Dynamics of Citation Collaboration Network

A thesis in partial fulfilment for the degree of

Doctor of Philosophy (PhD.)

in

Computer Science and Engineering

Submitted By

Dinesh Kumar Pradhan

NITD/PhD/CSE 2015/00580

under the supervision of

Dr. Subrata Nandi

Professor, Department of CSE, NIT Durgapur

Dr. Prasenjit Choudhury

Assistant Professor, Department of CSE, NIT Durgapur

and

Dr. Tanmoy Chakraborty

Assistant Professor, Department of CSE, IIIT Delhi

Department of Computer Science and Engineering

National Institute of Technology, Durgapur

17^{th} *June*, 2019

© by Dinesh Kumar Pradhan, 2019
All Rights Reserved.

Statement of thesis preparation

1. Thesis title: **Dynamics of Citation Collaboration Network**
2. Degree for which submitted: **Doctor of Philosophy (PhD.)**
3. The thesis guide was referred to for thesis preparation: **Yes**
4. Specifications regarding thesis format have been closely followed: **Yes**
5. The contents of the thesis were organized according to the guidelines: **Yes**

(Signature of Student)
Name: **Dinesh Kumar Pradhan**
Roll No.: **NITD/PhD/CSE 2015/00580**
Department: **Computer Science and Engineering**

This work is dedicated to my father...

sadly, we lost him forever.

Shakti Pada Pradhan
30. 11. 1942 - 29. 05. 2014

Declaration

I certify that,

1. The work covered in this thesis is innovative and has been done by me under the supervision of my supervisors.
2. The work has not been submitted to any other Institute for any degree or diploma.
3. I have followed the guiding principles given by the Institute in organizing the thesis.
4. I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.
5. Whenever I have used materials (theoretical analysis, data, figures and text) from other sources, I have given due credit to them by citing them it the text of the thesis and giving their particulars in the references.

Dinesh Kumar Pradhan

Certificate of Recommendation

This is to certify that the thesis entitled **"Dynamics of Citation Collaboration Network"**, submitted by **Dinesh Kumar Pradhan** for the partial fulfilment of the requirements for the award of the degree of **Doctor of Philosophy (PhD.) in Computer Science and Engineering**, is a bonafide research work under the guidance of **Dr. Subrata Nandi, Dr. Prasenjit Choudhury and Dr. Tanmoy Chakraborty**. The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma. In our opinion, this thesis is of the standard required for the partial fulfilment of the requirements for the award of the degree of **Doctor of Philosophy (PhD.)**.

Dr. Subrata Nandi
Professor, CSE Department
National Institute of Technology
Durgapur-713209, INDIA

Dr. Prasenjit Choudhury
Assistant Professor, CSE Department
National Institute of Technology
Durgapur-713209, INDIA

Dr. Tanmoy Chakraborty
Assistant Professor, CSE Department
Indraprastha Institute of Information Technology Delhi
Okhla Industrial Estate, Phase III, Delhi, 110020

Acknowledgements

First and foremost, I have to thank my research supervisor **Dr. Subrata Nandi, Dr. Prasenjit Choudhury and Dr. Tanmoy Chakraborty**. Without their assistance and dedicated involvement in every step throughout the process, this thesis would never have been accomplished. Their guidance helped me in all the time of research and writing of this thesis. I would like to thank them very much for the support, patience, and understanding over this duration of the research work.

I wish to acknowledge **Dr. Goutam Sanyal**, HoD, Department of Computer Science and Engineering, for providing us supporting environment to carry on our research work. I would also like to show gratitude and respect to all faculty members of our department for their support.

I thank my friends and students whoever is involved either directly or indirectly in this research project.

Last but not the least, I would like to thank my parents, my family and my friends for their unconditional support.

Dinesh Kumar Pradhan

Contents

List of Figures

List of Tables

Abstract

In the academic community with digitalization and wider visibility, there has been an exponential increase in the number of publication, author, venue, etc. Lower subscription charges and open accessibility option has opened multiple choices for authors to publish their article in different venues. Intuitively, an important research question that comes out is whether research quality is simultaneously withheld as exponential growth occurs in rate of all scientific entities? Thus, quality assessment is a critical issue and it is necessary to understand loopholes in existing quantifiers. Among others, issues such as re-defining author ranking strategy, analyzing citation behaviour of top articles, understanding the importance of venue and need to re-define its metric and making improvement towards fair and accurate review process needs to be addressed.

Research funding and promotion of an author is mainly dependent upon author ranking indexes such as h-index, g-index, i3-index, etc. These indexes consider only citation and publication count of an author for assessment. As a result of which researchers in their early career mostly get zero index value. Even after surpassing the zero index value, another issue that occurs is a tie in index values among them. Therefore, the question is, are current author ranking indexes consistent enough to assess new as well as seasoned authors ? If there exists a tie between authors, are citation and publication count the only two factors to judge an author's research calibre?

Citation is a widely accepted quantifier in all bibliographic affairs for measuring quality. With increasing rate in publication, the citation distribution is found to be highly right skewed. It implies that only a small count of papers collect large number of citation. Categorizing the citation trajectories of top articles into different classes and studying their nature in different time window might give good insights on what factors influence a paper to collect quality citation?

Assessment metric for journals such as 'Impact Factor' also shows some inconsistencies. Since 2009, Thomson Reuters indexing firm and a large volume of other works report several cases of anomalous citation such as excessive self-citation, citation stacking and cartel where a group of journals, authors, editors, and publishers mutually boost each other's Impact Factor. However, identifying exact feature set from past bibliographic data may be helpful for early detection and monitoring of such malicious instances.

Fair and accurate judgment to an article is directly dependent upon the entire review process. Out of which assignment of a paper to reviewer is the most challenging task for all venue organizers. In the case of conference, multiple issues are accounted such as time constraint, availability of reviewers, exact expertise matching, conflict of interest management, balancing the workload among reviewers, judgment of reviewer's interest, etc.

To address above described issues, we use multiple bibliographic data set collected from ArnetMiner, Google Scholar and Microsoft Academic Graph. In the thesis we tried to address four specific problems: **first**, a new author ranking metric, **second**, studied various citation profiles of top articles, **third**, reasoning different citation patterns in the context of journals and **finally**, an automated conflict of interest management approach to improvise reviewer assignment strategy in conferences.

For ranking authors, we propose a 3-layer based indexing mechanism defined as 'C3-Index'. It is capable enough to break ties and gives at least some score to all active authors. The score is calculated from three different dimensions as, their paper citation, collaboration with other researchers and from whom they get citations. C3-Index is based on a modified PageRank algorithm which is applied to all three layers in a different form. Further C3-Index can be used to detect future prospect of an author.

Towards profiling of citation trend, we find that the data is right skewed. There exist a few well-cited papers where the probability for getting 1000 citations is less than 10^{-6} . There is an increasing preferential attachment rate for papers from 2000 onward. We profile on time series citation acquired by an article and get some rarely seen trajectories and identify them as Sleeping Beauties, Discovery papers, Hot papers, etc.

In the context of studying anomalous citation practices as pointed out by a volume of existing reports, we try to find similar citation instances and how frequently they occur in a data set of Computer Science journal papers. We extract citation patterns in the form of common motif and further analyze their impact factor on a temporal basis. As our finding, patterns are one of the major, but it needs fine tuning with few other features that could

correctly identify a journal involved in anomalous citation practices.

For an automated reviewer assignment strategy, we reformulate *RAP* and propose a multi-criteria based greedy matrix approximation technique. Here, we consider the biasness factor which is derived from co-authorship relations instead of self-declared *CoI* (conflict of interest). As a proof of concept, we execute our approach on a real conference data set (ICBIM 2016) collected from EasyChair where there are 51 accepted papers, and the number of reviewers are 40. We find that our assignment algorithm gives a better overall assignment quality score in terms of maximizing field similarity, minimizing biasness factor and balancing reviewer's load.

In this thesis, we propose a new author ranking metric, studied various static paper citation profiles, investigated different citation patterns in journals as well as reasoned them and proposed an automated CoI management for reviewer assignment strategy in conferences. However, the findings of the thesis trigger a few more interesting issues which may be considered for further study like, C3-index can be improvised using machine learning algorithms for predicting future prospect of an author as well as a research project. The same can be used to dynamically categorize papers at their early stage. Identifying additional features to detect malicious cases in venue-based collaboration, etc. These might play an important role in monitoring the quality of all research entities.

Dynamics of Citation Collaboration Network

Chapter 1

Introduction

A citation is a key token of recognition that is used to pay tribute to pioneers for their breakthrough discoveries, credit to peers by acknowledging related works, proofreading and improving methodologies to correct one's own work or other related works to extend research in a new dimension. Besides, it is a fundamental quantifier of quality assessment of several scientific entities at an atomic level and recognizes an author's research potential in subsequent related works in the same or interdisciplinary field of research. Citation analysis in terms of measuring, computing, quantifying, comparing entities, analyzing and re-defining metrics is perhaps, a significant tool to depict the progress of science in society over the years[1]. Thus, a citation network represents research inclination and knowledge graph of science in which research output in terms of scientific publications, author or venue form knowledge sources or nodes, and their interlinked relationships mapped in terms of citation represents relatedness among various aspects of knowledge. The complexity of the citation network has increased ten-fold since the 1990's with an exponential increase in the node and edge count and its rapidly changing dynamics over time [2]. This complexity is getting so specialized that an exhaustive study is required to analyze trends and make crucial decisions to maintain the research standard. In citation network, when a paper (P_i) gets cited by another paper (P_j), key entity author (A_i) is in turn getting cited by an author (A_j).

In the last decade, science has technologically advanced at a breakneck pace, and an expedite growth in researchers, and their publications have occurred which has evolved heightened need to define bibliometric measures. Bibliometrics as a new discipline has gathered ardent attention of researchers around the globe due to impressive advances in

online accessibility, computation, and storage of massive data sets. With huge information overload and filter failure, there is a need to measure quality and assess possible manipulation at several stages of publication process through extensive citation analysis. Consequently, strict quality metrics can help planning committees decide on investing on research projects, promotions and tenure for researchers, research managers or board of governors assembling a research group, a publication house or an editor setting a standard for its competitors, etc. Moreover, it can closely help to monitor the whole evaluation procedure and an author's performance regarding quantity, quality, and consistency in upholding his relative position.

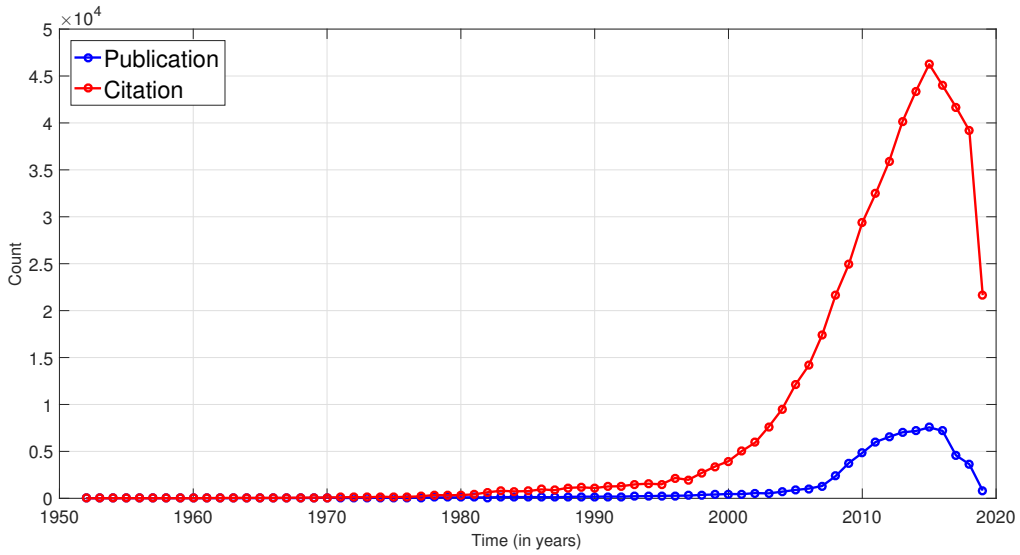


Figure 1.1: As per Microsoft academic graph, research on citation analysis starts from mid-1950's but actual growth is visible from the year 2000 on-wards. By the year of 2015, total publication count reaches 7,562, whereas citation count reached up to 46,255.

Advancement on citation analysis research: Analysis of citation network involve research on quantifying metric for better evaluation, analyzing qualitative aspects like, influence of social academic relation among scientific entities and other computationally efficient approaches. Ranking of author, journal, research group, institutions, evaluating top-cited article, tracing evolution trajectory of multi-disciplinary field etc are used to correctly evaluate research performance. Further, these have diverse application such as identifying expert reviewer, top author, institution ad hence, nation. Accuracy in data determines true

merit of an author which in turn helps in taking funding decision, tenure and promotion. Due to increase in competition for limited resources, strict evaluation is essential for maintaining the ethics of scientific community. Hence, the importance of study in this field has enormously increased.

Electronic archived data sets such as Microsoft academic graph, Google Scholar, Web of science consists of exhaustive statistics of publications, author and their citations which can highlight citation trends over time and help take crucial decisions while measuring publication process appropriately. With the easy availability of various interfaces such as Sciverse, Scopus, Scival diverse microscopic entities of research community such as author, publications, journals, research groups, institutions, cities, countries, domain, etc. can be compared and assessed in a variety of conditional aspects.

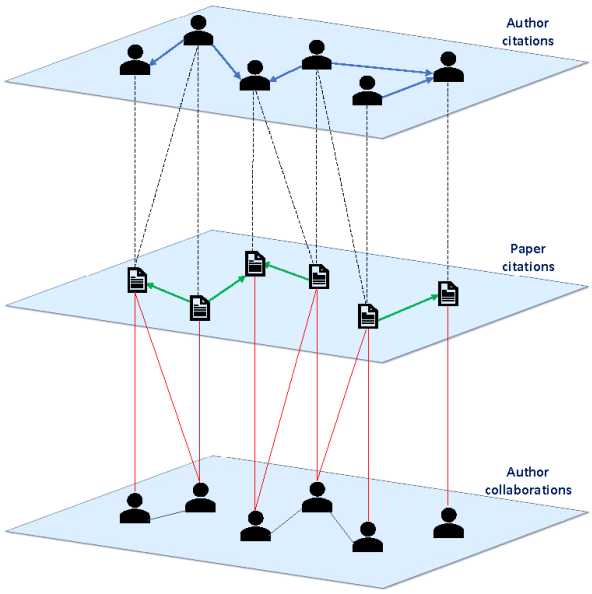


Figure 1.2: Complexity, as well as dynamicity, is visible in the above figure. By considering paper as vertices and citation as edges we are finding *paper citation network* (middle layer). Again paper is closely associated with authors who collaborate as co-authors to form another layer of the network, i.e., *author collaboration network* (lower layer). We can extract one more layer from the core layer is, if one author's paper is cited by another author's paper means that author is citing the previous author and forming a network is *author citation network* (upper layer). By this way, the entire 3-layered network is emerging as a complex network. If we include time factor here, then dynamic characteristics will also be observable.

