. These and other advantages motivate the study of the collaboration patterns of researchers. Such patterns can be observed directly in networks of manuscript co-authorship. In such a network, nodes represent authors and the links between them indicate that they have co-authored a paper.

Several researchers constructed and studied large-scale networks representing collaborations in Mathematics, Biology, Physics, and Neuroscience. Most studies have performed bibliometric analysis of scientific publications, evaluated and ranked scholars on their research performances, and studied structural characteristics of the collaboration networks. Certain studies on collaboration networks are focused to specific geographical regions (i.e., country or countries). Studies on longitudinal analysis of the collaboration

networks have helped to understand the publication trends of researchers.

Most studies have analyzed collaboration networks of either authors or institutions or country. To best of our knowledge there is no study that has analyzed collaboration networks on all the three levels. The increase in international collaboration is not only a trend of the 21<sup>st</sup> century, but one that has been noted in bibliometric studies. However, very few studies have examined this collaboration activity. Also, there is very little knowledge and understanding about the role and the nature of geographical proximity towards scientific collaborations.

In this dissertation, we analyze the collaboration networks at several levels (i.e., authors, institutions, and countries) in the field of Computer Science. We perform longitudinal analysis on publication trends and investigate collaboration patterns based on various geographical factors, such as distance and location. We investigate authors' affiliation trends and their average productivity. We investigate, if the size (i.e., number of authors) and subject diversity of a institute play any role towards average productivity of that institute. We also analyze if there is any correlation between scientific size of a country and the GDP of that nation. We then rank authors, institutions and countries to list the top collaborators and also rank authors, institutions and countries based on network metrics. Last, using visualization techniques we show how authors and institutions are distributed globally.

The results indicate that co-authorship networks in Computer Science have network properties similar to real-world networks and can be categorized as a scale-free network. The longitudinal analysis on the publication trends depicts a shift in the trend on number

of authors in a research publication. Our findings show that geographical proximity plays a vital role in the collaborations patterns of authors. We observed, a growth in the trend for international collaborations for institutions. We found that the scientific size of a country is correlated to the GDP of that nation.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	4
1.2	Our Approach . . . . .	6
1.3	Contribution of this Dissertation . . . . .	7
1.4	Overview of the Chapters . . . . .	8
<b>2</b>	<b>Background and Related Work</b>	<b>10</b>
2.1	Scientific Collaboration . . . . .	10
2.2	Geography in Other Networks . . . . .	19
<b>3</b>	<b>Research Methods and Data</b>	<b>27</b>
3.1	Concepts from Graph Theory . . . . .	27
3.2	Concepts of Network Analysis . . . . .	29
3.3	The Dataset . . . . .	40
<b>4</b>	<b>Network of Authors</b>	<b>48</b>
4.1	Introduction . . . . .	48
4.2	General Network Characteristics . . . . .	49
4.3	Collaboration Pattern . . . . .	57

4.4	Community Analysis and Area Diversity . . . . .	66
4.5	Ranking of Authors . . . . .	72
4.6	Geographical Distribution . . . . .	74
4.7	Conclusion . . . . .	79
<b>5</b>	<b>Network of Institutions</b>	<b>80</b>
5.1	Introduction . . . . .	80
5.2	General Network Characteristics . . . . .	81
5.3	Collaboration Pattern . . . . .	86
5.4	Ranking of Institutions . . . . .	97
5.5	Geographical Distribution . . . . .	99
5.6	Conclusion . . . . .	103
<b>6</b>	<b>Network of Countries</b>	<b>105</b>
6.1	Introduction . . . . .	105
6.2	General Network Characteristics . . . . .	106
6.3	Collaboration Pattern . . . . .	108
6.4	Ranking of Countries . . . . .	116
6.5	Geographical Distribution . . . . .	118
6.6	Conclusion . . . . .	120
<b>7</b>	<b>Conclusions</b>	<b>121</b>
<b>A</b>	<b>ACM Computing Classification System</b>	<b>124</b>
<b>B</b>	<b>Network of Subjects By Country</b>	<b>130</b>
<b>C</b>	<b>Country Code</b>	<b>137</b>





**Dedicated to My Family**

*For their endless love, support, and encouragement*

# Acknowledgements

I owe a tremendous debt of gratitude to the many people who made this dissertation possible.

Most importantly, I would like to thank my father, Divakara Murthy, my mother, Shashikala, my brother, Prashant, my sister-in-law, Kavya, and my nephew, Aahan, who encouraged and helped me at every stage of my personal and academic life. I thank them for their sacrifices and faith in me and allowing me to be as ambitious as I wanted. I would like to thank my wife, Swathi. She has always been there to cheer me up and she has stood by me through the good and bad times. I also thank my extended families, including my parents-in-law, Omprakash and Leela, my siblings-in-law, Shruthi and Darren, and beloved Iestyn and Eeto, for always supporting me.

I would like to express my great appreciation and gratitude to my friend, mentor, and advisor, Dr. Ronaldo Menezes, for his excellent guidance, caring, patience, enthusiasm, and constant encouragement throughout this research. Without his supervision, the dissertation would have been impossible. I could not have imagined having a better advisor and mentor for my Ph.D. study.

I would also like to thank Dr. Eraldo Ribeiro, Dr. Gerald Marin, Dr. Subasi Munewer, and Dr. Scott Tilley for guiding my research. Their insightful comments, understanding, and interest in my work were extremely encouraging.

I am also thankful to my past and current colleagues for their inspiring support and presence as a great comfort to me. Their friendship makes my life a wonderful experience.

Finally, many thanks to the Department of Computer Science for financially supporting my study and providing me with an excellent atmosphere for doing research.

# **Declaration**

I declare that the work in this dissertation is solely my own except where attributed and cited to another author. Most of the material in this dissertation has been previously published by the author. For a complete list of publications, please refer to Appendix D.

# Chapter 1

## Introduction

Interest in networks, and particularly in social networks, has flourished recently [66]. A social network [64, 83] is considered to be a set of actors, mostly individuals or organizations, and their relationships are represented as ties between these individuals. A tie may represent a casual acquaintance, close family bonds, or any common interest between the individuals linked.

Networks play an important role in disciplines such as computer systems [40], transportation [74], artificial intelligence [42], global economy [75], and biology [45]. Network analysis techniques have been used to understand structures and patterns of various kinds of real-world networks. The collaboration network of film actors from IMDB<sup>1</sup> is a class of social network. In this network, two actors are connected if they have acted in a movie together. The strength of their ties represents the number of movies in which they have both acted. Based on all movies since the 1980s, the network has over 400,000 actors as its nodes and  $2.5 \times 10^7$  links between the

---

<sup>1</sup><http://www.imdb.com/>

nodes [6, 64]. A number of biological systems can be represented as networks. Food webs are networks of species linked by predator-prey interactions. These networks have been mapped out in a few habitats by ecologists who use them to understand the interactions between species [69, 87, 76, 13]. Power grids are networks of generators, substations, and transformers that are connected to each other by high-voltage transmission lines spanning entire country [85, 5]. The World Wide Web is an example of an information network, which is a vast network with approximately  $2 \times 10^8$  nodes of Web pages containing information, linked together by hyper-links from one page to another [12, 71, 2, 3].

In the field of informetrics [29], the study of collaboration patterns is concerned with co-authorship and citation networks [15, 51]. Co-authorship networks form a class of social networks consisting of researchers who are connected to one another only if they have co-authored a paper – the assumption for the “social” relationship is due to the fact that if individuals have worked together on a paper, then one should safely assume that they are acquainted with each other. Co-authorship networks are used to determine the structure of collaboration and measure the status of an individual researcher [37]. Co-authorship is a common practice in the research community and it has been shown to improve both the quality and quantity of an individual’s researcher output [44, 57]. There are several factors to co-authorship that may have a positive affect on an individual or the entire group’s productivity; the productivity of a team is generally better than any individual’s. Another advantage is the reduction of time an individual has to devote to a single project, giving him or her an opportunity to work on multiple projects with other authors [16]. Collaboration also contributes to knowledge spread, particularly if the individuals belong to different fields [47]. Since collaboration can help promote research

productivity, research institutions, encourage and support scientific collaboration among individual researchers [55].

Co-authorship networks are generally undirected and weighted. They are undirected because it is hard to impose a directionality on the collaboration; all the individuals listed as authors of the papers are connected to each other forming a clique (i.e., fully connected graph) for each paper. However, it is common that individuals collaborate more than once, leading to weights being needed to express the strength of the relations.

Network analysis is a common approach for understanding patterns of interactions and relationships that exist between the actors. It is used to discover structural features such as: hubs, highly connected groups, and patterns of how the individuals interact with each other. Several researchers constructed and studied large-scale networks representing collaborations in Mathematics [38], Computer Science [59], Physics [65], Biology [65], and Neuroscience [8]. Studies on longitudinal analysis of the collaboration networks have helped to understand the publication trends of researchers. Some studies have evaluated and ranked scholars based on their research performances. Certain studies on collaboration networks are focused to specific geographical regions (i.e., country or countries). These studies are aimed to understand the national scientific productivity and identify potential areas for research and development of the nation.



The dataset used in most of the studies on scientific collaboration networks are usually obtained from existing Web services, such as Google Scholar<sup>2</sup>, DBLP<sup>3</sup>, CiteSeer<sup>4</sup>, and Microsoft Libra<sup>5</sup>.

## 1.1 Motivation

Measuring collaborations has for many years been of importance to the research community. One example is the concept of the *Erdős Number* that has permeated the mathematical research community for more than thirty years [14]. Paul Erdős was a prolific mathematician; he wrote at least 1,525 research papers in many different areas, mostly in collaboration with other mathematicians. During his lifetime, Erdős collaborated with 511 different researchers. As a tribute to his contributions, people started to measure their “distance” to Erdős. Hence, Erdős’ co-authors have an Erdős number equal to 1, while individuals who have written a paper with someone with Erdős number 1 have Erdős number 2, and so on. If there exists no chain of co-authorship connection to Erdős, then that individual’s Erdős number is infinite.

There are very few studies on the collaboration networks in Computer Science. Our dataset is more representative of the Computer Science community because it includes many conferences and journals for nearly 60 years. Previous studies on collaboration networks have focused either on network of authors or institution or country. However, to best of our knowledge there is no study that has analyzed the collaboration networks on all the three levels (i.e, authors, institutions, and country).

---

<sup>2</sup><http://scholar.google.com/>

<sup>3</sup><http://www.informatik.uni-trier.de/~ley/db/>

<sup>4</sup><http://citeseerx.ist.psu.edu/>

<sup>5</sup><http://libra.msra.cn>

Analyzing publication trends and collaboration patterns at every level, will give an overall view of the collaboration networks in Computer Science. There is very little understanding on the relation between subject diversity and the output performance of authors and institutions. Better visualization techniques show how authors and institutions are distributed globally.

Recent developments in information and communication technologies has seen the role of spatial distances and territorial boundaries to have diminished with respect to information exchange [78]. The internet has enabled scientists to communicate much faster with their collaborators and has made internationalization possible. Also, with the recent developments in roads and transportation, commuting costs are much more affordable and it has reduced the effective distance between people. Recent studies on mobile phone communication networks [53, 50] and blogs [58] have revealed that the probability for a social tie to occur between agents decays with a power of their distance. Some of the obstacles in long distance collaboration could be due to linguistic, cultural, and institutional differences [46]. Understanding how collaborations vary with respect to distance is vital as it would suggest scientists on how to choose future collaborators and maximize their productivity. There is very little knowledge about the role geographical proximity play towards scientific collaboration and its effect has not been fully understood yet.

## 1.2 Our Approach

Social Network Analysis (SNA) is a technique to analyze and measure network properties. From SNA, one can understand the relationships between people, teams, groups, companies, and other entities, if they are represented as a network. SNA techniques have been used earlier to understand and study collaborations in co-authorship networks [81], directors of companies [23], etc. Characteristics of the network can be described on two levels: global network properties and individual node properties.

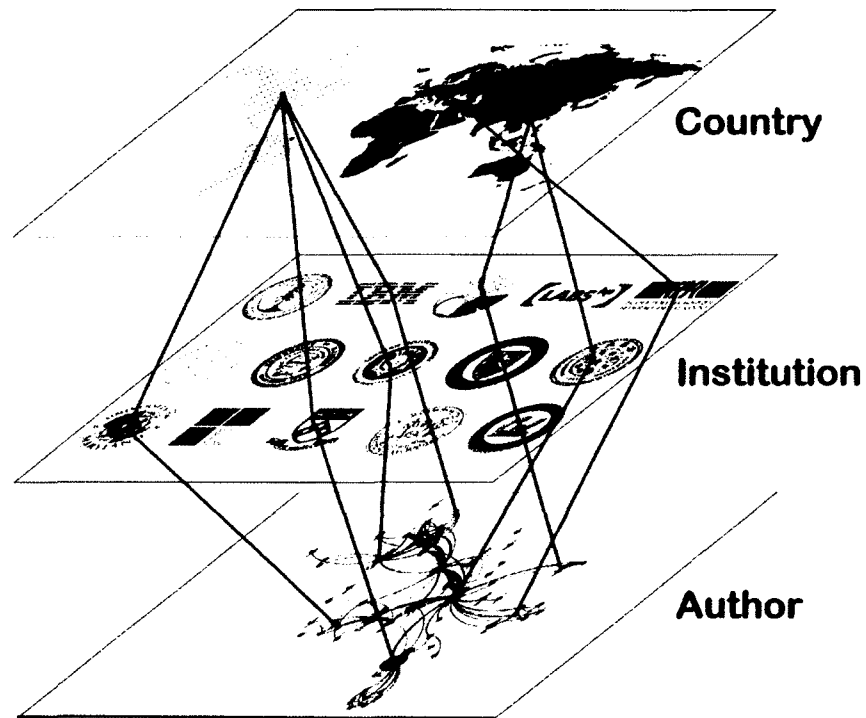


Figure 1.1: Multi-Level analysis on collaboration of authors, institutions, and countries.

In this research, we perform a multi-level analysis on collaboration of authors, institutions, and countries for the Computer Science discipline. Figure 1.1 shows the structure of our multi-level analysis and how each of them is connected to the other. Our dataset contains the authors who have published papers, indexed by the ACM Digital Library<sup>6</sup>. For each network, we measure and compare its network characteristics with other similar networks. We depict the publication trends and collaboration patterns of authors, institutions, and countries. Using visualization techniques, we depict the distribution of authors and institutions. Finally, we rank the top collaborators in the field of Computer Science, using metrics derived from the network structure.

### 1.3 Contribution of this Dissertation

This dissertation presents the results of our multi-level analysis on collaboration networks in Computer Science. We analyze network properties for a network of authors, a network of institutions and a network of countries. We give an in-depth analysis of the collaboration trends and patterns and how geographical factors such as distance and location play a role on collaboration. The contributions of the dissertation are as follows:

- **Network of Authors:** We do a micro-level analysis mainly focused on co-authorship networks; our analysis helps advance the existing knowledge of collaborations in the field of Computer Science. Our analysis on collaboration trends explains the evolution of collaborations and collaboration strategies. Understanding how physical distances between authors play a role helps authors to choose future collaborators and improve their productivity and quality of research. A general understanding of authors in Computer Science and on co-

---

<sup>6</sup><http://dl.acm.org/>

authorship strategies is expanded by additional knowledge on: (1) the number of co-authors to have in order to produce quality work, (2) an authors' most productive period in his or her career, (3) typical number of affiliations associated to authors in Computer Science. Visualization techniques help understanding how the Computer Science authors and papers are distributed globally.

- **Network of institutions:** Our meso-level analysis on collaboration networks is focused on a network of institutions. Our analysis helps in understanding the collaboration trends and strategies of these institutions. The size of a institute (i.e., in terms of number of authors) and research diversity (i.e., number of different research areas) can improve the productivity (i.e., number of papers) and the quality (i.e., number of citations) of these organizations. The productivity and the quality of these organizations can attract prolific researchers and quality Ph.D. students as well as receive research grants.
- **Network of Countries:** Our macro-level analysis are focused on a network of countries. The study helps in understanding of collaboration trends and strategies of these countries. Ranking countries based on various network metrics gives us a better understanding of the roles each country plays in the development of science, in our case, computer science.

## 1.4 Overview of the Chapters

The remainder of this dissertation is organized into the following chapters:

In Chapter 2, we provide a background and review the related literature. Our review is focused on two areas. First, we provide an overview of results of studies relevant to

collaboration networks. Second, we provide an overview of studies on networks which consider geographical factors.

In Chapter 3, we discuss concepts of graph theory and network analysis as well as the data collection techniques and some of the characteristics of our dataset.

In Chapter 4, we construct and investigate the network of authors. We analyze the network characteristics and investigate publication trends and collaboration patterns of authors. We rank the top authors in the field of Computer Science by network metrics. Using visualization techniques, we depict the distribution of authors, publications, and citations globally and also take a deeper look at a few countries.

In Chapter 5, we construct and investigate the network of institutions. We analyze the network characteristics and investigate publication trends and collaboration patterns of institutions. We rank the top institutions in the field of Computer Science by their collaboration level. Using visualization techniques, we depict the distribution of these organizations for the entire world and for United States.

In Chapter 6, we construct and investigate a network of countries. We analyze the network characteristics and investigate publication trends and collaboration patterns of these countries. We rank the top countries in the field of Computer Science by their collaboration level.

In Chapter 7, we review the contributions made in this dissertation, highlight weaknesses in the work as it stands, and suggest future areas for consideration and ways in which the current work could be extended.

## **Chapter 2**

# **Background and Related Work**

In this chapter, we provide a review of the literature related to the work described in this dissertation. The chapter is organized in two main sections. The first section provides an overview of results of studies relevant to collaboration networks. The second section provides an overview of literature related to studies of networks that consider geographical factors.

### **2.1 Scientific Collaboration**

Newman [65] studied the scientific collaboration networks for three disciplines, namely: Biomedical Research, Physics and Mathematics using bibliographic data from 1995-1999 for Biology and Physics, and 1940-2001 for Mathematics. This study was performed to understand the collaboration patterns, how they vary between subjects studied, and also how they vary temporally.

Table 2.1: Network Statistics for the three co-authorship networks analyzed by Newman [65].

	<b>Biology</b>	<b>Physics</b>	<b>Mathematics</b>
Number of authors	1,520,251	52,909	253,339
Number of papers	2,163,923	98,502	–
Papers per author	6.40	5.10	6.90
Authors per paper	3.75	2.53	1.45
Average collaborators	18.10	9.70	3.90
Largest component (Average distance)	4.60	5.90	7.60
Largest distance	24.00	20.00	27.00
Clustering coefficient	0.06	0.43	0.15
Assortativity	0.13	0.36	0.12

Table 2.1 lists some network metrics and statistics for the three co-authorship networks. The study revealed some interesting facts. The Biology network was the largest with 1.5 million authors over a five-year period. The Mathematics network, which covered a period of 60 years, had 250,000 authors. Clearly, the Biology community has more researchers compared to other fields. A similar pattern can be seen for the number of papers. The Mathematics database covers a longer time period (i.e., 60 years); this indicates that mathematicians are producing fewer papers than those in other fields. The number of authors per paper varied considerably among the subjects, with Mathematics having the smallest number and Biology having the largest. Biologists had significantly more co-authors when compared to mathematicians or physicists, a result that reflects the way research and experiments are done in these fields. Biologists work in larger groups, generally in a laboratory, whereas mathematicians tend to do more theoretical work and work alone or in pairs. This also explains the possibility of the lower productivity of mathematicians in terms of papers published per unit time (i.e., 60 years). Figure 2.1 shows the distribution of the number of co-authors that scientists have for the three subjects after the network statistics have been analyzed. All three subjects



have similar distributions, although the distribution for Biology (circles) has a fat tailed (i.e., the distribution region marked in yellow), which reflects a higher mean number of collaborators. Also, in each case, the distribution is fat-tailed, with a small fraction of scientist's having a very large number of collaborators; up to thousands.

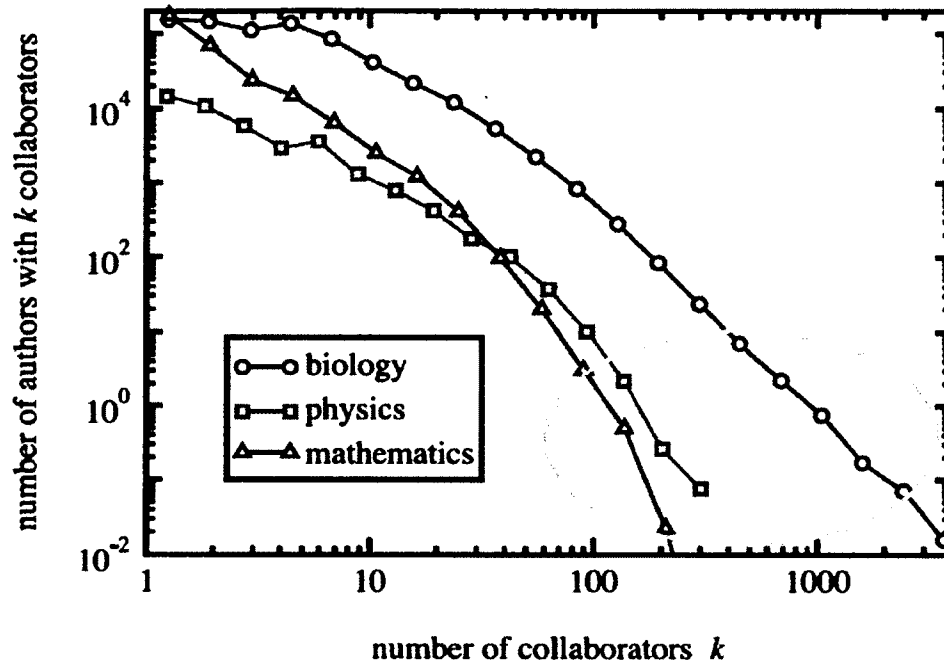


Figure 2.1: The distribution of the number of co-authors that scientists have for Physics, Biology, and Mathematics. The fat-tailed distribution region is marked in yellow [65].

A similar type of study was performed on the scientific collaboration of authors of the Pacific Asia Conference on Information Systems (PACIS) [17]. The data included all conference research papers published at PACIS from 1993 to 2008. The results showed that the percentage of co-authored papers grew from 50% in 1993 to 80% in 2008. Since the establishment of the conference in 1993, the number of papers

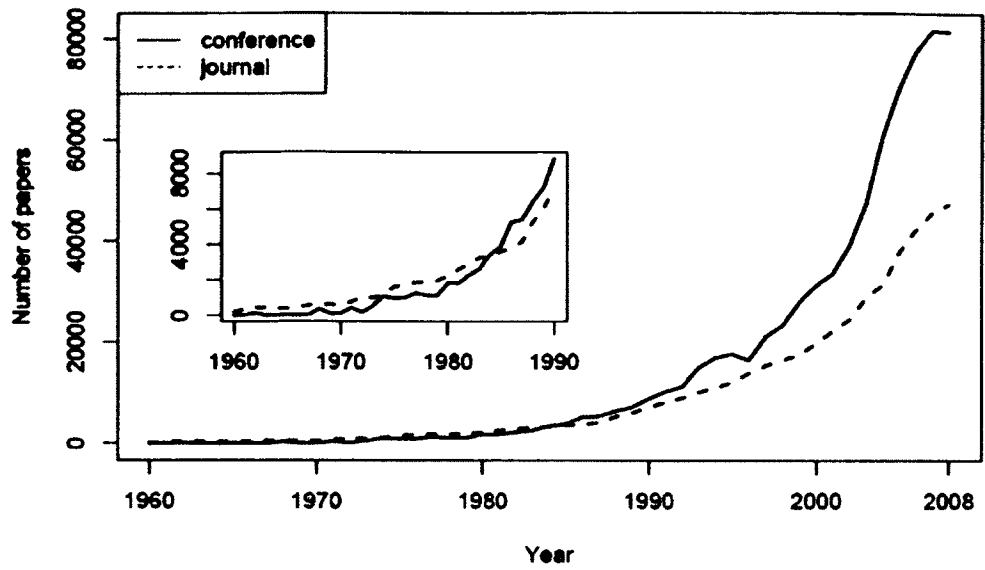
presented had grown significantly. Using visualization techniques and SNA metrics the researchers revealed the structural characteristics of the PACIS community and were able to identify influential members. The network contained a significantly large main component. The positive evolution of the main component in the PACIS network and the presence of a number of key individuals indicated a healthy status of the PACIS community. The presence of key individuals in the PACIS community help in attracting new research members to the community and to produce star (influential) researchers for the new generation. Star researchers play an important role in the community, but other researchers are also important, as without them, there would be no PACIS community.

Franceschet [34] analyzed how collaboration in Computer Science evolved in the last half-century using the publications in the DBLP dataset from 1936 to 2008. His study included large-scale properties of the collaboration network. Specifically, he investigated the temporal evolution of the properties of the collaboration network: number of publications, number of active researchers, network clustering, connectivity of the network, and average separation distance among researchers. He showed how these properties have changed in the last 50 years of Computer Science. Figure 2.2a shows the growth of the number of conference papers and number of journal papers published each year in the Computer Science discipline since 1960. It indicates that the computer scientists have become more productive and collaborative over time. Until 1983, the volume of journal papers appears to slightly dominate conference papers. However, since then, conferences are the preferred venue of publication in the Computer Science. Figure 2.2b shows the temporal evolution of the number of active authors and their productivity. It can be seen from the figure that both the variables are growing over time, but during the 1960s the authors' productivity shows some variations. The

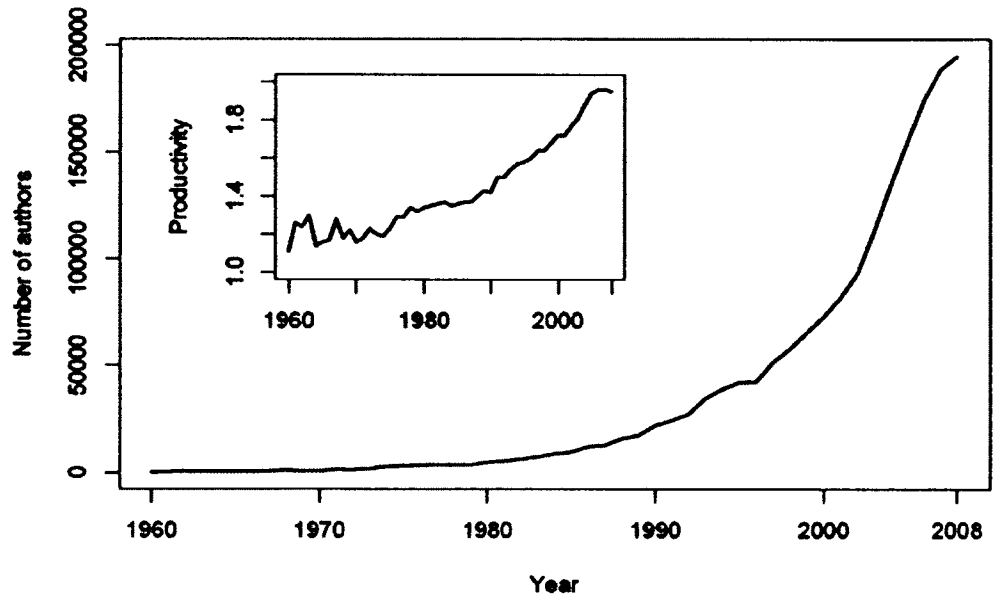
growth of the number of papers is due to the increase in the number of active authors and the rise of author productivity in terms of papers published. Figure 2.3a shows the collaboration network as of 1964. It shows a network composed of small clusters of nodes with high internal connection, but which are disconnected from the core of the network and other small groups. Over the years, new authors are added to the network as nodes when they publish a paper or journal article. New collaborations with existing authors add new edges to the network. Figure 2.3b shows the largest component of the collaboration network until the year 1980.

Ovalle-Perandones [67] in 2009 analyzed and measured the scientific co-authorship networks of Spanish research centers (companies, universities and hospitals) in the areas of Physiology and Pharmacology. The network contained about 470 institutions whose researchers co-authored 760 papers on pharmacology published during the period 1995-2005. Some of the institutions co-authored with organizations of different types, whereas some research centers exclusively collaborated with other research centers. A good hub is an organization that links to many others and a good authority is an organization that is linked by many different hubs. The university hospitals located at Barcelona, Zaragoza, and Seville proved to be good authorities and they have co-authorship ties with universities located in Barcelona and Oviedo, while the working relation is much less intense with private enterprise.

In 2011, Yu [90] analyzed and extracted the research groups from the co-authorship network of oncology in China. The study revealed that researchers from the regions of Beijing, Guangdong, and Shanghai cooperate most closely with each other and the

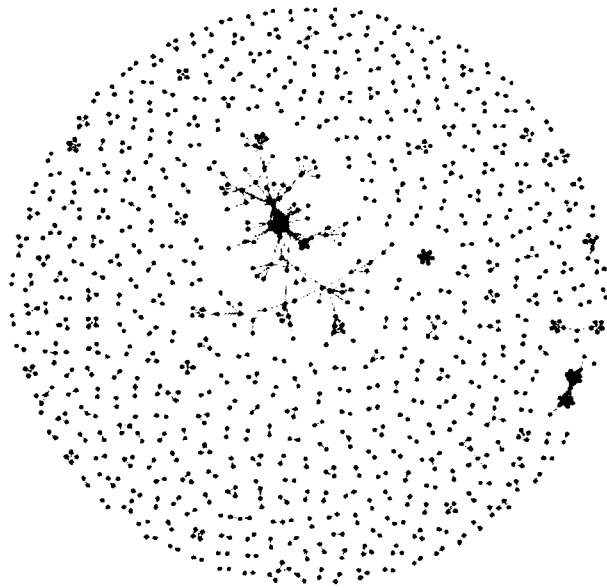


(a) Number of papers per year in Computer Science discipline.

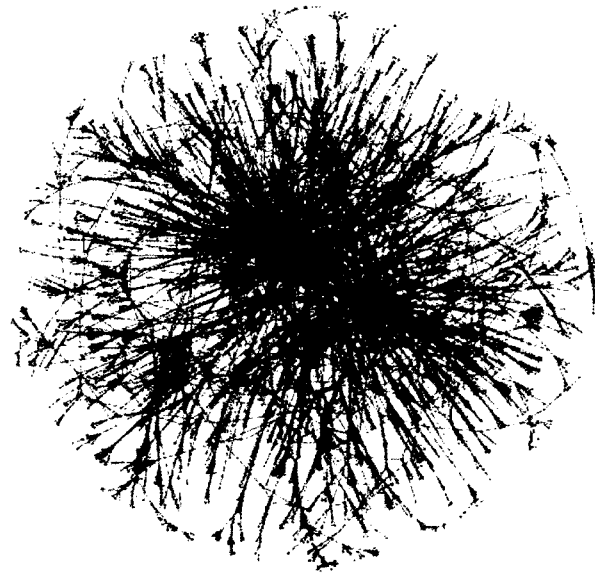


(b) Number of authors per year in Computer Science discipline.

Figure 2.2: A longitudinal analysis on number of authors and number of papers from 1960 to 2008 [34].



(a) The Computer Science collaboration networks in 1964.



(b) The Computer Science collaboration networks in 1980  
(largest component).

Figure 2.3: Visualization of the Computer Science collaboration network from DBLP dataset [34].

collaboration of researchers within “Administrative Divisions”<sup>1</sup> is much closer than the collaboration of researchers from different “Administrative Divisions”. The economic and educational levels within or between “Administrative Divisions” play a major role in collaboration activity. Their findings suggest that encouraging scientific cooperation among the regions with different economic levels will contribute to the economic progress in these underdeveloped areas.

As in most large organizations, individual and team performance is measured on a set of metrics that pertain to tasks performed. Similarly in academia, the performance of scholars and scientists is evaluated based on their academic activities such as teaching evaluations, number of grants, and research output [66, 43, 59]. Such evaluation of researchers is not only required by the department, but also for achieving a high reputation within the research community and for governmental fund allocation. High reputation attracts federal funding and also attracts highly qualified students around the world, which elevates research standards. Hence, it is important to identify key scholars and collaboration areas within universities for maximizing the research output. Cotta et al. [21] gathered data from DBLP, comprising more than 610,000 articles authored by several thousand computer scientists to identify the central actors in the network and what makes them important. After calculating the network metrics, they listed the top 10 authors based on their degree, betweenness, and closeness.

Co-authorship networks created by different kind of papers (i.e., technical reports, conferences papers, journals papers) might be different due to the kind of collaboration

---

<sup>1</sup>China has 34 “Administrative Divisions”, including 23 provinces, 4 municipalities, 5 autonomous regions directly under the Central Government, and 2 special administrative regions.

they imply. For instance the collaboration between authors for a technical report may be weak as it is written in a hurry and presents very preliminary results. However, the collaborations between authors for conferences papers are usually long term so they are stronger, and journal papers require a long committed scientific relationship. Journal papers take a very long time to be published and they involve several iterations of revision. Abbasi et al. [1] evaluated the performance in the scholarly collaboration network. They selected publication data between 2000 and 2009 of the top nine journals in “Information Science & Library Science.” The results showed that research performance in terms of number of publications is directly associated with a researcher’s position within the collaboration network. Scholars who have more connections with other authors and those who lie more on the shortest path between pairs of authors in the network show a better research performance.

There are studies that conduct a longitudinal analysis of co-authorship networks to understand the publication trend of researchers. Uddin et al. [79] studied the dataset that spanned over 20 years in an attempt to understand the trend and efficiency of co-authorship networks. Their primary data source is Scopus. They focused on papers published from 1990 to 2009 on steel structures. The probability of co-authoring can differ over time across different disciplines. Co-authorship is quite common in natural sciences when compared to social sciences, but it has been steadily increasing across all fields [30, 32, 41].

## 2.2 Geography in Other Networks

Certain studies on collaboration networks are focused on specific geographic regions (country or state) [89, 28, 54, 56]. The main purposes of these studies is to understand the national scientific productivity and identify potential areas for research and development to improve the economy of the nation. In this section, we also review some of the studies on other networks such as economic networks, transportation networks, and social networks in which geography plays an important factor.