Collaboration mechanics and dynamics of citation structure: An exhaustive analysis of scholarly communication concerning citation network is further, used to detect the impact of collaboration, map evolution of inter-disciplinary fields of research, cross-domain knowledge transfers and hence, assess research impact. Several factors play along to influence citations. Often, it is seen that a group of researchers from the same or different institutions collaborate together (peer-peer collaboration) and publish articles to share resources and enhance research standard. Besides, junior faculty or research beginners collaborate inherently with their Ph.D. Guide or senior researchers. Such kind of underlying collaboration mechanics largely impact citation dynamics and quality assessment of authors and their publications from several dimensions.

Paper, author, venue (3 major entities) and citations collected by them are used to form various networks such as paper-paper citation network, author citation network, author co-citation network, author co-author network, editor co-author network, reviewer co-author network, journal co-citation network where, nodes refer to author sometimes, specifying their scientific status, papers, and venue whereas, edges refer to count by which each has been co-authored, cited or co-cited [3].

Dynamic nature of citation depicts that corresponding edges and their weights in the respective paper, author co-author or venue co-citation sub-graphs vary over time. On a temporal scale, citation and strength of collaboration both exhibit large deviations and hence, a direct impact on the author's performance and paper quality. A consensus reveals that citation graph of published articles initially attracts citations for the first two to three years followed by a constant saturated peak and then, a final decay of citations for rest of the publication age. Consistency in citations collected over publication age, duration of active citation age followed by a decline of success, the influence of publication venue and collaboration impact are factors changing dynamics of citation profiles over time [1].

1.1 Terminology and definitions

Academic paper or scientific article: In academic publishing, a paper is a research work that is usually published in an academic journal. It contains original research conclusion with shreds of evidence or results or can be extension or reviews of the existing literature. Such a paper is also referred to as an article. Paper is a generic term and is often used for a published piece of writing that is, letters, communications, reports, reviews, and full publications. A quality article indirectly determines an author's eminence.

Citation: Broadly, a citation is a reference to a published or unpublished node (paper, author or venue) to acknowledge the contribution made by the cited node. More precisely, a citation is an abridged remark added to an intellectual research work produced by an author in the form of direct entry in the bibliographic reference section of current work at hand hence, acknowledging its significance. In academic publishing, it is used as a token of recognition to pay homage to pioneers or identify work of peers or others concerning a topic of discussion where citation appears. In general, a citation is a combination of in-body citation and bibliographic entry.

Considering directed citation mapping for a collection of articles can be represented in the form of the citation graph. Derek J. de Solla Price first demonstrated this in his 1965 article "Networks of Scientific Papers".

Citation network: A citation graph (or citation network) is a directed graph in which each vertex can represent either a research document (academic papers) or its authors (single or multiple authors) or venue of publication (journal or conference) and in which each edge refers to frequency or number of times each vertex has been cited, co-cited or co-authored by other vertices in the graph.

A typical application of the citation graph is to measure the impact of individual vertices in the graph and make comparative studies ultimately, to enhance research quality. Citation analysis is thus, used as the basis for calculating measures of an author's impact such as h-index, and for studying the structure and evolution of different fields of academic inquiry. The edges of the graph are not only directed, but they are also acyclic, that is there are no loops in the graph.

Citation analysis: Citation analysis refers to statistical as well as a calculative examination of frequency and pattern study in citation graphs of documents. It helps in analyzing generalized trends in citation links over time occurring between one or more vertices arranged into several layers and corresponding inter/intra-layer citation relationship. For instance, when a document cites another document becoming a part of the paper-paper citation graph, it reveals characteristic properties about the documents. The main objective behind doing this is to identify the top cited article and research contribution made by such articles. Other features such as citation collected by paper, citation collected by authors of paper till date, publisher, other social academic relation (such as, contribution made by research groups and other institution etc) are crucial. The field of analyzing diverse citation trajectories of paper is known as *bibliometrics*. Using citation analysis; a comparative study

measuring the impact of diverse scholarly articles without taking into account other factors which may affect citation patterns is required to be done. Among these, a recurrent one focuses on "field-dependent factors."

Author: An individual who formulates a unique problem, designs experiment and then, analyzes data and validates his finding is an author.

Author collaboration network: Starting from the late 17th century to 1920s, single authorship was followed, and the one-paper-one-author model worked well for distributing credit. Today, shared authorship or one-paper-multiple-authors model is widely practiced among all academic disciplines to enhance research standard and share resources.

The author-collaboration network is formed as a result of collaboration between an author and other authors from the same field, research group or affiliation. Two authors collaborate when they are both listed as authors in the same scientific article and hence, become co-authors.

Example includes a famous paper in high-energy physics illustrating the Large Hadron Collider, a 27-mile-long (43 km) particle accelerator that crosses the Swiss-French border; the paper included 2,926 authors collaborating from 169 research institutions.

Author – author citation / co-authorship network: When a paper is cited by another paper, the corresponding author of the cited article is getting a citation from an author of citing the article. Hence, in addition to the paper-paper citation network, it adds another author-author citation layer. Moreover, if a paper is authored by more than one author (multiple authors) then, the above-mentioned citation link is formed between groups of authors and to keep a record of that another layer referred to as author-co authorship is formed.

Publication venue: It is the source of existence of paper, i.e., conference or journal.

1.2 Issues

Quality under question? *Need for fair evaluation process:* Citation and publication of an author are considered most credible indicators to measure quality and authenticity of an author and his research work without considering the influence of location and semantics behind including it in reference of the citing article. A citation is a spontaneous process, and several factors involve such as biasness of an author to self-cite his own articles. However, peripheral is the relationship of his previous works to a current topic of interest at hand [4]. Moreover, some articles are cited so often that they become standard references for authors to cite them randomly. Thus citation, the fundamental quantifier of assessment is highly

susceptible to manipulation. Collaboration dynamics can also, negatively impact the evaluation of the publication process. Before publication, reviewer and co-author biasness result in an unfair degree of assessment in a review process. Authors overcome growing citation pressure following the easiest way through self-citations. A multi-authored paper experiences more self-citation than a research article. Peer-peer collaboration also influences an irrelevant exchange of citations. Junior faculty or new researchers are often biased to cite publications of their guide and senior researchers from the same affiliation or collaboration group. Collaboration impact in editor co-authorship networks results into adding a coercive citation to papers of their own journal. Thus, authors are inherently biased to cite papers of co-authors, collaborators or belonging to the same journal, conference, affiliation or language. Such a negative downturn can jeopardize the whole publication process and hence, progress trajectory of science [5].

Contradiction in citation based author's scoring strategies: With the exponential rise in author and publication rate; lowering subscription charges and open online multiple choices of publication venue since 2000; it has become essential to measure the performance of key entity of academic community, author by defining strict metrics. In recent years with increasing competition for limited funding resources, it is important to measure correctly the true merit of an author on comparison to the entire scientific community. Researchers in their early career require critical criticism as well as encouragement for their work. Performance measurement of authors can be broadly classified into two categories– *indexing* and *ranking strategies*. Most of the popularly used author ranking indexes such as, h-index, g-index and its other variants measure author's research impact through publication and citation count. These two parameters are highly susceptible to manipulation and hence, it needs to be avoided.

Also, not all paper by an author have equal contribution. Sometimes, it is seen in paper [6] that authors having varying bibliographic information have attained same h-index value. It is also reported by Kosmulski [7] that h-index fits for those mature author who have attained an h-index of at least ten and have published 50 papers. The problem with h-index is that new or medium ranked authors very less publication and citation count due to which majority of them obtain h-index of 0. These indices are incompetent in breaking ties among the low profile authors who are majority in number in reality. Popularly used h-index and its variants are deficient in breaking multiple ties between junior and medium ranked researchers. Ranking authors using *PageRank* based algorithm is another approach proposed

to measure an author's performance and even though, it allows multiple features to be added into the model however it adds up to further, complexities in computation. Thus, the performance of a new researcher is not adequately quantified. Therefore, the popularly used h-index measure is under question?

Lack of exhaustive citation graph analysis in large bibliographic databases: A quality research publication is a prized possession of an author which targeted at the right audience could lead to a genuine hike in quality metrics and author's relative performance. Over the last decade, visibility of papers has increased tenfold leading to an increase in the volume of the citation network. There is a need for an exhaustive study on the structure of citation pattern over time to develop a predictive model which could further, determine the nature of future citations. Through our dynamic study, we observe that citations are not collected uniformly.

Ambiguities in journal-based quality metrics: Colossal expansion of citation network since 2000 due to increased author and publication rate, digital accessibility, visibility, peer-peer collaboration propensity, multiple open choices of publication venue, journals are competing with each other to benchmark their journal, sometimes overcoming citation pressure through unethical means. A good journal should be claimed as a good journal because of publishing quality manuscripts and maintaining research standard; thus, helping an author decide where to publish. Thomson Reuters journal performance indicator, impact factor is widely used to measure research performance and compare between journals. Recent case studies reveal that in order to stand off competition, editors are forcing authors especially, junior researchers to add coercive citations to recently published articles in their journal as a condition of publication thus, artificially inflating impact factor and leading to several anomalies.

In annual citation indexing reports published by Thomson Reuters, many journals are being suppressed and blacklisted due to excessive self-citations and citation stacking, a phenomenon that reflects sudden outbreak of mutual citation exchange between two journals. In 2013, Thomson Reuters identified Brazilian citation cartel using its detection algorithm where more than one journal influenced citation statistics of each other due to the presence of underlying editor-author networks. With growing visibility and expedite growth in citation data set, there is an awareness in the research community where several cases of citation manipulation [8–11] are reported. Early detection of such anomalies and thus, redefining existing metrics is required. Strict quality assessment in the evaluation procedure is

an important issue now which was not in the pre-2000 era. Earlier, with less publication count, restricted accessibility and visibility, author or journal self- citation was a normal phenomenon, but with advances in computation, online availability, more extensive range of visibility and growing enormous network of authors and papers, excessive self-citation comes up with an intention to boost impact factor.

Peer-peer collaboration also influences the irrelevant exchange of citations[12]. In order to inflate journal metrics, editors and authors mutually participate in adding forceful citations to papers of their own journal [13]. Editor co-authorship relations also seem to bias citations for mutual benefit. Besides to boost the journal impact factor in a shorter duration of time, editors sometimes accept a large number of trivial manuscripts thereby, increasing publication volume [10]. Sometimes, editors are found to coerce authors especially target associate professors or research beginners to add forceful journal self-citations as a condition of publication. Mostly, they target papers with fewer authors so as to accost fewer researchers. Moreover, review articles are added to artificially inflate citations sometimes hidden from peer review under editorial review. Review articles receive more citations than original research publications. Commercial and profit targeted journals show these trends more in common especially in marketing, social sciences, and finance fields [8].

Review process anonymity: Peer review is the first stage of evaluation of a research article. In recent years, several conference management systems such as *Toronto Paper Matching System (TPMS)* combined with *Microsoft's Conference Management Toolkit (CMT)*, *EasyChair.com* etc are used. Toronto Paper Matching System is a standalone automated paper to reviewer recommender. It mostly generates reviewer matches based on self-declared text documents to derive field expertise and conflict of interest by both authors and the Technical Program Committee (TPC) members. However, recent practices focus more on field matching and reviewer load balancing and solely depend on self-declaration of conflict of interest. Ruling out the case of double-blind reviews, in single blind reviews, it'll be better if a reviewer has a high value of collaborative distance with all authors of a submitted paper. Also, in a real scenario in 'not so reputed' conferences, biasness is a challenging issue.

1.3 Motivation and objective

Overview: A broad objective of our research is to study the impact of several factors in quantifying author's performance (one who publish), publications (one that is published) as well as the reviewing process (fairness of selection) of accepting a paper for publication.

Scientific evaluation plays an important role in determining the development trajectory of science. A robust metric for ranking authors is essential for giving constructive criticism as well as encouragement to new authors. Such diverse and accountable evaluation metrics are much needed by the government/ non-government funding agencies, university faculty recruiters, administrators for determining promotion and tenure period of existing faculty members, departmental heads etc for evaluating return on research investment. Research grant and award recipients are also selected based on these metric. Such measures determine the current standard of research, impact of publication venues such as conferences and academic journal for augmenting research progress, significant field of study in recent years etc. Such factors help to discover new innovation and technologies thus, progressing the trend in positive direction [14].

Uniformity in author's contribution measure: To measure an author's impact in the academic community, citations should not be considered the only source of measurement. Social bonding of authors through multi-facet relationships like co-authorship, co-publishing (publishing to the same venues), co-chairing (in conferences), co-editorship (of journals), co-affiliation (in academic/research institutes), co-citation (of papers), and so on also bears significant weight to determine a fair ranking strategy for an author — besides, h-index and g-index capture either growth or saturation of scientific success of authors. They, therefore, fail to capture a decline of success. Analyzing decline of success is also simultaneously, essential for unfolding several author-centric aspects of research such as whether an author is still active in the community, how 'worthy' her recent publications are, or whether her older papers stand the test of time. Also in collaborative networks, it is seen that an author getting a citation does not always mean that he has contributed equally as his co-authors.

Citation profiling of papers into common patterns: For different fields of science, finite patterns from citation analysis have been detected. The motivation of our current research begins from a fundamental question that 'Are distributions in citations very identical for different fields of science or rather contrasting?' [15] A common accord in the literature reveals that citation dynamics seems to follow a generic universal pattern. Modelling heterogeneous citation patterns on a temporal scale, it is found that citation distributions seem to fit some of the generic scaling laws like exponential power law model and log-normal behavior. For an accurate estimation of predicting future citations, we evaluate some of the subtle aspects like initial attractiveness or increase in likelihood factor generating to linear preferential attachment models. A common trend is observed in case of applied research domain such as

Computer Science where papers published in Conferences quickly start to acquire a large volume of citation and gain active popularity and visibility within initial few years after publication, and likewise, such papers also result in experiencing an abrupt decay in citations and thus, popularity. But in the case of journal papers; though they take a considerable amount of time to get published and become noticed in the scientific network; most journal papers consistently continue to collect citations for a long time. We bring out a comparative analysis to study the influence of venue in evolving citation trajectories.

Can citation anomalies occur in Computer Science domain also? Many existing work report cases of anomalous citation instances in a domain such as business, sociology, psychology, multi-disciplinary sciences, etc. Mostly, it is reported that such anomalies are time specific and occur due to citation manipulation by editors, authors and publishers [8, 9, 11, 12, 16]. Since 2009, 'Clarivate Analytics' (Thomson Reuter's indexing firm) has started to blacklist their indexed journals involved in excessive self-cites and citation stacking cases [17, 18]. Recent literature has created awareness of unethical publication and citation practices. Here, we use a comparatively large bibliographic data set in Computer Science domain to see if we could get similar instances. In general, we try to see how citation patterns could be defined in journal-journal citation network? Further, it is challenging to dig out microscopic features that might be used to define anomalies on a macroscopic scale.

Automated paper to reviewer assignment strategy considering biasness factor: Biasness is a challenging issue specifically in 'not so reputed' conferences and single-blind review process. It is also an important factor that should be included in an automated paper to reviewer assignment strategy. However, the algorithm should also then, be computationally efficient. Top tier conferences strictly follow publication ethics and contribute the highest in maintaining research quality. But in 'not so reputed' conferences, instances of biasness may exist. Due to the lack of complete data of conference assignments, we could not study such instances. Mostly if paper submissions occur through 'EasyChair', the assignment quality would solely depend upon the self-declaration of conflict of interest by authors and TPC members.

Our main objective is to redefine the Reviewer Assignment Problem (RAP), considering the biasness factor along with field matching and reviewer load balancing. We aim to propose a basic realistic algorithm and test run our algorithm as a proof of concept on real conference data set to validate two objectives- First, whether assignment quality is improved more than manual assignments? Second, whether the algorithm is computationally efficient

or not including biasness factor? As mentioned in [19], several types of conflict of interest or biases may exist. However, while automating it is not possible to derive all such types from bibliographic data sets. We have defined 'biasness factor' in terms of collaborative distance in chapter 7.2.1. Here, we mainly derive biasness from co-authorship relations. Finally, automated assignment output can be only treated as a guideline for the program chair.

Thesis objective: In the review process, a field expert is assigned to invigilate, revise and verify the results of a research article and make necessary comments for modification of the work. Based on the comments of a reviewer, the article is accepted or rejected or sent for further modification by editorial board members of the publication venue. Hence, it is a decisive stage prior to publication, and strict quality monitoring is necessary. Several related works have proposed strategies for fair reviewer assignments considering maximum field matching between the reviewer's area of specialization and author's research work as well as the load balancing factor of a reviewer. However, our primary objective of the study is to measure biasness due to several reasons are seen in a review process. Reviewer co-author, collaborator or co-affiliation relations where, reviewer belonging to same institution or same research group is biased to accept publications blindly, however, peripheral and trivial are research works to editorial publication policies of the publication venue. Such biasness can also exist in peer-peer collaboration. Besides, junior researchers have a hand on for their acceptance of research work when the reviewer is a research guide or senior researcher from the same institution or collaboration group.

1.4 Problem formulation

Developing uniform author ranking strategy: The author is a vital entity of the research community. Based on studying finite citation patterns, a question arises that can we quantify with heuristics quality of an author? The widely accepted approach is to check the bibliographic record of a researcher that is, number and impact of publications in active citation age. Understanding the need to do a quality check on the author's performance, researchers from citation analysis and other domains proposed strategies for ranking authors mostly using publications made by corresponding authors and citations received by those publications.

Significant citation trends in bibliographic databases: Our primary objective is to study topical diversity (citation distribution, the influence of citation venues, etc.) of the publications in the Computer Science field? We try to answer a fundamental question that are there

some finite patterns in the nature of citations collected or whether citations are randomly distributed? Elementary insight in a research community is more the frequency with which an article collects citations, the more popular, visible and impactful is the research work.

Deriving anomalies in journal-journal citation network: Citation is a key unit of assessment. It can easily tamper which can jeopardize the entire research community [20, 21]. Although instances of unethical citation practices are reported in literature [8, 9, 16], general models of such citations are not defined in the form of common motifs. Here, we aim to derive some patterns where, along with self-loops, pairwise and group mutual citation occur between two or more than two journals.

Biasness measure, a key factor that needs to be considered while solving RAP: We redefine the objective function for RAP such that maximum field matching, minimum biasness and balanced workload for all reviewers occur. For maximizing the value of the objective function, we propose a greedy algorithm. First, we form two matrices which are *field matching* and *collaborative distance* matrix. Initially, those papers which have minimum average field matching percentage are assigned reviewers. Hence, papers are assigned reviewers in increasing order of their combined field similarity percentage and collaborative distance value. After each iteration, the algorithm ensures that the workload of a reviewer is balanced. Collaborative distance can be directly calculated from the conflict of interest self-declared by authors and reviewers in any CMS. However, for cross verification, we calculate collaborative distance value by extracting co-authorship relations for both reviewer and author set explicitly from a bibliographic data set.

1.5 Contribution

C-3 Index - a *PageRank* based author's performance and prospects measuring index: A fair ranking can be assigned to an author based on inter-linked citation relationships which are developed depending on the significance of his prior works. One approach for finding the research contribution of an author might be through systematically combining performance scores of the research publications they have authored. Majority of scientific research articles are written by multiple authors nowadays [22], which makes it even difficult to find individual contribution under such a shared author regime. A significant modification to the above naive strategy for combining credits from individual publications needs to be devised so that the exact contribution of an author in a multi-authored paper may be judged in an undisputed manner. Democratic view to citations may not be a decent way for computing

impact and author contribution. Citation from a Nobel Laureate or fields medallist should be given more weight-age than a citation from a research beginner. In our study, we use *PageRank* based approach where we can add multiple features and model it as a network but on the other hand, adding too many features can also increase computational complexity. *PageRank*-based methods for ranking authors use one or more form of a variety of features: citations from the peers, co-authorship with the peers, co-citations for the entities, and so on. To overcome the problem of multiple ties, we choose selective and influential features such as author-author citation impact, author co-author collaboration impact, and paper-paper citation impact. We propose a *PageRank* based multi-featured author indexing strategy where we use multiple features in a multi-layered model to assign a uniform ranking to an author.

While devising an efficient ranking strategy for measuring an author's performance, we propose a multi-layered *PageRank* based score using multiple features derived from author -author citation network, author co-author collaboration network and paper-paper citation network constructed from a massive bibliographic data set related to Computer Science domain. Given the chosen features and adoption of multi-layer *PageRank* algorithm, which can be considered as a measure of prestige, as well as a measure of significance can we model a uniform ranking strategy which can overcome the deficits of the existing metrics? Another question of curiosity arises that, are *PageRank* based ranking strategy really efficient, authentic and credible than simple citation-count based approaches? The chosen features are prominent as the structure of author co-authorship network reveals a good amount of relevant information depicting the nature and dynamics of collaboration mechanics which changes from time to time.

Citation profiling of scientific articles: While studying the paper-paper citation network, we try to evaluate using traditional models such as preferential attachment and cumulative citation distribution model alongside holistically analyzing through statistics aging factors that affect the growth of citation network and how a paper stands the test of time marking declining success in Computer Science domain. We try to derive homogeneous characteristic behavior and see examining such distributions, what factors govern how papers collect future citations in their active publication age. Through a comprehensive study of citation distributions; we bring out a comparative analysis to study the impact of venue in evolving citation trajectories. Authors take a vital interest in the journals or conferences they publish. A high impact factor journal or conference results in heightened author's per-

formance. Following publication, a general trend follows that there is an initial growth in citations till reaching a certain peak for two to three years followed by a stagnant phase, when the incoming citation rate becomes nearly constant and hence, experiences a gradual decay after which no further activity could be traced down. We further do profiling of citation trajectories into five categories based on peak detection- Early peak, late peak, multiple peaks, monotonically increasing and monotonically decreasing peak. Through an extensive literature search, we find an interesting line of well-cited papers with significant citation age that exhibits unique citation trajectories. We aim to study if unique citation trajectories such as Sleeping beauty (late recognition), Discovery papers and Hot papers are exhibited and if so in what proportion, to get a deeper insight in understanding exceptions which are difficult to model using conventional predictive models. Articles classified under 'Sleeping Beauty' or 'Revived Classics' exhibit exceptional phenomenon known as delayed recognition. Such papers remain unrecognized for several years after publication and suddenly experience citation spike with a huge volume of incoming citations thus, waking up from a long sleep also coined as hibernation period. We perform analysis on studying the behavior of well-cited papers divulging unique citation histories, i.e., the characteristic behavior of 'Sleeping Beauties', an evolution of a paper as a breakthrough discovery with altering factors like timing, chance, publishing journal and venue playing a key role. We also, study in depth citation profiles of another category of highly cited papers 'Hot Paper Publications' that accurately models heuristics of citation distributions gauging activity while taking into account both the volume of citations alongside the consistency with which they have been received. Such bibliometric indices could be further used in varied fields like measuring the impact factor, productivity and open new scope for bringing out productive researches.

We find in the Computer Science field that a total number of conference papers is more than papers published in journals. Several citation based statistical studies done on Physical Review journals indicates that in theoretical fields of science such as Quantum theory, Nuclear Physics, etc, authors tend to publish more in journals rather than conferences whereas in an applied domain such as Computer Science the scenario is entirely vice versa. Though conferences make a significant mark in this domain due to quick visibility yet while studying citation trajectories of unique well-cited papers, it is found that journals have also left a considerable impression and continue to remain a predominant choice of publication due to its steady citation growth.

Exploring citation patterns in journal-journal citation network: In this chapter, we ex-

tract citation pattern between journals in the form of a common motif from a Computer Science bibliographic data set which contains more than 2,500 journals. Apart from *self-loop*, we mostly focus on finding several patterns between two or more journals. For instance, *pairwise mutual citation* which either occur between two journals or more than two journals in the form of a chain and *group mutual citation* which occur between 3 or 4 journals and *uni-directed citations*.

On a macroscopic scale, the objective is to understand the nature of these pattern through a weighted directed graph that is, how a journal mutually interacts through citations. On a microscopic level, we try to understand time series change in impact factor. We see whether such pattern prevails for a long period or during any specific time duration. On time series data, we mostly study the occurrence of sudden peaks. Our study can narrow down the sample size by deriving such patterns. Such patterns are uniformly distributed in reputed journals also. However, we do not claim any malicious intent behind a citation as that would require statistically rigorous experiments which we tend to do as future work. We tag all such pattern with respect to six major publication houses and hence, see its influence. We devise a coupling algorithm for all such pairwise and group mutual citation instances and measure coupling strength and impact factor change on a temporal scale. Thus, in general, such grouping in journal-journal citation network during a particular period is influenced due to an abrupt increase in paper count, presence of self-referential paper, author self-citation, author co-author network, author-editor network, publication house, etc. Although a majority of the research community strictly follow a code of ethics. However, for such cases, there is an urgent need to redefine the existing bibliometrics.

Solving RAP using greedy approach by considering automated CoI: We redefine the objective function for *RAP* considering a biasness factor. We try to optimize the assignment output using a greedy matrix approximation approach. Biasness measure is calculated by finding the collaborative distance value using the co-authorship relation between reviewer and author. For reviewers, their field of expertise and co-author list is directly extracted from the hybrid data set. Similarly, for authors of submitted papers, the co-author list is extracted. Unlike other automated approaches, here all three factors are automatically derived from academic databases rather than depending on a manual declaration. For field similarity, keyword matching technique is used where a keyword from submitted papers is matched with the keyword list of reviewers. For biasness measure, the collaborative distance value is calculated (refer to chapter 7) for detailed methodology. We obtain a time complexity of

$O(m^2 * n^2)$ where m is the available number of reviewers and n is the number of submitted papers. Also, as a proof of concept, we test run our algorithm on a real data set *ICBIM 2016* collected from 'EasyChair'. Assignment quality is seen to improve as compared to manual assignments. Further, we can test run this algorithm for more data set. In the future study, we aim to use other optimization techniques and machine learning algorithms to improve the heuristics. In our current work, co-authorship relation is only used to define the biasness measure. However, other types of conflict of interest can also be derived from citation databases, where, information like author's affiliation etc is also available.

1.6 Conclusion

Throughout our study, we have tried to measure biasness and modify and redefine existing metrics. Citation forms an atomic quantifier of assessment and can be easily tampered in several ways. To understand the right credential of an author as well as paper, we need to free the entire evaluation of the publication process from biasness; otherwise, it can jeopardize the entire academic community. In our experimental outcome, we propose a *PageRank* based multi-layered author ranking metric. C-3 index takes into account three key features in determining the performance of an author: (i) the impact of citation received by papers of an author (ii) the impact of collaboration with his co-authors (iii) the impact of other authors from whom the considered author got cited. Performing statistical analysis on Computer Science citation data set, we retrieved a structured outlook in determining uniform patterns of citation distributions.

Further, we analyzed several citation patterns in journal-journal citation network. The outcome of our study may be used to narrow down the search space for finding malicious journals. Also, before publication, a fair reviewer assignment strategy is proposed. Throughout the thesis, several aspects of how an author can be rightly judged based on his merit in research career are studied. Starting from giving rank to an author in his/her early career, identifying success trajectories of paper published by the author, understanding nature of citation patterns in journal as publication venue plays a vital role in determining author's success and a bias free reviewer assignment strategy.

1.7 Summary

Citation analysis refers to statistical as well as a calculative examination of frequency and pattern study in citation graphs of documents. It helps in analyzing generalized trends in

citation links over time occurring between one or more vertices arranged into several layers and corresponding inter/intra-layer citation relationships. These have various applications, from identification of expert referees to review papers and grant proposals, to providing transparent data in support of academic merit review, tenure, and promotion decisions. This competition for limited resources may lead to ethically questionable behavior to increase citations.

Issues

* Quality under question? Need for fair evaluation process.
* Contradictory citation based author ranking strategies.
* Lack of exhaustive citation graph analysis in computer science bibliographic databases.
* Impact factor biased journal citation patterns.
* Review process anonymity, biasness measure is a challenging factor that needs to be included in an automated solution to RAP.