Laender et al. [52] in 2008 conducted a study to estimate the quality of the top Computer Science (CS) graduate programs in Brazil. Since 1977, every three years the Brazilian Ministry of Education's agency, CAPES<sup>2</sup>, has been assessing graduate programs in all fields in the country and generating a ranking based on the quality of their programs as per the recommendations of committees appointed for this task. The main aim of this evaluation was not just to measure the quality of these existing programs but also to establish a ranking to assist the government with the spending on advanced research and education. To evaluate the maturity of the top Brazilian CS graduate programs, they performed a comparative analysis with reputed CS graduate programs from other countries. They considered the top 8 graduate programs from Brazil, 16 graduate programs from North American (i.e., Canada and the US), and 6 graduate programs from European (i.e., England, Switzerland, and France). Results showed that the top Brazilian CS programs performed comparatively well when evaluated with the North American and European programs.

---

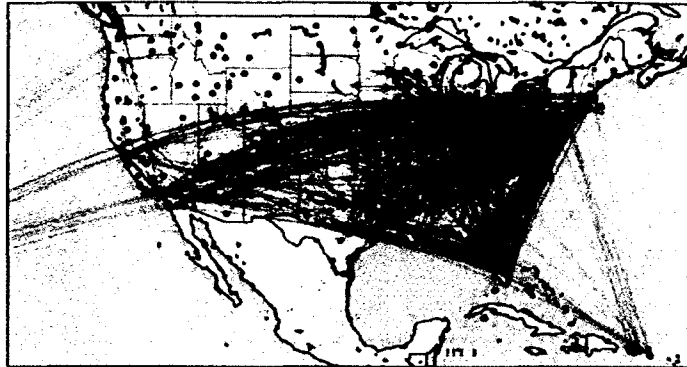
<sup>2</sup><http://www.capes.gov.br>



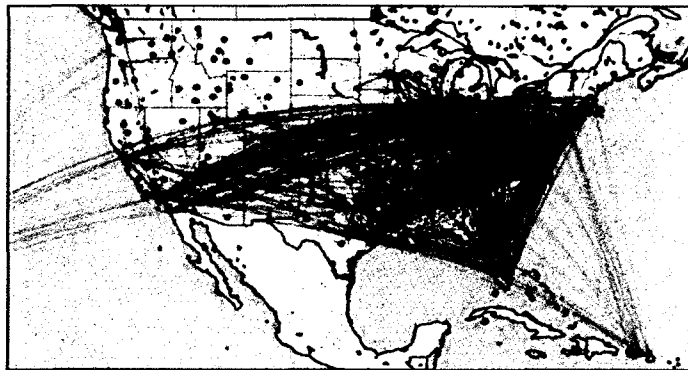
Air transportation systems are usually described as graphs, where the vertices represent the airports and the links represent the flight path between the airports. Such graphs are usually called airport networks. Studies on airport networks vary, based on the geographical scales, to a single country or single/multiple continent to world-wide. These networks show high heterogeneity in the distribution of connections per airport and in the traffic sustained by each connection. Fleurquin et al. [33] downloaded the data from the Bureau of Transport Statistics (BTS) to study the US airport network. This data contained 6,450,129 scheduled flights operated by 18 different carriers connecting 305 different commercial airports. The information per flight includes real and scheduled departure (arrival) times, origin and destination airport, an identification code (tail number) for each aircraft, airline, etc. With this dataset they analyzed the spreading of delays in an air-traffic network with the focus on the US airport network of 2010. They introduced a measure for the level of network-wide extension of delays by defining when an airport is considered as congested and studying how congested airports form connected clusters in the network. Figure 2.4 shows the maps with the congested airports and the connections between them for different time periods. The congestion dramatically changes from day to day: some days a large cluster would cover one-third of all the airports, while in others only one or two airports cluster together. Figure 2.5 shows how the size of the largest congested cluster varies from one day to the next.

Venugopal et al. [82] looked at organ transplantation data and investigated the current organ allocation process. For this study, they considered all transplants in the US since 1987, with the locations representing states or zip codes in the US. They show that visualization and the use of techniques from network science help in identifying and understanding the current problems in the organ donation process. Constructing

**April 4, 2010**



**March 9, 2010**



**March 12, 2010**

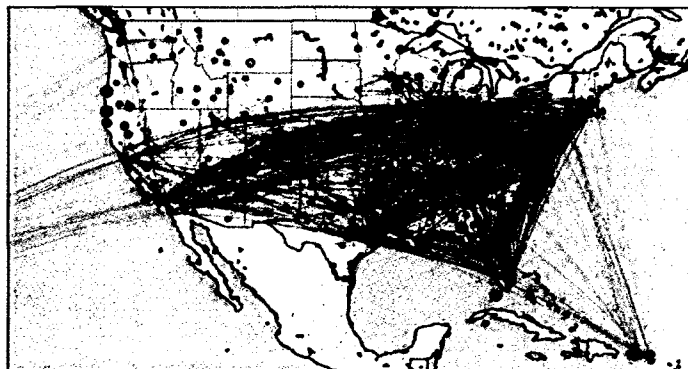


Figure 2.4: Maps of the congested airports also showing connections between them over different periods of the year. Red, orange, and green indicate to which cluster the airport belongs [33].

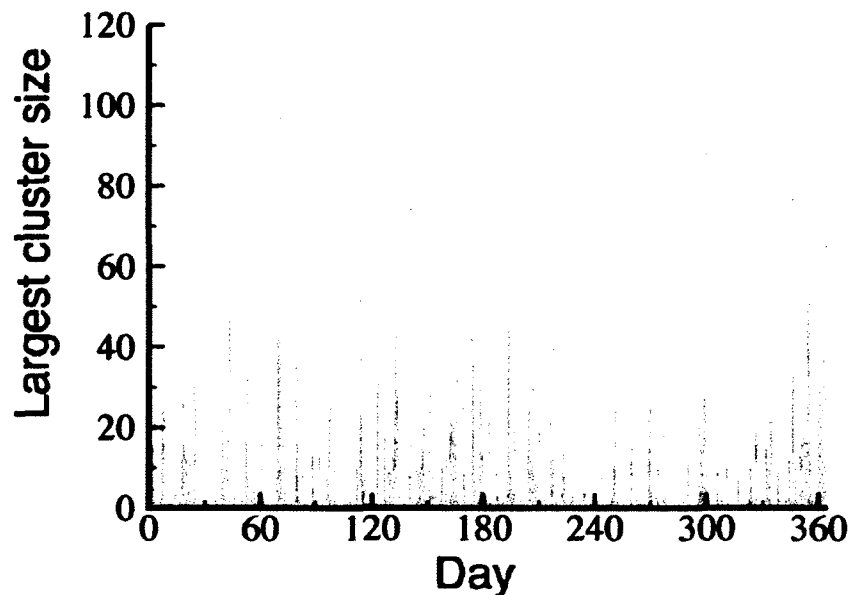


Figure 2.5: Daily size of the largest cluster as a function of time [33].

the network of locations (geographical social network, GSN), they emphasize that the distance travelled by organs is crucial to the health of the organ. Figure 2.6 shows the community analysis of GSN of various organs at the zipcode-level. They found that the heart network is not as well organized at the zipcode level as it appeared at the state level, whereas the liver network preserved its organization. The kidney network was denser and revealed community divisions with areas in the country dominated by African-American groups. The hotspots of the organ donors and organ recipients showed that urban-area hotspots tend to be related to organs received. Also, the data seemed to indicate that the current allocation policies benefit people in urban areas of the US.



(a)



(b)



(c)

Figure 2.6: The community analysis of organs of GSN at zipcode level: (a) heart, (b) liver, (c) kidney [82].

In the field of economics, studies have used networks to better understand the relations between companies or countries. There was lot of attention given to economic networks for the recent 2008-2009 global economic crisis, which was due to a large number of factors. In today's global economy, there are strong economic relationships between countries. Garas et al. [35] investigated how a crisis propagates from the country of origin to other countries in the world. For generating the economic network, they used two datasets obtained from the *Bureau Van Dijk*<sup>3</sup> in order to avoid any bias due to the network selection. The two global networks describe strong interaction patterns in the world economy. The first network is the Corporate Ownership Network (CON), which is based on the world's largest companies and all their subsidiaries and links 206 countries; the second network, International Trade Network (ITN), is created using aggregated trade data linking 82 countries. Their model showed that when a crisis is triggered with a controlled magnitude, it propagates from one country to another with a probability that depends on the strength of the economic ties between the countries involved and on the strength of the economy of the target country. Using the k-shell method, they were able to identify the role of the different countries in a world crisis. Figure 2.7 shows the 12 most effective countries for crisis spreading.

Many studies have used geography to understand social networks. Crandall et al. [22] investigated the extent to which social ties between people can be inferred, given that two people have been in approximately the same geographic locale at approximately the same time, on multiple occasions. To perform this study, they used a large scale dataset from the popular photo-sharing site Flickr. Most photos uploaded to Flickr include the time at which the photo was taken and many photos are also geo-tagged with

---

<sup>3</sup>Bureau Van Dijk Electronic Publishing (BvDEP) <http://libra.msra.cn>

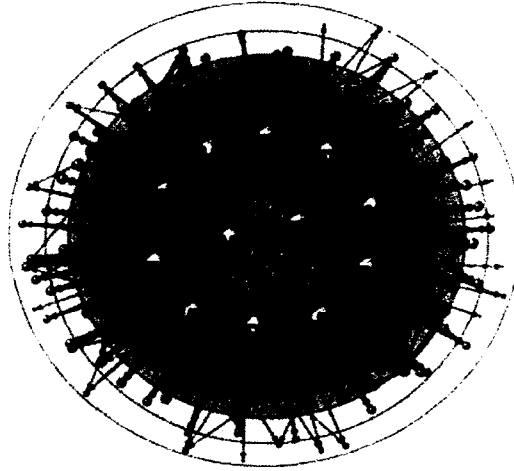


Figure 2.7: A layered structure of the global economic network of 206 countries of the world using the large corporation subsidiary relations. The outer layers are the loosely connected countries, while at the center, the nucleus contains 12 strongly connected countries. [35].

a latitude-longitude coordinate indicating where on Earth the photograph was taken. They defined spatio-temporal co-occurrence between two Flickr users as an instance in which they both took photos at approximately the same place and time. To understand their model, Figure 2.8 shows how spatio-temporal co-occurrences are counted for some sample time-stamped observations of individuals A and B. The surface of the earth was divided into grid-like cells, whose side lengths span  $s$  latitude-longitude degree. Two people co-occurred in a given  $s \times s$  cell  $C$ , at temporal range  $t$ , if both took photos geo-tagged with a location in cell  $C$  within  $t$  days of each other. The number of distinct cells are counted in which they had a co-occurrence at temporal range  $t$ . The researchers found that a large number of friendships involving these high-activity users exhibit spatio-temporal co-occurrences; for example, approximately 22% of all such friendships have one co-occurrence in a  $1^\circ$  latitude-longitude cell when the temporal

range is considered as 1 day.

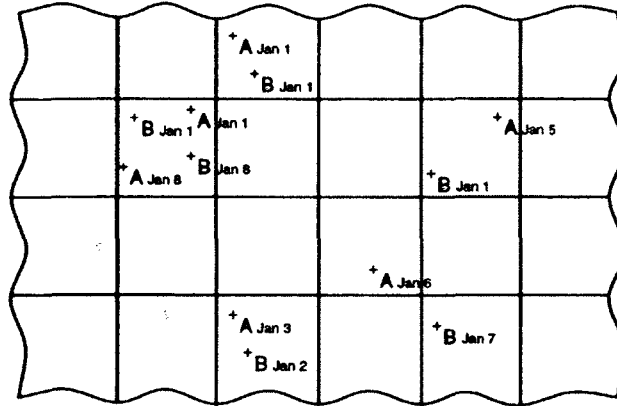


Figure 2.8: Illustration of the spatio-temporal co-occurrences between the individuals A and B [22]. The world is divided into discrete cells of size  $s \times s$ . For this example, there were 5 co-occurrences at a threshold range of  $t$  days.

Network concepts have also been used by scientists to understand long-distance relations between seismic activities. Ferreira et al. [31] studied the worldwide seismic events by generating a network of sites of earthquakes around the world. They used data from the world-wide earthquake catalog for the period between 1972 and 2011. For their investigation they only considered earthquakes with magnitude  $m \geq 4.5$  on the Richter scale. They found that the data on seismic events showed small-world characteristics and a particular geographical site with an event appeared to be related to many other sites around the world and not only to other events near to it.

# Chapter 3

## Research Methods and Data

### 3.1 Concepts from Graph Theory

A graph is a collection of vertices and edges (i.e., connections between vertices). From the point of view of Network Science, a graph is called a network when nodes represent real-world objects and edges represent the link between them. Nodes and links may have a variety of properties associated with them [80].

In graph theory, a  $n$ -partite graph is one in which the nodes can be divided into  $n$  sets so that no link in the graph exists within each set individually [86]. Hence a bi-partite graph is one in which the nodes can be divided into two sets  $U$  and  $V$  so that every link in the graph connects a node in  $V$  with a node in  $U$  [7]. In networks, the bi-partite concept refers to cases in which the network contains nodes of two types but the relationship is always between these different nodes [39]. What is interesting about these  $n$ -partite networks is that we can generate uni-partite projections of them. For instance, from the actor-movie network, we can create a social network of actors and relate them if they are



linked to the same movie in the bi-partite version of the network. Bi-partite networks are important for us because in our study we have a bi-partite network of papers and authors; we concentrate on the author projection. Figure 3.1 shows an example in which we have a bi-partite network and its projection as an author network. The value two for the link between A2 and A3 indicate the number of times the authors have collaborated.

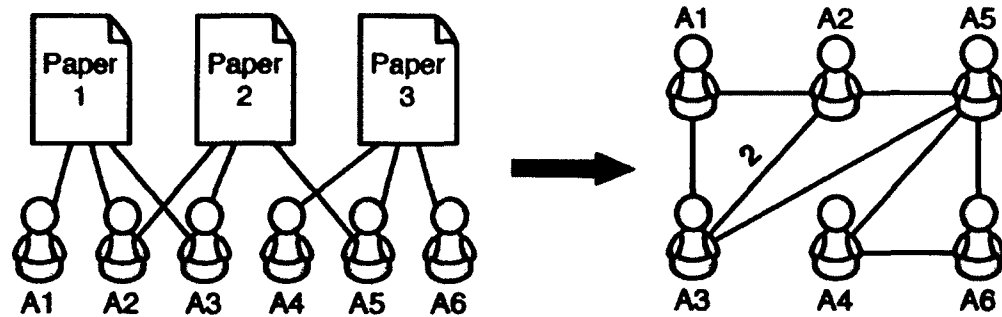


Figure 3.1: Projection of a bi-partite network of papers/authors into a network of authors.

Another way to understand the formation of an author network is to realize that since we are discussing collaborations, every published paper forms a clique where all authors of the paper are connected to each other. Hence the full author network is created out of an overlapping of these cliques. This explanation makes it easier to understand where the weight of the links come from. Figure 3.2 depicts the situation in which we have three papers with three authors each.

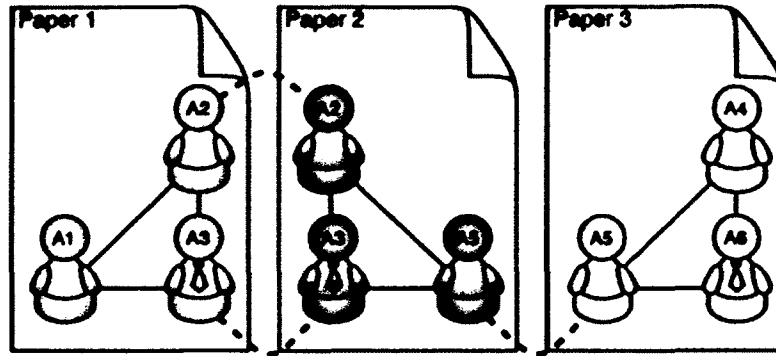


Figure 3.2: Each paper forms a clique of authors. The full network is formed from the overlap of these cliques. The resulting network is the same as the one depicted in Figure 3.1.

## 3.2 Concepts of Network Analysis

For any social network, we can describe its properties on two levels: global network metrics and individual node properties [73]. Global network metrics describe the characteristics of the entire social network, such as the networks's diameter, mean node-node distance, number of components, cliques, clusters and small-world phenomena. Individual properties relate to the analysis of the properties of network nodes, e.g., degree, betweenness, closeness or position in a cluster. Most measurements defined in this section are from [64].

The size of a social network is denoted by the number of nodes (authors in our case). A disconnected network contains a number of sub-networks called components. The degree of connectedness of a network is given by the density measure, which is the percentage of the number of actual connections over the total number of possible connections.

The status of a node is usually expressed in terms of its centrality, i.e., a measure of how central the node is to the network. Centrality is a structural characteristic of an individual in the network; the centrality score of each individual tells you something about how it fits within the overall network. Central nodes have a higher degree of influence, spread information, and are more likely to receive new information in the network. Individuals with low centrality are usually in the periphery of the network. One of the advantages of being in the periphery is that they are less likely to be influenced. There are, however, many variations of centralities among these; degree, closeness, and betweenness centralities are the most commonly used. In this dissertation, the following concepts are used in our analysis.

### ***Directionality***

Networks can be classified as directed and undirected networks. In directed networks, links have directionality, while in undirected networks, links have no directionality. More precisely, links in directed networks form an ordered pair  $(a, b)$ , where  $a$  and  $b$  are nodes in the network. In contrast, in undirected networks, the links are unordered pairs  $\{a, b\}$ . Given the above, we have that  $(a, b) \neq (b, a)$ , while  $\{a, b\} = \{b, a\}$ . Directionality plays a very important role in many kinds of networks. One example of a directed network is the food web, where direction of links indicates who eats whom. Scholarly co-authorship networks are undirected interactions, as co-authoring a paper involves actions that do not involve directionality. On the other hand, citation networks have directionality, where the directional activity occurs when a paper cites another paper.

### ***Link Weight***

In networks, links may have a weight. The weight of an link indicates the strength or the extent of interaction between the two nodes. For example, in a co-authorship networks a low link weight would indicate a weaker collaboration bond (i.e., few publications) between the authors, while a high link weight would indicate a stronger collaboration bond (i.e., high number of publications) between the authors. In network visualization, weights are often represented by link widths, whereby heavier weights have wider and more marked lines.

### ***Components***

A connected component is a set of nodes where everyone has a path to every other node in the component. In large networks, it is normal to have more than one connected components, i.e., partitioned into disconnected groups of nodes. Many real and artificial networks feature a *giant component*, i.e., a large connected component that contains the majority of the network's nodes.

### ***Giant Component***

The giant component is a connected component in a large network. A giant component contains the majority of the network's nodes. Moreover, when a network contains a giant component, it almost always contains only one. Let  $N_1$  be the size of the connected component  $C$  in a network of size  $N$ , then a giant component is a  $\frac{N_1}{N}$  fraction of the network [18]. Generally, all the network analysis are performed on the giant component. From here on we refer to  $N$  as the size of the giant component in this dissertation. Figure 3.3 shows a size distribution of connected components in a network.

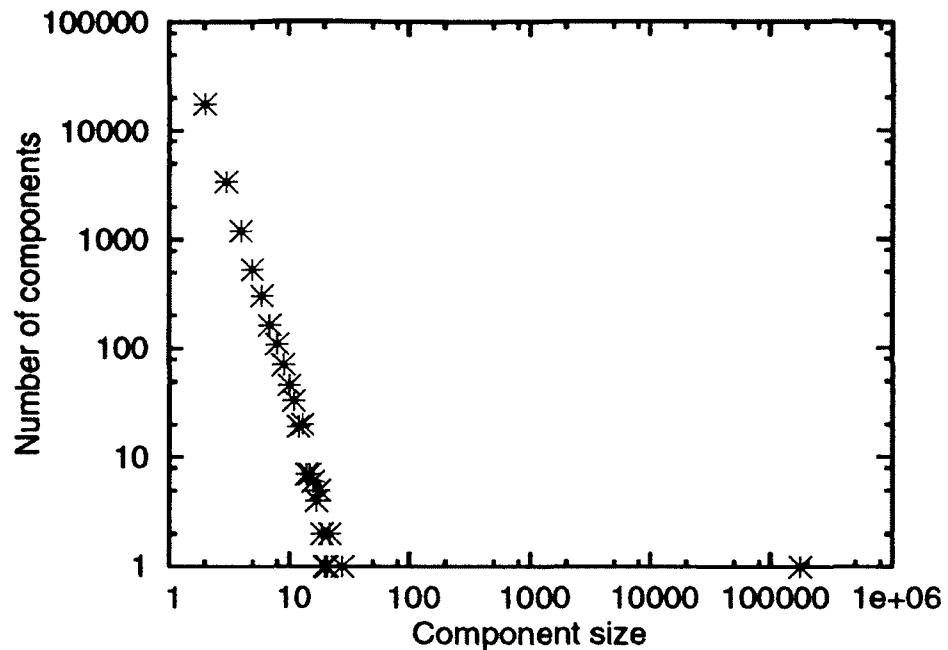


Figure 3.3: A size distribution of connected components [18].

### ***Geodesic Path***

The shortest path through the network from one node to another is considered as a geodesic path. Often there may be more than one geodesic path between two nodes.

### ***Diameter***

In a connected network, any two nodes can be connected by different paths running along the links of a network. A *geodesic path* is the shortest of these paths. The diameter of a network is the length (i.e., in number of links) of the longest geodesic path between any two nodes. If a network contains multiple components, then the diameter of the network is always calculated on the giant component.

### ***Degree Centrality***

The degree centrality of a node is defined as the total number of links that are adjacent to the node. It only measures how many connections authors have in the network. Nodes with a high degree in the network are usually called *hubs* and they are one of major actors who control the entire network. Figure 3.4 shows a network where the hubs are colored with red. The degree of a node  $i$ , is denoted by  $C_D$  and can be written as:

$$C_D = \sum_j m_{ij}, \quad (3.1)$$

where,  $m_{ij} = 1$  if there exists an link between nodes  $i$  and  $j$ , and  $m_{ij} = 0$  if there is no such link.

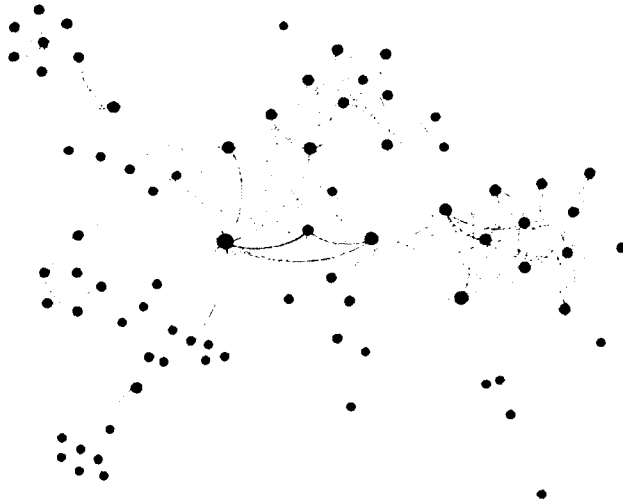


Figure 3.4: A node with a high degree of centrality is called a hub.

### ***Closeness Centrality***

The closeness centrality measures how close a node is to all other nodes. Individuals may be well connected to their immediate neighbours but they can be part of an isolated clique. Although such a node is locally well connected, its overall centrality is low. The closeness centrality of node  $i$  can be written as:

$$C_c(i) = \frac{1}{\sum_{j=1, j \neq i}^N d_{ij}}, \quad (3.2)$$

where,  $d_{ij}$  is the length of the shortest path between node  $i$  and node  $j$  and  $N$  is the number of nodes in the network. Closeness is an inverse measure of centrality, where a larger value indicates that the node is less central while a smaller value indicates a more central node.

Figure 3.5 shows a network where the nodes colored in red have the smallest closeness values. Nodes who have high closeness centrality are important influencers within their local network community. They may not be public figures to the entire network, but they are often respected locally and they can spread information faster since they have shorter paths to other nodes in the network.

### ***Betweenness Centrality***

The betweenness centrality of a node determines how often the node is found on the shortest path between a pair of nodes in the network. These nodes are usually very different from those with high closeness. Nodes with high betweenness often do not have the shortest average path to everyone else, but they have the greatest number of

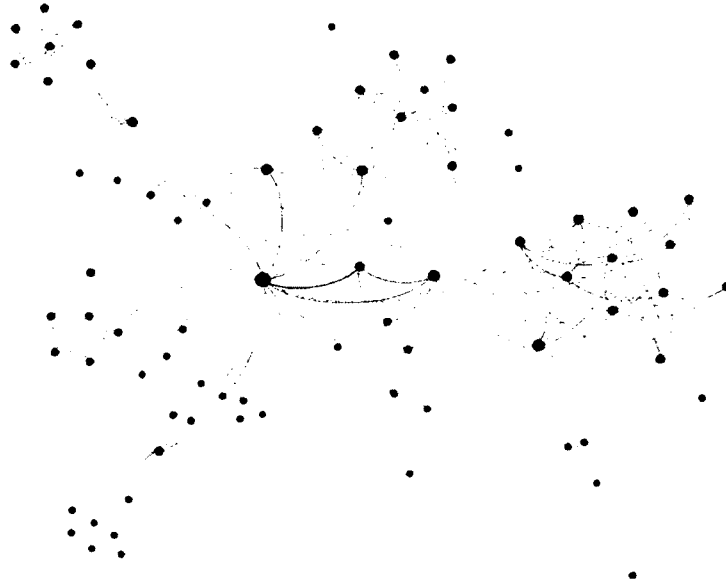


Figure 3.5: Nodes who are highly connected to others within their own cluster have a high closeness centrality.

shortest paths that necessarily have to go through them. In networks, nodes with high betweenness are often found at the intersections of more densely connected network communities. Figure 3.6 shows one such network where nodes with high betweenness are colored red. Because of their locations between network communities, they act as bridges for collaboration and information exchange. The betweenness centrality of a node  $i$  is written as:

$$C_B = \sum_{s \neq i \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}}, \quad (3.3)$$



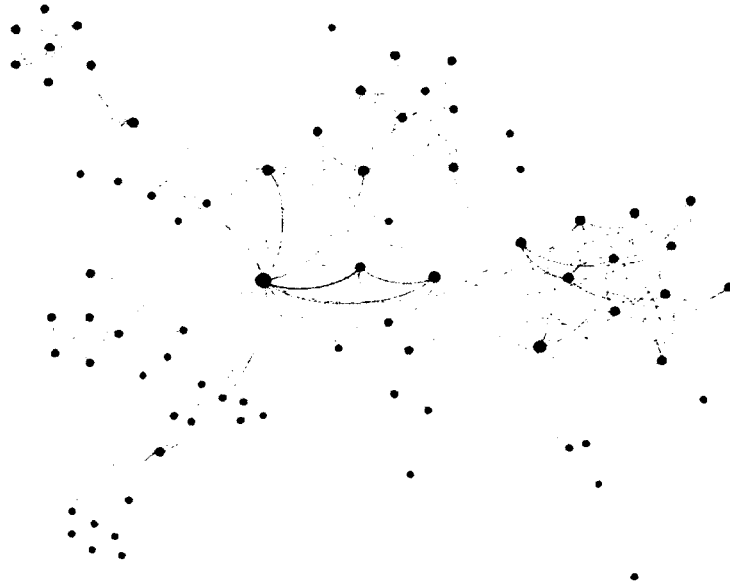


Figure 3.6: A node with high betweenness centrality acting as a bridge between clusters.

where,  $\sigma_{st}$  is the total number of shortest paths from node  $s$  to node  $t$  and  $\sigma_{st}(i)$  is the number of those paths that pass through  $i$ .

***Degree distribution and preferential attachment***

The most studied metric in networks is the degree distribution, i.e., the frequency distribution of degrees of nodes in a network. A degree distribution is normally displayed as a plot of node degrees on the  $x$ -axis and their cumulative frequency on the  $y$ -axis. Real-world networks have a highly skewed degree distribution, with the majority of nodes having a low degree and a small number of nodes having a high degree. A network whose degree distribution follows a power law is commonly known

as a *scale-free network* [9], that is, the probability that a network will have nodes of degree  $k$ , denoted by  $P(k)$  is given by:

$$P(k) \sim k^{-\lambda}, \quad (3.4)$$

where,  $\lambda$  is a positive constant.

A power-law distribution has been observed in many social networks [20]. One of the predominant generative mechanisms behind the formation of power-law degree distribution is *preferential attachment*, i.e., the notion that, as the network grows, there is a higher probability of new nodes to attach to nodes that have a higher degree.

### ***Cliques***

Cliques are subsets of a network where all nodes are connected to each other. From a sociological perspective, cliques are interesting network structures as they represent tight-knit groups of interconnected nodes who exclusively share specific characteristics and patterns of behavior.

### ***Clustering Coefficient***

Nodes in many real systems exhibit a tendency to form tightly connected subgraphs. This property can be quantified by the clustering coefficient [84], which is a measure of the degree to which the neighbours of a particular node are also connected to each other. In networks where relationships between nodes are represented by a link, transitivity represents a situation where node A is linked to B, node B is linked to C, and A and C are also connected; in other words, we have a clique of size 3. The clustering coefficient

of node  $i$  is defined as:

$$C_i = \frac{2m_i}{k_i(k_i - 1)}, \quad (3.5)$$

where,  $m_i$  is the number of links between the  $k_i$  neighbours of node  $i$ .

### ***Average Path Length***

The average path length  $\ell$  is the average number of steps along the shortest paths for all possible pairs of network nodes. For an undirected graph of  $N$  nodes, the shortest path length between two nodes, averaged over all pairs of nodes, is defined as:

$$\ell = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij}, \quad (3.6)$$

where,  $d_{ij}$  is the length of the shortest path between nodes  $i$  and  $j$ . In real networks, the average path length is expected to be:

$$\ell \approx \log N, \quad (3.7)$$

where,  $N$  is the number of nodes in the network.

A high clustering coefficient coupled with a short average path length indicates that the network exhibits *small-world* properties. A *small-world* is a network in which two nodes are only a few steps apart. Individuals are not necessarily all connected to each other, but they are all easily reachable from one another via a short path.

### ***Community Structure***

A considerable amount of research in Network Science is done around the study of community structures [70]. In networks, structure refers to high-level topologies that are separate from individual small-scale interactions. Most real-world networks have natural subdivisions. In social networks, people organize themselves along the lines of interest, language, age, occupation, and so forth. In networks, communities are defined as groups of densely interconnected nodes that are only sparsely connected with the rest of the network [36]. The interactions between the nodes within the same community are usually stronger compared to the interactions with nodes outside its cluster.

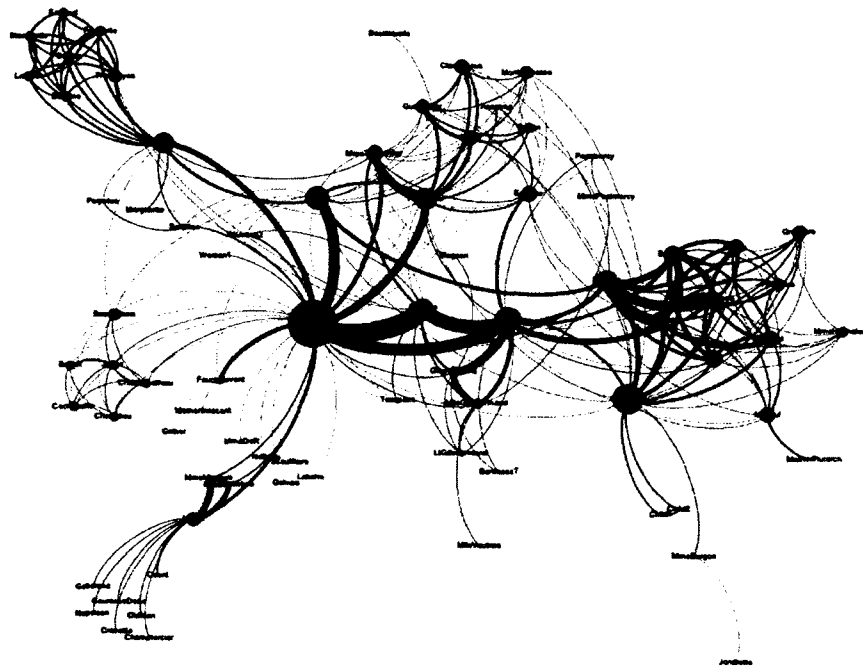


Figure 3.7: Social network of the characters in the novel, *Les Misérables*.

Figure 3.7 shows a visualization of social network of the characters in the novel, *Les Misérables*<sup>1</sup>. A link is drawn between two characters (nodes) if they co-appear in the novel. The node size reflects the number of connections each character has to other characters. Jean Valjean, the novel’s central character, is the largest node. Using community detection algorithms, various subgroups are identified to which each character belongs. The node color indicates which community they belong to.

For weighted networks, the techniques are based on algorithms that partition a network into *structural communities*. Structural communities are “cliquish” sub-graphs composed by groups of nodes that are highly connected to each other, but weakly connected to other nodes [19]. In networks it is important to study the community structure because it might display local properties that can be completely different from the properties of the entire network. Network clustering is sometimes confused with the technique of data clustering, which is a way of detecting groups of data-points in high-dimensional data spaces [4]. Since they both have some common features, algorithms from one can be adapted to another, and vice-versa. In real-world networks, nodes often belong to more than one community, and such a property is termed as a community overlap.

### 3.3 The Dataset

The Association for Computing and Machinery (ACM) is the primary society for Computer Science professionals. It includes many scientific journals, magazines, conference papers, and books in Computer Science. In order to perform our study,

---

<sup>1</sup><https://wiki.gephi.org/index.php/Datasets/>

we have gathered information about publications available in the ACM Digital Library and constructed a dataset to represent a bi-partite network. In the ACM website, each individual author has a profile page (URL) with details about his or her bibliographic data, affiliations, subject areas, URL to author's colleagues, and list of publications. Similarly, every paper has a webpage with details such as title, year of publication, citations, authors, publisher, subject classification<sup>2</sup> and list of references.