Motivation

* Uniformity in author's contribution measure.
* Citation profiling of papers into common patterns.
* Explore citation pattern in journals of Computer Science domain.
* Automated paper to reviewer assignment deriving biasness and other factors directly from academic databases.

Problem Formulation

* Developing uniform author ranking strategy.
* Significant citation trends in bibliographic databases.
* Extracting common citation motifs and their microscopic features such as time series change in impact factor.
* Maximum optimized assignment such that maximum field matching, minimum biasness and balanced workload for reviewers.

Contribution

* C-3 Index - *PageRank* based author's performance and prospects measure.
* Categorically citation profiling of scientific articles based upon feature sets.

∗ Identifying and justifying different citation patterns among journals.

∗ Automated CoI and greedy algorithmic approach to solve the reviewer assignment problem (RAP).

Chapter 2

Literature Survey

2.1 Introduction

Derek J. de Solla Price in his 1965 publication first emphasized inherent citation links between a network of scientific publications. Around the same year, worldwide growing dynamic citation networks is described in one of the works by Ralph Garner. Besides, Garfield and Sher validated that citation analysis could be used for generating topological maps depicting evolutionary growth of scientific topics. Later, in around 2002 an automated software was developed 'HistCite' with combined efforts of E. Garfield, A.I. Pudovkin and V.S. Istomin.

In order to keep a record of which recent works cite previous articles in the same field and establish a relationship between papers, indexing of citations in the form of bibliographic databases was started being used. E. Garfield in 1960, first introduced citation index for papers that is, 'Science Citation Index (SCI).' In 1997, CiteSeer published first automated citation indexing including only Computer Science and Information Science fields. The most widely used citation indices in academic domain include Google Scholar, Microsoft Academic CiteSeer, Scopus published by Elsevier, Web of science by citation indexing giant Thomson Reuters, etc. While some of them are available online, some can only be collected by paying subscription charges. Other than information retrieval, such citation indexing led to the advent of a new field 'Bibliometrics' where researchers became increasingly focused on measuring research quality. It also led to popularly used journal performance metric impact factor. These bibliographic databases help to extract citation graphs; thus, supplementing and detecting patterns that lead to breakthrough technological discoveries. With growing author, paper and citation rate, citation network is turning into a complex network, and

such citation index eases the task of segregating into multiple layers and inter-layer mapping between diverse entities.

In 1965, a milestone work published by Irving Sher demonstrated a correlation between the frequency of citations collected and eminence where noble prize winners and their works were cited comparatively 30-50 times more than average researchers. Such evidence claim that these bibliographic data could be used to measure the impact of not only journals or papers but also, individual authors, research groups, collaboration influences, institutions, and countries contributing to the academic community. Further, such citation impact indicators could help in ranking authors measuring based on their true credentials.

We do an extensive survey on existing literature done with citation analysis and divide it into three categories- Author centric, paper-centric and venue centric related works.

2.2 Author centric

A remarkable contribution is made by *J. E. Hirsch* who has quantified widely used author ranking metric *h-index* [23] by considering both citation and publication count of an author in equanimity. Considering several limitations of the above proposed metric, Hirsch again redefined the metric considering multiple co-authors of a given paper as an evaluation parameter. Another contribution made in this direction is by *Egghe et al.* who proposed the g-index. Moreover, a significant work is done by *Jin et al.* who proposed the *AR-index* which in some cases tend to decrease with an increase in citation of an author thus, overall credit for those researchers decrease who stand on the shoulder of their giants for a longer period. In this context, an observation is made by *Giovanni Abramo et al.* that authors who collaborate more with international research community attract more success than others but the converse may not be true.

In contrast, other PageRank based methodologies have been proposed for author and paper ranking purposes. For instance, *Chen et al.* [24] measures the impact of scientific papers considering data set of Physical Review journal. It is reported by them that choice of 0.5 as the damping factor is preferably better than conventionally used 0.85 in hyperlink network formed using web pages [25].

The ranking distribution of metrics for all authors follow power-law behavior. However, inadequate filtering for medium and low ranked authors using such metrics which are mostly based on citation-count based ranking cannot lead to fair analysis. They are also inadequate to correctly predict the future prospect of an author. Every author starts his jour-

ney with low h-index and gradually the metric should be competent that it tracks the overall evolution of an author's research over time. However, using such metrics only a handful of authors attain high h-index or g-index while others remain nearly static or at 0 h-index over a long period of time.

2.2.1 Clashes in contribution measure

Researchers working in author ranking domain consider publication count of an author and citation received by those publications as major evaluation parameter. A significant contribution is made by *J. E. Hirsch* who proposed the *h-index* [23] quantifying using both citation and publication count of an author. These metrics are in wide usage in the academic community due to ease of computation however, several limitations are also reported [26–28]. Popular alternatives to h-index include $\bar{h}-index$ (pronounced as $hbar$*-index*) [29] proposed by Hirsch, g-index proposed by Egghe et al. [30] and AR-index proposed by Jin et al. [31]. Interesting phenomenon of AR-index is that it may decrease with more citations acquired by author for those researchers who rely on their laurels for long.

H-index and its variants are precisely mathematically quantified [32] and have lower range bound [1]. However in recent times, number of researchers in every field has grown to millions. This infers that each h-index value is assigned to a large number of authors. Consequently, only few authors could attain a high h-index value and the distribution for all authors follow power law behavior (refer to Figure 4.1(a) and 4.1(c)). Moreover, there is inadequate filtering for medium and low ranked authors and for majority of such authors h-index remains static or at 0 for a long period of time.

Apart from publication and citations received by an individual author, there are other factors that impact an author's career trajectory such as venue of publication, affiliation of authors, country of origin etc. A detailed survey on Italian university system by Giovanni Abramo et al. report that authors who collaborate more with international community tend to attain more success than those who do not [33]. Also, another issue is unresolved that is, how to divide credit among authors in a multi-authored paper. In this context, Trueba et al. [34] use a concept where credit is divided among authors based on relative positioning of names of author in the paper. However, such a concept may not hold in case where the venue or domain has strict regulation on using alphabetical sequence [35] [2]. Another work

[1] http://www.webometrics.info/en/node/58
[2] https://en.wikipedia.org/wiki/Academic_authorship

by Tscharntke et al. [36] report several schemes in the context of credit contribution and its indecisiveness.

For author ranking, Ma et al. [37] reports that PageRank based calculation might be a better option replacing citation count as a major evaluation parameter of a paper. Ding et al. [38] defines two new concept popularity and prestige of a researcher and differentiates between them. Popularity of a researcher plainly refers to the number of citation received by an author whereas prestige is defined by the number of incoming citations received by an author from other high ranking authors.

Layers included in PageRank based calculation involve citation received from authors of same affiliation, co-authorship relation between authors of same affiliation, co-citation etc. Radicchi et al. [39] calculates the division of credit among author by forming a weighted citation network between authors. Barabási et al. [22] studied dynamic change in co-authorship relation of researchers over time. Ortega [40] by analyzing Microsoft Academic Search data report that the co-authorship graph of a researcher contain potential information about an author to evaluate his overall performance. Liu et al. [41] defines the concept of influence and status by studying the Digital Library academic community. Using weighted directed graph they represented collaboration networks of an author and used it as a parameter to determine an author's rank. Ding et al. [42] presented a study where they built a author co-citation graph and by varying damping factor from 0.05 to 0.95 observed its effects on the network. A detailed analysis on several variants is presented by Michal Nykl et al. [43].

How can we use more than one evaluation parameter at the same time in a PageRank based methodology? A more recent literature [44–46] proposes a new index p-index that takes two features citation network formed between paper-paper layer and co-authorship network formed between author-author layer. The cumulative scores from paper-paper citation layer is calculated using a PageRank based approach where weights are assigned to each author of a paper according to their order of authorship in a multi-authored paper. Finally, the division of credit among co-authors is calculated in form of percentile score. However, several inconsistencies exist such as order of authorship in a paper is not a valid method to determine precisely author contributions. The social network properties formed by the author co-author network are also not taken into consideration.

Modeling the impact of research considering multiple evaluation parameters at the same time in form of multi-layer network is first proposed by Cui et al. [47]. S. Boccaletti et al. [25] discusses several properties and application of multi-layer networks. It depict the inconsis-

tencies of a single layered network and how multi-layered networks can be useful in this. Halu et al. [48] defines a new term 'centrality' using random walks in multi-lared network. Domenico et al. [49] proposes tensor calculation instead of individual calculation of information from individual nodes. Calculating information from individual layer can also sometimes lead to misleading results.

Existing work also uses the concept of heterogeneous network which is very similar to multiplex networks. Zhou et al. [50] uses the same concept of heterogeneous network in form of two-layered graph for ranking author-author and paper network simultaneously. Their framework consist of random walks on graph. In a large scale based system, querying such as finding expert authors in a particular domain is similar to objectives of author ranking. Deng et al. [51] tried to model the constraints of a network as regularization constant. Similarly, Yan et al. [52] proposes a three-layer network model as heterogeneous network for author ranking purpose.

Major issues:

∗ Citation count and the number of publications that form a key parameter in h-index and related metrics can be easily manipulated. Also, only a few authors are found to obtain high h-index while a majority of the authors lie in zero h-index range. New researchers and medium ranked researchers cannot be fairly assessed.

∗ With existing metrics, often performance tie occurs between low ranked authors. A consistent conflict resolution metric is required.

∗ Credit sharing among co-authors in a multi-authored scientific paper is still an unresolved issue to determine the actual contribution of a particular researcher.
∗ While based strategy is an efficient step towards ranking authors, yet adding too many features can increase the complexity of computation.

∗ Adding multiple features in a single layer can make the whole ranking system clumsy.

Are PageRank based method more efficient in extracting information from network than

other citation-count based method? A more recent work by Fiala et al. [53] report that there is no such full proof that PageRank based method outperform other simple citation based methods. Main objective of our work draws from here whether Page Rank based method could be used in multi-layered network to extract information instead of simple calculation from citation count based methodologies.

2.3 Paper centric

The first significant contribution in this direction was made by Price [54] in 1965 where he introduced a cumulative advantage model to describe how citation distributions widely follow power-law hypothesis. Shockley in the year 1957 claimed that the rate of publication of scientific articles is governed by a log-normal form of distribution. An extensive study on scientific citation profiles to understand how citation count widely varies as a function of time has been performed by many scientists. Chakraborty et al. have claimed that most papers did not follow the traditional citation models rather they identified six different types of citation profiles based on count and position of peak depicting temporal behavior of citations received by a paper in a profile [1]. They are peakInit, peakMul, peakLate, MonDec, MonIncr and Oth. Raddichi et al. claimed that any science field could be characterized by the same citation distributions apart from the changing scaling factors. On the other hand, Waltman et al. observed to have non-universal citation distributions in quite a number of fields. Research fields including Engineering Sciences, Social sciences and Material sciences with a comparatively low average citation count per publication are found to follow such trends. Wang et al. calculated a paper's inherent fitness to measure the impact of a publication [55]. Redner's 2005 paper [56, 57] performed an exhaustive statistical analysis with 110 years of citation data of Physical Review journals. Covering a very large data set, Redner also proposed for the definition of 'Revived classics' and tried to analyze the papers that exhibited delayed recognition phenomenon [58] in Physical Review journals. The study of growing citation network models also gives evidence that the behavior and interaction of nodes strongly vary depending on its age[59].

2.3.1 Citation graph analysis and profiling

Studies are going on for a long time to understand the elementary mechanisms that drive citation dynamics. The first significant contribution in this direction was made by Price in 1965 where he introduced cumulative advantage model [54] to describe how citation distributions widely follow power-law hypothesis [56]. However more recently, results shown

by Brzezinski [60] report that power-law in citation distributions accounts for only a small range of the published articles, i.e., less than 1% for most of the science fields [61]. Several debates had undergone already among past physicists about whether citation distribution follows power-law [62], stretched exponential or log-normal form. Shockley [63] in the year 1957 claimed that the rate of publication of scientific articles was governed by a log-normal form of distribution. Redner [56] concluded that log-normal distribution fits a better range of citation distributions than power-law form [64].

Cumulative advantage [65] refers to the probability of citations that a paper receives depending on the number of citations it has already collected [66]. This concept by Barbasi and Albert has got popularized as 'Preferential attachment model' [67] which is being widely used to form topological connections in complex networks. But the traditional preferential attachment model fails to take into account the age structure of accumulated citations while analyzing citation trajectories. There are others who have tried to explain the behavior of citation distributions taking into account both these factors [68, 69]. For example, Krapivsky et al. [70] and Redner [57] introduced a new concept called 'redirection mechanism' where a randomly selected paper is likely to get cited with a probability $(1-r)$, and for one of its references the probability becomes r, where r is a randomly selected paper. However, even these models are unable to detail with the sudden burst in citations [56, 71]. Here, we extensively study the behavior of citation distributions taking into account the influence of age as well as the venue of publication and see how they affect the citation activity of publication over its total publication age.

Overall, no one could sufficiently justify the fact that the probability of received citations usually varies as a function of time and age [55, 56]. Although citation-based metrics do favor players who are present for a longer time in the field [72]; nevertheless, empirical studies show that just after initial few years of publication, papers start to collect a large number of citations which tends to decay exponentially with time [73, 74]. An extensive study on scientific citation profiles to understand how citation count widely varies as a function of time has been performed by many scientists [1, 75]. More recently, the study of growing citation network models [59, 76] also give evidence that the behavior and interaction of nodes strongly vary depending on its age [77]. However, there are others who have modulated citation dynamics with a better approach. For example, Wang et al. calculated a paper's inherent fitness to measure the impact of a research work which can be interpreted as the response of a scientific community to a particular research field [55]. But the limitations that were drawn

with the arbitrary identification and definition of sleeping beauty [58] could not justify the pattern.

The key questions that we ask in this regard are – How diverse is the Computer Science field? How have citations on this data set been distributed over time? How publication venues, i.e., papers being published in journal or conferences significantly influence the distribution of citations in an applied domain of research such as Computer Science and thus, impact citation trajectories? What is the behavior of the graph depicting the fact that 'Wealth attracts more wealth' or 'Popularity is attractive' leading to drawing possible conclusions with respect to Preferential attachment model? [78] Given a certain period of time, what are the most impacting papers that evolved as a breakthrough discovery? What is the characteristic behavior of Sleeping Beauty (SB) phenomenon with sudden citation bumps after being dormant for a long period in Computer Science? We define here the criteria for 'Hot publications' and thus, study the citation activity of hot papers in Computer Science domain.

Major issues:

∗ Paper quality measurement also requires biasness check as a paper is an elementary entity for assessing author's performance which further needs insight on characteristic growth in citation pattern (citation distribution, the influence of publication venues, etc.) which is not yet studied in the field of Computer Science.

∗ A cross-domain comparative study to scale generic behavior of citations is lacking in the literature.

2.4 Venue centric

Impact factor (IF) is a widely used evaluation parameter for journals. It is also used by Thomson-Reuters indexing firm for ranking them. Accordingly, author and editors have key interest in publishing their articles in a high impact factor journal. In late 1990's, first cases are reported where editors are asking author to cite papers of the same journal.

In the context of conference venue, the most imperial task of all is paper assignment to reviewers. In top tier conference, thousands of such submission occur annually and multi-

ple assignments need to be done within given time constraints to corresponding hundreds of reviewer belonging to Program Chair (PC). Extensive literature survey done in this direction reveal that many works and recommender system exist extracting expertise and diverse reviewer set for a given reviewer. For this, reviewer profile is constructed by extracting topic set of reviewers [79]. Then, a group assignment of reviewers to papers, which is a generalization of the classic Reviewer Assignment Problem (RAP), considering the relevance of the papers to topics as weights. A unique case study shows reviewers who are found for just one article (Journal Assignment Problem) and propose an exact algorithm which is fast in practice, as opposed to brute-force solutions [80]. For the general case of having to assign multiple papers, which is too hard to be solved exactly, a greedy algorithm that achieves a 1/2-approximation ratio compared to the exact solution.

2.4.1 Ambiguities in impact factor measure of journals

Several works have largely studied the kinetics of journal self-citations and how they are manipulated to inflate impact factors artificially. Chorus and Waltman [81] proposed a measure IFBSCP (Impact Factor Biased Self-Citation Practices), a ratio between share of self-citation count to papers published particularly in the years taken into consideration for impact factor calculation (i.e., the last two years) to the relative share of self-citations to papers published in the journal for the preceding five years. If the IFBSCP came greater than 1, it indicated an unbalanced proportion of self-citations to papers published in the journal in the impact factor years. Results reveal a majority of journals with IFBSCP > 1, though there can be legitimate reasons also for overly self-cited journals. Studying self-citation activity on a time basis [9] reveal that self-citations usually are peaked during post-publication years 0 to 7 and then, results in a sharp decay. Also, IFBSCP higher than set thresholds (1,1.5,...,3) got relatively stable from 1987 to 2004 and increased since then. Impact factor (two years) [21] has recently gained attention and is considered a prime parameter for assessing the scientific world. For the Web of Science database, [82] has shown the proportion of scholarly publications under the 'impact factor' field considerably increased during the mid-1990's. Also, [82],[81] shows that rise in impact factor is field specific. After a thorough study, observations reveal that Life Sciences have high IFBSCP values than journals in other fields taken for study, i.e., Physical and Social Sciences. We carry out our analysis to study self-citation dynamics in Computer Science domain.

Lately apart from self-loops, several anomalous citation behaviors could be tracked down

where a group of researchers frequently share an increased number of interlinks within its own community than links outside of this community in the same research field. Citation networks are continuously changing thus, leading to the identification of complex interconnections between authors, co-authors, collaborators or even third party thus, bringing out the hidden relationships. One such newly identified phenomenon is the citation cartel where a group of authors influences citation dynamics of each other to artificially boost their own performance in the scientific community by manipulating the quantifying metric of research merit, impact factor to a large margin. In 1999, this phenomenon first came into light in an essay published by G. Franck [83] who defined it as a class of journals and editors coming together in a close association for a short period to mutually inflate impact factor and influence each other's reputation in the scientific community. Fair assessment and acknowledgment to a genuine research contribution form the basis for further development in science. A far overlooked problem has emerged where through unethical citation means, the basic quantifying tool is being manipulated and left astray.

Istok Fister Jr. et al. [11] approached towards identifying cartels by finding inter-linked relationships in multi-layer networks, i.e., paper-paper and author-author citation networks using triplets "subject-predicate-object" logic in RDF (Resource Description Framework) format. Here, a particular author (subject) is connected to his/her own paper (object) by any relation say, paper p_i gets cited by paper p_j or an author A_i cites an author A_j and in turn author A_j cites A_i or paper p_k co-authored by A_i and A_j together etc. thus, a high citation flow to recently published articles can be observed between them leading to suspected citation cartels. Heneberg identifies a case on citation stacking where three journals of physics published by the same publisher Editura Academiei Romane are continuously giving a high number of cites to each other. Dramatic inflation in impact factor and JCR ranking is seen from the year 2009 to 2014 with one journal among them Romanian Reports in Physics increasing its publication count to almost double in the years taken for study. Also, it is observed that this citation network grows dense during post-publication year 1-2 and gradually, is invisible for the post-publication years 0 and 4-7. Though there was no documentation earlier, the first broadly classified case of unethical citation mingling came to light in the year 2012 when Thomson Reuters blacklisted three journals for citation stacking along with 48 journals for excessive self-citations. Out of the three medical journals, two journals 'Medical Science Monitor' and 'The Scientific World' journal never cited the third journal 'Cell Transplantation' until 2010. Each of the two journals published a review article

in its 2010 issue. For the journal 'Medical Science Monitor', the review paper cited a total of 490 articles out of which 445 were to papers published during the year 2008 and 2009 in the 'Cell Transplantation' journal and the remaining 44 were given as self-citations to papers published in its own journal in the same impact factor year window. For the journal 'The Scientific World', the review paper cited a total of 124 articles out of which 96 were to papers published in the 'Cell Transplantation' journal for the two preceding years and the remaining 26 were to papers published in its own journal in the same year window.

Major issues:

∗ No empirical study has been earlier performed considering large data set to track down anomalies in complex citation networks.

∗ There are many instances of papers published which are not at par whereas many quality research contributions get rejected. Before publication, a review process also needs strict scrutiny especially in conferences where huge manuscript submissions need to be reviewed in a limited time frame. There can be suspected manipulation in publication if there is any biasness between a reviewer and author of a paper.

2.4.2 Review process anonymity in conferences

Dumais and Nielsen [84] first considered the Reviewer Assignment Problem (RAP) as a retrieval problem, which attempts to find the reviewer who is qualified in regard and have a matching field as that of the submitted paper. The similarities were calculated using Bayesian probabilistic matrix factorization and linear regression; root mean square of bids and vector space models, interval fuzzy ontologies, but this lead to uneven distribution of workload among the reviewers [85]. For better outcome, there may be an ordered a priority among the eligible reviewers for the evaluation of a paper, but this would be a tedious task as individual assignment would be involved for each paper and would lead to a drawback stated in [86], where a hill climbing algorithm is stated to find the optimal desired value without considering the automatic reviewers group construction. Long et al. [19, 87] formulated and solved RAP as a Set-coverage Group-based Reviewer Assignment Problem (SGRAP). It tells us that if the reviewer is an expert in each and every topic included in that paper, the paper is said to

be well reviewed.

Long et al. [19, 87], consider in his process all the topics equally weighted, i.e., equally important which is not the case practically. In this method, the expertise of reviewers and the content of papers are converted into set of topics and the quality of assigning a group of reviewers $(r_i, ..., r_j)$ to paper p are assessed by the set coverage ratio, i.e.,

$$[(T_{r_i} \cup .... \cup T_{r_j}) \cap Tp]/[Tp]$$

where T_{r_i} is the set of topics in the expertise of r_i whereas Tp is the set of p's topic. Since the importance given to the topic was not practical, another approach to solving this issue was Weighted-coverage Group-based Reviewer Assignment Problem (WGRAP). In WGRAP [80], firstly focus is given on equal load distribution of the reviewers and the papers. Then secondly, on assigning papers to the reviewers which is based on reviewer's expertise as well as on the importance of the topics. The general WGRAP where multiple papers are given to multiple reviewers is an NP-hard problem and finding the exact solution for that is an expensive affair; hence, a polynomial time algorithm is proposed for that. Another problem stated in this work is Maximum Topic Coverage Paper-Reviewer Assignment (MaxTC-PRA) [88, 89], where a maximum assignment of paper to suitable reviewers is emphasized in order to cover the maximum number of distinct topics in all the papers following three constraints: 1) Paper Demand Constraint: A certain number of reviewers reviews each paper. 2) Reviewer Workload Constraint: Each reviewer reviews at least a certain number of papers. 3) Conflict of Interest (COI): There should not be any difference between the taste of the reviewer and the content of the paper.

Another significant discovery in this field is GRAPE (Global Review Assignment Processing Engine) [90, 91], which is supposed to be embedded in CMS where the reviewers will be assigned the papers of their interest. It will also judge the result of the reviewer based on quality and efficiency. These were the few important related works over the issue of the Assignment process. If we could draw out the demerits from all these works, then obviously we may lead to new ideas with a better result.

Recent work use machine learning algorithms [92–94] and recommender systems [95] for automated assignments. However, all three factors such as field expertise, conflict management, and reviewer's workload are considered taking self-declaration from both reviewer's and author side.

Major issues:

∗ Exactness in the review of a scientific article ensures better article fabrication.

∗ Depth of knowledge in the subject of a reviewer may ensure better evaluation.

∗ Partiality may occur when an article is reviewed by known persons.

∗ Excessive load on a reviewer may degrade the evaluation, so it is also important.

2.5 Conclusion

Analysis of the citation pattern of a scientific article is extensively used to measure the impact of research. All those explorations were carried out to measure the impact of three major entities, named as the paper; it's author/authors and the venue. So, we also divide our study into three broad categories for better understanding and finding.

2.6 Summary

∗ H-index is widely accepted author indexing method, but it is not fruitful for new authors. Chances are high of a tie among multiple authors.

∗ Considering citation network as a complex network and using the algorithm to calculate the importance of nodes (authors, papers or venue) is a good idea, but implementation challenge is there.

∗ Impact of a venue (journals/conferences) is derived by the collaborative impact of papers they accepted. So, understanding the pattern of citation count of a paper may be interesting.

∗ It can be possible to dilute the actual impact of the venue in different ways. Few researchers put light on those areas also.

Chapter 3

Experimental set-up

3.1 Introduction

Experiments conducted on scientometric research and analysis require a data set. We use openly available bibliographic data sets for different problems. Further for some experiments, we use standalone systems. Moreover, for few computationally exhaustive experiments, we used a distributed cluster environment. Details of lab setup are given in section 3.3.

Many bibliographic data sets are available of which most frequently used includes Microsoft Academic Graph (MAG), ArnetMiner data set, DBLP, Web of Science, Scopus, Google Scholar, etc. Some of these data sets are freely accessible online whereas; some can only be downloaded by payment. Our major issue in data set collection and processing includes funding restrictions and limitation of resources. For related reasons, we collect freely available citation data sets and extract only those fields and its related information as per our problem formulation. Some of these data sets are incompetent with incomplete information; thus, not fulfilling our research requirements. As a result of which, we crawl information about missing fields from the website and manually complete it.

Collection and processing of data play a pivotal role in undergoing any research work. Due to the presence of diverse key entities such as paper, author, journal, domain, discipline, author's affiliation, institution, research group, cities and countries in our data set; a large citation network is formed which conclusively could not be represented as a directed graph, but forms a dynamic complex network. Over the years, there has been a progressive increase in paper and author rate. As a result of which, bibliographic data sets have also

enormously grown in volume. Besides, the citation relationship (edges) between several entities (nodes) mentioned above are changing dynamically over time. Not all data sets keep a detailed record. Consequently, to handle and process such data sets, advanced computation and processing speed of systems are required which becomes a drawback in many cases. We propose to do an extensive study of such a complex network from several dimensions by computing, analyzing and using single and multi-layered modelling of the network.

Initially, when we started our research we could not find complete data and required fields in free online available data set ArnetMiner V5 [1] as required for our first experiment which consists of 1,572,277 papers and 2,084,019 citation relationships. So, we crawl incomplete information from Microsoft Academic Search (MAS) website and form our hybrid data set. Our hybrid data set completes several incompetencies of ArnetMiner V5 data set and is 4 GB in size with publications till 2012. An even more stable and complete version was released later, ArnetMiner V7. Compared to above two data set, Microsoft Academic Graph (MAG) which was freely accessible for a short period of time is enormous (100 GB) in size and consists of many additional fields with a complete record. MAG data set was collected in 05/02/2016 but mostly after addressing three of the major problems addressed in this thesis. Also, due to the limitation of resources we could use this data set only for one of our works. Till date, no conference data set gives detailed information about year wise instance of reviewer list and records including assignment of papers to reviewers for different publication venue that is, journals or conferences. As a proof of concept for our proposed algorithm and automated conflict management in conferences, we collect data set of *ICBIM 2016* from *'EasyChair.'*

This chapter is arranged as, section 3.2 describes data set collection, filtering, and characteristics of four different chosen data set. Section 3.3 describes lab setup and resource specification.

3.2 Data set collection, filtering, and characteristics

All proposed research is to be performed on Computer Science bibliographic data set. Data set are either collected online with free access or crawled and manually completed from different sources. Though there are several drawbacks in each of them such as citations to cross-domain publications are missing, incomplete author and venue details for some articles, missing author co-citation or co-authorship records, irrelevant dangling references to

[1] https://aminer.org/billboard/citation

articles that do not exist in our data set. Combining all data set collected, we work with following five entities: paper, author, publication venue (journal and conference series), event (instances of different conference series) and field of study of a paper. For performing our research on author and venue centric related issues, we use four different data set that include- Hybrid data set, ArnetMiner (V7) data set, Microsoft Academic Graph (MAG) and data set including several event conference instances with a detailed list of reviewer and assignment records between reviewer and paper list.

3.2.1 Hybrid data set

When we begin our study, we download the *ArnetMiner V5* data set and find it incomplete. We use manual web crawling to fill in the incomplete field and their details. We directly crawl data from 'Microsoft Academic Search (MAS)' website which contains latest update of publication details. The paper ID's are extracted from the rank list provided by MAS. Next, meta data of papers are extracted corresponding to unique paper ID's. Tor is used to distribute the load to multi-processor system. It took six weeks to crawl all details related to 6 million papers [96, 97].