Since the dataset with the information was not readily available, we wrote a Web crawler to extract the required information. For the crawler to begin, we used the URL of all Turing award winners<sup>3</sup> between 1966 and 2012, as our seed list and they were assigned as depth 1 in our database. All the collaborators of depth 1 authors were assigned as depth 2. There is always a possibility that an author from depth 2 could have more than one connections with depth 1 authors. We therefore made sure that no duplicate data was included. We crawled up to depth 4 and extracted information of authors who had published a paper by 2013. Each author in the ACM is assigned a unique identifier (UID). Additional information about the authors, such as, name, number of papers published, number of citations, publication years, author UID and affiliations were also extracted to our database from the authors' webpages. Similar to author UID, every paper is also assigned a paper UID by ACM. While the author's details were gathered, his paper UIDs and paper URLs were stored in a separate file from the author's publication list. Once we had extracted all the required information about authors, we re-crawled to gather paper details. Using a Web crawler, we extracted the information and processed the bibliographic data available for each

---

<sup>2</sup><http://www.acm.org/about/class/1998>

<sup>3</sup><http://amturing.acm.org/byyear.cfm>

paper found; information such as published year, title, authors, citation, and ACM subject classification were stored as part of our dataset.

Once we had extracted all the required information, our local ACM dataset included 241,700 authors who had published 994,714 papers spanning approximately 62 years (i.e., works available in the ACM Digital Library from 1951–2013), although the core of the dataset is from 1981 to 2010. Table 3.1 shows the distribution of authors based on the depth they belong to in our dataset. Depth 1 consists of Turing award winners (1966–2012). Depth 2 authors have collaborations with depth 1 authors, and so on.

Table 3.1: ACM Dataset Depth Distribution

<b>Crawler Depth</b>	<b>No. of Authors</b>
Depth 1	58
Depth 2	2,832
Depth 3	43,929
Depth 4	194,881

Our dataset contains publication information for 994,714 papers. Table 3.2 shows the distribution of papers from 1951–2010. It can be clearly seen that the number of publications in the field of Computer Science has grown over time.

Before performing any kind of analysis on the network. We used tools such as Gephi and Cytospace to calculate and find the giant component in our network. We performed all our analysis on the giant component. Table 3.3 shows the number of nodes in the complete network and the number of nodes in the giant component for that network.

Table 3.2: Paper Distribution

<b>Year</b>	<b>Papers Published</b>
1951 - 1960	427
1961 - 1970	3,378
1971 - 1980	15,734
1981 - 1990	76,828
1991 - 2000	234,896
2001 - 2010	587,568

Table 3.3: Giant Component in our Collaboration Networks

<b>Network Type</b>	<b>Complete Network</b>	<b>Giant Component</b>	<b>Fraction (%)</b>
Authors	241,700	195,084	81
Institutions	10,963	10,805	98
Country	143	143	100

The field Computer Science is very diverse and has many specializations. The ACM has introduced a classification system for its publications. When submitting a paper to ACM, authors are required to specify the subject for the classification. The classification divides the Computer Science field into 11 main areas: General Literature (A), Hardware (B), Computer Systems Organization (C), Software (D), Data (E), Theory of Computing (F), Mathematics of Computing (G), Information Systems (H), Computing Methodologies (I), Computer Applications (J) and Computing Milieux (K). These areas in turn are subdivided into specific fields (e.g., Artificial Intelligence (I2), Computer-Communication Networks (C2), Software Engineering (D2) (See Appendix A for details). Table 3.4 shows the distribution of the papers in our dataset based on subject. Clearly, it can be seen that a few areas in Computer Science have a higher publication count than others. It indicates that there are more research activities in certain areas than in others. Figure 3.8a shows the distribution of papers based on ACM subjects from 1951 to 2010. Figure 3.8b shows a longitudinal analysis of the papers



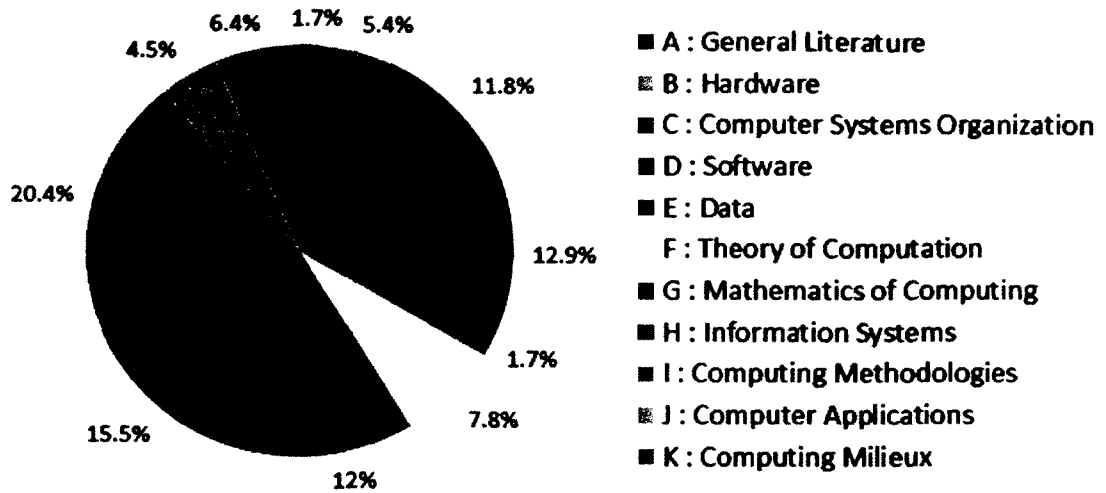
based on ACM subjects for every year from 1951 to 2010. It shows that certain research areas in Computer Science, such as Software, had comparatively more publications than others. As the Computer Science field grew, publications in other areas of Computer Science increased.

Table 3.4: Paper Distribution By Subject Classification

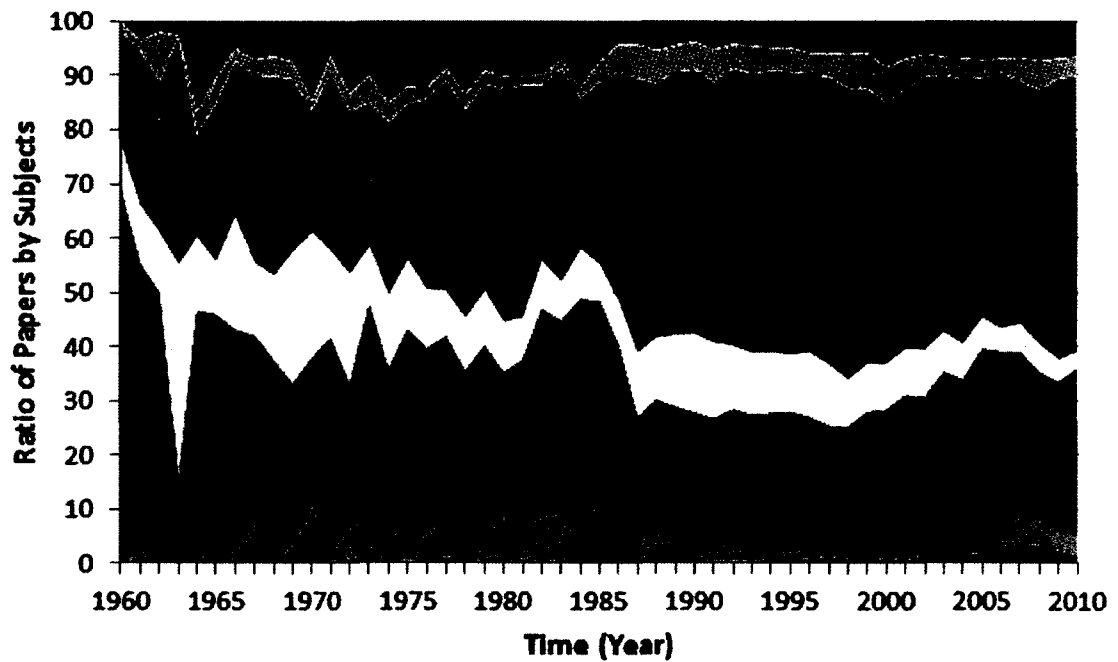
<b>Subject</b>	<b>Papers Published</b>
A. General Literature	7,101
B. Hardware	21,785
C. Computer Systems Organization	48,142
D. Software	52,332
E. Data	6,733
F. Theory of Computation	31,848
G. Mathematics of Computing	48,687
H. Information Systems	63,066
I. Computing Methodologies	82,779
J. Computer Applications	18,123
K. Computing Milieux	25,865
No Data	440,514

With the dataset available locally, we used the Google Geocoding API<sup>4</sup> to convert affiliations into geographic coordinates (i.e., latitude, longitude, country, state, etc.); this is important for us because we use the geographical location of authors to analyze the behavior of distances between collaborations. Geographic coordinates can also be used to show how the authors and papers are distributed geographically across the world. Along with the information of author and paper in our dataset, we also created another table to hold paper–author information. The information for this table was added while we were crawling for paper details. This information is important to be stored, as from this, we can generate the author–author list with their collaboration count. There are

<sup>4</sup><https://developers.google.com/maps/documentation/geocoding/>



(a) Total percentage of papers published in each ACM subject classification from 1951 to 2010.



(b) A longitudinal analysis of papers by ACM subject classification from the year 1951 to 2010.

Figure 3.8: Paper Distribution By Subject Classification.

many open-source softwares available for visualization and for analysing large network graphs, such as gephi<sup>5</sup>, cytoscape<sup>6</sup> and tulip<sup>7</sup>, just to name a few.

To understand the growth of the Computer Science community, we analyze the following bibliometric properties for Computer Science: number of papers and number of authors, for each year from 1960 to 2010. Figure 3.9a shows the size of the Computer Science discipline in terms of number of papers published each year since 1960. The red dashed line indicates the best fit function. The Computer Science discipline is steadily expanding in terms of number of published papers. There were very few publications between 1960 to 1970. Around 300 papers were published in 1960 and around 500 in 1970. However, since 1985-86, the number of published papers has grown exponentially. During the year 2010, the Computer Science community published about 90,000 papers. The Computer Science community published about 20% more papers in 2010 when compared to 2009. To understand this growth of the Computer Science discipline, we performed a longitudinal analysis on number of active authors per year.

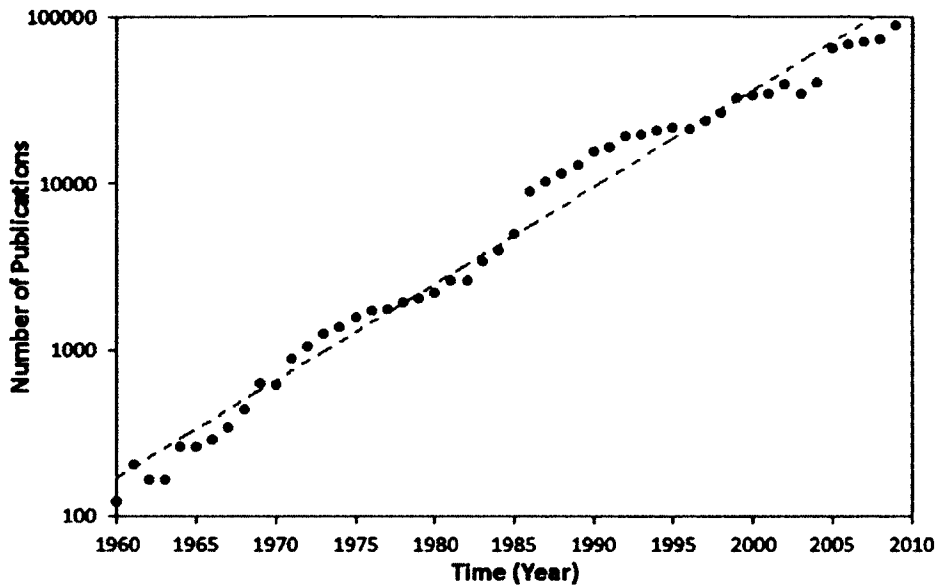
For each year, we studied the number of active authors i.e., the number of authors, who published at least one paper in that year. Figure 3.9b shows the temporal evolution of number of active authors in the Computer Science community. The Computer Science community has grown over time, although there were slight oscillations during the early 2000's. There were around 100 authors in 1960 and by 2010, there were around 65,000 active authors.

---

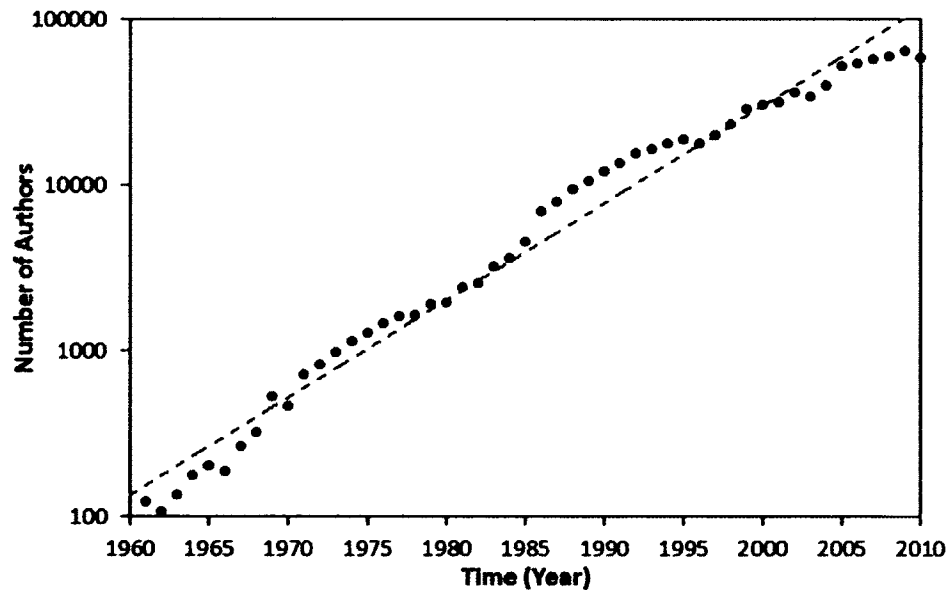
<sup>5</sup><https://gephi.org/>

<sup>6</sup><http://www.cytoscape.org/>

<sup>7</sup><http://tulip.labri.fr/TulipDrupal/>



(a) The number of papers published each year since 1960.



(b) The size of the Computer Science community in terms of number of active authors for each year since 1960.

Figure 3.9: Growth of the Computer Science community.

# Chapter 4

## Network of Authors

### 4.1 Introduction

Interactions between researchers is well known to be the essence of any research practice [61]. Collaboration is an intense form of interaction. Most phases of any research activity consist of discussion, analysis, exchange of results, and writing a paper together – in short they communicate to collaborate. The first collaborative scientific paper was published in 1665 [10] and the number of collaborative papers has increased ever since. With this dramatic increase of co-authored papers between individual researchers as well as among research institutes, it can be assumed that collaboration has become a prerequisite for modern research. To large extent, the structure of this scientific collaboration appears to be organized by the scientists themselves [49]. The increase in the number of international conferences has also fostered research collaborations. As a consequence of the need for coordination and joint funding of costly experiments, there is a general tendency towards internationalization in various

fields [77].

In this chapter, we investigate the Network of Authors (NOA). A co-authorship network is a social network consisting of a collection of researchers. Two authors are considered connected if they have authored a paper together. We analyze the characteristics of collaboration networks and compare them with other social networks. With the support of a collaboration network, we investigate bibliographic properties such as the average number of papers per author and the average number of authors per paper. We study how the network has evolved and how these properties have changed over time in the last 50 years (since 1960). We investigate the publication trends and the collaboration patterns in Computer Science and, using visualization techniques, we understand the distribution of authors, publications, and citations for the entire world and also take a deeper look at a few countries. Using a community detection algorithm, we identify communities in Computer Science. Finally, we rank the top collaborators in the field of Computer Science.

## **4.2 General Network Characteristics**

The properties of a social network can be described on two levels, global network metrics and individual node properties. Global network metrics describe the characteristics of an entire social network, for example, the network's diameter, mean node distance, number of communities, number of cliques, cluster-coefficient of the network, and small-world phenomena. Individual properties relate to the analysis of the properties of network nodes, e.g., distance to all the nodes in the network, betweenness centrality, closeness centrality, degree of the node, and its position in a

cluster. These network metrics help us understand the structure of the entire network and the importance of individual nodes in the network. These metrics are also very useful in comparing similarities and differences between networks of the same kind.

With the network constructed, we used tools such as Gephi<sup>1</sup> and Cytoscape<sup>2</sup> to visualize the network. We first performed a visual analysis to understand the structure of the ACM author network. Figure 4.1 shows a visualization of the NOA per decade from 1960 to 2010. Each node represents an author and the link represents co-authorships. The size of a node in the network represents the number of links that one node has to other nodes. Figure 4.1a shows that during 1960, the Computer Science community was very small, with small clusters and very few nodes acting as hubs. However, it can be seen that the network becomes increasingly denser with time, as new researchers join the community and new links are created when existing researchers in the network collaborate. Researches who have worked together once tend to re-collaborate [60] and this strength of their collaboration is shown by weights to the links in the network. The structure of the network is dynamic, so it continues to change with time. Researchers who continue to make new collaborations tend to move towards the centre of the network [48]. Prolific researchers (i.e., hubs in the network) tend to have many collaborations; they are usually situated in the centre of the network. The nodes on the periphery of the networks are those researchers who have fewer connections.

We performed Social Network Analysis (SNA) on the NOA to identify the kind of network we are dealing with. The characteristics of the NOA are shown in Table 4.1

---

<sup>1</sup><http://gephi.github.io/>

<sup>2</sup><http://www.cytoscape.org/>

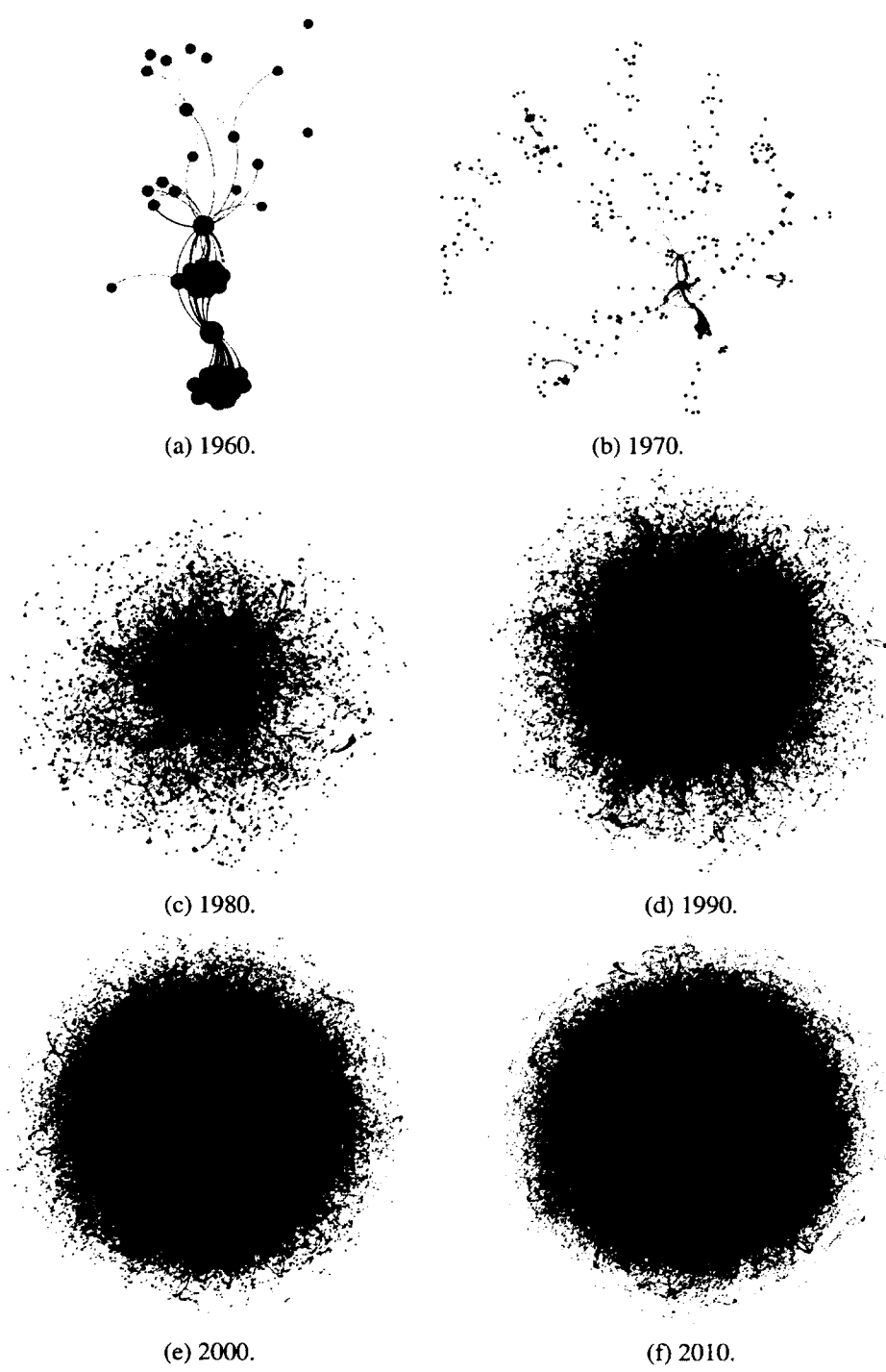


Figure 4.1: Visualization of the NOAA per decade starting in 1960 (top-left) to 2010 (bottom-right). Note that the network becomes increasingly denser with time.



Table 4.1: Network Statistics: A comparison between the NOA and the measurements take from the Film-Actors Networks [6].

<b>Measure</b>	<b>NOA</b>	<b>Film-Actors Network</b>
Nodes ( $n$ )	195,084	449,913
Links ( $m$ )	929,576	25,516,482
Mean Degree ( $z$ )	8.97	113.43
Exponent Power Law ( $\lambda$ )	2.37	2.30
Average Clustering Coeff. ( $C$ )	0.68	0.20
Average Path Length ( $\ell$ )	5.28	3.48

and are compared to other social networks available in the literature [6]. The network statistics for the NOA show that there are about 195,084 authors ( $n$ ) in the community who had about 929,576 connections ( $m$ ) between them. The average number of collaborators per author is  $z = 8.975$ , which gives a measure of the density of collaboration. The average path length  $\ell = 5.284$ , tells that any author in this community can reach another author in the same community with fewer than six connections between them. Comparing the two networks, the NOA demonstrates a higher clustering, meaning this network has higher collaborations that form triads (cliques of degree 3). Also, the high clustering coefficient of the network indicates that the network is organized in groups of highly collaborative individuals with few connections outside of the group. Figure 4.2 depicts the probability density function of degree of an author,  $D$ , follows a fat tailed distribution, indicating a significant heterogeneity.  $P(D)$  is the probability that a randomly selected node in a network has degree  $D$ . While most authors have fewer connections, a few have a large number of connections. The power-law characteristics for this network is within the expected values for real-world networks with  $\lambda = 2.576$ . Most of the real-world networks have  $2 \leq \lambda \leq 3$ . Hence, the NOA can clearly be characterized as a small-world network. Figure 4.3a shows the probability density function of number of publications by authors,  $P$ , follows a fat tailed distribution.

While most authors have fewer publications, a few have a large number of publications. Figure 4.3b shows the probability density function of number of citations received by authors,  $C$ , follows a fat tailed distribution. While most authors have received fewer citations, a few have received large number of citations.

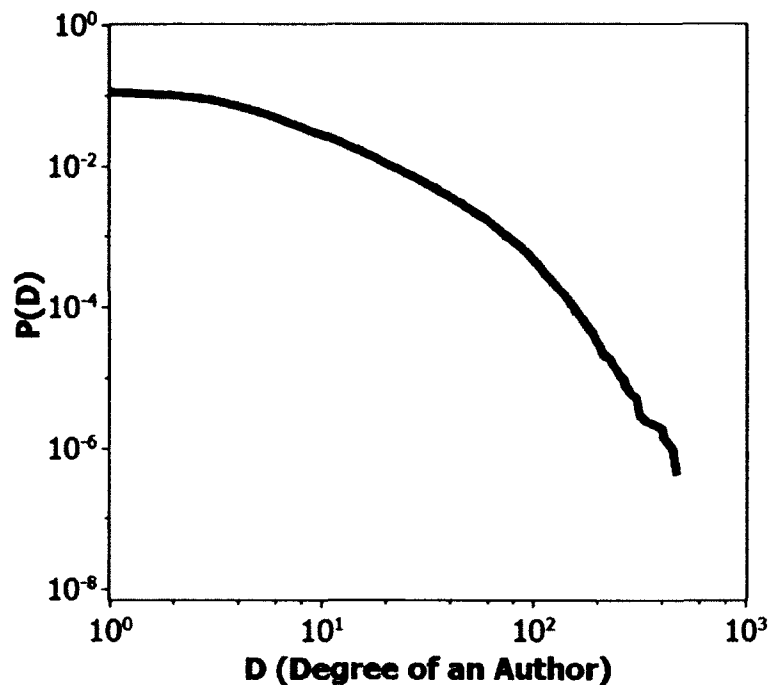
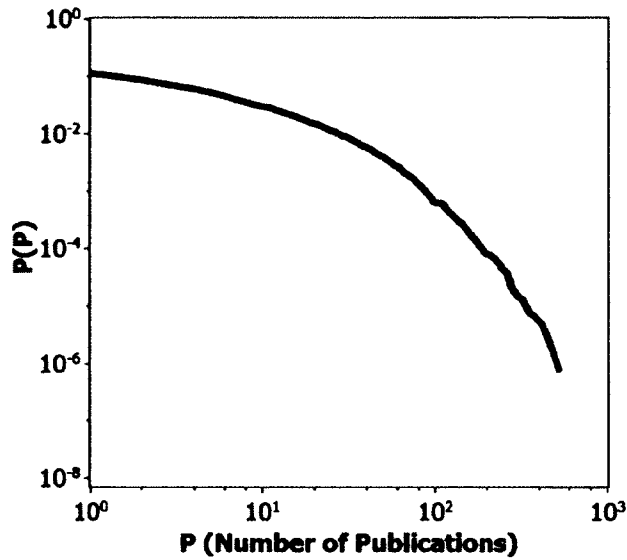
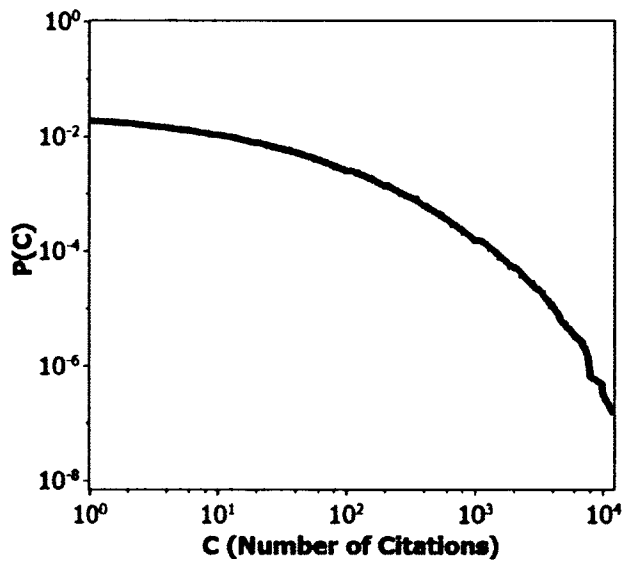


Figure 4.2: The probability density function of degree of an author,  $D$ , follows a fat tailed distribution, indicating a significant heterogeneity. While most authors have fewer connections, a few have a large number of connections. The ACM author network follows a power-law with  $\lambda = 2.576$ .

It is often of interest to analyze the statistical properties of networks to examine the entire distribution of a quantity, rather than just to look at the mean. Figure 3.9 shows that the Computer Science discipline has grown exponentially. What are the reasons for this exponential growth? Are individual researchers publishing more papers?



(a) The probability density function of number of publications by authors,  $P$ , follows a fat tailed distribution. While most authors have fewer publications, a few have a large number of publications.



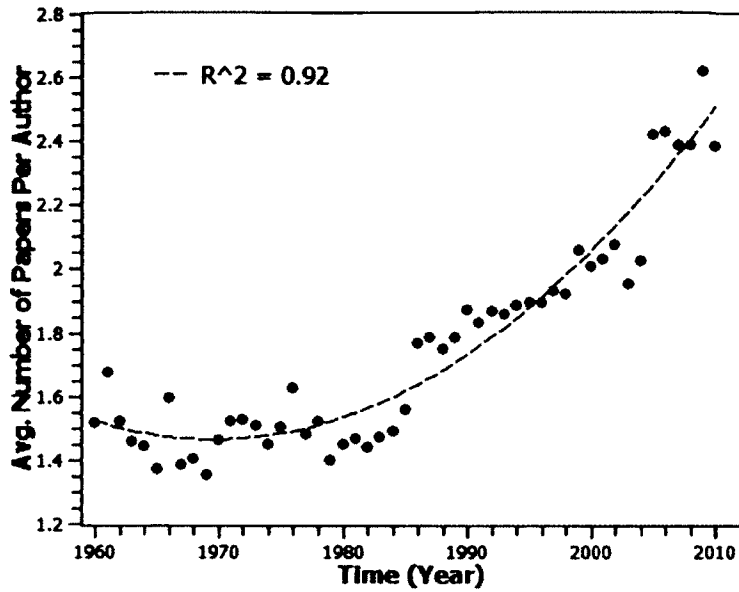
(b) The probability density function of number of citations received by authors,  $C$ , follows a fat tailed distribution. While most authors have received fewer citations, a few have received large number of citations.

Figure 4.3: The probability density function of publications and citations of authors.

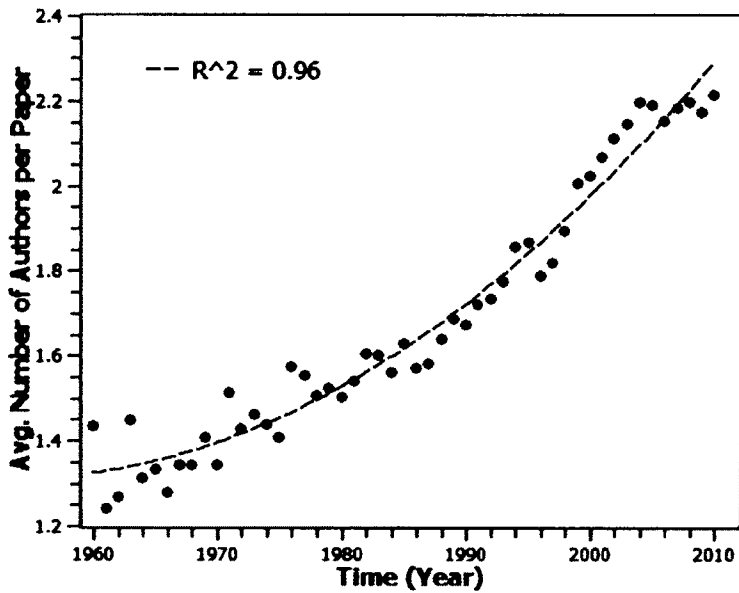
Are authors now collaborating more than before? To understand this growth for the Computer Science field, we performed other longitudinal studies. We took advantage of the network science tools and made a longitudinal study on the properties that emerge from co-authorship in Computer Science. For each year from 1960 to 2010, we analyzed the temporal evolution of the following bibliometric properties: average number of authors per paper and average number of papers per author.

Figure 4.4a shows a temporal evolution of an active authors' productivity in terms of the average number of papers published per author for each year since 1960. The red dashed line indicates the best fit function. It can be seen from the figure that author productivity has been growing incrementally over time. The average number of papers per author in 1960 was 1.5 papers and it had grown to 2.6 papers by 2010. The Computer Science discipline has grown substantially since there are more active authors and these authors have written more papers. Why is there an increase in the authors productivity over time? To understand this, we studied the average collaboration level in papers for each year.

Finally, for each year, we observed the average collaboration level in papers, i.e., the average number of authors per paper published in that year. In Figure 4.4b, we see that the collaboration in papers increased over time. The red dashed line indicates the best fit function for the data. The average number of authors per paper in 2010 was 2.2 authors, while the average in 1960 was 1.3 authors. Hence, we can conclude that the growth of the Computer Science discipline in terms of number of papers published is not only due to the growth of number of active authors in the community, but also due to the growth in the productivity of the authors and in the average collaboration level in papers.



(a) Average number of papers published per author for each year since 1960. The red dashed line indicates the best fit function given by:  $f(x) = ax^2 - bx + c$ , where  $a = 0.0007$ ,  $b = -2.63$  and  $c = 2596$ .



(b) Average co-authors per paper for each year since 1960. The red dashed line indicates the best fit function given by:  $f(x) = ax^2 - bx + c$ , where  $a = 0.0003$ ,  $b = -1.27$  and  $c = 1250$ .

Figure 4.4: Longitudinal analysis of average productivity and average collaboration level of ACM authors.

### 4.3 Collaboration Pattern

Given the importance of scientific collaboration for increased productivity and shared expertise, insights into the nature of such collaboration are of considerable value. Analysing the structure and the evolution of the network can provide useful insights about the nature of research and especially collaborative research in Computer Science. Using the dataset that spans over 50 years, we performed a longitudinal analysis on the publication trends. We analyzed the publication patterns for every year from 1960 to 2010. Also, using the publication information that was extracted for each paper, we investigated the collaboration patterns of researchers in the Computer Science community. We also studied the distribution of number of papers published to the number of co-authors in that paper.