Table 3.1: Hybrid data set characteristics

	Raw	Filtered
Number of valid entries	6,473,171	5,549,317
Number of entries with no venue	343,090	–
Number of entries with no author	45,551	–
Number of entries with no publication year	191,864	–
Number of authors	4,186,412	3,186,412
Avg. number of papers per author	5.18	5.04
Avg. number of authors per paper	2.49	2.67
Number of unique publication venues	6,143	5,938

After crawling all paper with insufficient details such as lack of unique paper ID's, publication year, author list, venue etc are removed. Few spurious forward citation to papers published after the source paper cited have also been obtained and removed. Papers published in the range between 1950-2012 are only considered that have at least one incoming or outgoing citation. Remaining subset of data consists of 5 million papers and the data size

is 4GB. Spurious references to papers not present in the data set are removed. Exact data set information is shown in Table 3.1.

3.2.2 ArnetMiner

With its next release which is more stable and complete, we downloaded ArnetMiner (V7) data set freely available online. We are using data set developed by Tang et al. in connection with [98] and which is available freely in ArnetMiner Project[2]. The original data set has 2,244,021 papers and 4,354,534 citation relationships. We observed some inconsistencies in the data set in the form of one or more missing data fields. Those data elements that have missing fields (except the abstract field of a particular publication) are not usable in our current study and hence, are filtered out from the data set. One particular information component missing in the above data set is the field/domain of research a particular paper may be associated, which is an important consideration in our present work. data set prepared by Chakraborty et al [96] is used for finding the field of interest of research for a particular author and augment the same for the authors in our data set. After initial processing and filtering, details about the fields of ArnetMiner data set is as below:

ArnetMiner citation data set (V7)

PIndex : Unique paper id for each paper

PTitle : Paper title

PAuthor: Comma separated list of authors

PYear : The year of publishing of paper

PConf : Venue of publishing (not used yet)

PRef : References of the paper (comma separated)

This data set consists of publications ranging from the year 1936 to September 2013 (i.e., 77 years). There are in total 17,524 invalid authors, 4,78,428 invalid citations, three invalid years and 66 multiple invalid fields. Maximum multiple author count for a single paper is 119, and maximum citations to a paper are 51,373, a maximum publication by an author is 1,100 and that, the author is Wei Wang, maximum citation for an author is 120,637, and that author is David E. Goldberg, maximum co-authors for an author is 1,575, and that author is Wei Wang, maximum publication in a year is 1,46,030 that, is in the year 2011. Average publication per author is 1.670856565, and average citation per paper is 22.91131.

[2]https://aminer.org/billboard/citation

Table 3.2: ArnetMiner data set features

Attribute	Count
Number of publications	1,747,995
Number of authors	1,046,167
Number of conferences	8525
Average publication per author	1.670856565
Average citation per paper	22.91131

3.2.3 Microsoft Academic Graph (MAG)

We downloaded Microsoft Academic Graph (MAG) data set which was freely available only for a very short duration of time in 05/02/2016, and the zipped file is 28.2 GB in size [99]. It consists of total 126,909,021 publications, 114,698,044 authors and 528,682,289 citations. Microsoft Academic Graph (MAG) contains a detailed bibliographic history with recorded list of each version of publication volume, issue, year, full citation list, complete author's profile (affiliation, address, website and publication list) and list of journals / conferences including details of publication list, year range, citation and paper count. Almost every procedure following scholarly communication and its entities are given in details in the form of a heterogeneous graph in this data set, and after unzipping, we get a total of 12 fields. Citation relationships or edges between diverse entities or nodes are intuitive and justified as a given paper is published by a single or multiple authors and gets published in journal/conference [100]. Although this data set has a vast scope of a research study; due to the limitation of resources, we only used paper-paper citation network and performed data processing and statistical analysis to determine finite citation patterns over time. Out of the extracted 12 entities, we only used paper, paper references, journals, conferences, and paper authors files. After unzipping and initial processing of MAG data set, we get the data dictionaries as below:

MAG data dictionaries

Affiliations (288 KB)

Authors (1.3 GB)

Conference instances (3.3 MB)

Conferences (30 KB)

Fields of study (695 KB)

Field of study hierarchy (2.12 MB)
Journals (356 KB)
Paper authors/affiliations (4.01 GB)
Paper keywords (1.8 GB)
Paper references (3.38 GB)
Papers (9.05 GB)
Paper URL's (summary, abstract, pdf etc) (8.69 GB).

Table 3.3: MAG data set physiognomies

Attribute	Count
Number of publications	1,269,909,021
Number of authors	114,698,044
Number of journals	23,404
Number of journal papers	51,900,106
Number of conference papers	75,008,519
Number of citations	528,682,289
Avg. number of papers per author	1.1064619463
Avg. number of authors per paper	0.9037816469

3.2.4 Data set for reviewer assignment problem

We collect a real conference data set of *ICBIM 2016* from 'EasyChair.' It is used to test run our proposed algorithm for doing automated assignments. The details of the data set are given in chapter 7. Further, we extract factors required to solve *RAP* from 'hybrid data set.'

Further, while performing venue centric study and measuring the credibility of widely used impact factor metric, we propose a single layer study of journal-journal citation network coupling together paper-paper citation network where journals represent nodes and citation edges correspond to cite/being cited relationship between journals.

3.3 Resource configuration and experimental set-up

In this section, we describe the hardware and network setup used to run our experiments. For most of the experiments, we use a standalone system whose detailed specification are as given in section 3.3.1. Some experiments which include implementing *PageRank* algorithm

are resource hungry. In section 3.3.2, we describe the steps to set up a distributed MPI cluster, for which we used mpich 3.2 library.

3.3.1 Standalone system

PROCESSOR
Intel® Xeon® Processor E7-4850 v4
* 40 MB Cache
* 16 Core
* 32 Threads
* 2.80 GHz Max Turbo Frequency

MEMORY
* 192 GB via 6 DIMMs at 2666 MHz

OPERATING SYSTEM
* Red Hat® Enterprise Linux® 7 Server Standard (1-2 Sockets) (Up to 1 Guest)

STORAGE
* 4 x 2.0TB SATA 6.0Gb/s 7200RPM - 2.5" - Seagate Exos 7E2000 Series (512e)

ETHERNET
* Intel® 10-Gigabit Ethernet Converged Network Adapter X520-DA2 (2x SFP+)

3.3.2 Cluster platform

We deployed parallel computing concept in an existing network of our computer laboratory. The laboratory consists of 60 nodes interconnected in *star topology*. Entire process of setting up the cluster environment to execute our experiments are described in different steps as below:

3.3.2.1 Resource configuration

1. **Node:**
   OS- Linux (Ubuntu 16.04 LTS)
   No. of cores- 2

Main memory- 2 GB RAM
Processor speed- 2.67 GHz

2. **Switch:**
DLINK
24 ports
100 BASE TX
DAX

3.3.2.2 Pre-requisites before setup

Before setting up MPI cluster, we need to identify several characteristics of network such as topology, latency, and bandwidth which directly determine the performance of several MPI collective functions. Further, we also try to propose a framework for doing all the experimental study.

Identifying topology of MPI cluster: We aim to find network topology of the cluster using MPI_Send function. We keep the number of nodes constant that is, 60. We assign a number of ranks or processes per node as 2. Thus, the total number of MPI processes are 120. Next, we run 60 experiments by sending messages from a fixed rank 1 to all even ranks (0, 2, 4...120). Simultaneously, we vary message size from 800 bytes to $8 * 10^8$ bytes. We find that ideally for given message size, transmission time from a fixed source to all even ranked destination processes are same. Therefore, we can conclude that topology of this network is *star topology* where the cost for any node to communicate with any other node in the network is same for fixed message size. All nodes are equidistant connected centrally by a switch in LAN.

Finding latency and bandwidth of MPI cluster: Latency and bandwidth are fundamental characteristics of the network. Simply, latency is the time it takes for a message to travel from its point of origin to the point of destination. Latency is measured point to point. It is the sum total of all delays that is, propagation delay, transmission delay, processing delay and queuing delay. We calculate mean latency by sending 0-byte messages using MPI_Send between a pair of processes over multiple iterations. If we send a 0-byte message, then the consumed time will be the latency. We get an average latency of 3 microseconds.

Bandwidth refers to a maximum number of bytes that can be sent through a physical or logical communication channel in 1 second. For bandwidth calculation, we use iperf tool. Using command $iperf - s - w1024k - i2$, we run one node as server and with $iperf - c <$

$IP > -i2 - t20 - w1024k - P4$ we run all other nodes as client and measure bandwidth one by one. Here, p refers to the number of processes which is changed as 1, 2, 4, 8, etc. We collect bandwidth readings from all client nodes and calculate average bandwidth over multiple iterations as 93.875 Mbits/second.

Proposed framework Using MPI_Wtime() function which returns elapsed time on the calling process, we time several MPI point to point operations that is, blocking send and receive (MPI_Send/MPI_Recv) and MPI collective operations (MPI_Bcast, MPI_Scatter, MPI_-Gather, MPI_Reduce). Keeping the number of processes per node as 2 (constant) initially, we vary the number of nodes as 8, 16 and 24. We vary message sizes from 800 bytes to $8 * 10^8$ bytes and run 10 experiments for each message size. At last, we calculate the average time from 10 iterations for each message size and MPI function.

3.3.2.3 Steps for MPI cluster setup

Pre-requisite to set up MPI cluster is that all nodes should be installed with Linux. It is the most commonly used OS in HPC community. For our cluster, we install Ubuntu (16.04) on a total of 60 nodes. Out of which, we chose one node as a 'Master' node and all remaining nodes as 'Client' node. In a typical small HPC cluster, any node can be the login node that is, a node which is used to write or run parallel MPI codes. All nodes should have root permissions and be able to ssh to each other without passwords. Following steps should be followed for setting up an MPI cluster:

1. *Configure hosts file:*

   While running parallel programs, we need nodes communicating between them, and it'll not be appropriate to type ip addresses so often. Instead, we give a name to all nodes connected over the network. 'hosts' file is used by the devices operating system to map hostnames to ip addresses.

   Master ip - 10.0.22.31, Master host name - ub31.

   Client ip - 10.0.22.2 - 10.0.22.61, Client host name – [ub02 – ub61].

   For defining the hostnames, we open /etc/hosts file and mention server and client's ip as well as hostname.

   sudo gedit /etc/hosts

   127.0.0.1 local host

10.0.22.31 ub31

10.0.22.32 ub32

...

2. *NFS Installation:*

   NFS allows the system to share directories and other contents of the shared "mirror" folder with all other nodes in the network. Using NFS, we could share a directory via NFS in master which the client mounts to exchange data.

   (a) Master side configuration:

      i. We have to install nfs server in Master node by using the following command:
      $root@Master$ **sudo apt-get install nfs-server**

      ii. Next, we create one folder 'mirror' in all nodes (on the system directory) to store our allocated task and programs accessible by all nodes.
      $root@Master$ **sudo mkdir /mirror**
      $root@Client1$ **sudo mkdir /mirror**
      $root@Client2$ **sudo mkdir /mirror**

      iii. Now, to share all the data in a folder (mirror) of master node with all other nodes, we have to synchronize this folder with all nodes editing /etc/exports file on the master node. We set all permissions (read/ write) for this folder so that it becomes accessible from all the client nodes by folder.
      $root@Master$ **sudo gedit /etc/exports**

      In the exports file, * for making /mirror folder accessible by all nodes. We can also mention particular ip addresses. We add this line in the exports file in the master node.
      /mirror *(rw, sync)

   There are several options we can use among the following-

      i. rw: This is to enable both read and write option. ro is for read-only.

      ii. sync: This applies changes to the shared directory only after changes are committed.

iii. *no_subtree_check* : This option prevents the subtree checking. When a shared directory is the subdirectory of a larger filesystem, nfs performs scans of every directory above it, in order to verify its permissions and details. Disabling the subtree check may increase the reliability of NFS, but reduce security.

iv. *no_root_squash* : This allows root account to connect to the folder.
root@Master **sudo chmod –R 777 /mirror**

(b) Client side configuration

i. Required packages for the client machine is being installed.
*root@Client*1 **sudo apt-get install nfs-client**
*root@Client*2 **sudo apt-get install nfs-client**

ii. Mounting the shared 'mirror' directory from master to all clients. Mounting means making directories or files available to the operating system for further use. For permanent mounting (on each reboot) we have to edit in /etc/fstab file.
root@Master **sudo mount ub31:/mirror /mirror**
root@Master **cd /etc/fstab**

iii. Edit this line and save it in all slave nodes
Master:/test /test nfs

iv. After making above changes, we need to restart nfs server.
root@Master **sudo service nfs-kernel-server restart**

v. We can check whether the folder is mounted using
df –h

3. *Setting up passwordless SSH* Our nodes will be talking over the network via SSH and share data via NFS. To communicate with the process running on different nodes, the system requires certain authentication or login mode. So, we require passwordless login between master and slave nodes. The functionality of SSH will be helpful in doing this. In SSH, we have to share the key between master and client nodes. For that, we use RSA or security algorithm to generate keys. The steps are as follows:

(a) We install open ssh server on all nodes.
root@Master **sudo apt-get install openssh-server**

(b) In ssh configuration file $/etc/ssh/sshd\_config$, we change password authentication to 'No'.

(c) Next, we generate RSA key.

$root@Master$ **ssh-keygen –t rsa**

(d) Next, we make a directory .ssh on the client system from the master.

$root@Master$ **sudo ssh client name@clientip mkdir –p .ssh**

(e) After creating this key on home directory /.ssh at master node , we have to copy this public key to another client nodes $authorized\_keys$ file safely using this command:

$root@Master$ **cat .ssh/id_rsa.pub | sshclient name@clientip**
'cat >> .ssh/ authorized_keys'

(f) We set permissions for the above folder on client machine.

$root@Master$ **ssh client name@clientip 'chmod 700.ssh;**

chmod 640 .ssh/ authorized_keys'

(g) Next, we test whether the passwordless ssh is setup or not by login password less from server to client.

$root@Master$ **ssh bcr@10.0.22.32**

4. *Setting up a machine file* We create a file /etc/machinefile and mention hostnames of the nodes followed by a colon which refers to the number of processes to spawn on each node. Our processors in the cluster are all dual-core machines thus; minimum we can have 2 processors on a single node.

   ub31: 4
   ub32: 3
   ub33: 4
   ....

5. *Installing gcc compiler* As we know that the cluster is designed for high-performance computing and for that, we have to compile special purpose code by some compiler (SGI compiler, FORTRAN) at the master node which can convert source code into executable code. These codes divide into several tasks, and this task will be distributed

among all slave nodes for parallel processing. Hence, we install GCC to compile all code on the master node and other necessary stuff by installing the build-essential package:

$root@Master$ **sudo apt-get install build-essential**

6. *Installing mpich library (version 3.2)* Now at the end of the installation steps lastly we have to install important package of a cluster to pass message among the nodes to perform the task, allocate the task and submit output back to master node, so we have to install MPI library.

   $root@Master$ **sudo apt install mpich**

   To test,

   which mpiexec

   which mpirun

7. *Compiling and running mpi codes* Next, we compile and run the mpi codes using the following commands:

   $root@Master$ **mpicc –o mpi_hello mpi_hello.c**

   $root@Master$ **mpiexec –n 8 –f machinefile ./mpi_hello**

3.4 Conclusion

Citation network is formed by considering papers, authors and venues as node and relationship among them as an edge. Different kinds of relationship occur among them; at the same time numbers of nodes can also be numerous. So, we consider it as a complex network and try to justify different complex network algorithms at different stages.

Various citation network data set is available online to explore, among them ArnetMiner is quite popular for its open access but comparatively smaller than Microsoft Academic Search Bibliography data set. According to our requirement, we develop two different data set also. One is for reviewer assignment problem as no such data set available for open access, and the other is pure citation data set with some extra fields according to our requirement.

3.5 Summary

∗ Keeping in mind the complexity, we divide the citation network into three different layers. Layers are interconnected, so we executed experiments in the entire network but at times in different layers individually also.

∗ Total four data set we used to carry out the complete process because of their different characteristics and relationship.

∗ Microsoft Academic Graph bibliographic data set is the largest in terms of volume and entries and reviewer assignment data set is the smallest as it is almost manually completed.

∗ Our hybrid data set is also created by crawled data from Microsoft Academic Graph, Google Scholar and a few other websites.

∗ We filtered ArnetMiner data set as there were few invalid entries then extract required fields to create data dictionaries according to our usefulness.

∗ For most of the experiments, we use a standalone system. Some experiments which include implementing *PageRank* algorithm are resource hungry. We have used MPI platform to setup cluster and do parallel processing.

Chapter 4

Influential factors behind an author's performance and prospect

4.1 Introduction

In order to have a proper track of information about the advancement of science and the most prominent researches being done in various emerging fields over the years, it has become extremely crucial to measure the quality of science. Now a day's publications of research papers have increased in quantitative terms, but it might happen that only a few among them may actually contribute to the cause of solving a major drawback and thus, leading to becoming a groundbreaking discovery sometimes. Moreover, scientists around the world from different research backgrounds are competing hard for research funding. So, it has become extremely important to recognize and acknowledge high-quality researches and thus, promote committees, funding agencies, national research and academic universities and government to acknowledge the work of prominent researchers. Although it is not possible to standardize a metric as an optimal and relevant measure of science yet the widely accepted approach to judge qualitatively is to check the bibliographic record of a researcher, i.e., not only to focus on the number of publications but also its impact in terms of application and validation. Researches with varied bibliographic credentials may have the same h – index [6]. Kosmulski correctly pointed out that h- index is more suitable for assessment of those scientists who have published at least 50 papers and had h- index of at least ten [7]. To cite an example, suppose two researchers have two publications each. The publications of the first researcher have 100 citations each, whereas the second researcher has only four

citations, 2 for each publication. Unfortunately, both researchers receive h-index of 2. This example points out that h-index performs poorly for young researchers with few publications, which has been well-documented [101].

Careful analysis and assessment of research work are much required these days. Most certainly, because it'll help to visualize and track the present day construct of science, how some of the leading research fields have exhibited patterns of growth over the years, what are the trending and authoritative research areas that have led to new discoveries, how new fields have evolved which might have remained dormant over the years and so on. It also helps to create a significant background about the impact of various academic journals. Also, such an assessment could guide the beginners who are new in this field with valuable feedback and thus, comment on their proficiency in the depth of particular research work and also areas of improvement in future. It also gives them an aggregate position in the scientific world which would indicate their current performance. Such analysis can be beneficial for recruiting committees in various research institutes and national academies while hiring new researchers as a performance graph will be in front of them, and this would ease up the sorting of the most deserved candidate. In the case of recruiting for departmental positions, such metrics may lead to accession and allocation of prerequisite resources. For some cases, it might also be used for selecting fitting candidates for promotion and tenure. While the selection of award recipients in scientific societies, such metrics could pave a headway for acknowledging and felicitating prominent researches who have been involved in impactful and dynamic researches. An accountable measure of science is required by various governments, legislative bodies, political organizations and board of directors so that they can evaluate by documenting performance based on its effective application, scientific productivity, and range of implementation and thus, calculating the net returns expected on making a research investment. Those agencies which are involved in research funding; such metrics of analysis could direct them towards funding in more emerging areas of researches.

Analysis of publication trajectory of authors from the faculty of Psychology during their first 7-postdoctoral years revealed that the publication rate increased annually only after the completion of their doctorate program [102]. It is obvious that the publication count preceding the tenure period doesn't tell the whole story. Though, it was observed that the publication rate gradually increased reaching at the peak during the first four years rather in the two years that was immediately before tenure. Another examination done by Long [87] on measuring scientific productivity based on gender differences revealed that in the initial ten

years of their career, women publish fewer articles than men, but this anomaly is reversed later on in their careers. According to research done on ISI Database, John Ridley Stroop could only publish three papers during his career. One of the articles among them was cited for 3,810 times, whereas the other two received less than 1% of the citations as compared to the former paper [103]. Although total citation count is necessary and can be used as a metric to evaluate an author's scientific excellence, but in this case, Stroops involvement in contributions to Psychology was only limited to the period prior to the completion of his Ph.D. It comes as news to no one that if a candidate for tenure has no or very few scientific publications, the assessment of the candidate's science is likely to be bleak.

Several studies have been conducted by formulating scientific progress in terms of networks (such as citation network, co-authorship/collaboration network) [40, 41]. Studies on co-authorship networks focus on network topology and statistical network mechanics [45]. Although our research also deals with citation and collaboration networks, we take a different approach by studying micro-level network properties, with the aim of applying centrality measures such as PageRank for impact analysis [39, 104–106]. Collaboration is a fundamental aspect and to determine one's pragmatic attitude towards collaboration [107] is one of our prime motives.

While grading a research work based on citation analysis, the impact or effectiveness of a scientific publication is only measured depending upon the number of citations. The drawback that arises here is that even when an obscure paper is manipulated with a large number of citations, it receives the same weight as to one which has genuinely done some groundbreaking research as for both the citations received are measured to be equal.

Corresponding to the bibliometric area, Pinski and Narin were the first to mark the differences between popularity and prestige [108]. They proposed a similar algorithm like PageRank where the principalEigen value corresponding to the Eigenvector of the journal citation matrix was used to denote the journal prestige. Also, Bollen et al. [109] designed a weighted PageRank algorithm and defined the popularity and prestige of journals recursively. According to his speculations, popular journals were those which had been cited more frequently by less prestigious journals whereas, on the other hand, prestigious journals were defined as those which had been cited more by the highly- prestigious journals. More recently, Ding and Cronin[38] have revised and protracted this approach by using weighted citation count to measure an author's prestige. Here, they redefined popularity as the total number of times an author is cited and prestige as the number of times an author is cited by highly cited pa-

pers. The major focus while calculating the prestige of an author was given to the fact that use simple citation count method but the highly cited papers must receive more weights.

Being a scientific author is brilliant, but becoming a star author is marvelous. Devising a mechanism to find the star of the present is great: may be useful for selecting the best candidates for recruitment in a university, or, for choosing the best person in a discipline for best researcher award. But formulating a method to find a future star is overwhelming, and would be very useful to a research funding agency for awarding the research grants among the applicants. And to come up with a ranking strategy that would describe the present as well as the future; is simply outstanding. The challenges are manifold – (i) to find categorically the features that influence a scientific author in present and in future, (ii) to design a unified strategy to combine those factors in a meaningful way that is universally acceptable, (iii) to devise an efficient and effective way of computation to present the *'generic rank'* of an author in a systematic way, and (iv) to capture the dynamic behaviour of a scientific authors' productivity. With the advent of web-based publication of scientific articles, the volume of scientific literary components, in the form of number of publications, number of authors, number of venues of publication (journals/conferences) and so on, expand manifold in all the discipline of scientific research, which compels the introduction of computerized ranking mechanisms and tools for ranking of papers, authors, journals, conferences etc. In this article, we shall study the feasibility and scope of constructing such a ranking mechanism, and will propose a ranking strategy that may serve as a flag-bearer towards this direction.

Our main objective is to determine whether a better author ranking can be obtained using journal values, was achieved. The best of our author ranking systems were obtained by using journal impact values in PageRank, which was applied to a citation network of publications. The effectiveness of the ranking system was confirmed after calculations were carried out involving authors who were awarded after the final year used in our data set or who were awarded in selected categories.

In our paper we take a completely different approach and try to depict all the possible linkages between author-author, author - co-author and paper-paper citations and see that analysing such a network topology provides a better alternative to measure the impact and quality of research work not only among the highly cited authors but also between medium and low profile authors who are majority in number. Also, such a networking analysis proves to be an effective tool while ranking papers, authors or journals in qualitative terms. The recent progression in using PageRank algorithm in scholarly data articles has proven itself to

be quite effective. Although weighted PageRank has been used quite effectively in measuring the prestige of journals, the concept has very rarely moved to the direction of extending to authors as well. The idea extended upon the contributions made by Ding and Cronin implements how PageRank or weighted PageRank algorithms could be used in author citation networks to measure the two parameters of impact analysis, i.e., popularity and prestige [38].

Large scale network analysis using PageRank has proved to be of great success. Weighted PageRank methodology [109, 110] is found helpful in determining the impact of journal. However, only a few research work exist which have applied the same concept to author ranking domain. Major issues encountered in this problem are discussed in the following set of work Fiala et al., Ding, Radicchi, et al., Życzkowski, etc. The following set of work follows after a paper is published by Ding and Cronin [38] which discuss the issue of use of author citation network for measuring the impact of researchers. In this work, we construct a network from a large bibliographic data set in Computer Science domain between several entities such as authors, publications, journals etc Our main objective is to do rank author based on the quality of their publication correctly using implementation of PageRank algorithm.

In this work, we quantify a new author ranking metric $C^3-index$ mainly based on quality of publication, citation and collaboration diversity of an author. The components are considered in a multi-layer network model and it is seen that one of the components (ACI-score) has 98% correlation with the widely used metric h-index. Proposed metric containing other two component score *PCI-score and AAI-score* brings more insight than h-index. We also obtain a set of authors attaining high AAI-score during 1998-2008 tend to attain high h-index during the later years. Thus, this metric is capable enough to predict future performance of an author. In a time window of 4-5 years the proposed metric can capture the growth of an author's performance for them later the h-index value also tends to increase.

4.2 Motivation and objective

Existing study uses the following feature for ranking authors such as, (a) author-author citation [38], (b) author-author co-citation [42], (c) author-author co-authorship [40, 41], (d) author-author collaboration [1] [111] and (e) paper-paper citation network [23, 29, 30].

[1] A super-set graph constructed from author-author co-authorship network which takes into account social relation between authors other than co-authorship. For instance, authors from same affiliation, committee members participated in same conference, mutual friends in any social media group, editors of same journal, etc. However, this feature could not be used frequently due to lack of suitable data set.

Citations received by a paper possibly determine how influential is the research work of an author in bibliographic domain. As a result of which a plethora of work exist which only use the *paper-paper citation network* to rank authors. However, not all papers of an author have equal contribution. Hence, it is challenging to combine individual paper scores to get a consistent author score. Alternatively, other author network such as, *author-author citation network* or *author-author co-citation network* could be used to rank authors. Here, *author-author co-citation network* refer to two authors being connected by an edge if they cite the same paper. An intuitive justification might be that authors researching on the same topical domain usually read and refer to the same set of papers.

Several flaws add up if citation-based scoring technique are directly used for ranking author. Initially, after publication it is seen that it takes some time to get noticed by the academic community and collect citation [1]. Due to which recently published articles are wrongly evaluated if they are ranked only on the basis of citation count. Moreover, the same limitation is seen in case of novice author. They may not collect citation and acquire attention in the early days of their career. Also, the time required between communication of a paper to venue and its actual publication is also not short. Due to above limitations in citation-based scoring technique, other features such as *author-author co-author network and author-author collaboration network* [112] need to be explored. Intuitive reason behind this might be that higher ranking author usually collaborate with other high ranking author present either as co-author or social collaborator.

Consequently, co-author network and social collaboration into research group has a major impact in determining author's success. An eminent author belonging to top institution, appointed as a Technical Program Committee (TPC) member in a good conference, editor in a prestigious journal etc influence largely research of his/her co-author or collaborator. Hence, it is quite intuitive to exploit such social relationship to derive the influence of a given author with respect to entire author network. However, author-author citation network can be directly derived from paper-paper citation network but the latter could not be replaced by the former. Hence, we cannot completely remove the paper-paper citation network. A better approach would be to devise a strategy which scores an author on the basis of above mentioned feature.

Each feature in our data set can be represented as a complex network. Combining more than one feature leads to a multi-layer complex network and applying Page Rank like algorithm into it is quite cost-inefficient. Hence, redundancy and multi-collinearity in feature

set should be removed. Considering only the most required features, we try to reduce the dimensionality of feature set.