Using the publication year and the list of authors available for each paper, we investigated the publication trends of researchers in Computer Science. Figure 4.5 shows the percentage of papers published by one-author (Blue), two-authors (Red) and three or more authors (Green) from 1960 to 2010. The graph shows that in the early period, there was a general tendency of researchers to publish single author papers. About 80% of papers published in 1960 were single author papers. This infers that back then researchers were less collaborative. There could be many reasons, such as lack of transportation or maybe delays in communication between researchers; but it is very hard to determine the main reason. However, since then the trend of single author papers has been reducing over time. Over time, Computer Science researchers have been increasingly collaborative. By the year 2010, one-author papers were reduced by 50%. The graph shows that the trend of two-authors papers remains somewhat steady

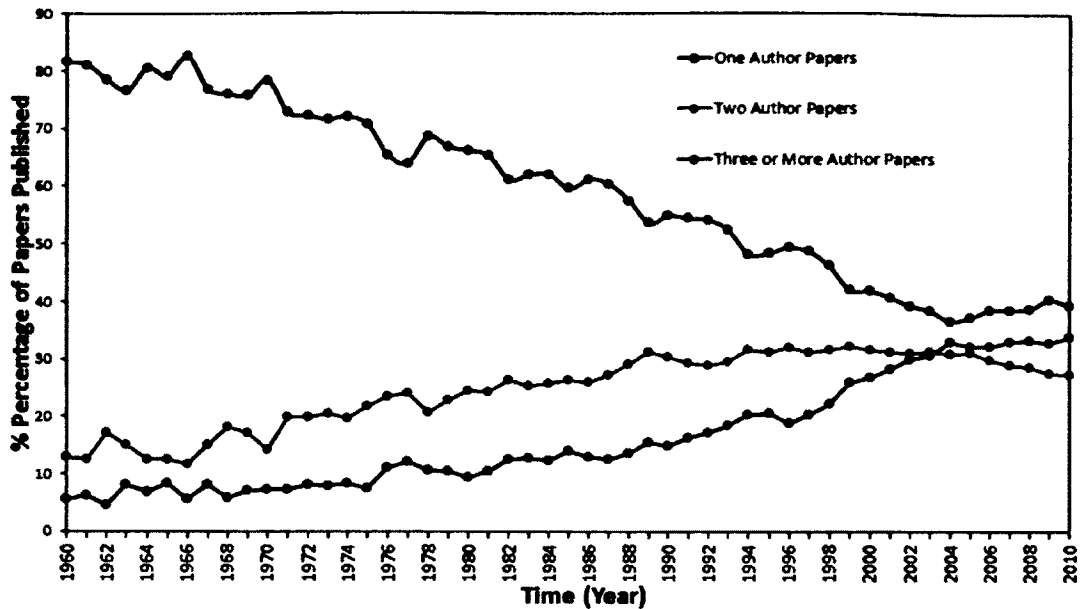


Figure 4.5: A longitudinal analysis of the publication trend from the year 1960 to 2010. The graph shows the percentage of papers published by one-author (Blue), two-authors (Red), and three or more authors (Green). Around the year 2004, there is a shift in trend with researchers publishing three or more authored papers.

with very gradual growth during 1980's. The graph also depicts a shift in the trend with papers with three or more authors becoming dominant after 2004. This gives an indication that researchers now prefer to collaborate on publications that have multiple authors. To further understand this publication trend, we investigated the effect of distance between the researcher collaborators in our network.

The improvements in transportation and communication should have enabled researchers to collaborate more and those improvements would have also made internationalization possible. Despite globalization, other studies have shown that many interactions still occur between researchers with closer geographical proximity [49]. It

helps researchers avoid or minimize many problems that arise during their research work – meeting other researchers, writing papers together, exchanging results, supervising co-workers, etc. To understand the role of physical distance, we measured the average distance between collaborators using the affiliation information of each author from their publications. We used the “haversine” Equation 4.2 to calculate the distance between two latitude/longitude point. Using the Equation 4.3, we calculate the average collaboration distance for an author. We then calculated the average collaboration distance using the Equation 4.4 for every year from 1980 to 2010.

$$a = \sin^2(\Delta\theta/2) + \cos\theta_1 \times \cos\theta_2 \times \sin^2(\Delta\lambda/2), \quad (4.1)$$

where,  $\theta$  is latitude and  $\lambda$  is longitude

$$\text{distance} = 2 \times R \times \text{atan2}(\sqrt{a}, \sqrt{1-a}), \quad (4.2)$$

where, R is earth’s radius (mean radius = 6,371 kms) and a is given by Equation 4.1.

$$\text{collabdist} = \frac{\sum_{i=1}^k \text{distance}_i \times \text{linkweight}_i}{\sum_{i=1}^k \text{linkweight}_i}, \quad (4.3)$$

where, k is the number of collaborators associated to an author,  $\text{distance}_i$  is the distance in kms calculated by Equation 4.2, and  $\text{linkweight}_i$  is the number of collaborations with the  $i^{\text{th}}$  author.

$$\text{Avg. collaboration distance} = \frac{\sum_{i=1}^n \text{collabdist}_i}{n}, \quad (4.4)$$

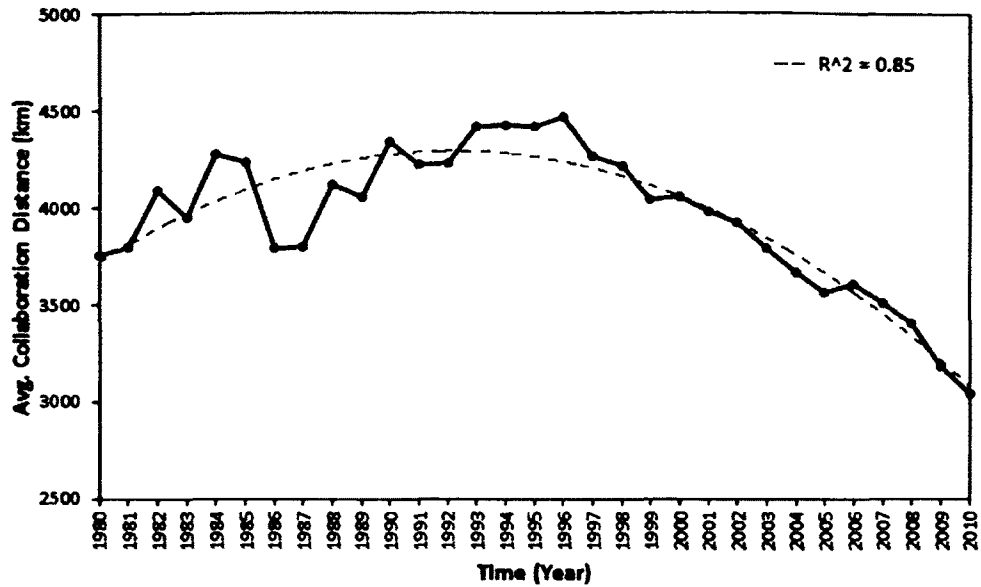


where,  $n$  is the number of authors and  $collabdist$  is the average collaboration distance for an author, given by Equation 4.3.

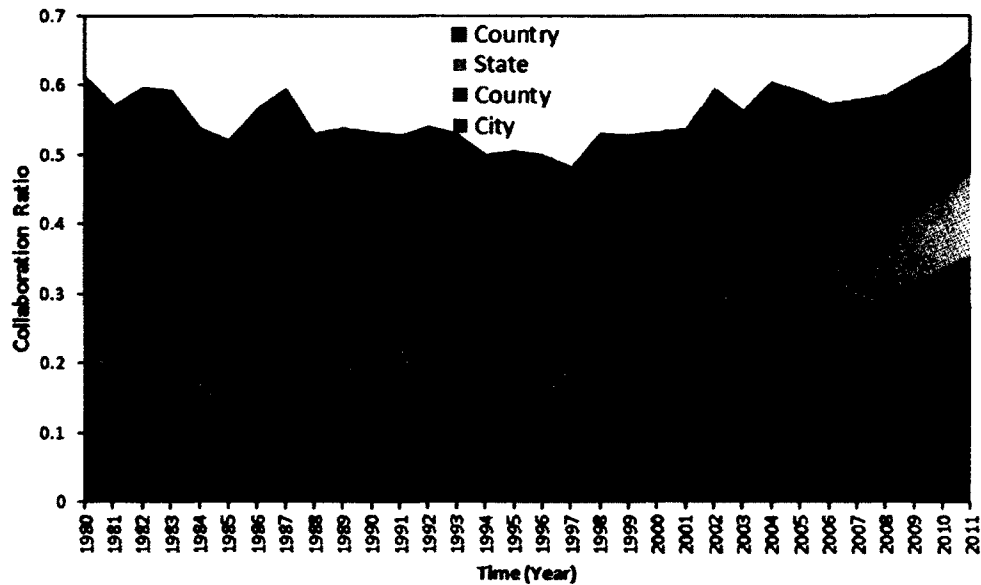
Figure 4.6a shows the average collaboration distance for a year and the red dashed line is the best fit function. During the early period, the graph shows slight oscillations between 1980 and 1990. It can be seen that the average collaboration distance has decreased significantly starting around the year 1996. In the year 2004, the average collaboration distance takes a further steep decrease, which is about the same year that papers with three or more authors started to be published more.

We performed an analysis on the collaborations and measured the fraction of local collaborations in relation to the total number of collaborations. Figure 4.6b shows the evolution of local collaboration. To measure the level of collaboration, we divided them into four different levels: city level collaborations (which would generally be for collaborations within the institution or institutions nearby), county level collaborations (collaborations with institutions in the same county), state level collaborations, and country level collaborations. The results show that researchers seem to have increased their collaborations within their own institution but not with other institutions in the same state. All the increase at the state level and county level is a reflection of city level increase. This analysis indicates that strong collaborators probably tend to move closer to enhance their research productivity. This is just an assumption and we would like to confirm this hypothesis with other results.

Figure 4.7 shows the probability density function of number of co-authors in a paper,  $CA$ , follows a fat tailed distribution.  $P(CA)$  is the probability that a randomly



(a) Average collaboration distance for a year. The red dashed line indicates the best fit function given by:  $f(x) = ax^2 + bx + c$ , where  $a = -3.77$ ,  $b = 99.55$  and  $c = 3632$ .



(b) Evolution of local collaborations.

Figure 4.6: Analysis was performed by considering authors' affiliation. Figure 4.6a shows the average distance between collaborators. Since 1996 we have seen a decrease of nearly 40% of the average distance, which leads us to argue that collaborations are becoming more local with time. Figure 4.6b confirms our hypothesis that collaborations are more local.

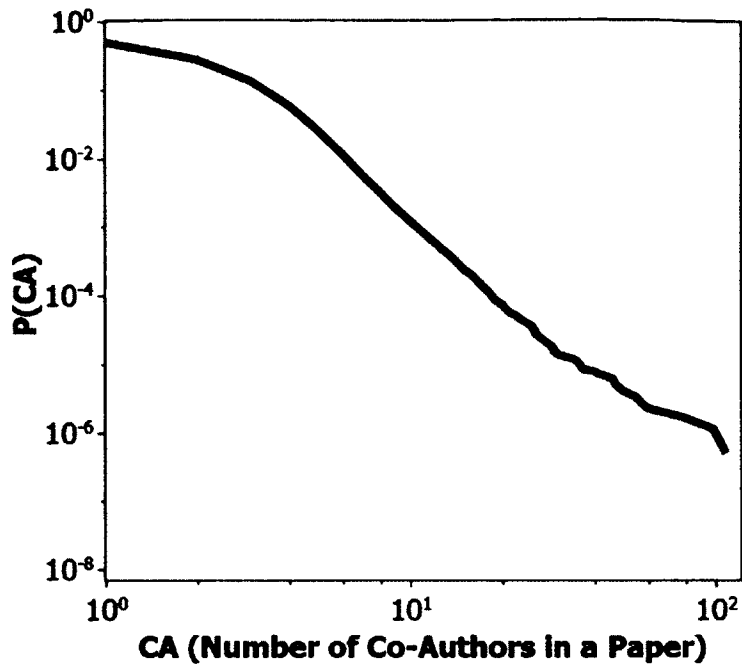


Figure 4.7: The probability density function of number of co-authors in a paper,  $CA$ , follows a fat tailed distribution. While most papers have fewer co-authors, a few have large number of co-authors.

selected node in a network has,  $CA$  number of co-authors in a paper. While most papers have fewer co-authors, a few have large number of co-authors. The Computer Science community has many single author papers and very few research publications containing many authors. Fewer than 1% of the papers in our dataset have 11 or more co-authors in the paper.

From our previous analysis, we learned that authors are collaborating now more than before to increase their productivity and improve their position in the Computer Science community. However, does collaboration improve the quality of research? Does the number of citations to a research paper correlate to the number of co-authors in that

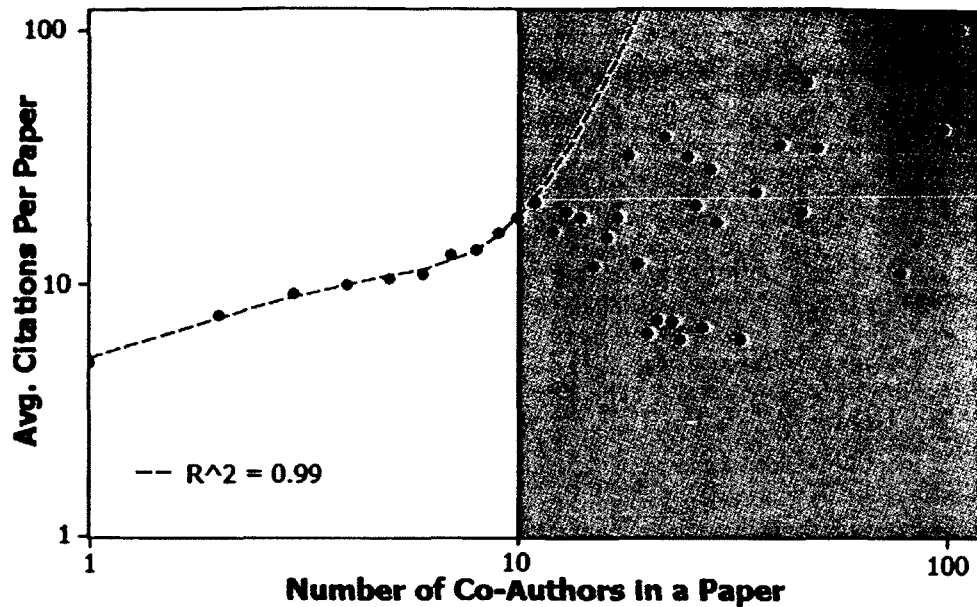


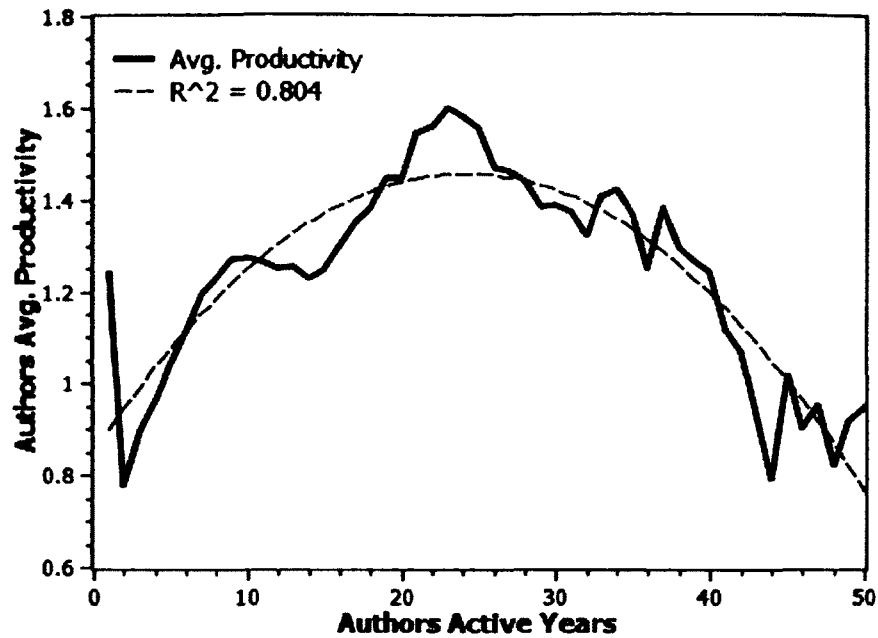
Figure 4.8: Average citations per paper to number of co-authors in that paper. The red dashed line indicates the best fit function given by:  $f(x) = ax^3 + bx^2 + cx + d$ , where  $a = 0.03$ ,  $b = -0.46$ ,  $c = 3.23$  and  $d = 2.32$ .

research paper? If so, what would be a good number of co-authors to have to produce a quality work? Figure 4.8 shows the correlation between the average citations per paper to the number of co-authors in that paper. Pearson's correlation was calculated between the two variables; it tells us that they are positively correlated (0.67). Since there are fewer papers published with 11 or more authors, we see slight oscillations in the values of average citations per paper. The grayed region in the graph indicates this information.

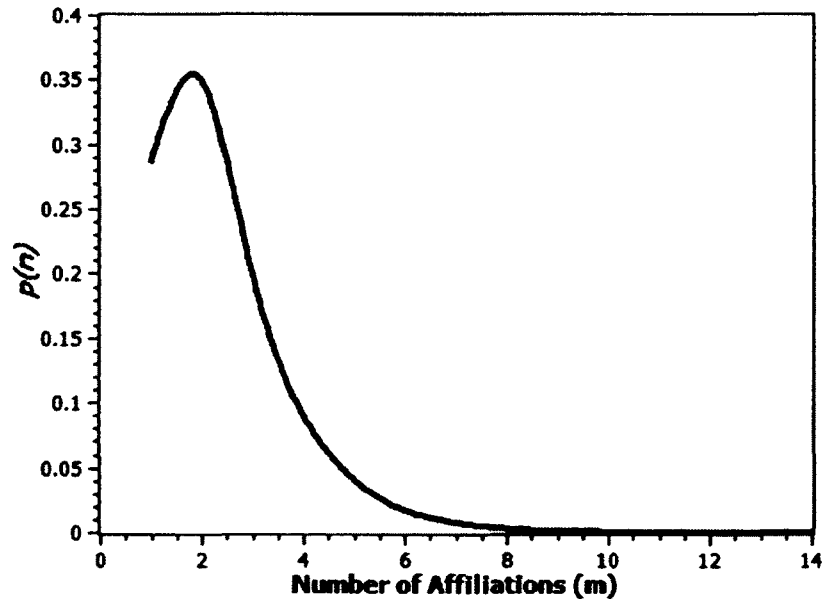
Every individual author's career varies from that of other authors. Some authors' careers are short lived and some authors stay in the field for a long timer. We analyzed our data to find if there is a certain period in an author's career (active years) when they produce more publications. Figure 4.9a shows the average publications of all the

authors with respect to their publication year. We consider an author's active years from his or her first publication year to last publication year. Even when there are no publications during this period, the author is still considered to be active. It can be seen from the graph that most authors are productive during the first year, i.e., more than one publication per author. Although their productivity drops initially, thereafter it increases and their most productive periods are between 15 - 30 years of their active years. After this, an author's productivity starts to drop. From the graphs, it can be seen that there are authors who have been active for close to 50 years, but they are very few.

Changing institutions is an integral part of an academic life and mobility is often important in furthering a professional career [72]. Using the affiliations of authors from each of their publications, we investigated their career movements. We considered only the established researchers, i.e., authors who have been active in the field of Computer Science for five and more years. We computed  $P(n)$ , the probability for a researcher to have  $n$  different affiliations associated during his career as shown in Figure 4.9b. Figure shows that authors are mostly associated with one, two, or three affiliations and there are very few authors with eight or more affiliations associated with them.



(a) Authors average productivity and the red dashed line indicates the best fit function given by:  $f(x) = ax^3 + bx^2 + cx + d$ , where  $a = -0.001$ ,  $b = 0.05$  and  $c = 0.851$ .



(b) The probability density function  $P(n)$  of number of affiliations associated to an established authors who have been active for 5 and more years. Authors are mostly associated with one, two or three affiliations,  $P(n)$  decays quickly as  $n$  increases.

Figure 4.9: Analysis on authors active years and affiliations.

## 4.4 Community Analysis and Area Diversity

To understand how the communities of researchers are structured, we used a detection algorithm to find communities. In network sciences, the concept of community has been studied as a way to find tightly connected groups of nodes whose number of links within the community is significantly higher than the number of links to nodes outside the community. The one we used was proposed by Palla et al. [68], which is generally referred to as the clique-percolation algorithm for overlapping communities. This algorithm differs from others because nodes may belong to more than one community, hence the term “overlapping.” The size of the overlap between the communities, given by the number of nodes that are in the overlap, identifies how connected the two communities are in the network of communities. Figure 4.10 shows the size distribution of communities found using the best configuration in the clique-percolation algorithm.

We looked at how these communities are organized and the effect of area diversity in the community to the overall publication record of the community. The community detection algorithm found many communities but we considered only the largest ones: 30 communities in total. They range in size (number of authors) from 7,044 for the largest community to 101 for the smallest. Here we like to understand whether the diversity of areas of publication in these communities related to how prolific the authors are.

ACM has introduced a classification system for publications. The classification divides the Computer Science field into 11 main areas: General Literature (A), Hardware (B), Computer System Organization (C), Software (D), Data (E), Theory of

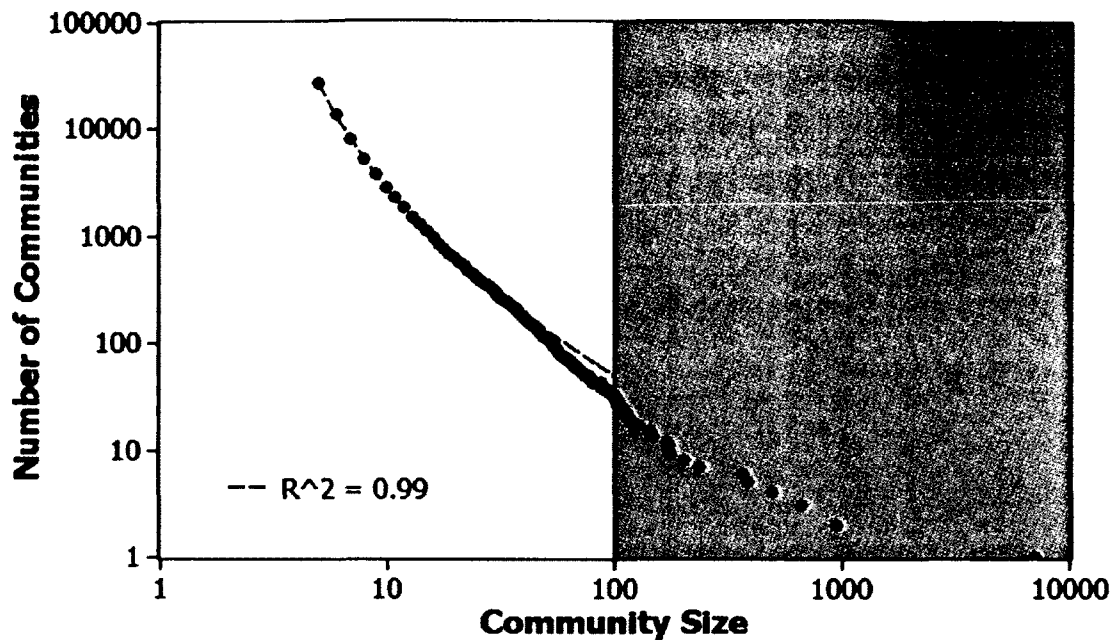


Figure 4.10: Community size distribution follows the scaling law with  $\gamma = 1.66$ . The red dashed line indicates the best fit function given by:  $f(x) = e^{a+\frac{b}{x}+c\log x}$ , where  $a = 9.18$ ,  $b = 14.5$  and  $c = -2.7$ .

Computation (F), Mathematics of Computing (G), Information System (H), Computing Methodologies (I), Computer Applications (J), and Computing Milieux (K). These areas are in turn subdivided into specific fields, for example, D.3 relates to Programming Languages (See Appendix A for details). In total we considered 80 second-level subdivisions in the ACM classification.

Figure 4.11 shows a chart in which rows represent the communities (the number inside the parenthesis indicates the number of authors that belong to that community) and the columns represent all 80 second-level ACM subdivisions. Each cell represents the number of papers published by a particular community in the specific area. Red indicates a high number, yellow indicates a low number, and white indicates no data.





Figure 4.11: Map of field diversity in communities of Computer Science researchers. Each row represents a community and each column an area of research in Computer Science. The color intensity is a count of how many authors of that community have published in the specific area.

The rows (communities) have been sorted from top to bottom based on number of authors in that community. The top row is the largest community with 7,044 authors who published 54,284 papers, which have 834,532 citations in 80 different areas. The bottom row is the smallest community considered in our study, which contains 101 authors who published 645 papers which have 6,032 citations in 39 areas only. To understand if diversity does influence research prolifically, we considered the most diverse and least diverse communities. The least diverse community has an average of 7.87 citations per paper and the most diverse community has an average of 15.37 citations per paper. We calculated Pearson's correlation between the average number of publications per author and diversity, as well as the correlation between the average number of citations per author and diversity. In both cases, the numbers are not correlated or are very low positively correlated (0.18 and 0.09), leading us to conclude that the diversity of a collaborative network does not follow trends in the area of computation.

As researchers grow in their career, they tend to publish papers in different areas of Computer Science. We consider an established researcher to be one who has been active for five years or more. Generally, most of the established researchers have a majority of their publications in either one or two areas of Computer Science. To understand how subjects are connected to each other, we analyzed a network of Computer Science subjects. Figure 4.12 shows the visualization of the Computer Science subject network. In this network, a node represents a subject and the size represents the number of papers published in a particular subject by all the researchers in the Computer Science community. Two subjects are connected if an author publishes a paper in both the subjects. The link weight represents the number of authors who have published papers in both the subjects. From the graph, we can see that C.2 (Computer-Communication

Networks) and I.2 (Artificial Intelligence) appear to be the largest nodes, i.e., there are many papers published in these areas. The links between the nodes C.2 and D.4 (Operating Systems) and nodes H.2 (Database Management) and H.3 (Information Storage and Retrieval) appear to be strong. This tells us that there are many authors who have published in C.2 and also published in D.4. Likewise, most of the authors who published in H.2 have also published in H.3. We can infer that authors generally publish papers in the area of their specialization. However when they publish in another area, it is often very closely related to the author's main area of research. We also performed a similar analysis on the network of subjects for a few countries (See Appendix B for details).

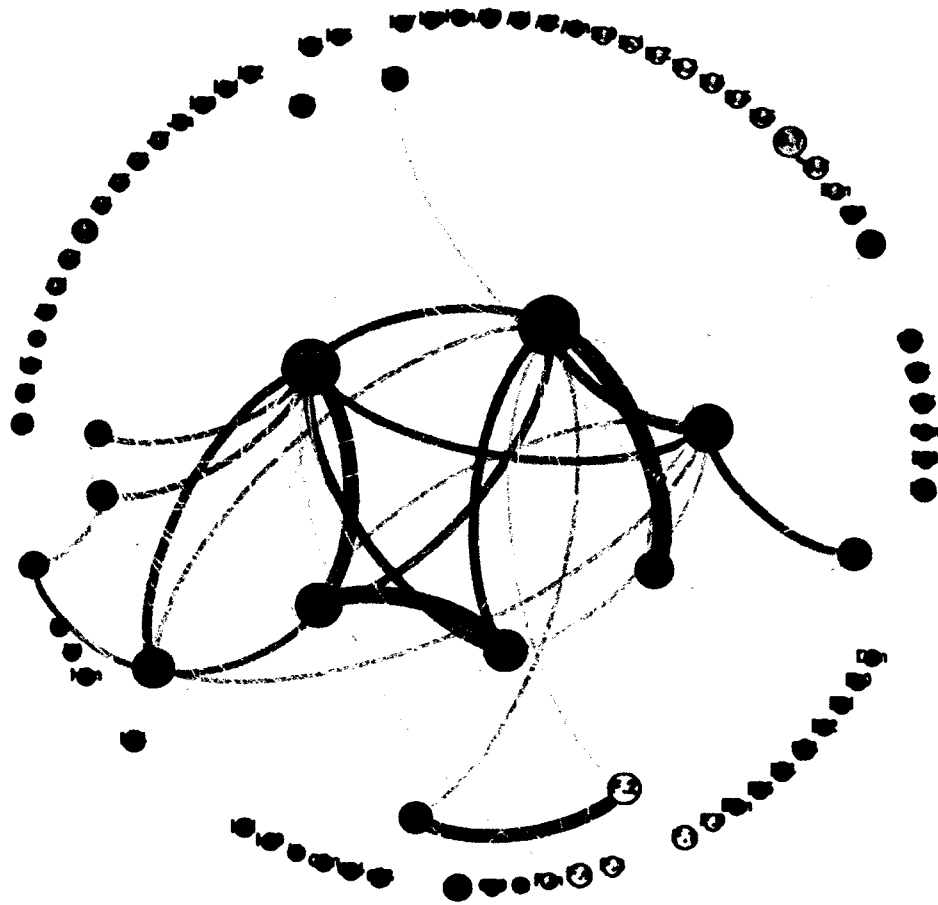


Figure 4.12: Visualization of the Computer Science subject network. Each color represents a subcategory in Computer Science (See Appendix A for details). Every node represents a subject (categorized by ACM) and the size represents the number of papers published in that subject. A link weight would represent the number of authors who have published papers in both the subjects. Link weight is represent in the graph by its thickness and color density.

## 4.5 Ranking of Authors

The performance of scholars and scientists can be evaluated based on publications, citations, and the number of grants they have received. Such evaluation of researchers is not only required by the department, but also for achieving a reputation within the research community and for governmental research fund allocation. A high reputation attracts federal funding and also attracts highly qualified students around the world which elevates research standards and goals. Hence, it is important to identify key scholars in the Computer Science community. Since the work is related to collaboration, we ranked based on degree centrality, which in network terms corresponds to the number of author's collaborations. Table 4.2 shows the top most collaborative authors in Computer Science according to the ACM dataset. The top collaborator in the list is Ian Forster, who is an American computer scientist and the Director of the Computation Institute, a joint institute of the University of Chicago and Argonne National Laboratory. Jack Dongorra, who is second on the list, is an American University Distinguished Professor of Computer Science in the Electrical Engineering and Computer Science Department at the University of Tennessee. Looking at the table you can also notice that Ewa Deelman, a Research Associate Professor in the Computer Science Department at the University of Southern California, has only 97 publications but she still appears in the top 10; this is due to the fact that she has participated in papers with tens of authors leading to an inflation of her rank.

Other network metrics, such as betweenness centrality and closeness centrality, also indicate how important an author is in the network. Table 4.3 shows the top individuals based on betweenness centrality and closeness centrality. Ranking the authors based

Table 4.2: Top CS researchers according to their level of collaboration (number of distinct collaborations).

<b>Author Name</b>	<i>degree</i>	<b>Publications</b>	<b>Citations</b>
Ian Foster	471	274	7,138
Jack J Dongarra	453	485	5,298
Li Feng Zhang	413	312	4,599
Mingqiang Li	408	396	2,850
Lei Zhang	342	273	1,182
Philip S Yu	324	525	7,916
Ewa Deelman	318	97	1,255
Geoffrey Fox	315	272	1,347
Edward Allan Fox	314	268	1,486
Jiawei Wei Han	311	416	11,775

Table 4.3: Top CS researchers ranked by Betweenness and Closeness Centrality

<b>Author Name</b>	<i>betweenness</i>	<b>Author Name</b>	<i>closeness</i>
Mingqiang Li	3.70E+08	Li Feng Zhang	3.7694
Li Feng Zhang	3.55E+08	Mingqiang Li	3.8590
Jack J Dongarra	3.28E+08	Christos Faloutsos	3.8709
Christos Faloutsos	1.85E+08	Ian Foster	3.9024
Lei Zhang	1.80E+08	Phillip S Yu	3.9025
Laurence Tianruo Yang	1.76E+08	Jiawei Wei Han	3.9033
Barry William Boehm	1.66E+08	Umeshwar Dayal	3.9046
Ian Foster	1.57E+08	Hector Garcia-Molina	3.9055
Phillip S Yu	1.57E+08	Jack J Dongarra	3.9129
Jiawei Wei Han	1.57E+08	Gerhard Weikum	3.9315

on these network metrics gives us a different perspective on the authors who tend to quickly receive and spread information respectively in the network. Table 4.4 shows the top 10 collaborators in Computer Science. For each collaborator we show the number of publications they had together and the citations they have received for that work.

Table 4.4: List of Top 10 Collaborators in Computer Science.

<b>Collaborators</b>	<b>Publications</b>	<b>Citations</b>
Sudhakar Mannapuram Reddy/Irith Pomeranz	307	1,027
Didier Dubois/Henri M Prade	219	2,931
Tomoya Enokido/Makoto Takizawa	184	151
Mahmut Taylan Kandemir/J Irwin	122	1,387
Leonard Barolli/Arjan Durresi	121	205
Jiajun jun Bu/Chun Hao Chen	117	219
Tharam Singh Dillon/Elizabeth J Chang	113	232
Leonard Barolli/Fatos Xhafa	112	63
Enrico Macii/Massimo Poncino	110	601
Takahiro Hara/Shojiro Nishio	110	200

## 4.6 Geographical Distribution

It is important to understand the research contribution of each country in the field of Computer Science. We used visualization techniques to understand how authors, publications, and citations are distributed per country and per county in the US. The color intensity indicates the value for that particular region. Bright red indicates a high value and light yellow indicates a very low value, whereas grey indicates that we do not have any data for that specific region.

Figure 4.13 shows the distribution of authors, publications, and citations per country. The correlation appears to be fairly high, meaning that the countries with a high number of authors have a higher publication count and higher citation count. Only a few countries have a citation level proportional to the number of publications; these include the US, Germany, and Great Britain. Researchers from China and France publish considerably well but in general, the works are not being cited at the same proportion. Note that we are not implying that the works of certain countries are less valuable due to their lack of citations – this cannot be derived from these results. Figure 4.14 shows

the distribution of authors, publications, and citations of countries only in Europe. This figure gives a better visual analysis of countries in Europe. Clearly, Great Britain and Germany appear to be the strongest contributors to the field of Computer Science.

Figure 4.15 shows the distribution of authors, publications and citations of counties in the US. Counties such as Santa Clara and Los Angeles in California, Middlesex in Massachusetts, and Allegheny in Pennsylvania are shown by dark red, indicating that there are many authors with a lot of publications and most of their works are well cited. This is quite expected given these locations are home to world-renowned scientists working at the top institutions in the country.





(a) Distribution of Authors.



(b) Distribution of Papers.



(c) Distribution of Citations.

Figure 4.13: Heatmap of number of authors, number of publications, and number of citations (top to bottom) for countries.



(a) Distribution of Authors.

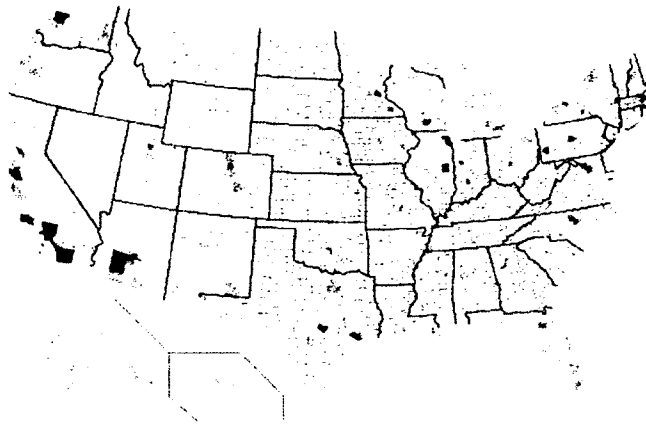


(b) Distribution of Papers.

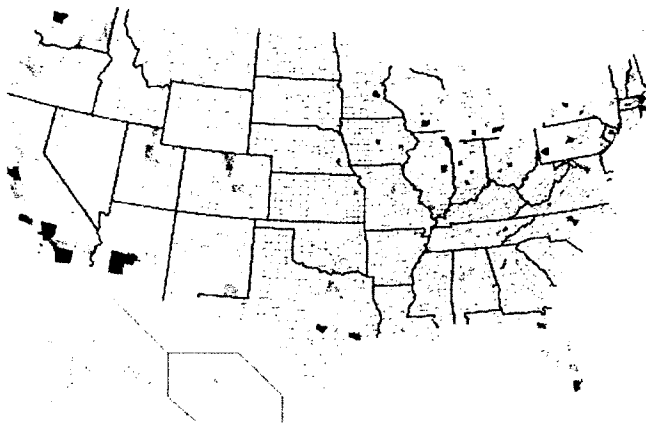


(c) Distribution of Citations.

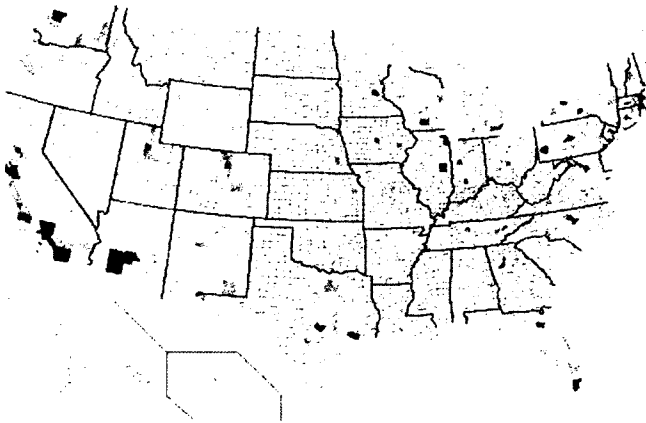
Figure 4.14: Heatmap of number of authors, number of publications, and number of citations (top to bottom) for European countries.



(a) Distribution of Authors.



(b) Distribution of Papers.



(c) Distribution of Citations.

Figure 4.15: Heatmap of number of authors, number of publications, and number of citations (top to bottom) for US counties.

## 4.7 Conclusion

In this chapter, we investigated the network of authors (NOA). We analyzed the characteristics of the collaboration networks and compared them with other social networks. We found that NOA follows power-law characteristics with  $\lambda = 2.576$  and since  $2 \leq \lambda \leq 3$ , it can be categorized as a scale-free network. Its average path length  $\ell = 5.284$ , i.e., any author in the community can be reached with fewer than six connections between them.

We also investigated the bibliographic properties and performed a longitudinal analysis to understand the publication trends and collaboration patterns of authors. Since 1960, there has been a gradual drop in single author papers and a steady increase in three or more authors per paper. We observed that since 1996, there has been a decrease of nearly 40% in the average collaboration distance, which leads us to argue that collaborations are becoming more local with time. Researchers in Computer Science are very productive during the 15 to 30 year period of their active careers and most researchers are associated with one, two, or three affiliations.

Finally, we ranked the top authors in terms of number of distinct collaborators and other network metrics in the field of Computer Science. Ian Foster and Li Feng Zhang are few of the authors who clearly appear to be the most influential in this network. Visualization techniques were used to understand the distribution of publications, citation, and authors for the entire world and for the United States.

# Chapter 5

## Network of Institutions

### 5.1 Introduction

In the previous chapter, we learned about collaboration patterns and trends of researchers in Computer Science. Researchers are generally affiliated with some university or a research lab. One could say that collaboration and exchange of knowledge happens not only between researchers but also between institutions (universities/labs), although collaboration between institutions is mainly driven by the collaboration of authors. Institutions are increasingly recognized as central actors in the production and delivery of new knowledge, and they play a unique role in supporting economic development [11]. Fostering closer ties between institutions is crucial. The interaction between them plays a critical role as a source of fundamental knowledge and innovation to new technologies [62].

In this chapter, we investigate the Network of Institutions (NOI). We analyze the characteristics of the collaboration networks and compare them with the network of

authors. We investigate bibliographic properties and do a longitudinal analysis on publication trends and collaboration patterns of institutions. Finally, we rank the top institutions in the field of Computer Science based on network metrics.

## 5.2 General Network Characteristics

The existing dataset for the NOA had to be restructured to perform our analysis on NOI. Here the bi-partite network, i.e., institutions-paper network, is represented by two sets of nodes (institutions and papers), and the links running from institutions to papers. A paper is associated with a institute if at least one of the authors from that institute has authored that paper. Since the study is focused on understanding the collaboration between institutions, we concentrate only on the institute projection. In NOI, every node represents a institute and two institutions are connected if the authors from those institutions have co-authored a paper.

Once the network was constructed, we performed an analysis to identify the kind of network we were dealing with and compared the network metrics with NOA. The characteristics of the NOI is as shown in Table 5.1 and is compared to NOA. The network statistics for the NOI shows that there are about 12,541 universities and research labs ( $n$ ) in the community, which had about 94,817 connections ( $m$ ) between them. The average number of collaborators per institute is  $z = 15.121$ , which tells that, on average, every institute collaborates with 15 other institutions. The average path length  $\ell = 3.41$ , tells that any institute in this community can reach another institute in the same community with fewer than four connections between them. The NOI demonstrates a high average clustering coefficient ( $C$ ) = 0.509, meaning this network has a high number

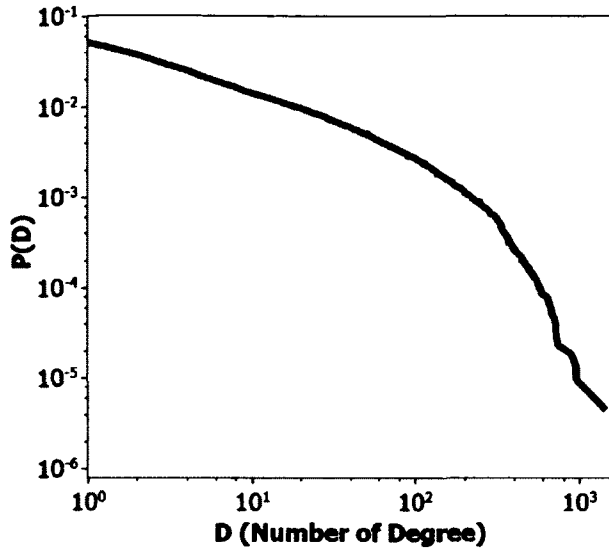
of collaborations that form triads. However, when compared to NOA, its average clustering coefficient is comparatively low. The clustering coefficient of this network indicates that the network is organized in groups of highly collaborative institutions with few connections outside of the group.

Table 5.1: Network Statistics: A comparison between the Network of Institutions and ACM Network of authors.

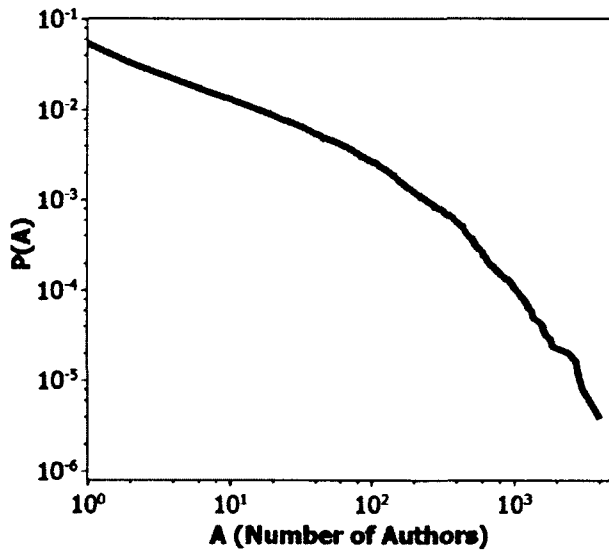
<b>Measure</b>	<b>NOI</b>	<b>NOA</b>
Nodes ( $n$ )	12,541	237,351
Links ( $m$ )	94,817	1,065,078
Mean Degree ( $z$ )	15.12	8.97
Exponent Power Law ( $\lambda$ )	1.22	2.37
Average Clustering Coeff. ( $C$ )	0.51	0.68
Average Path Length ( $\ell$ )	3.41	5.28

Figure 5.1a depicts a probability density function of degree of an institute,  $D$ . The NOI follows a power-law, i.e., there are very few institutions that have a high number of connections and there are many institutions that have very few connections. The power-law characteristic for this network is  $\lambda = 1.22$ . Based on the number of distinct collaborators, some of the top institutions are Carnegie Mellon University (1,155 collaborators), Massachusetts Institute of Technology (1,154 collaborators), and Stanford University (941 collaborators).

Figure 5.1b shows the probability density function of number of authors in a institute,  $P$ . This distribution follows a power-law, i.e., there are very few institutions to which many authors are affiliated and there are many institutions with very few authors affiliated to it. It is important to note that, in our analysis, we also consider all the authors who were previously associated with a institute in their career, towards the number of



(a) The probability density function of degree of an institute,  $D$ , follows a fat tailed distribution, indicating a significant heterogeneity. While most institutions have fewer connections, a few have a large number of connections. The NOI follows a power-law with  $\lambda = 1.22$ .



(b) The probability density function of number of authors in a institute,  $P$ , follows a fat tailed distribution. While most institutions have fewer authors, a few have a large number of authors.

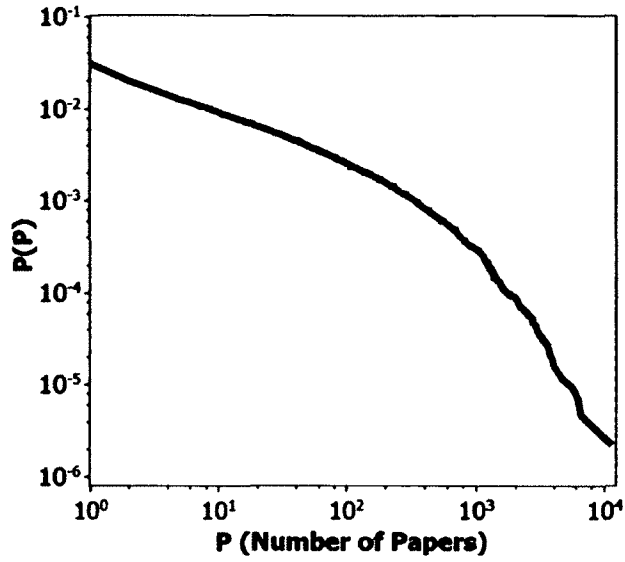
Figure 5.1: The probability density function of degree and authors of a institute.



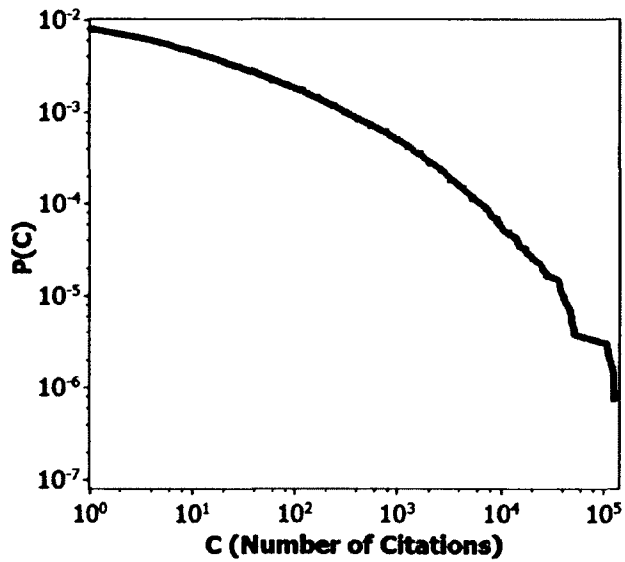
authors associated with that institute. Based on the number of authors associated with a institute, some of the top institutions are, Carnegie Mellon University (3,945 authors), University of Minnesota (3,035 authors), and Massachusetts Institute of Technology (2,809 authors).

Figure 5.2a shows the probability density function of number of publications by a institute,  $P$ . This distribution follows a power-law, i.e., there are very few institutions that have published a high number of papers and there are many institutions with very few publications associated with them. A paper is associated with a institute based on an author's affiliation at the time of publication and not his or her current affiliation. Based on the number of publications associated with a institute, some of the top institutions are Carnegie Mellon University (10,958 publications), Stanford University (6,602 publications), and Massachusetts Institute of Technology (6,325 publications).

Figure 5.2b shows the probability density function of number of citations received by a institute,  $C$ . This distribution follows a power-law, i.e., there are very few institutions that have received a high number of citations to the papers associated with them and there are many institutions with very few citations to the papers associated with them. Based on the number of citations received by a institute, some of the top institutions are Carnegie Mellon University (130,080 citations), Stanford University (126,709 citations) and Massachusetts Institute of Technology (113,395 citations).



(a) The probability density function of number of publications by a institute,  $P$ , follows a fat tailed distribution. While most institutions have fewer publications, a few have a large number of publications.



(b) The probability density function of number of citations received by a institute,  $C$ , follows a fat tailed distribution. While most institutions have received fewer citations, a few have received large number of citations.

Figure 5.2: The probability density function of publications and citations of a institute.

### 5.3 Collaboration Pattern

Using the publication year and author's affiliation information available for each paper, we investigated the publication trends of institutions in Computer Science. Figure 5.3 shows the percentage of papers published by one institute (Blue), two institutions (Red), and three or more institutions (Green) from 1960 to 2010. Comparing this graph with Figure 4.5 for NOA, we see some similarities in trends. The graph shows that in the early period, there was a general tendency for researchers to publish single author papers or publish with researchers from the same institute. More than 90% of the papers published during the 1960s were one institute papers. However, since then the trend of one institute papers has been deteriorating over time and by 2010 it had dropped to 70%. We can infer from this that, back then, researchers were less collaborative and if they collaborated, it was with researchers from the same institute. However, with time Computer Science researchers have been increasingly collaborative. The graph shows that the trend of two and three or more institutions per paper has been growing somewhat steadily over time. Fewer than 5% of papers published during the 1960s were two institute papers, but this number had increased to 20% by 2010. In contrast, the three or more institute papers, grew from 3% in the 1960 to 9% in 2010. This indicates that the majority of the papers are still published by a single institute, but recently there has been a growth in multiple institutions per paper. To further understand this publication trend, we investigated the average number of institutions per paper with respect to time.

For each year, we observed the trend of collaboration level per paper, i.e., the average number of institutions per paper published in that year. In Figure 5.4, the collaboration level for each year is indicated by a blue bubble with an error bar representing the

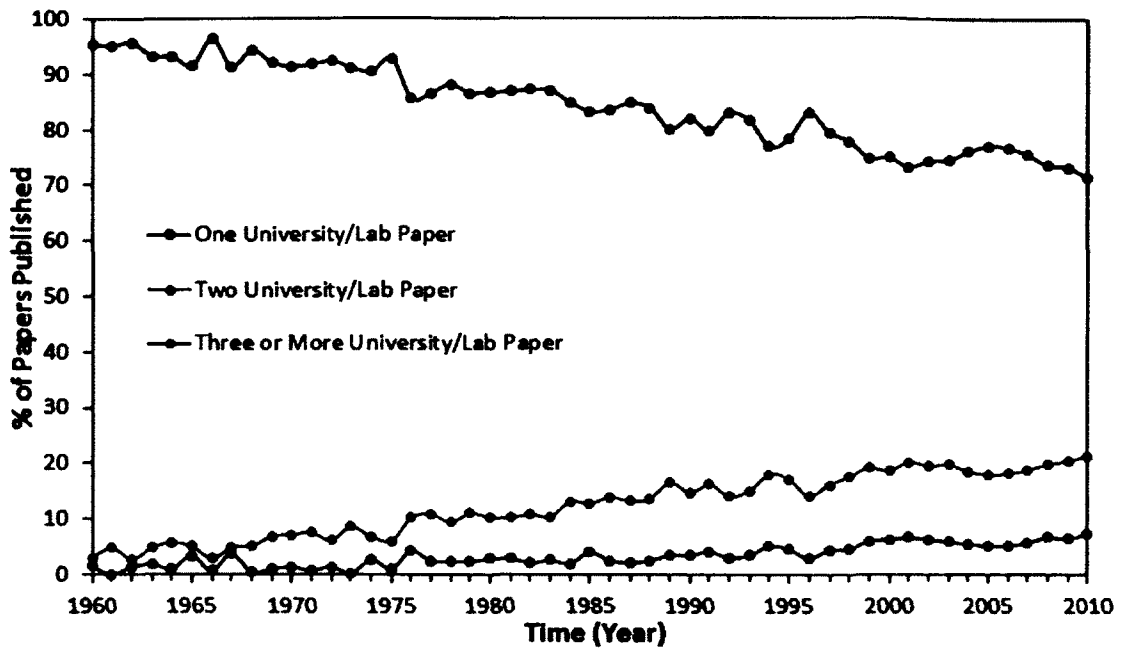


Figure 5.3: A longitudinal analysis of the publication trend from 1960 to 2010. The graph shows the percentage of papers published by one institute (Blue), two institutions (Red), and three or more institutions (Green).

variability of data. The red dashed line indicates the best fit function for the data. We see that the average number of institutions per paper has increased gradually over time. The average number of institutions per paper in 2010 was 1.2 institutions, while the average in 1960 was approximately 1.0 institute per paper. Hence, we can conclude that there has been a marginal growth in the Computer Science discipline in terms of the average number of institutions per paper. This is mainly due to the growth of papers being published with two and three or more institutions per paper. Another reason for this increase in trend could be the growth in the number of active authors in the Computer Science community, as this could lead to high number of collaborations between the institutions. However, it is very hard to infer anything about internationalization from this analysis. To understand the collaboration pattern, we investigated the collaboration

of institutions by considering their geographical aspect.

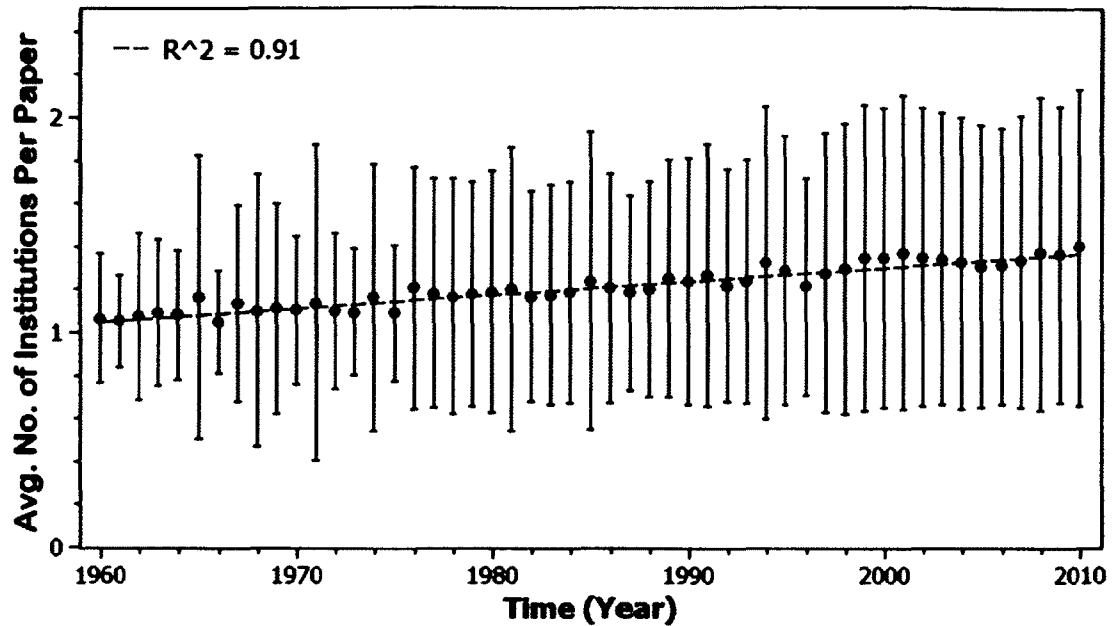


Figure 5.4: A longitudinal analysis of collaboration level from the year 1960 to 2010. The graph shows average number of institutions per paper with error bars for every year. The red dashed line indicates the best fit function given by:  $f(x) = ax + b$ , where  $a = 0.006$  and  $b = -11.16$ .

We know that collaboration between researchers improves their productivity and knowledge. In addition, collaboration indirectly improves the productivity and impact of the institute from where these researchers come. This is quite important in the Computer Science community as it attracts high profile researchers and new and aspiring PhD students. In our previous analysis, we analyzed the percentage of publications by one, two, and three or more institutions per paper. However, it does not tell you if these collaborations are national or international. Also, the one institute paper does not distinguish, whether they are single author or multiple authors from the same

institute. It would be interesting to see what percentage of collaborations happen with other institutions within the same nation and the percentage of collaborations with institutions outside the nation. Figure 5.5, shows a longitudinal analysis of the collaboration trend from 1960 to 2010. The graph shows the percentage of intra-institute collaborations (Blue), i.e., researchers from the same institute and single author papers are not included in this analysis, national collaboration (Red), i.e., collaboration between researchers from different institutions but within the same nation as well as international collaboration (Green), i.e., collaboration between researchers from different institutions from two or more nations. From the graph, we can see that intra-institute collaboration has gradually deteriorated over time. In 1960, around 80% of collaborations were intra-institute and it had dropped to 55% by 2010. In contrast, the national collaboration level remained somewhat steady, but with slight oscillations. What is very interesting in the graph is the growth in international collaboration. In 1960, international collaboration was almost 0%, but by the year 2010, it had grown to approximately 30%. We can infer that during the early period, researchers were generally collaborating within their own institute, but researchers now prefer to collaborate with researchers from institutions outside their country. One of the reasons for this growth could be the fact that researchers still continue to publish papers with colleagues from their previous affiliations.

Since the analysis is focused on the number of institutions per paper, Figure 5.6 shows The probability density function of number of institutions in a paper,  $CA$ . We can see that this distribution follows a power-law, i.e., there are very few papers published with many institutions associated with that paper and there are many papers published with one or few institutions associated with the paper. In our dataset, we find that there are about 364,608 papers published from a single institute and 91,554 papers published

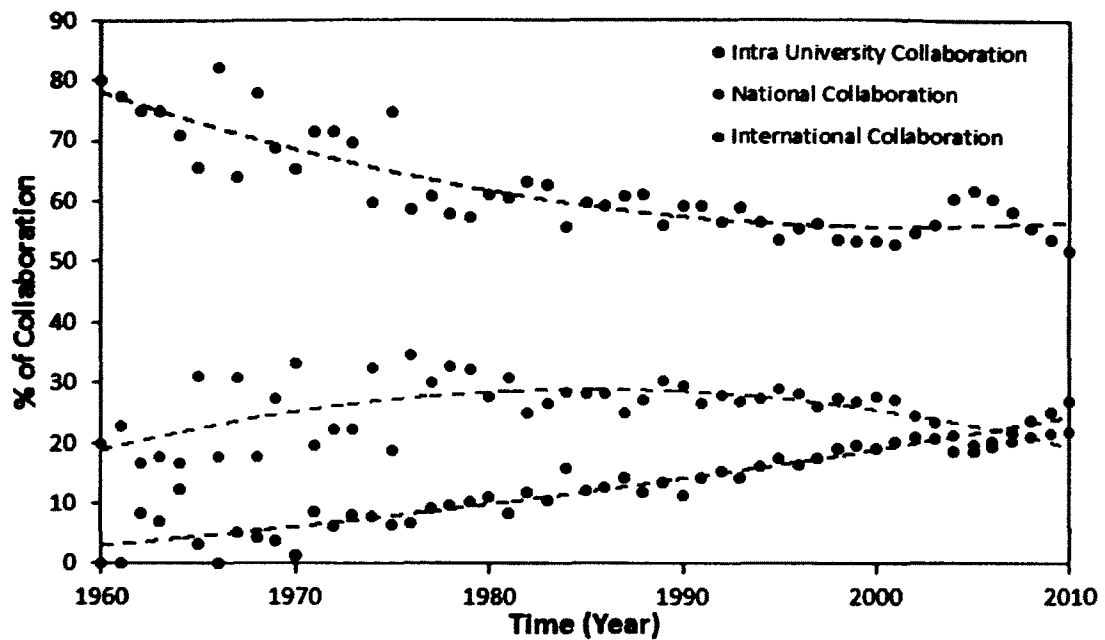


Figure 5.5: A longitudinal analysis of collaboration trend with a best fit function from the year 1960 to 2010. The graph shows the percentage of collaboration of researchers within the same institute (Blue), national collaboration i.e., collaboration between institutions from the same country (Red) and international collaboration i.e., collaboration between institutions from two or more nations (Green).

by two institutions. One of the papers titled “The Grid2003 Production Grid: Principles and Practice,” is associated with authors from 19 different institutions. The grayed portion in the graph indicates the noise in the data for this analysis.

Does the impact (citations received) of a paper increase with an increase in the number of institutions collaborating on a paper? From our previous analysis on NOA, we found that the average citations per paper to the number of co-authors on that paper are positively correlated. Figure 5.7 shows a correlation analysis of the average impact (i.e., average citations per paper) to the number of institutions per paper. The red dashed

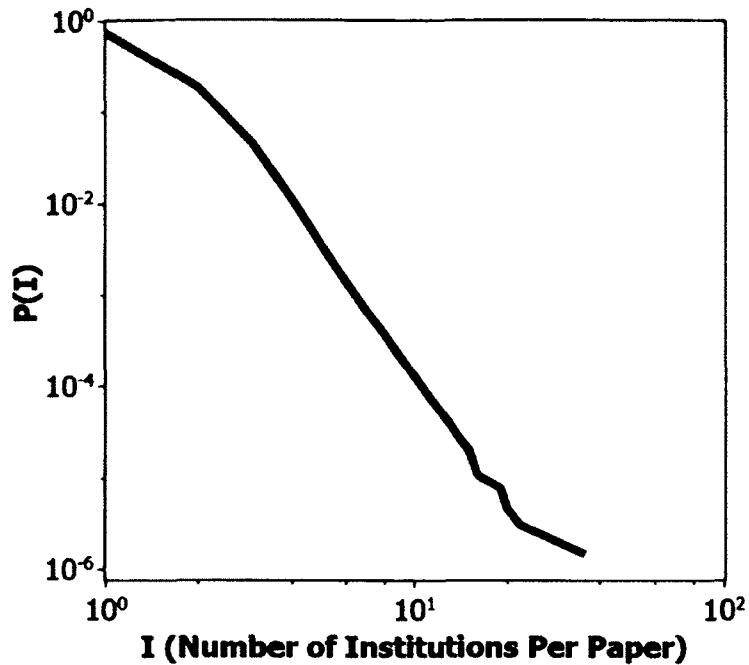


Figure 5.6: The probability density function of number of institutions in a paper, CA, follows a fat tailed distribution. While most papers have fewer institutions, a few have large number of institutions.

line shows the best fit function for the data. We calculated a Pearson's correlation and we found that they have low positive correlation (0.32). Since there are very few publications with more than 15 institutions per paper, we can consider them as noise in the data. So from this analysis we, infer that if a paper has more authors from many different institutions, then it can lead to high citations to the paper, although it may not be the case all the time, as we can see from the graph. Fewer than 1% of the papers in our dataset have 10 or more institutions in the paper and the grayed region in the graph shows this information.



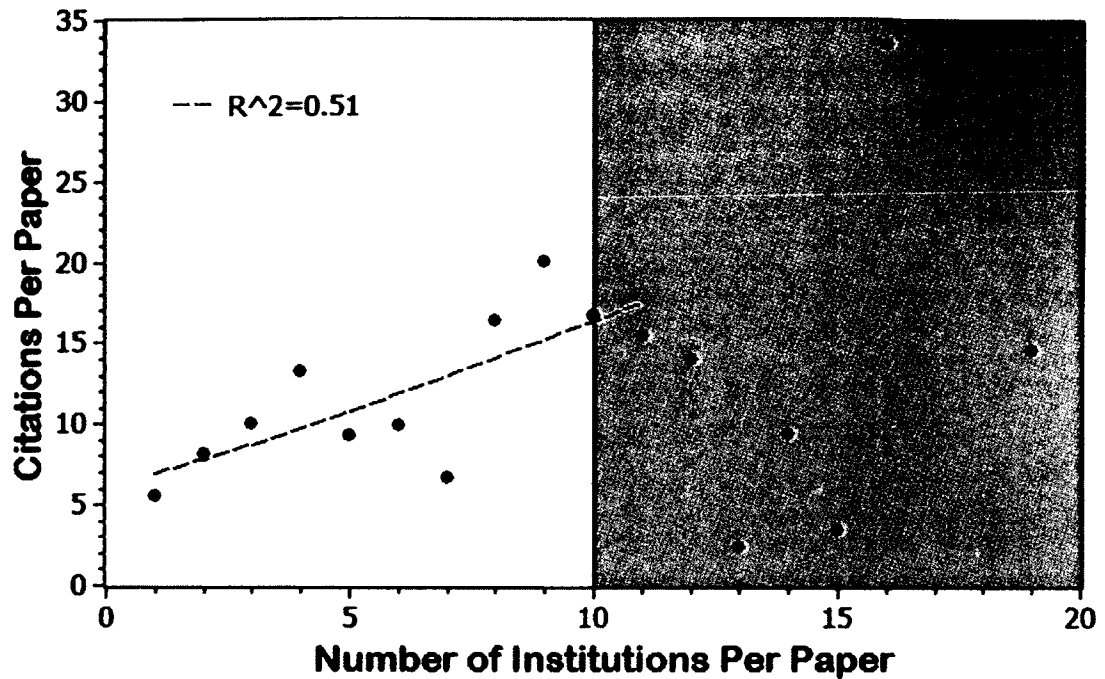
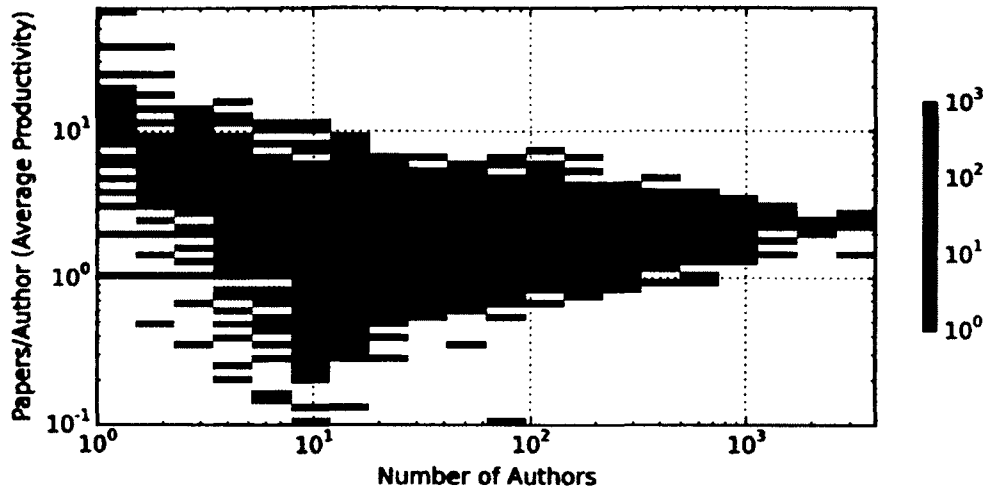


Figure 5.7: A correlation analysis of average impact (i.e., average citations per paper) to number of institutions per paper. The red dashed line indicates the best fit function given by:  $f(x) = ax^3 + bx^2 + cx + d$ , where  $a = 0.008$ ,  $b = -0.26$ ,  $c = 2.78$  and  $d = 3.31$ .

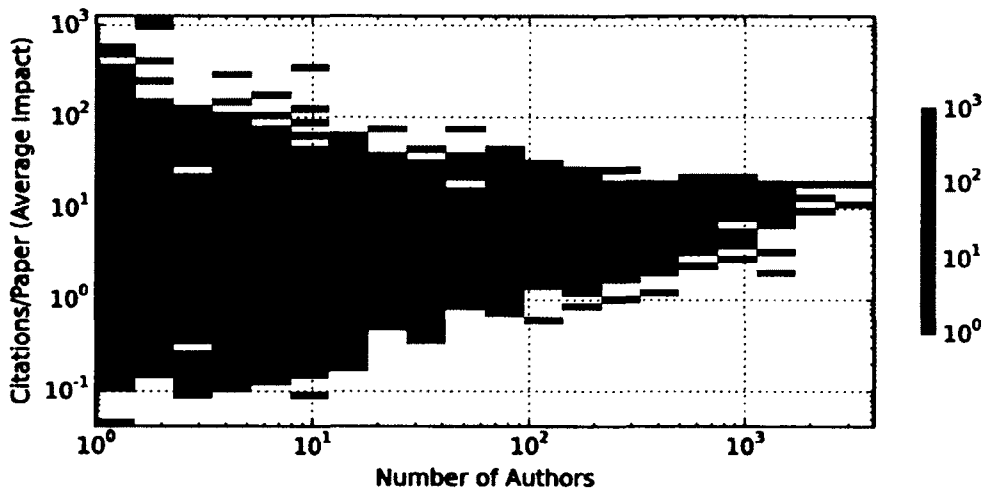
Research performance of a institute can also depend on the size of the institute, i.e., the number of researchers working in that institute. The size of a institute can have a direct impact on the number of papers published and the quality of the researchers' work. It can also lead to further collaboration between researchers from the same institute. To understand this factor, we did a distribution analysis of average productivity (i.e., number of papers published to number of authors from that institute) and average impact (i.e., number of citations received to the number of papers published from that institute) for all the institutions. Figure 5.8 shows a scattered distribution of institutions, with the x-axis representing the size of that institute and the y-axis representing the

average productivity/impact. The color intensity in the distribution indicates the number of institutions that have an average productivity/impact for a given number of authors from that institute. Blue implies very few institutions, red implies a range of 10 to 100 institutions, and green indicates 1000 or more institutions. The Figure 5.8a shows the distribution of average productivity for all institutions with respect to the number of authors associated with that institute. The Pearson's correlation calculated between the two variables tells us that they are not correlated or very low positively correlated (0.07). This tells us that the average productivity of a institute is not directly related to the number of authors from that institute. Just by observing the graph, we can say that there are lot of institutions in the range of 10 to 100 authors with an average productivity of 1 to 3. Figure 5.8b shows the distribution of average impact for all institutions with respect to the number of authors associated with that institute. The Pearson's correlation calculated between the two variables tells us that they are not correlated (0.01). Just by observing the graph, we can say that there are a lot of institutions in the range of 1 to 100 authors with an average impact of 1 to 10. From this analysis we can infer that the size of a institute does not really play any major role in productivity and impact.

Similar to the size of a institute, research performance of a institute can also depend on the number of different research areas (subject diversity) contributed by that institute. Subject diversity of a institute can have a direct impact on the number of papers published and the quality of the researchers work. It can also lead to further collaboration between researchers from the same institute. To understand this factor, we did a distribution analysis of average productivity and average impact for all the institutions with respect to subject diversity of that institute. Figure 5.9 shows a scattered distribution of institutions, with the x-axis representing the subject diversity



(a) Average productivity of a institute considering the number of authors from that institute. The color intensity in the distribution indicates the number of institutions.

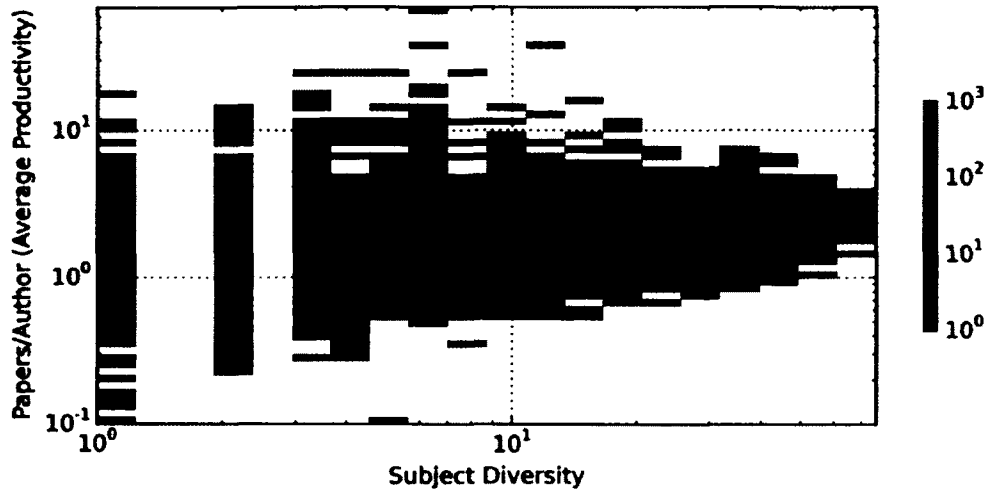


(b) Average impact of a institute considering the number of authors from that institute. The color intensity in the distribution indicates the number of institutions.

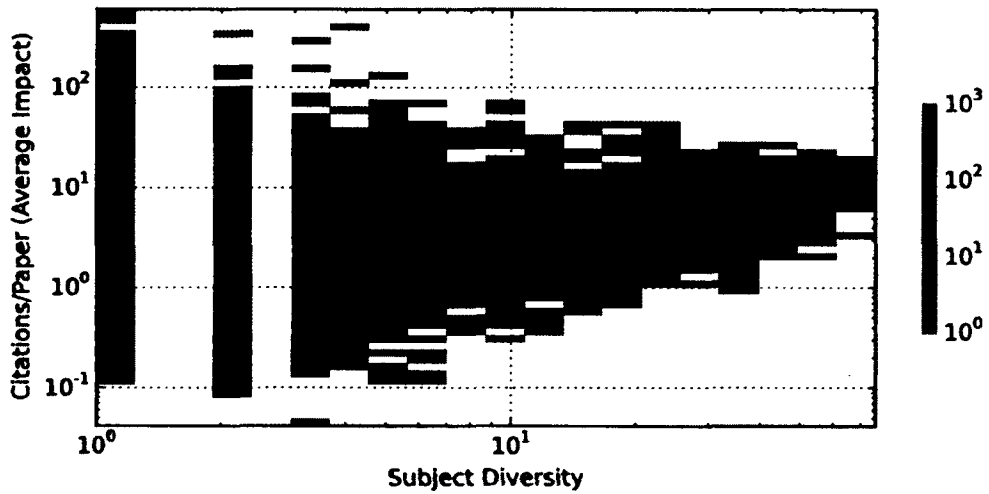
Figure 5.8: Research performance of a institute with respect to its size.

of that institute and the y-axis representing the average productivity/impact. The color intensity in the distribution indicates the number of institutions that have an average productivity/impact for a given subject diversity from that institute. Figure 5.9a shows the distribution of average productivity for all institutions with respect to subject diversity for those institutions. The Pearson's correlation calculated between the two variables tells us that they are not correlated or very low positively correlated (0.2). This tells us that the average productivity of a institute is not directly related to subject diversity of that institute. Just by observing the graph, we can say that there are many institutions with subject diversity in the range of 1 to 20, having an average productivity of 1 to 3. Figure 5.8b shows the distribution of average impact for all institutions with respect to subject diversity for that institute. The Pearson's correlation calculated between the two variables tells us that they are not correlated (-0.01). Just by observing the graph, we can say that there are a lot of institutions with subject diversity in the range of 1 to 20, having an average impact of 1 to 10. From this analysis, we can infer that subject diversity of a institute does not directly play any major role in productivity and impact.

A general assumption would be that the higher the number of authors associated with a institute, the higher the subject diversity for that institute, i.e., they are correlated. Figure 5.10 shows a scattered distribution of institutions, with the x-axis representing the size of that institute and the y-axis representing the subject diversity of that institute. The color intensity in the distribution indicates the number of institutions with a certain number of authors associated with that institute for a given subject diversity. The Pearson's correlation calculated between the two variables tells us that they are positively correlated (0.7). From this analysis we can infer that when the number of



(a) Average productivity of a institute considering the number of different research areas (subject diversity) contributed by that institute. The color intensity in the distribution indicates the number of institutions.



(b) Average impact of a institute considering the number of different research areas contributed by that institute. The color intensity in the distribution indicates the number of institutions.

Figure 5.9: Research performance of a institute with respect to subject diversity.

authors associated with an institute is high, the number of research areas contributed by that institute is also high. Also, we can see from the graph that, after a certain point, the number of research areas stabilizes as there are only a finite number of research areas in Computer Science.

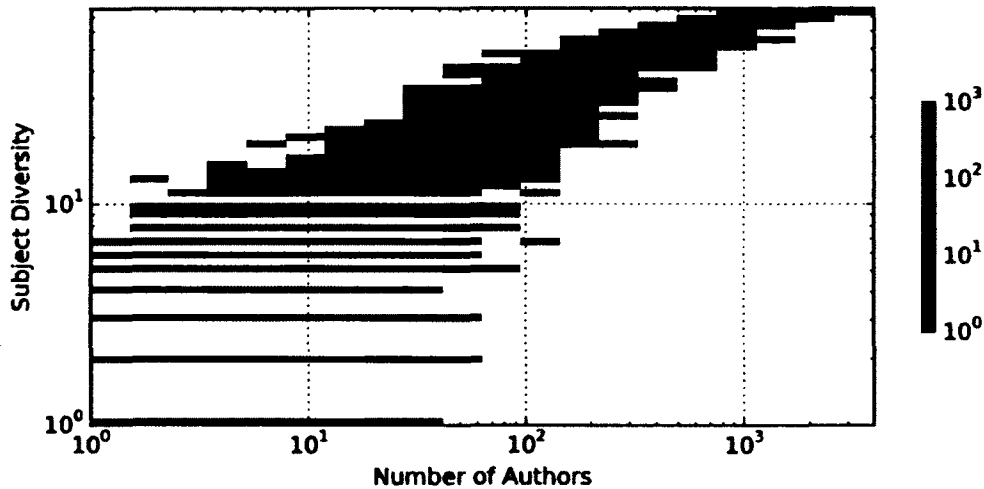


Figure 5.10: A correlation analysis of subject diversity and number of authors associated with that institute. The color intensity in the distribution indicates the number of institutions.

## 5.4 Ranking of Institutions

Similar to the ranking of authors, the performance of institutions can be evaluated based on publications, citations, and the number of grants they have received. Such an evaluation not only helps in identifying top institutions, but also for helping those institutions achieve a reputation within the research community and for governmental research fund allocation. A good reputation attracts federal funding and also attracts highly qualified students around the world which elevates research standards and goals.

Hence, it is important to identify top institutions in the Computer Science community. Since the work is related to collaboration, we ranked based on degree centrality, which in network terms corresponds to the number of distinct collaborations the institute has. Table 5.2 shows the top collaborative institutions in Computer Science according to the ACM dataset. Carnegie Mellon University is clearly the most collaborative with a high number of authors associated with the institute, who have published many papers and they are collectively very well cited.

Table 5.2: Top 10 Institutions according to their level of collaboration (number of distinct collaborations).

<b>Institute Name</b>	<b>degree</b>	<b>Authors</b>	<b>Papers</b>	<b>Citations</b>
Carnegie Mellon University	1,409	3,945	10,958	130,080
University Of California Berkeley	973	2,453	5,669	108,472
Massachusetts Institute of Technology	964	2,809	6,325	113,395
Stanford University	907	2,779	6,602	126,709
University of Maryland	745	1,373	4,049	42,478
University of Waterloo	740	940	3,276	17,583
University of Illinois	727	1,843	4,768	42,478
Georgia Institute of Technology	725	1,521	3,946	40,765
Princeton University	719	1,314	3,760	49,938
Purdue University	710	1,261	3,629	25,368

Other network metrics, such as betweenness centrality and closeness centrality also indicate how important a institute is in the network. Table 5.3 shows the top institutions based on betweenness centrality and closeness centrality. Ranking the institutions based on these network metrics gives us a different perspective on which institutions tend to quickly receive and spread information in the network.

Table 5.3: Top Institutions ranked by Betweenness and Closeness Centrality

<b>Institute Name</b>	<b>betweenness</b>
Carnegie Mellon University	3,457,008
Massachusetts Institute of Technology	1,597,913
University of California Berkeley	1,533,483
Stanford University	1,395,320
University of Waterloo	1,226,568
University of Maryland	1,107,020
Tsinghua University Beijing China	1,031,398
Georgia Institute of Technology	1,003,339
University of Pennsylvania	938,502
Purdue University	919,138

<b>Institute Name</b>	<b>closeness</b>
Carnegie Mellon University	2.00647
Massachusetts Institute of Technology	2.10310
University Of California Berkeley	2.10385
Stanford University	2.11199
University of Waterloo	2.14485
University of Illinois	2.15068
University of Maryland	2.15272
Georgia Institute of Technology	2.15577
Tsinghua University Beijing China	2.16160
Microsoft Research	2.16845

## 5.5 Geographical Distribution

Using visualization tools, we performed a visual analysis to understand the NOI. Figure 5.11 shows a visualization of the NOI superimposed over a world map. Each node in the network represents a university or a research lab. A link between them represents that authors from those institutions have worked together to publish a paper. The size of the node represents the number of immediate neighbours to that node. Using Newman's modularity algorithm [63], we identified communities in the NOI. Several iterations of the algorithm identified eight distinct communities in the network



and the nodes of those communities are represented in different colours. The figure shows that there are three main communities in Europe, which are in dark green, dark blue, and pink. Institutions in the dark blue community seem to be mainly from Italy and Switzerland, whereas institutions in the pink community appear to be largely from Spain. The green community is the largest in Europe and it appears to include Great Britain, France, Germany, and institutions from their neighbouring countries.

Similarly, there are three main communities in North America, which are in red, orange, and blue. In the blue community, most of the institutions seem to be located in the east of America. They also appear to be mainly formed by institutions with low average productivity (in terms of number of publications). Most of the hubs in North America appear to be in the orange community, which appears to contain fewer nodes; however, they appear to be strongly connected to each other. We can infer that institutions in California and the north-eastern states of America tend to collaborate more. Interestingly, institutions from Israel are part of this community. This indicates that they collaborate highly with institutions in America. Institutions in the red community are mainly from Canada and the north-eastern states of America.

This network is very interesting as it gives us very good understanding of the kind of collaboration the institutions from their countries have. Institutions in South America mainly collaborate with institutions in Europe, which probably has to do to with language, as researchers from Brazil and Portugal can communicate very comfortably. Institutions in China and Japan have fewer collaborations with other countries but have strong collaborations within their country. Figure 5.12 shows the network of institutions from the United States. Using Newman's modularity algorithm,

we identified communities in the NOI for the United States. The algorithm identified seven distinct communities in the network and the nodes of those communities are represented in different colours. From the figure, we can see that the purple community appears to be the strongest or well connected (in terms of number of connections between the nodes within the community). The institutions of this community appear to be mainly from Boston, New York, Seattle, San Francisco and San Diego. The biggest node (yellow) in the network is Carnegie Mellon University and it clearly appears to be the hub of this network. It also appears to be in a different community that has fewer hubs when compared to the purple community. Some of the other institutions in this community are: University of Texas, Austin and University of Illinois, Chicago. Table 5.4 shows the top institutions by country. Ranking the institutions by country gives us a different perspective on which institutions are the leading source of knowledge in their respective countries.

Table 5.4: Top institutions by Country

<b>Institute Name</b>	<b>Country</b>	<b>Publications</b>	<b>Citations</b>
Carnegie Mellon University	United States	7,810	117,482
University of Waterloo	Canada	2,549	16,045
IBM Software Group based in Beijing	China	2,433	9,377
University of Amsterdam	Netherlands	2,122	19,794
University of Tokyo	Japan	1,617	7,926
National University of Singapore	Singapore	1,608	9,958
Swiss Federal Institute of Technology	Switzerland	1,401	13,585
Israel Institute of Technology Haifa	Israel	1,356	15,781
University of Edinburgh	Great Britain	1,300	14,894
Technische Universitt Berlin	Germany	1,178	5,406

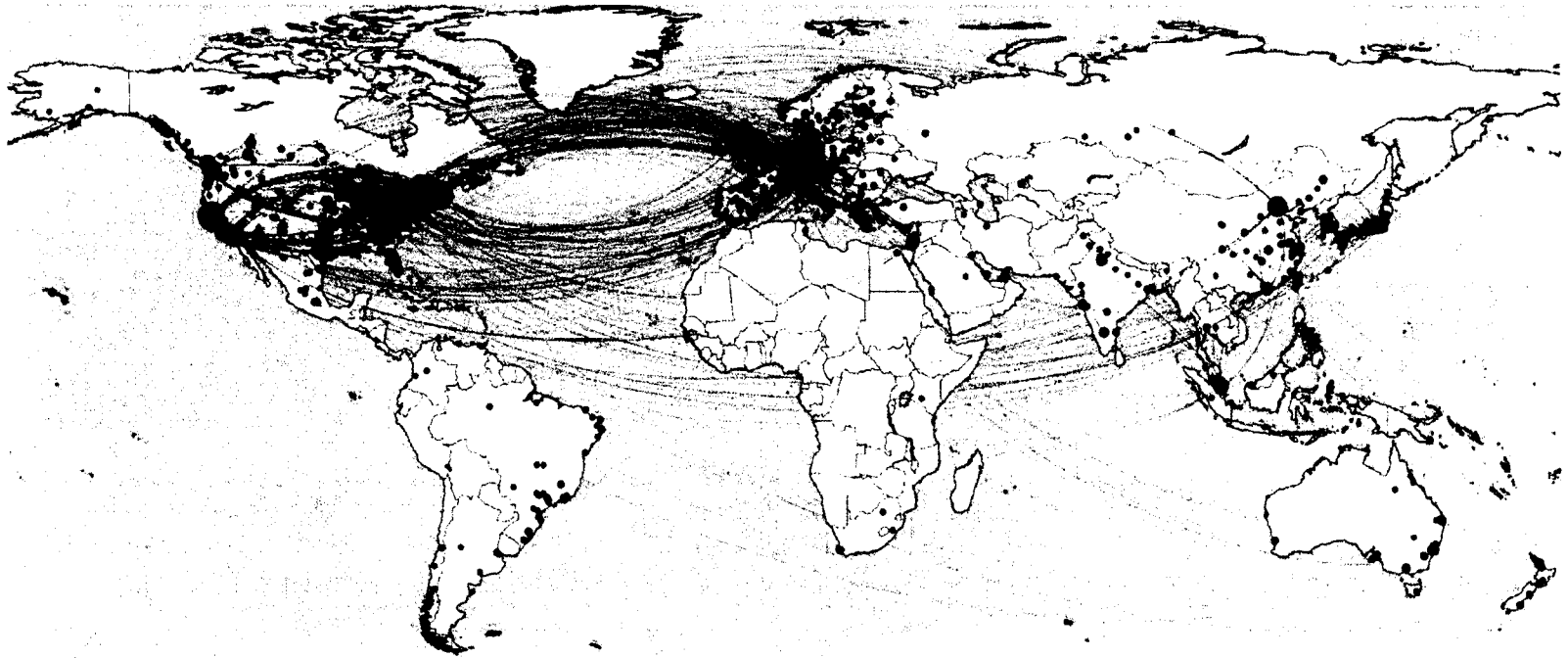


Figure 5.11: Visualization of the network of institutions superimposed over a world map.

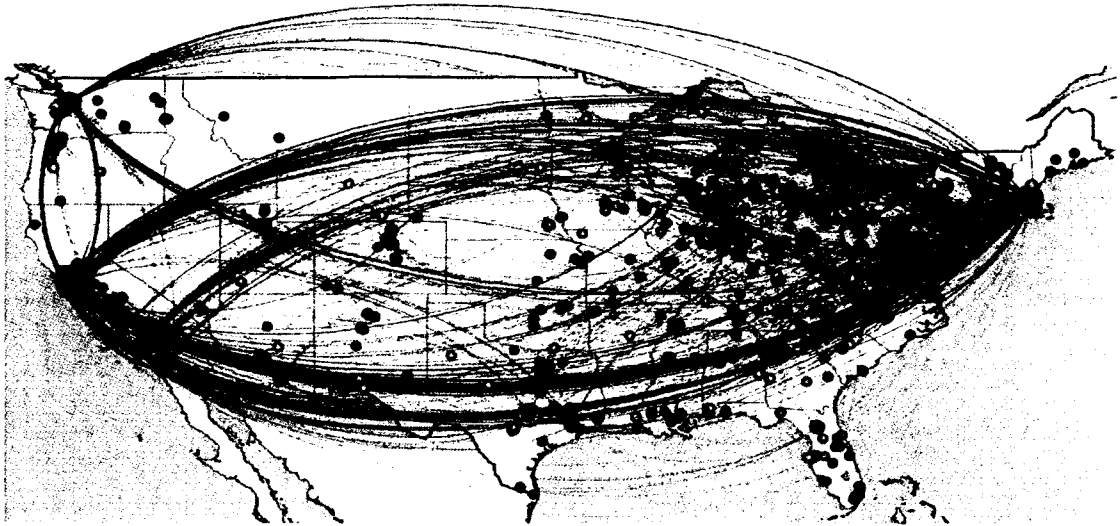


Figure 5.12: Visualization of the network of institutions superimposed over United States map.

## 5.6 Conclusion

In this chapter, we investigated the network of institutions. We analyzed the characteristics of the NOI and compared it with network of authors. We found that NOI follows power-law characteristics with  $\lambda = 1.22$  and an average path length  $\ell = 3.41$ , i.e., any institute in the community can be reached with fewer than four connections between them.

We also investigated the bibliographic properties and performed a longitudinal analysis to understand the publication trends and collaboration patterns of institutions. Since 1960, there has been a gradual drop in single institute papers and a steady increase in the average number of institutions per paper, an indication that institutions are now collaborating more than before. Similarly, we observed a growth in the trend for international collaboration. We found that the size (number of authors) and

subject diversity of an institute does not really play any major role towards the average productivity and average impact of that institute, i.e., they are not correlated or have a very low positive correlation.

Finally, we ranked the top institutions in terms of the number of distinct collaborators and other network metrics in the field of Computer Science. Carnegie Mellon University clearly appears to be one of the major actors in this network. Visualization techniques were used to understand this network; it gave us a better picture of where the hubs are located geographically and the communities in this network.

# Chapter 6

## Network of Countries

### 6.1 Introduction

In previous chapters, we learned about collaboration patterns and trends of authors and institutions in Computer Science. We showed how publications patterns vary between authors and between institutions. Similarly, publication practices differ widely within and between countries. Beaver and Rosen [24] noted collaboration across national borders as early as the nineteenth century. They found that international collaborations had increased towards the end of the century and has grown to importance in the present century. The most multi-authored scientific paper was published in Physics Letters B in 2010. This paper had 3,222 authors from 32 different countries, who contributed to a study of “charged-particle multiplicities” performed in the Large Hadron Collider at CERN [88]. Investigating the collaboration of countries helps to understand the scientific impact of the countries.

In this chapter, we investigate the network of countries (NOC). We analyze the characteristics of a collaboration network, and compare with the network of authors and network of institutions. We investigate bibliographic properties and do a longitudinal analysis on publication trends and collaboration patterns of countries. Finally, we rank the top countries in the field of Computer Science based on network metrics.

## 6.2 General Network Characteristics

Using the authors' affiliation, we were able to generate a country-paper bi-partite network for our analysis. In a country-paper network, there are two sets of nodes (countries and papers) with the links running from countries to papers. A paper is associated with a country if at least one of the authors from that country has authored that paper. Since the study is focused on understanding the collaboration between countries, we concentrate only on the country projection. In NOC, every node represents a country and two countries are connected if authors from those countries have co-authored a paper.

Once the network was constructed, we performed an analysis to identify the kind of network we are dealing with. The characteristics of the NOC is as shown in the Table 6.1 and is compared with NOA and NOI. The network statistics for the NOC shows that there are about 143 countries ( $n$ ) in the community, which had about 1,562 connections ( $m$ ) between them. The average number of collaborators per country is  $z = 21.84$ , which tells that, on an average, every country collaborates with 21 other countries. The average path length  $\ell = 1.95$ , tells that any country can reach another country with fewer than two connections between them. The NOC demonstrates a high

average clustering coefficient ( $C$ ) = 0.84, meaning this network has a high number of collaborations that form triads. Compared to NOA and NOI, the average clustering coefficient of NOC is comparatively high. The clustering coefficient of this network indicates that the network is organized in groups of highly collaborative countries with a few connections outside of the group.

Table 6.1: Network Statistics: A comparison between the Network of Countries, Network of Institutions, and Network of Authors.

<b>Measure</b>	<b>NOC</b>	<b>NOI</b>	<b>NOA</b>
Nodes ( $n$ )	143	12,541	237,351
Links ( $m$ )	1,562	94,817	1,065,078
Mean Degree ( $z$ )	21.84	15.12	8.97
Exponent Power Law ( $\lambda$ )	0.87	1.22	2.37
Average Clustering Coeff. ( $C$ )	0.84	0.51	0.68
Average Path Length ( $\ell$ )	1.95	3.41	5.28

Figure 6.1 depicts the probability density function of degree of a country,  $D$ , follows a fat tailed distribution, indicating a significant heterogeneity.  $P(D)$  is the probability that a randomly selected node in a network has degree  $D$ . While most countries have fewer connections, a few have a large number of connections. Based on the number of distinct collaborators, some of the top countries are United States (121 collaborators), Great Britain (88 collaborators), Germany (85 collaborators), and Canada (84 collaborators).



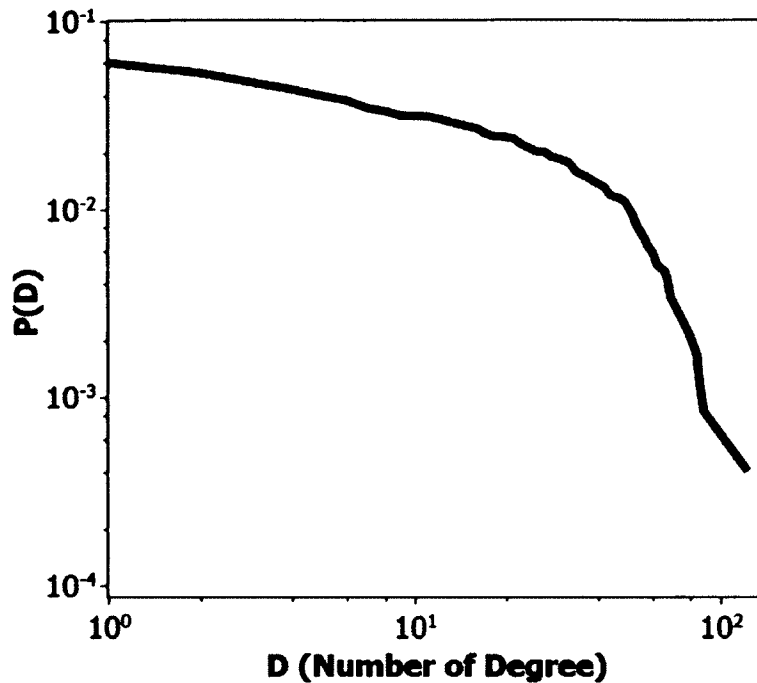


Figure 6.1: The probability density function of degree of a country,  $D$ , follows a fat tailed distribution, indicating a significant heterogeneity. While most countries have fewer connections, a few have a large number of connections.

### 6.3 Collaboration Pattern

Using the publication year and authors' affiliation information available for each paper, we investigated the publication trends of countries in the field of Computer Science. Figure 6.2 shows the percentage of papers published by authors from one country (Blue), two countries (Red) and three or more countries (Green) from the year 1960 to 2010. Comparing this graph with Figure 4.5 for NOA and Figure 5.3 for NOI, we see some similarities in trends. The graph shows that in the early period, there was a general tendency for researchers to publish single author papers or publish with researchers from institutions of the same country. Almost, 100% of papers published during the

1960s were one country papers. However, since then the trend of one country papers has been declining and by 2010 it had dropped to 75%. This further concludes that back then, researchers were less collaborative and if they collaborated, it was with researchers from the same institutions or researchers from other institutions of the same nation. However, with time Computer Science researchers have been increasingly collaborative. The graph shows that the trend of two countries per paper has been growing somewhat steadily over time. Fewer than 1% of papers published in the 1960 were two country papers, but this has increased to 20% by 2010. In contrast, the trend of three or more countries per paper still remains low and around 2010 these were only 3% of papers published. This indicates that majority of papers are still published by researchers from the same nation; however recently there has been a growth in two countries per paper. To further understand this publication trend, we investigated the average number of countries per paper with respect to time.

For each year, we observed the trend of collaboration level per paper, i.e., the average number of countries per paper published in that year. In Figure 6.3, the collaboration level for each year is indicated by a blue bubble with an error bar representing the variability of data. The red dashed line indicates the best fit function for the data. We see that the average number of countries per paper has increased gradually over time. The average number of countries per paper in 2010 was 2.2 countries, while the average in 1960 was approximately 1.2 countries per paper. Hence, we can conclude that there has been a marginal growth in the Computer Science discipline in terms of average number of countries per paper. This is mainly due to the growth of papers being published with two and three or more countries per paper. Another reason for this increase in trend could be the growth in the number of active authors in the Computer Science

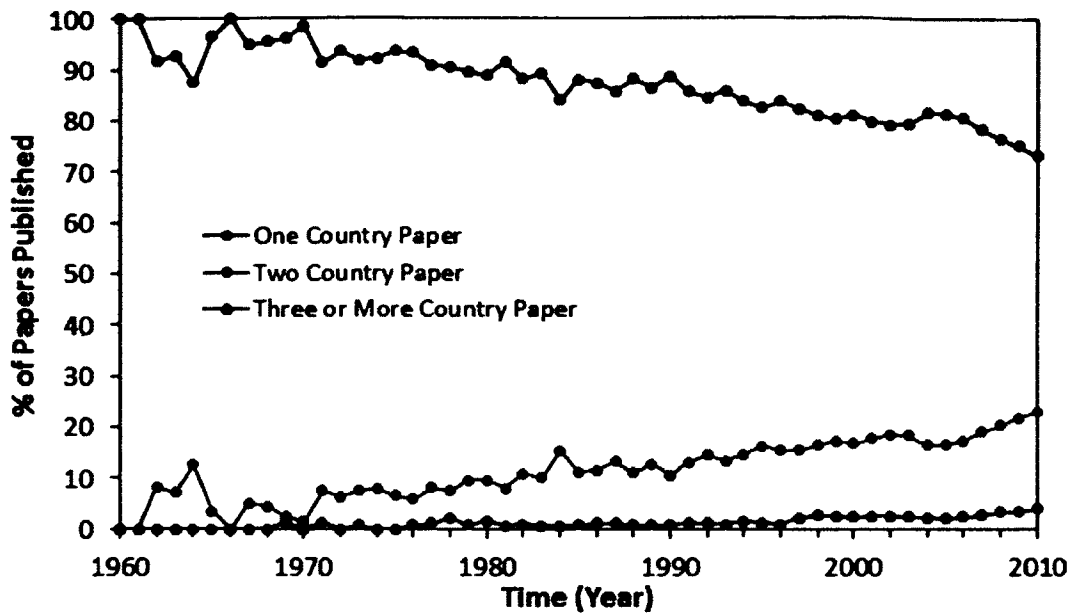


Figure 6.2: A longitudinal analysis of the publication trend from 1960 to 2010. The graph shows the percentage of papers published by researchers from one country (Blue), two countries (Red), and three or more countries (Green).

community, as this could lead to a high number of collaborations between the countries. However, it is very hard to infer anything about internationalization from this analysis. To understand the collaboration pattern, we investigated the collaboration of countries by considering their geographical aspect.

We know that collaboration between researchers improves their productivity and knowledge [16]. Also, institutions act as the source of new knowledge for any nation. The collaboration of authors and institutions indirectly improves the productivity and impact of the nation from where these researchers come. This is quite important in the Computer Science community as it attracts high profile researchers and new and aspiring Ph.D. students. In our previous analysis, we analyzed the percentage of publications

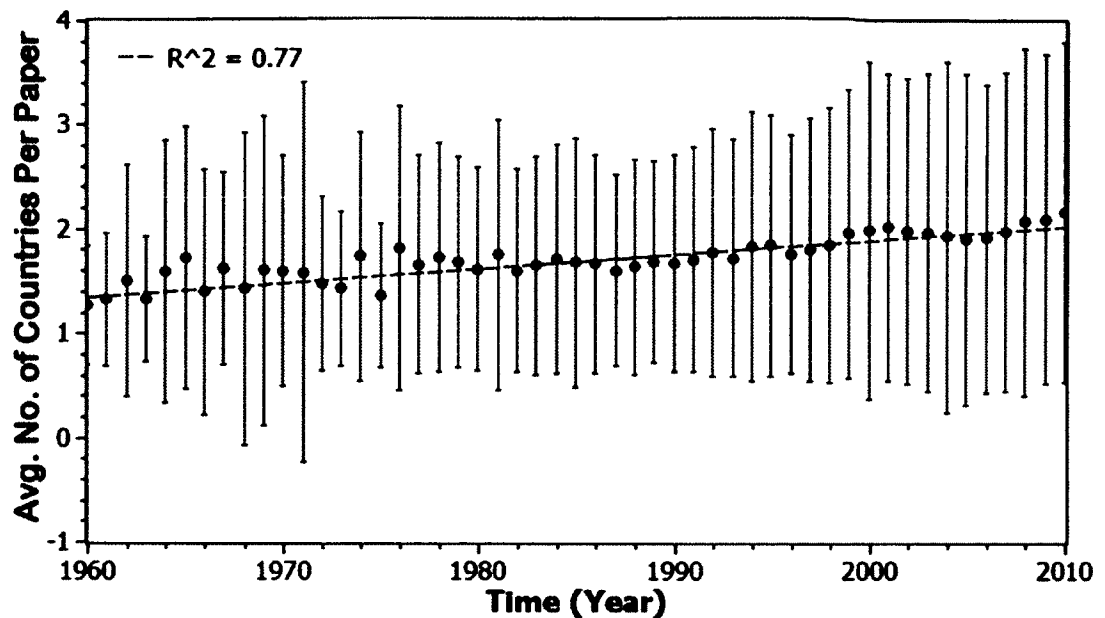


Figure 6.3: A longitudinal analysis of collaboration level from 1960 to 2010. The graph shows the average number of countries per paper with an error bar for every year. The red dashed line indicates the best fit function given by:  $f(x) = ax + b$ , where  $a = 0.013$  and  $b = -25.03$ .

by one, two, and three or more countries per paper. However, it does not tell you if these collaborations are national or international. Also, the one country paper does not distinguish if they are single author or multiple authors from the same country. It would be interesting to see what percentage of collaborations happens with researchers from other institutions but within the same nation as opposed to the percentage of collaborations with researchers from institutions outside the nation. Figure 6.4 shows a longitudinal analysis of the collaboration trend from 1960 to 2010. The graph shows the percentage of national collaboration (Blue), i.e., collaboration between researchers from the same or other institutions but within the same nation as well as international collaboration (Red), i.e., collaboration between researchers from other institutions from

two or more nations. From the graph we can see that national collaboration has gradually decreased over time. In 1960, almost 100% of collaborations were with researchers from the same nation and this had dropped to 75% by 2010. What is very interesting in the graph is the growth in international collaboration. In 1960, international collaboration was almost 0%, but by the year 2010, it had grown to approximately 25%. We can infer that during the early period, researchers were generally collaborating with researchers from the same nation, but researchers now prefer to collaborate with researchers from institutions outside their country.

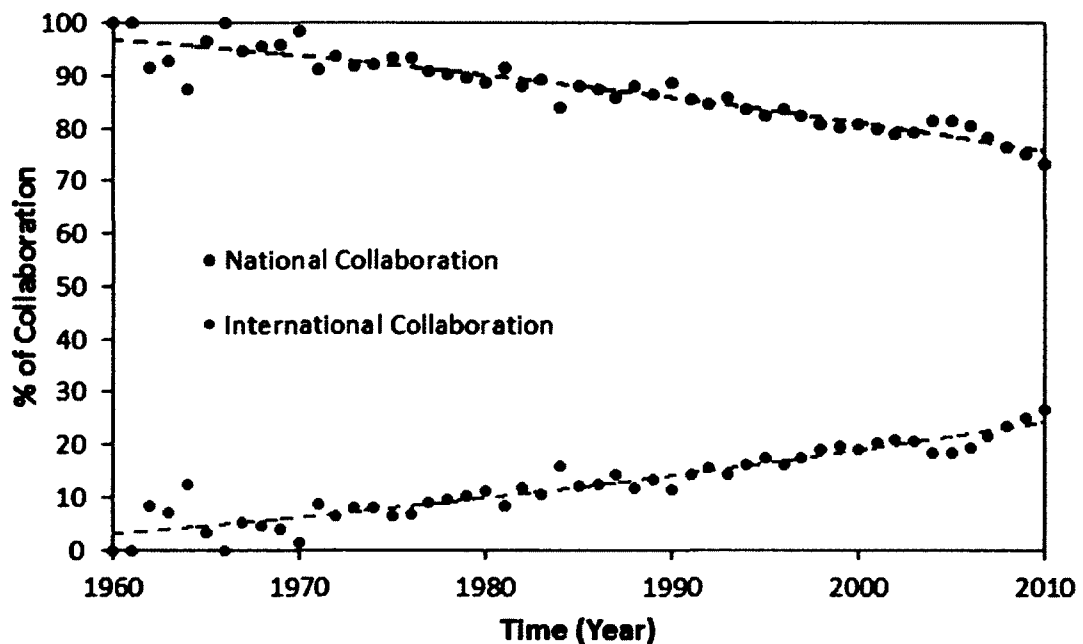


Figure 6.4: A longitudinal analysis of the collaboration trend with a best fit function from 1960 to 2010. The graph shows the percentage of the collaboration of researchers from the same country (Blue), i.e., national collaboration and collaboration between researchers from two or more nations (Red), i.e., international collaboration.

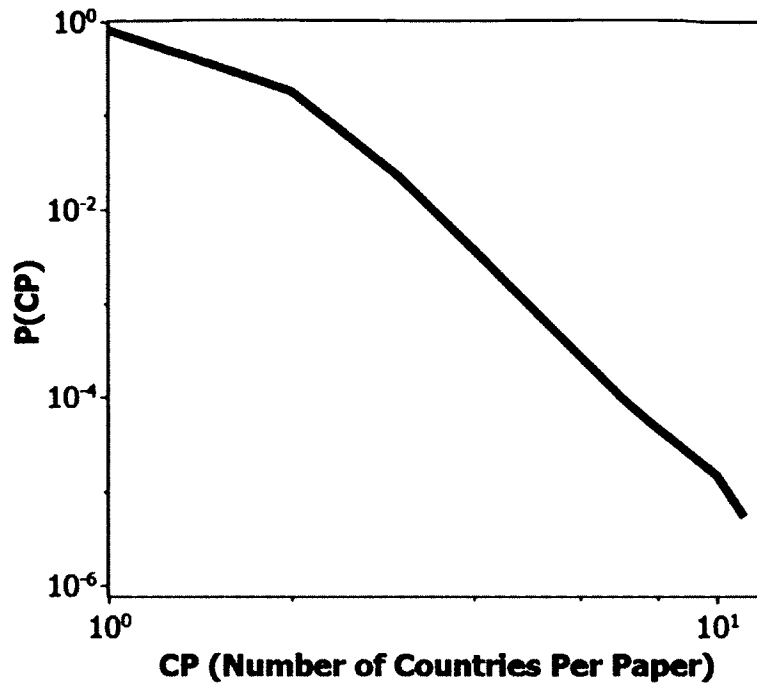


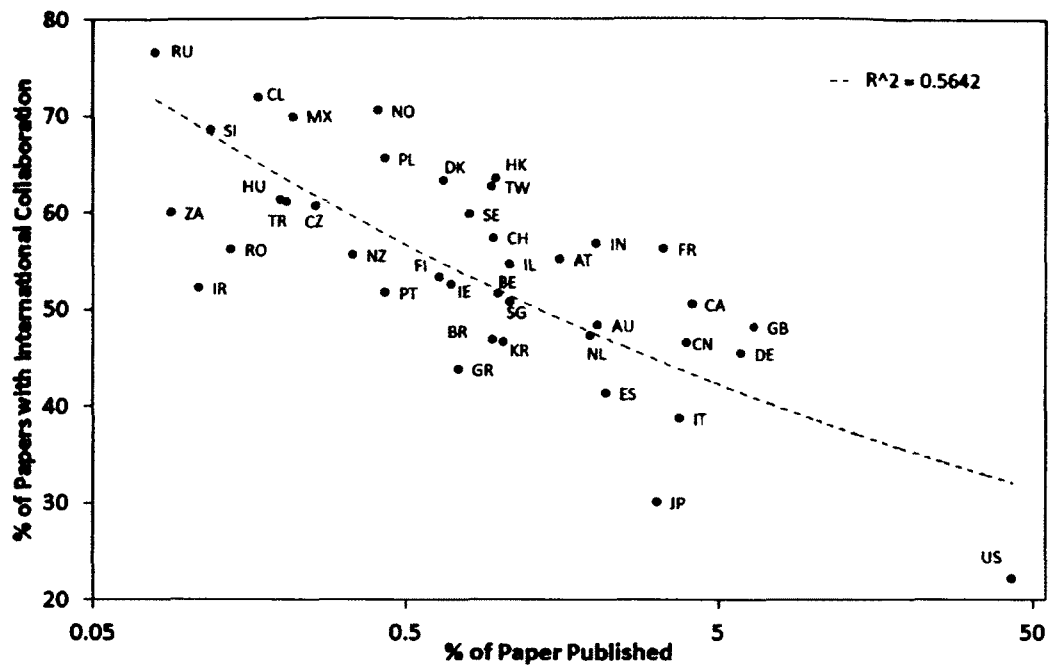
Figure 6.5: The probability density function of number of countries in a paper, *CA*, follows a fat tailed distribution. While most papers have fewer countries, a few have large number of countries.

The analysis is focused on the number of countries per paper. Figure 6.5 shows the probability density function of number of countries in a paper, *CA*. We can see that this distribution follows a power-law, i.e., there are very few papers published with many countries associated with that paper and there are many papers published with one or few countries associated with the paper. In our dataset, we found that there are about 212,549 papers published as a single country paper and 51,890 papers published as a two country paper. One of the papers titled “The Tomato Sequencing Project, the first cornerstone of the International Solanaceae Project (SOL),” is associated with authors from 11 different countries. Fewer than 1% of the papers in our dataset have 6 or more

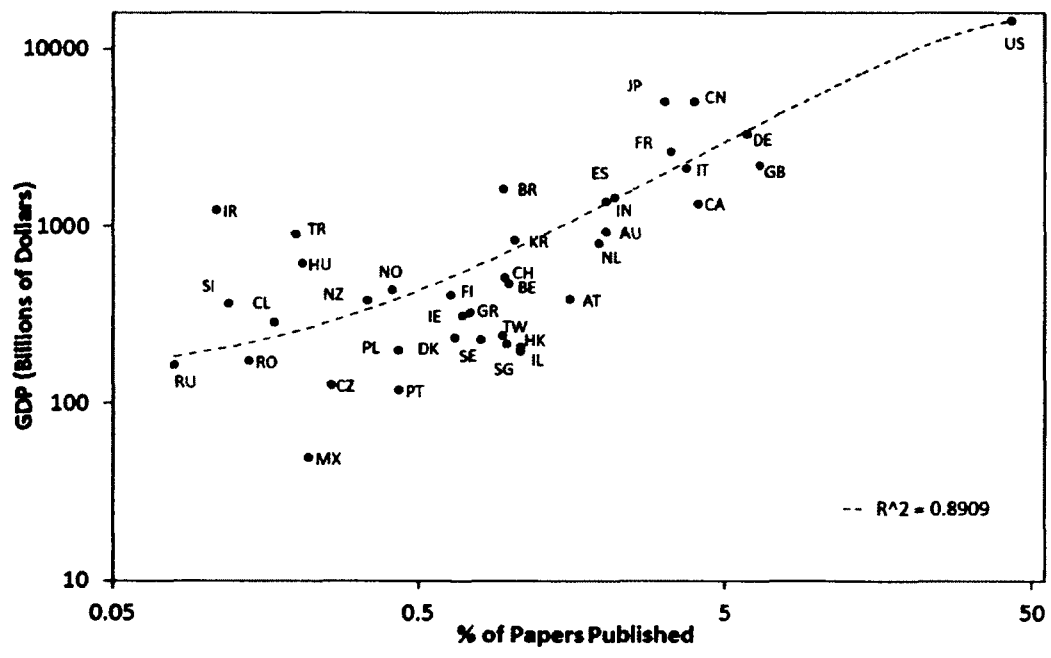
countries in the paper and the grayed region in the graph shows this information.

Figure 6.6a shows a scatter distribution of the scientific size of the country to the percentage of international collaborations for forty countries (See Appendix C for details). We define scientific size of a country as the percentage of papers published by that country to all the papers published. This measure tells the percentage of scientific production of a country in Computer Science. The scientific size of a country seems to be inversely related to the rate of international collaboration: with a decreasing volume of publications, the share of international collaborations grows. Nevertheless, there is a large scatter in the percentage of international collaborations, especially for the countries with small scientific output. The correlation analysis for the two variables shows that they are negatively correlated (0.7).

We think that there are many reasons for this relationship between scientific size of a country and the rate of international collaboration. A tendency towards specialization in Computer Science makes the research areas more narrowly focused. Researchers who come from scientifically peripheral countries are likely to find few, if any, collaborators in their own country. Hence, they have to look for collaborators from other countries. Another reason for their higher rate of international collaboration might be the need of cost sharing on the research. The GDP of a country also is a major factor in the scientific productivity of the country. Figure 6.6b shows a scatter distribution of the scientific size of the country to the GDP for forty countries.



(a) Scientific size of the country and its international collaboration. The red dashed line indicates the best fit function given by:  $f(x) = ax^{-b}$ , where  $a = 51.927$  and  $b = -0.128$ .



(b) Scientific size of the country and its GDP. The red dashed line indicates the best fit function given by:  $f(x) = ax^2 + bx + c$ , where  $a = -26$ ,  $b = 4 \times 10^7$  and  $c = 1 \times 10^{11}$ .

Figure 6.6  
115



## 6.4 Ranking of Countries

Similar to ranking of authors and institutions, we evaluated the productivity and performance of countries. A high reputation attracts funding for institutions and it also attracts highly qualified researchers and prospective students around the world, which further elevates the research standards of the nation. Hence, it is important to identify the research contributions and list the top countries in the Computer Science community. Since the work is related to collaboration, we ranked based on degree centrality, which in network terms corresponds to the number of distinct collaborations the country has. Table 6.2 shows the top collaborative countries in Computer Science according to the ACM dataset. For every county, we show their distinct collaborators (degree), total collaborations (weighted degree), percentage of global collaborations, main collaborator with the number of papers published between them, number of authors, number of publications, percentage of publication, and number of citations the country has received. The United States is clearly the most collaborative with a large number of authors associated with the country. These authors have published many papers and they are collectively very well cited. Canada is one of the main collaborators with the United States, with 4,285 collaborations between them. The United States also happens to be the main collaborator for many countries. Hong Kong and Ireland are the only exceptions, as their main collaborators are China and Great Britain respectively.

Other network metrics, such as betweenness centrality and closeness centrality, also indicate how important a country is in the network. Table 6.3 shows the top countries based on betweenness centrality and closeness centrality. Ranking the countries based

Table 6.2: Productivity and collaboration patterns of top 30 countries.

Country	degree	Wt. degree (W)	%W	Main Collaborator	Authors	Publications(P)	%P	Citations
United States	121	37,828	23.35	Canada (4,285)	67,018	668,982	42.93	5,652,424
Great Britain	88	13,360	8.24	United States (3,658)	7,834	101,513	6.51	678,080
Germany	85	12,469	7.69	United States (3,563)	8,640	92,216	5.91	433,632
Canada	84	10,437	6.44	United States (4,285)	5,810	64,332	4.12	392,836
France	80	6,888	4.25	United States (1,727)	3,884	52,289	3.35	246,912
Italy	76	8,387	5.17	United States (2,420)	4,482	58,702	3.76	252,509
Netherlands	72	5,092	3.14	United States (1,100)	2,479	30,492	1.95	158,452
Australia	69	4,224	2.60	United States (1,046)	2,122	32,151	2.06	145,613
China	67	7,563	4.67	United States (2,725)	5,518	61,767	3.96	183,986
Spain	67	4,838	2.98	United States (1,203)	2,684	34,173	2.19	117,977
Japan	66	3,836	2.36	United States (1,336)	4,425	49,805	3.19	137,179
India	62	2,979	1.83	United States (1,491)	2,462	32,034	2.05	274,378
Belgium	60	2,351	1.45	United States (536)	1,226	15,563	0.99	75,869
Sweden	60	1,793	1.10	United States (347)	975	12,495	0.80	63,276
Denmark	58	1,792	1.10	United States (396)	798	10,345	0.66	72,452
Switzerland	57	3,927	2.42	United States (1,098)	1,742	15,044	0.96	96,566
Austria	56	2,460	1.51	United States (553)	1,672	24,374	1.56	16,3794
Israel	54	3,607	2.22	United States (2,085)	1,704	16,982	1.08	122,413
Finland	54	1,474	0.91	United States (441)	777	10,035	0.64	46,698
Hong Kong	53	4,064	2.50	China (1,266)	1,044	15,190	0.97	62,839
Singapore	52	2,819	1.74	United States (947)	1,258	16,840	1.08	70,453
South Korea	52	1,607	0.99	United States (828)	1,204	16,097	1.03	63,075
Brazil	51	1,782	1.10	United States (575)	1,524	14,906	0.95	51,994
Greece	49	1,203	0.74	United States (542)	1,034	11,575	0.74	46,171
Norway	49	1,940	1.19	United States (267)	521	6,529	0.41	28,705
Poland	49	942	0.58	United States (201)	466	6,745	0.43	21,752
Ireland	47	1,464	0.90	Great Britain (234)	735	10,987	0.70	52,614
Russia	44	387	0.23	United States (105)	190	1,364	0.08	3,266
Portugal	43	993	0.61	United States (197)	587	6,733	0.43	19,275
Czech Republic	42	667	0.41	United States (201)	341	4,055	0.26	16,363

Table 6.3: Top countries ranked by Betweenness and Closeness Centrality

Country	betweenness	Country	closeness
United States	3,125	United States	1.14788
Germany	850	Great Britain	1.38028
Great Britain	719	Germany	1.40148
France	543	Canada	1.40843
Canada	533	France	1.43662
Spain	525	Italy	1.46478
Italy	455	Netherlands	1.49295
Netherlands	386	Spain	1.52816
Japan	277	China	1.52816
India	198	Japan	1.53521

on these network metrics gives us a different perspective on which countries tend to quickly receive and spread information in the network.

## 6.5 Geographical Distribution

Using visualization tools, we performed a visual analysis to understand the NOC. Figure 6.7 shows a NOC superimposed over a world map. Each node in the network represents a country (placed on the center of each country) and a link between them represents that authors from these countries have worked together to publish a paper. The size (and color) of the node represents, the number of distinct collaborators and the link thickness represents their collaboration strength. Clearly, the United States is the hub in this network and is connected to most of the countries. There are a few strong connections from countries in South America to the United States, but there are many connections from South America to the countries in Europe and one of the reasons for this could be a common language; researchers can communicate very comfortably between them.

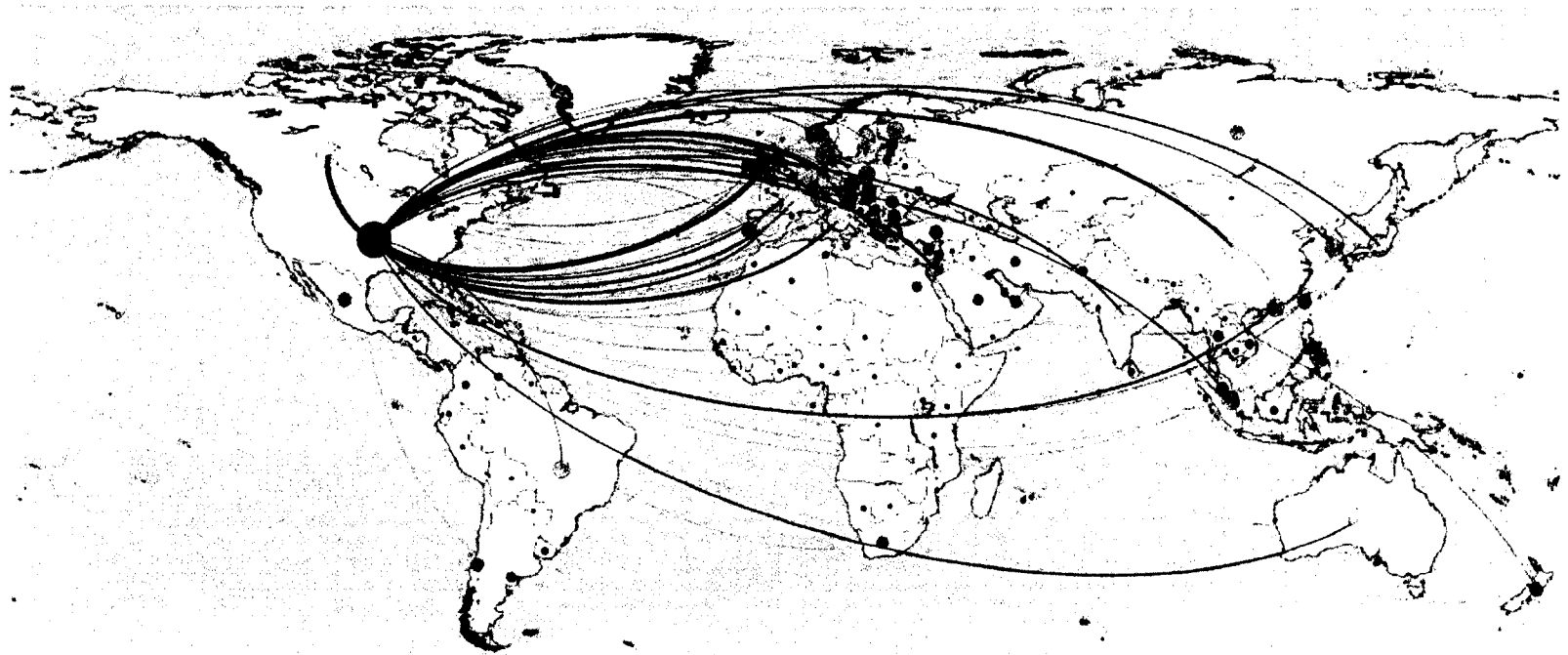


Figure 6.7: Visualization of the network of countries superimposed over a world map.

## 6.6 Conclusion

In this chapter, we investigated the network of countries. We analyzed the characteristics of the NOC and compared it with the network of authors and network of institutions. NOC has an average path length  $\ell = 1.95$ , i.e., any country in the community can be reached with fewer than two connections between them. We also investigated the bibliographic properties and performed a longitudinal analysis to understand the publication trends and collaboration patterns of countries. Since 1960, there has been a gradual drop in single country papers and a steady increase in the average number of countries per paper, an indication that countries are now collaborating more than before. Similarly, we observed a growth in the trend for international collaboration. We found that the scientific size of a country is inversely related to the rate of international collaboration but correlated with the GDP of the nation.

Finally, we rank the top countries in terms of the number of distinct collaborators and other network metrics in the field of Computer Science. The United States clearly appears to be one of the major actors in this network. The visualization techniques were used to understand this network. They gave a better picture of where the hubs are located geographically.

# Chapter 7

## Conclusions

In this dissertation, we investigated and analyzed collaboration patterns at several levels (i.e., authors, institutions, and countries) in the field of Computer Science from a dataset extracted from Association of Computing and Machinery (ACM) Digital Library. We discussed the characteristics of these networks and presented our analysis on the role geography plays towards collaborations. We investigated collaboration patterns and publication trends based on various geographical factors, such as distance and location. We ranked authors, institutions and countries to list the top collaborators and also ranked them based on other network metrics.

In our micro-level analysis focused on network of authors (NOA) we found that, NOA follows power-law characteristics with  $\lambda = 2.576$  and since  $2 \leq \lambda \leq 3$ , it can be categorized as a scale-free network. We investigated publication trends and collaboration patterns of authors. We found that since 1960, there has been a gradual drop in single author papers and a steady increase in three or more authors per paper. We observed that since 1996, there has been a decrease of nearly 40% in the average

collaboration distance, which leads us to argue that collaborations are becoming more local with time. Researchers in Computer Science are very productive during the 15 to 30 year period of their active careers and most researchers are associated with one, two, or three affiliations. We have ranked the authors based on the number of collaborations they have and other network metrics. Future work includes understanding the mobility patterns of authors based on their affiliation. It would also be very interesting to perform and compare the analysis at a national level, i.e., the network of authors in the United States.

Our meso-level analysis on collaboration networks is focused on a network of institutions (NOI). We analyzed the characteristics of the NOI and compared it with NOA. We found that NOI follows power-law characteristics with  $\lambda = 1.22$  and an average path length  $\ell = 3.41$ . We performed a longitudinal analysis to understand the publication trends and collaboration patterns of institutions. Since 1960, there has been a gradual drop in single institute papers and a steady increase in the average number of institutions per paper, an indication that institutions are now collaborating more than before. We also observed, a growth in the trend for international collaboration. We found that the size (number of authors) and subject diversity of a institute does not really play any major role towards the average productivity and average impact of that institute. Future work includes understanding if physical distances between institutions play any role towards collaboration. It would also be very interesting to perform and compare the analysis at a national level, i.e., network of institutions in the United States.

Our macro-level analysis are focused on a network of countries (NOC). NOC has an average path length  $\ell = 1.95$ , i.e., any country in the community can be reached

with fewer than two connections between them. We investigated publication trends and collaboration patterns of authors. Since 1960, there has been a gradual drop in single country papers and a steady increase in the average number of countries per paper, an indication that countries are now collaborating more than before. Similarly, we observed a growth in the trend for international collaboration. We found that the scientific size of a country is inversely related to the rate of international collaboration but correlated with the GDP of the nation. we rank the top countries in terms of the number of distinct collaborators and other network metrics in the field of Computer Science. The United States clearly appears to be one of the major actors in this network. Future work includes investigating the scientific and technological competitiveness of nations. It would also be interesting to investigate the knowledge flow between developed and developing countries.

This is a fascinating subject which deserves more attention. We plan to extend our dataset to include works from the IEEE Computer Society. Since our analysis is done on publications in Computer Science prior to 2011. It would be interesting to re-perform these analysis and verify our findings in a few years time.



# Appendix A

## ACM Computing Classification System

The ACM Computing Classification System is a subject classification system for computer science devised by ACM. The system is being used by the various ACM journals to organize subjects by area. There are 11 top-level categories and each is further sub-categorized.

- **A. : GENERAL LITERATURE**
  - **A.0 : GENERAL**
  - **A.1 : INTRODUCTORY AND SURVEY**
  - **A.2 : REFERENCE** (e.g., dictionaries, encyclopedias, glossaries)
  - **A.m : MISCELLANEOUS**
  
- **B. : HARDWARE**
  - **B.0 : GENERAL**
  - **B.1 : CONTROL STRUCTURES AND MICROPROGRAMMING**
  - **B.2 : ARITHMETIC AND LOGIC STRUCTURE**

- **B.3** : MEMORY STRUCTURES
- **B.4** : INPUT/OUTPUT AND DATA COMMUNICATIONS
- **B.5** : REGISTER-TRANSFER-LEVEL IMPLEMENTATION
- **B.6** : LOGIC DESIGN
- **B.7** : INTEGRATED CIRCUITS
- **B.8** : PERFORMANCE AND RELIABILITY
- **B.m** : MISCELLANEOUS
- **C.** : COMPUTER SYSTEMS ORGANIZATION
  - **C.0** : GENERAL
  - **C.1** : PROCESSOR ARCHITECTURES
  - **C.2** : COMPUTER-COMMUNICATION NETWORKS
  - **C.3** : SPECIAL-PURPOSE AND APPLICATION-BASED SYSTEMS
  - **C.4** : PERFORMANCE OF SYSTEMS
  - **C.5** : COMPUTER SYSTEM IMPLEMENTATION
  - **C.m** : MISCELLANEOUS
- **D.** : SOFTWARE
  - **D.0** : GENERAL
  - **D.1** : PROGRAMMING TECHNIQUES
  - **D.2** : SOFTWARE ENGINEERING
  - **D.3** : PROGRAMMING LANGUAGES

- **D.4** : OPERATING SYSTEMS
- **D.m** : MISCELLANEOUS
- **E.** : DATA
  - **E.0** : GENERAL
  - **E.1** : DATA STRUCTURES
  - **E.2** : DATA STORAGE REPRESENTATIONS
  - **E.3** : DATA ENCRYPTION
  - **E.4** : CODING AND INFORMATION THEORY
  - **E.5** : FILES
  - **E.m** : MISCELLANEOUS
- **F.** : THEORY OF COMPUTATION
  - **F.0** : GENERAL
  - **F.1** : COMPUTATION BY ABSTRACT DEVICES
  - **F.2** : ANALYSIS OF ALGORITHMS AND PROBLEM COMPLEXITY
  - **F.3** : LOGICS AND MEANINGS OF PROGRAMS
  - **F.4** : MATHEMATICAL LOGIC AND FORMAL LANGUAGES
  - **F.m** : MISCELLANEOUS
- **G.** : MATHEMATICS OF COMPUTING
  - **G.0** : GENERAL
  - **G.1** : NUMERICAL ANALYSIS

- **G.2** : DISCRETE MATHEMATICS
- **G.3** : PROBABILITY AND STATISTICS
- **G.4** : MATHEMATICAL SOFTWARE
- **G.m** : MISCELLANEOUS
- **H.** : INFORMATION SYSTEMS
  - **H.0** : GENERAL
  - **H.1** : MODELS AND PRINCIPLES
  - **H.2** : DATABASE MANAGEMENT
  - **H.3** : INFORMATION STORAGE AND RETRIEVAL
  - **H.4** : INFORMATION SYSTEMS APPLICATIONS
  - **H.5** : INFORMATION INTERFACES AND PRESENTATION (e.g., HCI)
  - **G.m** : MISCELLANEOUS
- **I.** : COMPUTING METHODOLOGIES
  - **I.0** : GENERAL
  - **I.1** : SYMBOLIC AND ALGEBRAIC MANIPULATION
  - **I.2** : ARTIFICIAL INTELLIGENCE
  - **I.3** : COMPUTER GRAPHICS
  - **I.4** : IMAGE PROCESSING AND COMPUTER VISION
  - **I.5** : PATTERN RECOGNITION
  - **I.6** : SIMULATION AND MODELING

- **I.7** : DOCUMENT AND TEXT PROCESSING
- **I.m** : MISCELLANEOUS
- **J.** : COMPUTER APPLICATIONS
  - **J.0** : GENERAL
  - **J.1** : ADMINISTRATIVE DATA PROCESSING
  - **J.2** : PHYSICAL SCIENCES AND ENGINEERING
  - **J.3** : LIFE AND MEDICAL SCIENCES
  - **J.4** : SOCIAL AND BEHAVIORAL SCIENCES
  - **J.5** : ARTS AND HUMANITIES
  - **J.6** : COMPUTER-AIDED ENGINEERING
  - **J.7** : COMPUTERS IN OTHER SYSTEMS
  - **J.m** : MISCELLANEOUS
- **K.** : COMPUTING MILIEUX
  - **K.0** : GENERAL
  - **K.1** : THE COMPUTER INDUSTRY
  - **K.2** : HISTORY OF COMPUTING
  - **K.3** : COMPUTERS AND EDUCATION
  - **K.4** : COMPUTERS AND SOCIETY
  - **K.5** : LEGAL ASPECTS OF COMPUTING
  - **K.6** : MANAGEMENT OF COMPUTING AND INFORMATION SYSTEMS

- **K.7** : THE COMPUTING PROFESSION
- **K.7** : PERSONAL COMPUTING
- **K.m** : MISCELLANEOUS

## **Appendix B**

### **Network of Subjects By Country**

Generally, countries usually specialize in only a few sub-fields of research be it in Computer Science, Physics, or Chemistry. There could be many reasons for this: high amount of research funding from the government or the general research demands in that country for those sub-fields. Using visualization techniques, we show the network of subjects in Computer Science for a few of the countries and a tree map distribution, showing the percentage of publications in the sub-fields of Computer Science. In this network, the color represents a sub-field of Computer Science (See Figure 3.8 for details), while a node represents a subject. The size represents the number of papers published in a particular subject by all the researchers in the Computer Science community. Two subjects are connected if an author publishes a paper in both the subjects. So, the link weight represents the number of authors who have published papers in both the subjects.

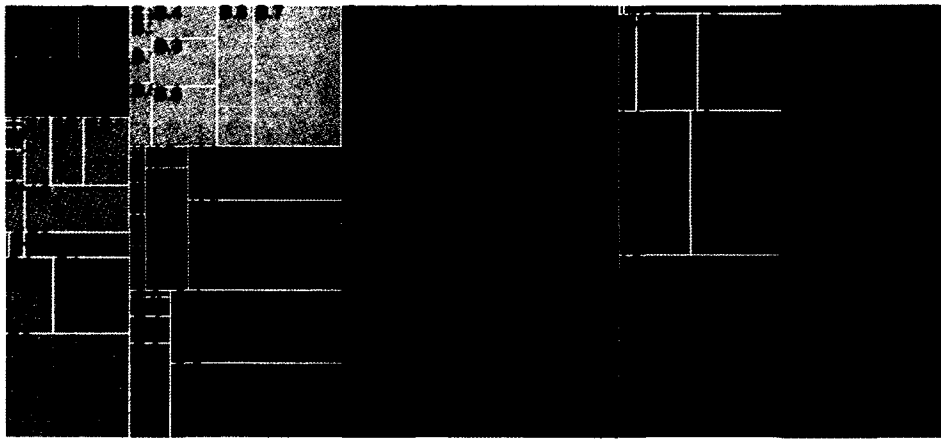
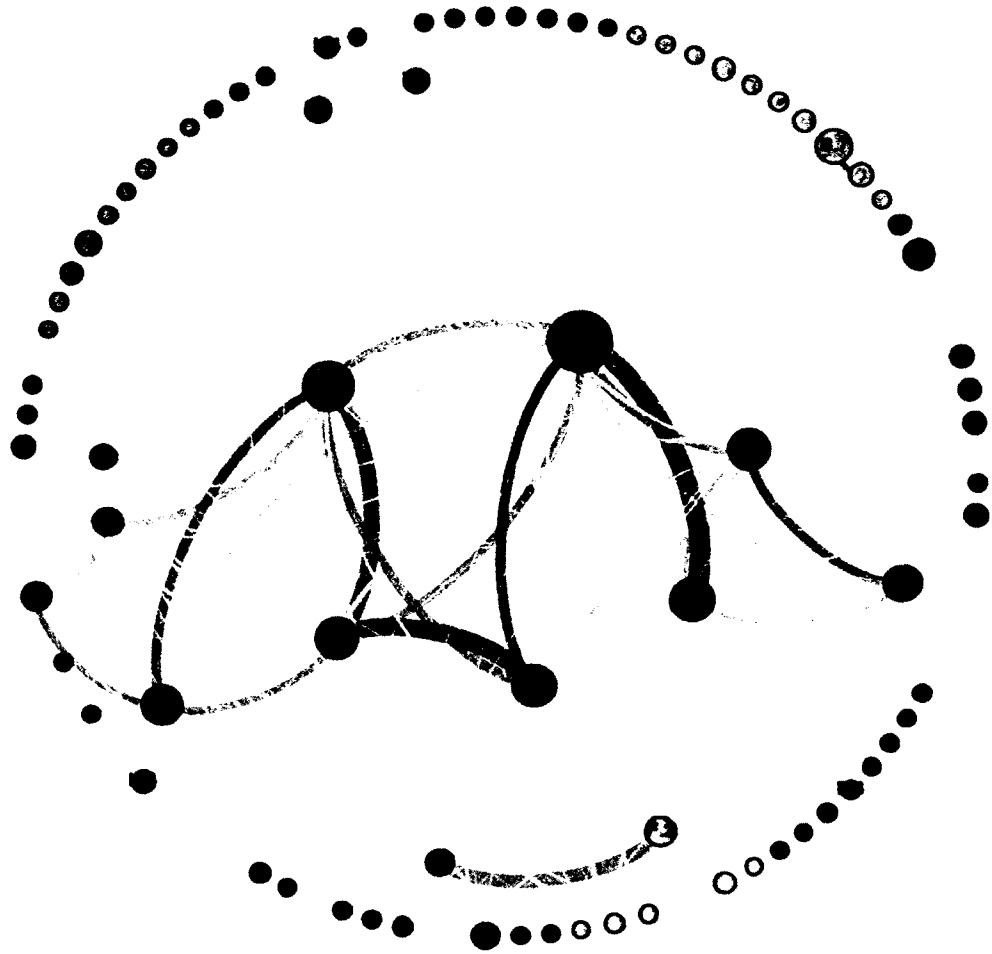


Figure B.1: Visualization of the network of subjects and a tree map distribution of papers published in different areas of computer science by the authors from the United States.



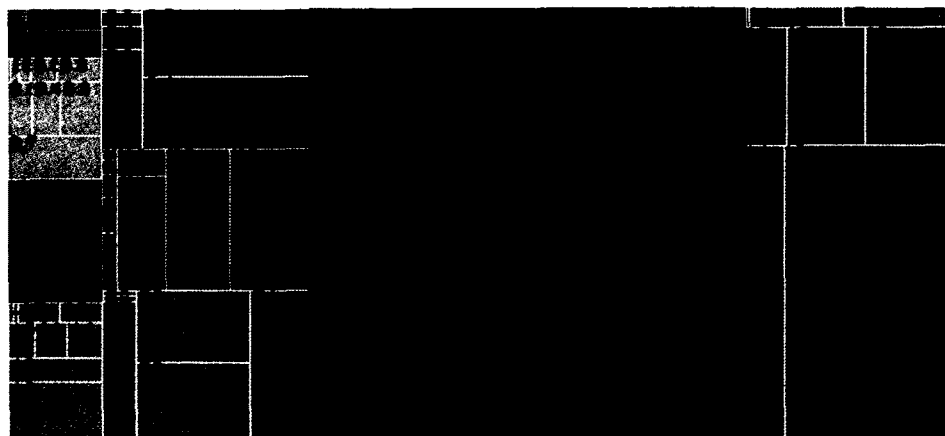
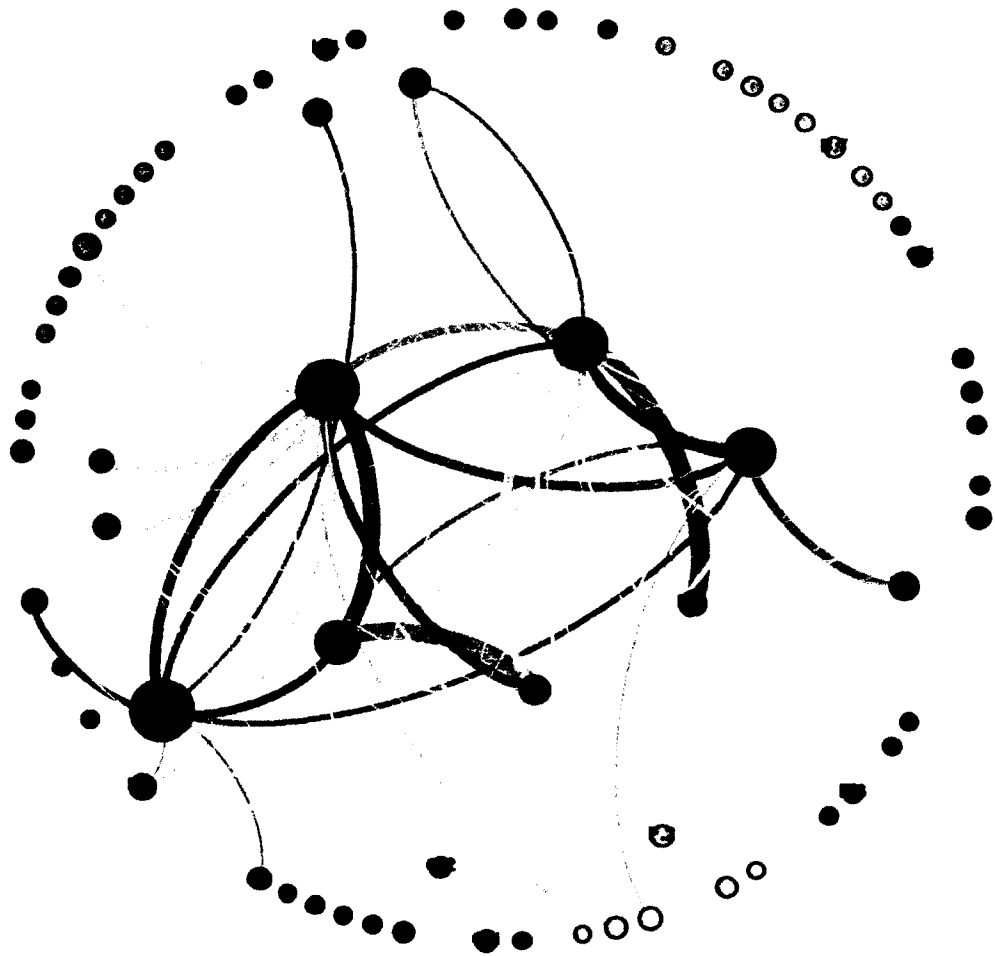


Figure B.2: Visualization of the network of subjects and a tree map distribution of papers published in different areas of computer science by the authors from Great Britain.

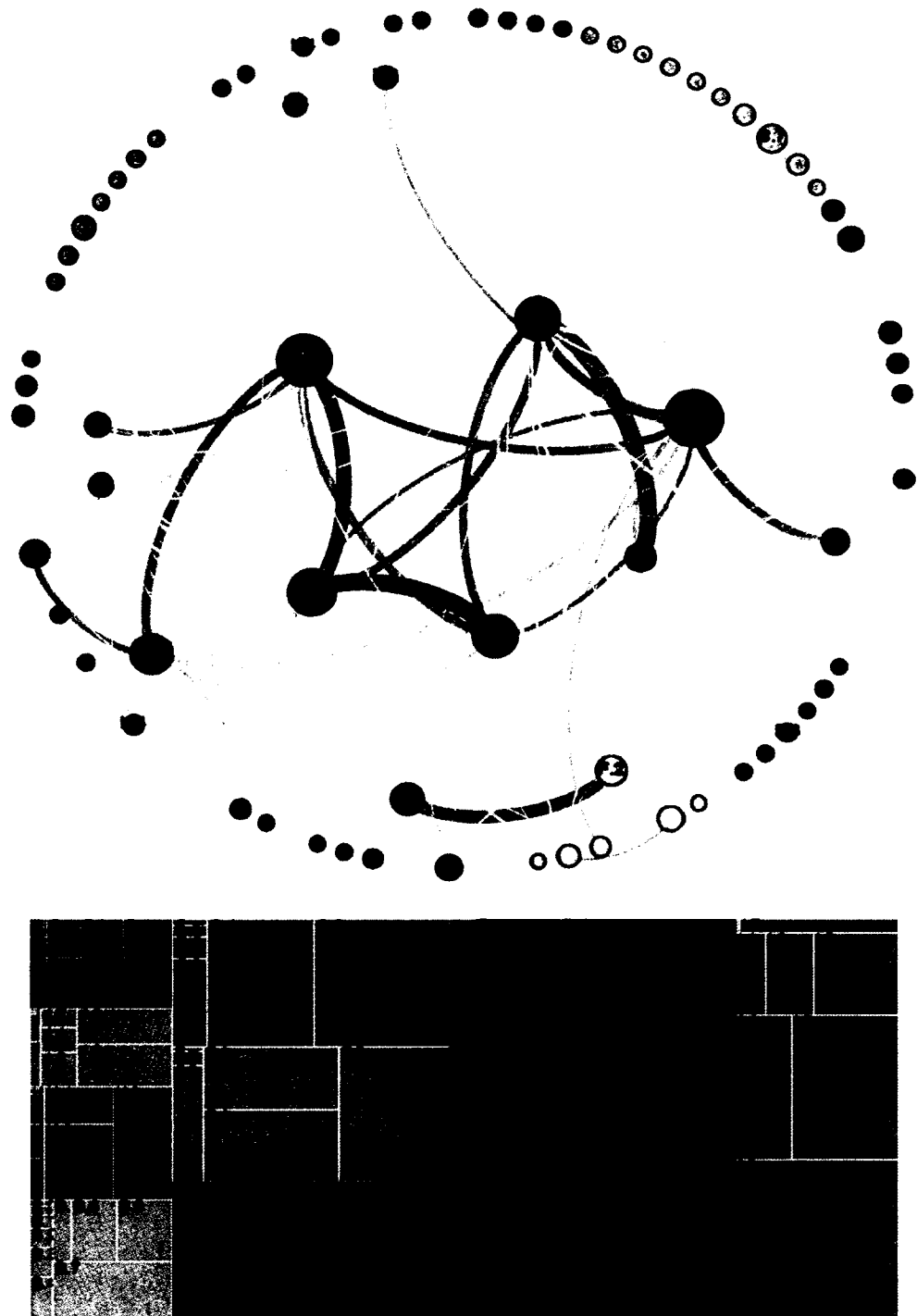


Figure B.3: Visualization of the network of subjects and a tree map distribution of papers published in different areas of computer science by the authors from Germany.

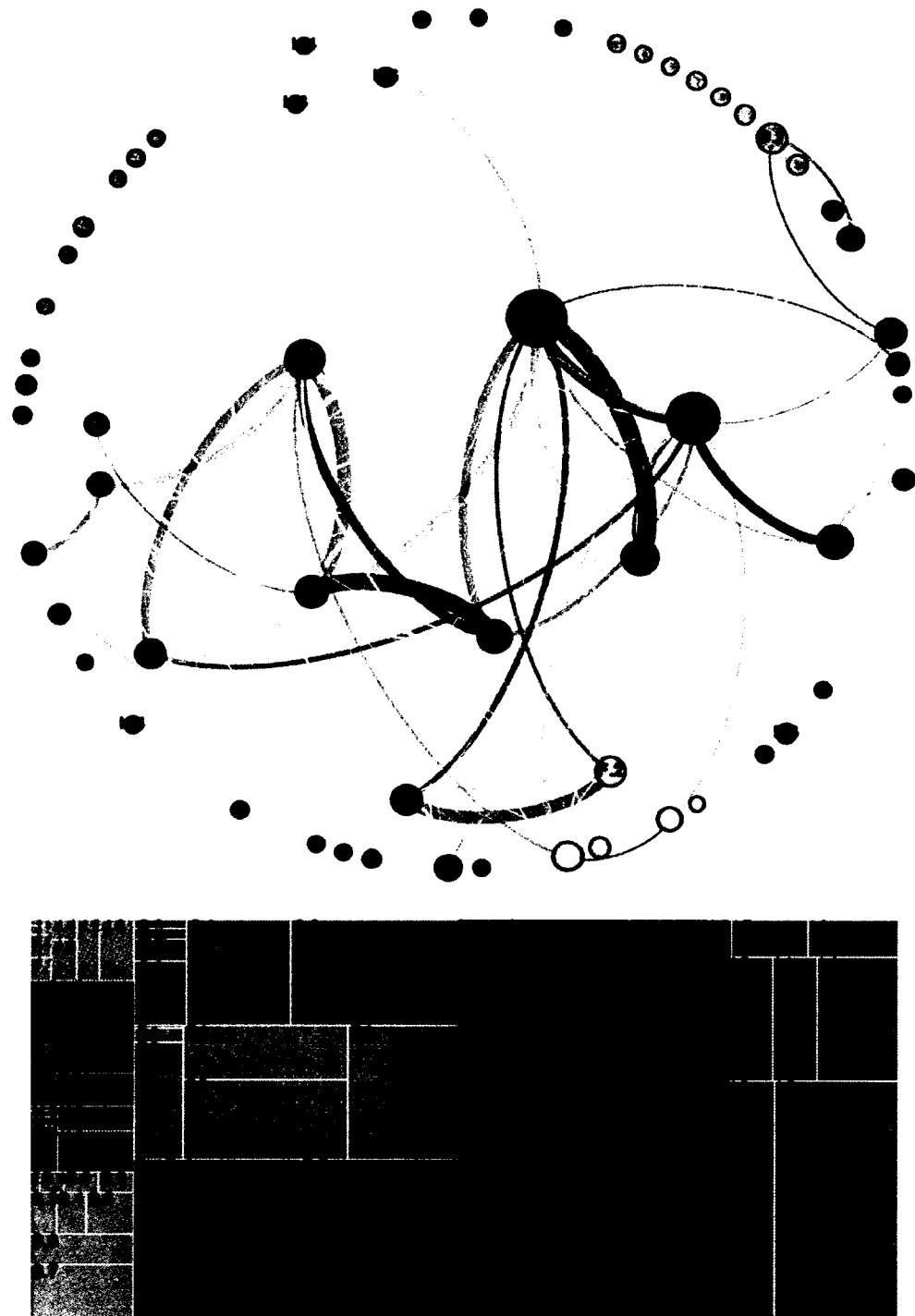


Figure B.4: Visualization of the network of subjects and a tree map distribution of papers published in different areas of computer science by the authors from France.

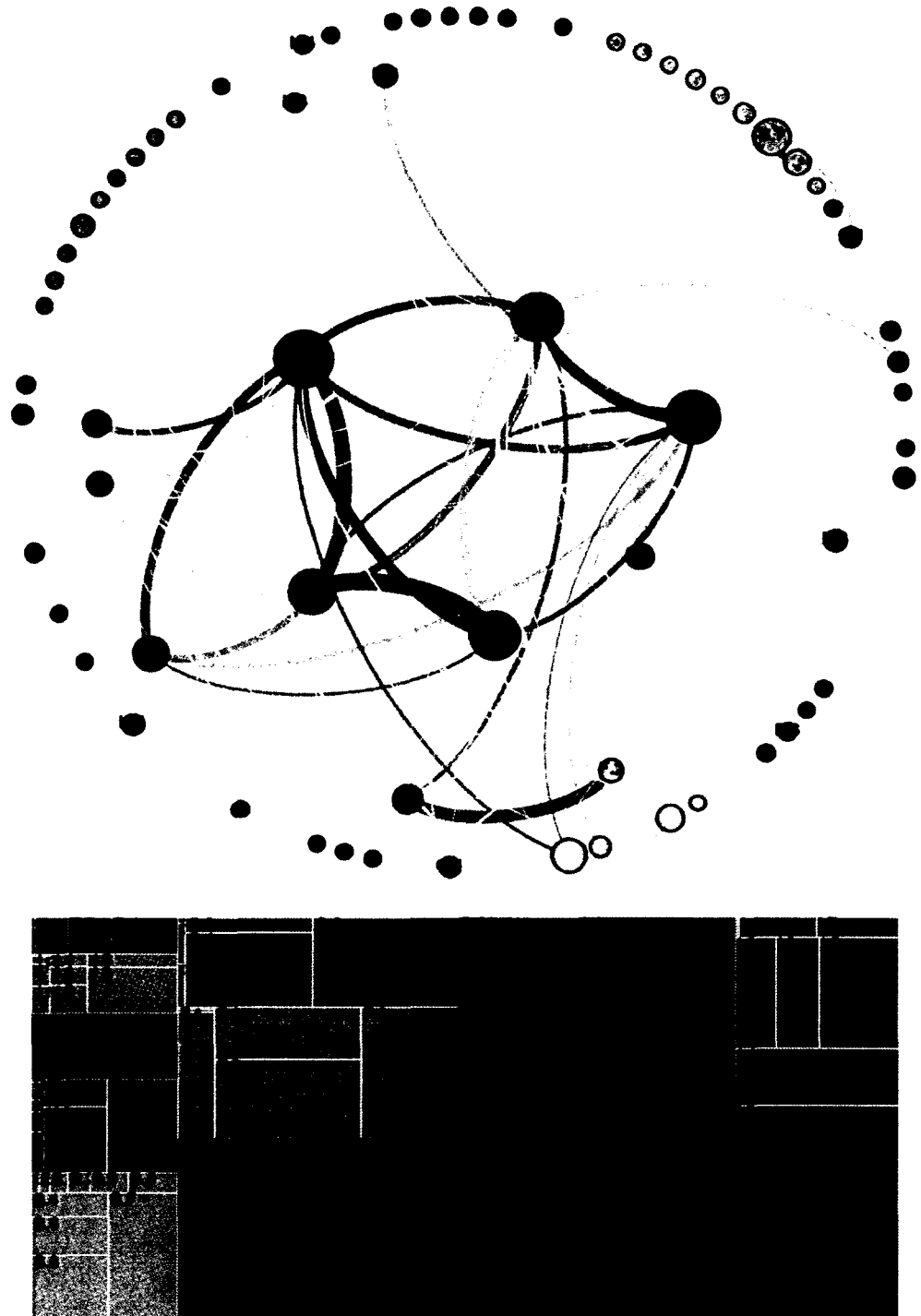


Figure B.5: Visualization of the network of subjects and a tree map distribution of papers published in different areas of computer science by the authors from Italy.

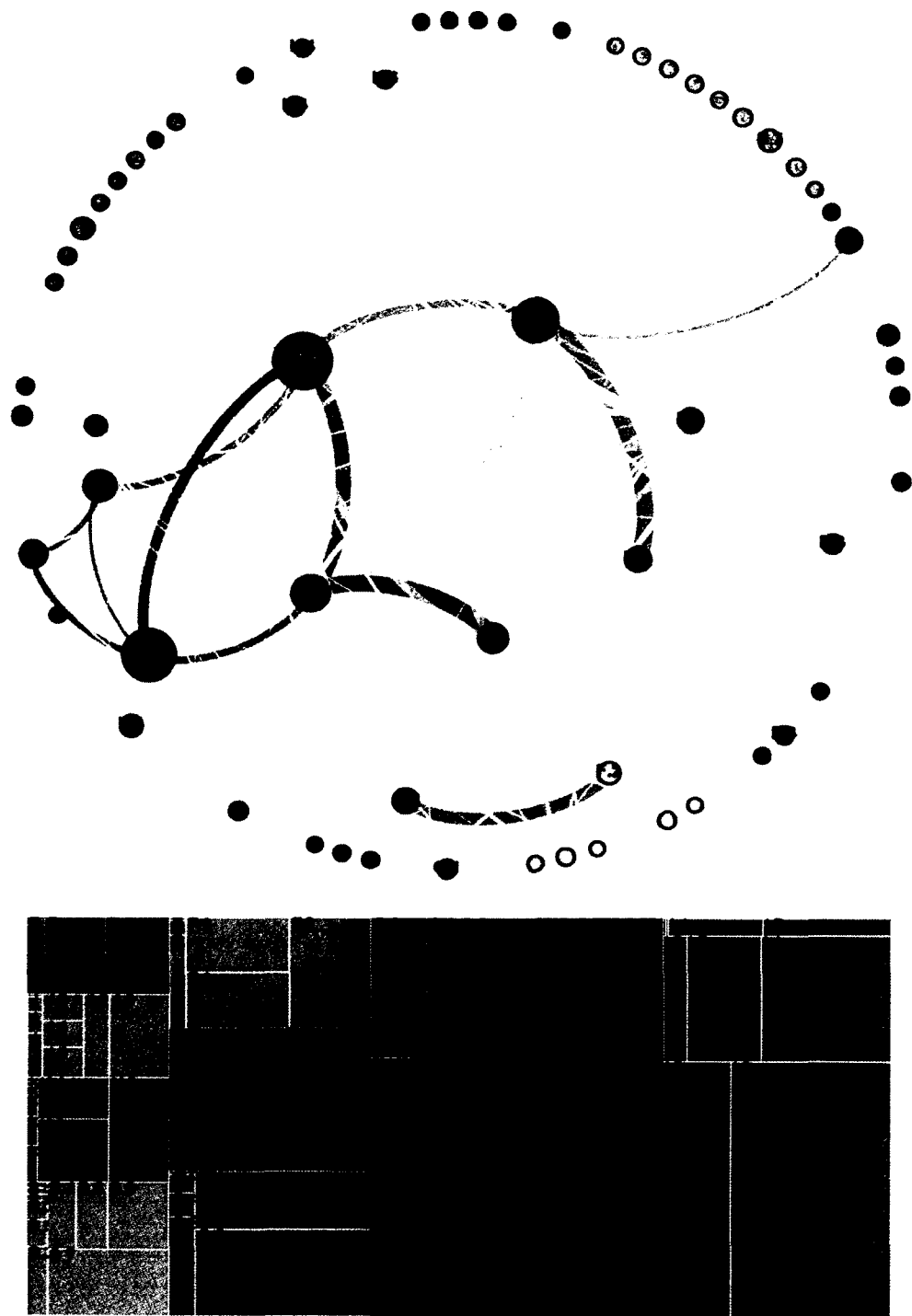


Figure B.6: Visualization of the network of subjects and a tree map distribution of papers published in different areas of computer science by the authors from Japan.

# Appendix C

## Country Code

ISO 3166-1 alpha-2 codes are two-letter country codes defined in ISO 3166-1, published by the International Organization for Standardization (ISO) to represent countries and special areas of geographical interest.

<b>Country</b>	<b>Code</b>	<b>Country</b>	<b>Code</b>	<b>Country</b>	<b>Code</b>
United States	US	Japan	JP	Singapore	SG
Great Britain	GB	India	IN	South Korea	KR
Germany	DE	Belgium	BE	Brazil	BR
Canada	CA	Sweden	SE	Greece	GR
France	FR	Denmark	DK	Norway	NO
Italy	IT	Switzerland	CH	Poland	PL
Netherlands	NL	Austria	AT	Ireland	IE
Australia	AU	Israel	IL	Russia	RU
China	CN	Finland	FI	Portugal	PT
Spain	ES	Hong Kong	HK	Czech Republic	CZ
Hungary	HU	Turkey	TR	Mexico	MX
New Zealand	NZ	Taiwan	TW	Chile	CL
Slovenia	SL	Romania	RO	South Africa	ZA

# Appendix D

## List of Publications

The following is the list of publications that have been produced during the course of this Ph.D. research.

### Papers published or accepted for publication

1. Pramod Divakarmurthy and Ronaldo Menezes. The effect of citations to collaboration networks. In *Complex Networks*, pages 177–185. Springer, 2013
2. Pramod Divakarmurthy and Ronaldo Menezes. Area diversity in computer science collaborations. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pages 2041–2042. ACM, 2012
3. Pramod Divakarmurthy, Pooja Biswas, and Ronaldo Menezes. A temporal analysis of geographical distances in computer science collaborations. In *Privacy, security, risk and trust (passat), 2011 IEEE third international conference on and 2011 IEEE third international conference on social computing (socialcom)*, pages 657–660. IEEE, 2011

# Bibliography

- [1] A. Abbasi, K.S.K. Chung, and L. Hossain. Egocentric analysis of co-authorship network structure, position and performance. *Information Processing & Management*, 2011.
- [2] L. Adamic. The small world web. *Research and Advanced Technology for Digital Libraries*, pages 852–852, 1999.
- [3] L.A. Adamic and B.A. Huberman. Power-law distribution of the world wide web. *Science*, 287(5461):2115–2115, 2000.
- [4] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. *Automatic subspace clustering of high dimensional data for data mining applications*, volume 27. ACM, 1998.
- [5] R. Albert, I. Albert, and G.L. Nakarado. Structural vulnerability of the north american power grid. *Physical Review E*, 69(2):025103, 2004.
- [6] Luis A Nunes Amaral, Antonio Scala, Marc Barthélemy, and H Eugene Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences*, 97(21):11149–11152, 2000.



- [7] Armen S Asratian. *Bipartite graphs and their applications*, volume 131. Cambridge University Press, 1998.
- [8] A.L. Barabási, 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):590–614, 2002.
- [9] Albert-László Barabási et al. Scale-free networks: a decade and beyond. *science*, 325(5939):412, 2009.
- [10] D deB Beaver and Richard Rosen. Studies in scientific collaboration. *Scientometrics*, 1(1):65–84, 1978.
- [11] Paul Benneworth and Stuart Dawley. Managing the university third strand innovation process? developing innovation support services in regionally engaged universities. *Knowledge, Technology & Policy*, 18(3):74–94, 2005.
- [12] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Computer networks*, 33(1):309–320, 2000.
- [13] J. Camacho, R. Guimera, and L.A.N. Amaral. Analytical solution of a model for complex food webs. *Physical Review E*, 65(3):030901, 2002.
- [14] Rodrigo De Castro and Jerrold W. Grossman. Famous trails to paul erds. *MATHEMATICAL INTELLIGENCER*, 21:51–63, 1999.
- [15] Chaomei Chen. Visualising semantic spaces and author co-citation networks in digital libraries. *Information processing & management*, 35(3):401–420, 1999.

- [16] Tzeng-Ji Chen, Yu-Chun Chen, Shinn-Jang Hwang, and Li-Fang Chou. International collaboration of clinical medicine research in taiwan, 1990-2004: A bibliometric analysis. *Journal of the Chinese Medical Association*, 70(3):110–116, 2007.
- [17] F. Cheong and B.J. Corbitt. A social network analysis of the co-authorship network of the pacific asia conference on information systems from 1993 to 2008. 2009.
- [18] Fan Chung and Linyuan Lu. The volume of the giant component of a random graph with given expected degrees. *SIAM Journal on Discrete Mathematics*, 20(2):395–411, 2006.
- [19] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.
- [20] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.
- [21] C. Cotta and J.J. Merelo. The complex network of evolutionary computation authors: an initial study. *arXiv preprint physics/0507196*, 2005.
- [22] David J Crandall, Lars Backstrom, Dan Cosley, Siddharth Suri, Daniel Huttenlocher, and Jon Kleinberg. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, 107(52):22436–22441, 2010.
- [23] G.F. Davis, M. Yoo, and W.E. Baker. The small world of the american corporate elite, 1982-2001. *Strategic organization*, 1(3):301–326, 2003.

- [24] DB De Beaver and R Rosen. Studies in scientific collaboration. *Scientometrics*, 1(2):133–149, 1979.
- [25] Pramod Divakarmurthy, Pooja Biswas, and Ronaldo Menezes. A temporal analysis of geographical distances in computer science collaborations. In *Privacy, security, risk and trust (passat), 2011 IEEE third international conference on and 2011 IEEE third international conference on social computing (socialcom)*, pages 657–660. IEEE, 2011.
- [26] Pramod Divakarmurthy and Ronaldo Menezes. Area diversity in computer science collaborations. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pages 2041–2042. ACM, 2012.
- [27] Pramod Divakarmurthy and Ronaldo Menezes. The effect of citations to collaboration networks. In *Complex Networks*, pages 177–185. Springer, 2013.
- [28] I.N. Durbach, D. Naidoo, and J. Mouton. Co-authorship networks in south african chemistry and mathematics. *South African Journal of Science*, 104(11-12):487–492, 2008.
- [29] Leo Egghe and Ronald Rousseau. Introduction to informetrics: Quantitative methods in library, documentation and information science. 1990.
- [30] J.W. Endersby. Collaborative research in the social sciences: Multiple authorship and publication credit. *Social Science Quarterly*, 77(2):375–392, 1996.
- [31] Douglas SR Ferreira, Andrés RR Papa, and Ronaldo Menezes. Small world picture of worldwide seismic events. *Physica A: Statistical Mechanics and its Applications*, 408:170–180, 2014.

- [32] B.S. Fisher, C.T. Cobane, T.M. Vander Ven, and F.T. Cullen. How many authors does it take to publish an article? trends and patterns in political science. *PS: Political Science and Politics*, pages 847–856, 1998.
- [33] Pablo Fleurquin, José J Ramasco, and Victor M Eguiluz. Systemic delay propagation in the us airport network. *Scientific reports*, 3, 2013.
- [34] M. Franceschet. Collaboration in computer science: A network science approach. *Journal of the American Society for Information Science and Technology*, 62(10):1992–2012, 2011.
- [35] Antonios Garas, Panos Argyrakis, Céline Rozenblat, Marco Tomassini, and Shlomo Havlin. Worldwide spreading of economic crisis. *New journal of Physics*, 12(11):113043, 2010.
- [36] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.
- [37] Wolfgang Glänzel and András Schubert. Analysing scientific networks through co-authorship. In *Handbook of quantitative science and technology research*, pages 257–276. Springer, 2005.
- [38] J.W. Grossman and P.D.F. Ion. On a portion of the well-known collaboration graph. *Congressus Numerantium*, pages 129–132, 1995.
- [39] Jean-Loup Guillaume and Matthieu Latapy. Bipartite graphs as models of complex networks. *Physica A: Statistical Mechanics and its Applications*, 371(2):795–813, 2006.

- [40] Fred Halsall and Data Links. Computer networks and open systems. *Addison-Wesley Publishers*, pages 112–125, 1995.
- [41] L.L. Hargens. *Patterns of scientific research*. American Sociological Association, 1975.
- [42] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.
- [43] H. Hou, H. Kretschmer, and Z. Liu. The structure of scientific collaboration networks in scientometrics. *Scientometrics*, 75(2):189–202, 2008.
- [44] J. Hudson. Trends in multi-authored papers in economics. *The Journal of Economic Perspectives*, 10(3):153–158, 1996.
- [45] Björn H Junker and Falk Schreiber. *Analysis of biological networks*, volume 2. John Wiley & Sons, 2008.
- [46] J Sylvan Katz. Geographical proximity and scientific collaboration. *Scientometrics*, 31(1):31–43, 1994.
- [47] J Sylvan Katz and Ben R Martin. What is research collaboration? *Research policy*, 26(1):1–18, 1997.
- [48] Gueorgi Kossinets and Duncan J Watts. Empirical analysis of an evolving social network. *Science*, 311(5757):88–90, 2006.
- [49] Robert Kraut, Carmen Egido, and Jolene Galegher. Patterns of contact and communication in scientific research collaboration. In *Proceedings of the 1988 ACM conference on Computer-supported cooperative work*, pages 1–12. ACM, 1988.

- [50] Gautier Krings, Francesco Calabrese, Carlo Ratti, and Vincent D Blondel. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(07):L07003, 2009.
- [51] Luka Kronegger, Franc Mali, Anuška Ferligoj, and Patrick Doreian. Collaboration structures in slovenian scientific communities. *Scientometrics*, 90(2):631–647, 2012.
- [52] A.H.F. Laender, C.J.P. de Lucena, J.C. Maldonado, E. de Souza e Silva, and N. Ziviani. Assessing the research and education quality of the top brazilian computer science graduate programs. *ACM SIGCSE Bulletin*, 40(2):135–145, 2008.
- [53] Renaud Lambiotte, Vincent D Blondel, Cristobald de Kerchove, Etienne Huens, Christophe Prieur, Zbigniew Smoreda, and Paul Van Dooren. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317–5325, 2008.
- [54] M. Leclerc, Y. Okubo, L. Frigoletto, and J.F. Miquel. Scientific co-operation between canada and the european community. *Science and public Policy*, 19(1):15–24, 1992.
- [55] Sooho Lee and Barry Bozeman. The impact of research collaboration on scientific productivity. *Social studies of science*, 35(5):673–702, 2005.
- [56] G.A. Lemarchand. The long-term dynamics of co-authorship scientific networks: Iberoamerican countries (1973–2010). *Research Policy*, 2011.

- [57] Chien Hsiang Liao. How to improve research quality? examining the impacts of collaboration intensity and member diversity in collaboration networks. *Scientometrics*, 86(3):747–761, 2011.
- [58] David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11623–11628, 2005.
- [59] X. Liu, J. Bollen, M.L. Nelson, and H. Van de Sompel. Co-authorship networks in the digital library research community. *Information processing & management*, 41(6):1462–1480, 2005.
- [60] Terttu Luukkonen, Olle Persson, and Gunnar Sivertsen. Understanding patterns of international scientific collaboration. *Science, Technology & Human Values*, 17(1):101–126, 1992.
- [61] Göran Melin and Olle Persson. Studying research collaboration using co-authorships. *Scientometrics*, 36(3):363–377, 1996.
- [62] David C Mowery and Bhaven N Sampat. Universities in national innovation systems. *The Oxford handbook of innovation*, pages 209–239, 2005.
- [63] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.
- [64] M.E.J. Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

- [65] M.E.J. Newman. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5200–5205, 2004.
- [66] Evelien Otte and Ronald Rousseau. Social network analysis: a powerful strategy, also for the information sciences. *Journal of information Science*, 28(6):441–453, 2002.
- [67] M.A. Ovalle-Perandones, A. Perianes-Rodriguez, and C. Olmeda-Gomez. Hubs and authorities in a spanish co-authorship network. pages 514–518, 2009.
- [68] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.
- [69] S.L. Pimm. *The balance of nature?: ecological issues in the conservation of species and communities*. University of Chicago Press, 1992.
- [70] Mason A Porter, Jukka-Pekka Onnela, and Peter J Mucha. Communities in networks. *Notices of the AMS*, 56(9):1082–1097, 2009.
- [71] A. Reka, H. Jeong, and A.L. Barabasi. Diameter of the world wide web. *Nature*, 401(9):130–131, 1999.
- [72] Quirin Schiermeier. Career choices: The mobility imperative. *Nature*, 470(7335):563–564, 2011.
- [73] John Scott. *Social network analysis*. Sage, 2012.
- [74] Yosef Sheffi. Urban transportation networks: equilibrium analysis with mathematical programming methods. 1985.



- [75] David A Smith and Douglas R White. Structure and dynamics of the global economy: network analysis of international trade 1965–1980. *Social Forces*, 70(4):857–893, 1992.
- [76] R.V. Sole and M. Montoya. Complexity and fragility in ecological networks. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1480):2039–2045, 2001.
- [77] Krishnappa Subramanyam. Bibliometric studies of research collaboration: A review. *Journal of information Science*, 6(1):33–38, 1983.
- [78] Christian Thiemann, Fabian Theis, Daniel Grady, Rafael Brune, and Dirk Brockmann. The structure of borders in a small world. *PloS one*, 5(11):e15422, 2010.
- [79] S. Uddin, L. Hossain, A. Abbasi, and K. Rasmussen. Trend and efficiency analysis of co-authorship network. *Scientometrics*, 90(2):687–699, 2012.
- [80] Maarten Van Steen. Graph theory and complex networks. *An Introduction*, 2010.
- [81] T. Velden, A. Haque, and C. Lagoze. A new approach to analyzing patterns of collaboration in co-authorship networks: mesoscopic analysis and interpretation. *Scientometrics*, 85(1):219–242, 2010.
- [82] Srividhya Venugopal, Evan Stoner, Martin Cadeiras, and Ronaldo Menezes. Understanding organ transplantation in the usa using geographical social networks. *Social Network Analysis and Mining*, 3(3):457–473, 2013.
- [83] S. Wasserman and K. Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

- [84] D. Watts and S. Strogatz. The small world problem. *Collective Dynamics of Small-World Networks*, 393:440–442, 1998.
- [85] D.J. Watts. *Small worlds: the dynamics of networks between order and randomness*. Princeton university press, 2003.
- [86] Douglas Brent West et al. *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River, 2001.
- [87] R.J. Williams, E.L. Berlow, J.A. Dunne, A.L. Barabási, and N.D. Martinez. Two degrees of separation in complex food webs. *Proceedings of the National Academy of Sciences*, 99(20):12913–12916, 2002.
- [88] James Wilsdon et al. *Knowledge, networks and nations: Global scientific collaboration in the 21st century*. The Royal Society, 2011.
- [89] R. Yousefi-Nooraie, M. Akbari-Kamrani, R.A. Hanneman, and A. Etemadi. Association between co-authorship network and scientific productivity and impact indicators in academic medical research centers: A case study in iran. *Health Research Policy and Systems*, 6(1):9, 2008.
- [90] Q. Yu, H. Shao, and Z. Duan. Research groups of oncology co-authorship network in china. *Scientometrics*, 89(2):553–567, 2011.