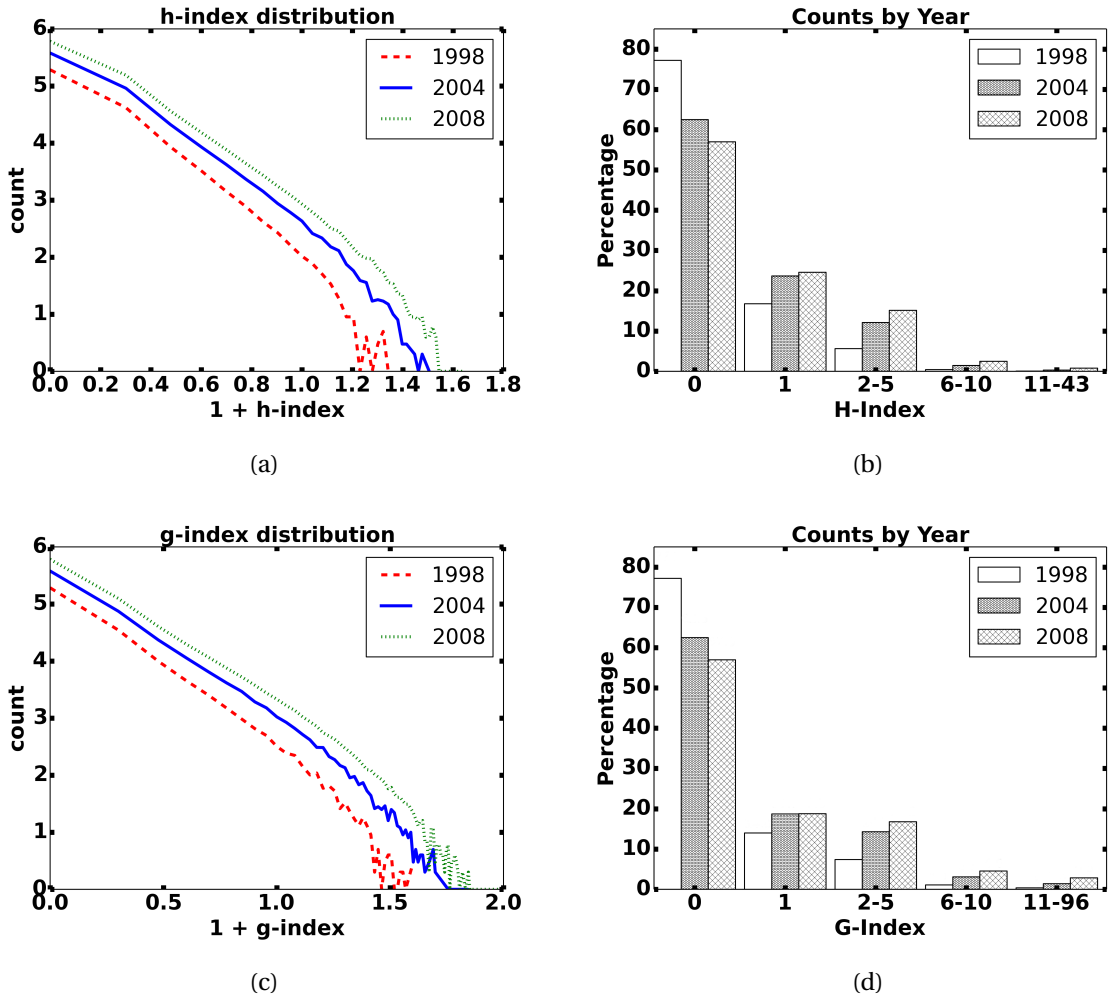


Figure 4.1: **(a)** Number of authors are plotted against different h-index values (plus one) in log-log scale for three different years: 1998, 2004 and 2008. **(b)** Percentage of authors as distributed across five different h-index bins for the year 1998 (left bars), distribution of the same set of authors in 2004 (middle bars), and in 2008 (rightmost bars). The figure reflects that a limited fragment of authors attains high h-index over the years, but majority remains unimproved. **(c)** and **(d)** show the similar plots against g-index. The near straight line nature of all the curves in **(a)** and **(c)** ensures power-law behavior of both h- and g-index. **(b)** and **(d)** suggest that a small fragment of authors having low index values gradually improve over the years, whereas the majority remain unchanged. It is necessary to characterize as well as to predict, in well advance, the fragment of authors that have prospect of improvement.

Concluding, we find that *author-author citation* and *author co-author network* are two indispensable feature which are mainly required for ranking an author. Due to lack of data, we could not include the *author-author social collaboration* relation into this study. Also, we have removed the *author-author co-citation network* from our study as it is less intuitive and becomes collinear with other features. Summarizing, in this research work, we explore three relation that is, paper-paper citation, author-author citation and author co-author relation for ranking an author based on his true merit.

The main idea behind using both paper-paper citation and author-author citation relation is that when a given paper is cited either by recent publication of a highly ranked eminent researcher or a citation by unknown author; then considering author-author citation relation, the credit of citation from the former will be more than the latter.

4.3 Terminology and definitions:

h-index: It is a standard author ranking metric. It is defined as maximum h value such that an author has collected at least h citation in each of his paper for publishing h papers. [29].

g-index: It is another author ranking metric which is improved based upon h-index. Initially, for a given author a set of articles are sorted in decreasing order of their collected citation. g-index is the largest number such that top g-articles receive together at least g^2 citations [31].

PageRank: PageRank is a link analysis algorithm which assigns a numerical weight to a page based upon its vote of support by all other pages on the web. It is commonly used by Google for ranking websites [113].

4.4 Materials and methodology

In this section, we outline the strategy and underlying network model used for ranking authors. Before it, we briefly describe our data set.

For this experiment, we used the data set developed by Tang et al. [98] which is available freely. The original data set had 2,244,021 papers and 4,354,534 citation relationships. We observed some inconsistencies in the data set in the form of one or more missing data fields. Those data elements that have missing fields (except the abstract field) is not usable in our current study and hence are filtered out from the data set. One particular information component missing in the above data set is the field/domain of research a particular paper may be associated, which is an important consideration in our present work. Data set prepared

by Chakraborty et al [114] is used for finding the field of interest of research for a particular author and augment the same for the authors in our data set.

4.4.1 Network model and proposed strategy

In this chapter, we propose a PageRank based author ranking strategy considering multiple feature that is, paper-paper citation, author-author citation, and author-author collaboration defined as, $C^3 - index$. The index may help to accurately score lower ranking authors. Since the proposed indexing strategy consider most significant factor determining an author's success, it is consistent and helps to resolve ambiguity and uncertainty between low-profile author. Also, the strategy may be used to predict rising star author in the early stage of their career. The final calculated $C^3 - index$ score is obtained by the sum of three individual component score obtained from three layers, scores being normalized in such a way that the sum of scores of all the authors is unity. The individual component score from different layers are computed using 'PageRank' based strategies for respective layers of the network. The strategy is elaborated in detail in Section 4.4.

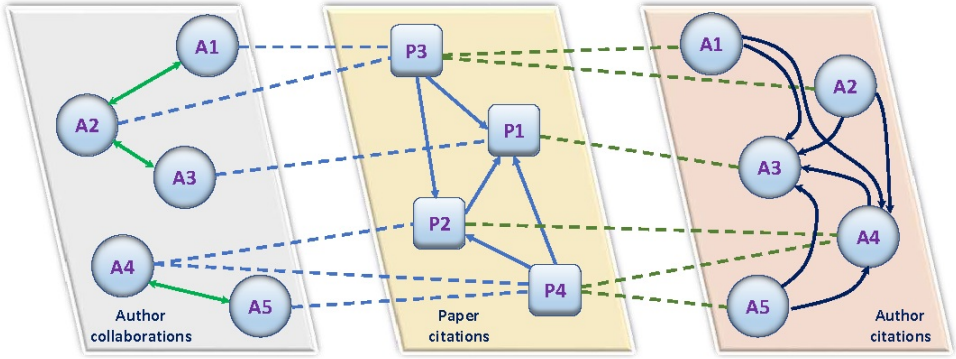


Figure 4.2: Three-layer network model used in $C^3 - index$ for ranking authors. Individual layers are: (i) Author citation layer – a weighted directed network, where vertices are the authors, and weighted edges are drawn from vertex A_j to A_i if author A_j cites the papers of author A_i, the weight of the edge being the number of papers of author A_i being cited by author A_j; (ii) Author co-authorship layer – a weighted undirected network where vertices are authors, and undirected weighted edges are given between authors who jointly published papers, the weight of the edge being the number of papers the pair co-authored; (iii) Paper citation layer – a directed network where vertices are the papers, and edges are drawn from paper P_j to paper P_i, if paper P_j cites paper P_i. Lastly, there are inter-layer edges from author A_i to paper P_j, if one of the authors in paper P_j is A_i.

From data set mainly using the first three data set - Hybrid, ArnetMiner and Microsoft

Academic Graph (MAG); as required we form different data dictionaries and store them in the form of JSON files. On a macroscopic view, these data dictionaries represent citation relationships (edges) between chosen entities under study and hence, forms a complex citation network. These dictionaries represent mappings such as paper id - year, paper id - references, author id - paper id, paper id - list of author id, author id - author name, field id - field name, author id - list of co-authors id, conference id - conference name, journal id - journal name, paper id - conference/journal id, etc. Thus, we propose modelling such citation mappings in the form of single or multi-layer network models as suited analyzing problems. For author-centric study addressing deficits of present performance metrics using PageRank based algorithm, we use a multi-layered model with multiple features added onto each layer. Using such citation mappings, we create a network using I-Graph module.

The proposed indexing strategy $C^3 - index$ is constructed using multiple layer of citation-collaboration network as described in Figure 4.2. The three layers correspond to author-author citation network, author-author co-authorship network, and paper-paper citation network from left to right respectively. Three layers are briefly described as below:

1. Paper-Paper citation network: It is a directed un-weighted graph. Each vertex represent paper and edge represent the number of cites from one paper to another.

2. Author-Author citation network: It is a directed weighted graph. Each vertex represent an author and weight of directed edge from node A to node B is the number of papers authored by A where he cites paper authored by B. However we ignore the author's self-citation.

3. Author-Author collaboration network: It is an un-directed weighted graph. Each vertex represent an author and weight of the edge between A and B represents the number of papers co-authored by them.

The score calculated from three different layer that is, **P**aper **C**itation **I**ndex (PCI), the **A**uthor **C**itation **I**ndex (ACI) and **A**uthor co-**A**uthorship **I**ndex (AAI) are given in table 4.1 for eight selected authors for year 1998. The corresponding $C^3 - index$ score is compared with their respective h-index and g-index. The eight authors are selected from results shown in Figure 4.4. For better visualization, the $C^3 - index$ final score and its score obtained from individual layer are multiplied by the number of authors in data set for a given year, so that *the average* $C^3 - index$ *score for a given year is always unity.* A strong correlation is observed between h-index, g-index and ACI component of the proposed C^3-index, but a weak correlation with the other two components. The above correlations suggest that the widely used citation-based author ranking strategies like h-index and g-index may capture the effect of

ACI component of the C^3-index nicely, but fail to capture the effects of the other two components.

Table 4.1: Indexing score comparison for randomly selected authors

Author	h-index	g-index	C^3-index	ACI, PCI, AAI
B. Bollobas (E)	1	1	7.88	0.45, 4.68, 2.54
B. Shneiderman (A)	13	20	54.86	23.12, 18.12, 13.42
G. Rozenberg (F)	4	5	26.81	2.94, 14.44, 9.21
H. V. Jagadish (B)	11	16	17.66	6.50, 5.70, 5.24
M. S. Hsiao (G)	4	5	2.21	0.78, 0.64, 0.58
Ronald L. Rivest (C)	9	27	79.02	39.58, 28.07, 11.17
S. Shelah (H)	2	3	15.17	0.44, 8.29, 6.24
Tova Milo (D)	7	11	6.06	2.26, 1.74, 1.86

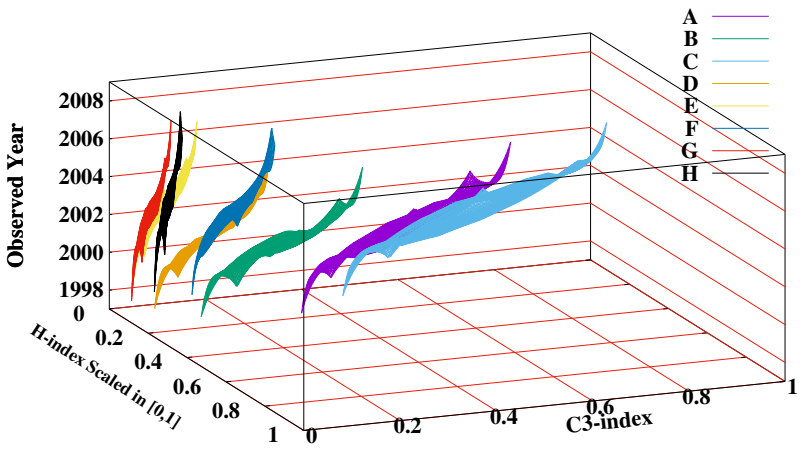


Figure 4.3: The 3D plot in the figure alongside shows the changes in h-index and $C^3 - index$ over the years for all the selected authors mentioned in the table 4.1 during the year range 1998 to 2008. The h-index values are scaled within the range [0,1] by dividing actual h-index of the corresponding author by the observed maximum h-index value for an author in the data set to maintain clarity of the figure. Authors having higher $C^3 - index$ in 1998 show steeper growth both in h-index and in $C^3 - index$ as they progressed over the year, which may be an indication that $C^3 - index$ somehow captures the future success in advance.

In Figure 4.3 (3D plot), we see that authors whose surface has higher $C^3 - index$ score in initial years of their research career have steeper line of progress over the years for both indices that is, h-index and $C^3 - index$. This clearly reflects the fact that $C^3 - index$ can predict the future success of an author in the initial days of his/her career. This might be due to consideration of the AAI and PCI components in the $C^3 - index$ score (see Section 4.4 for a detailed description). The AAI component captures the co-author influence to a given author for whom $C^3 - index$ score is calculated. On the other hand, the latter captures credit collected by a given author due to its co-author influence. The rest of the paper is devoted to characterize $C^3 - index$ and to critically analyze whether it could be used for the purpose of predicting future prospect.

4.4.2 Network construction

The multi-layer network is constructed from the filtered data set. The construction of paper-paper citation network is briefly described – paper are considered as vertices and if a paper (P_i) cites another paper (P_j), a directed edge is connected from P_i to P_j.

Further, the construction of other two layers require that corresponding to a paper all author information be extracted. A major issue faced here is *author name disambiguation*. For removing ambiguity from author name, we use 'RankMatch' algorithm proposed by Liu et al [115]. This algorithm is referred as it uses an unsupervised learning approach. Moreover, it proves to be efficient when same type of bibliographic data set are considered in multiple discipline. Initially, a unique author ID is assigned against each author name in the data set. Next, it follows two step – (i) Firstly, It clusters the author index ID for all possible variation of same author name. (ii) Secondly, other paper feature set such as venue of publication, title of paper, publication year are used to distinguish the clusters and pull out the real candidate pool. Finally, unique author list is extracted and corresponding to it author indexes are added for further reference.

A vertex is added in the author co-author network and author-author citation network for unique author connections in the given list. An un-directed weighted edge is connected between two vertices A_i and A_j if both authors correspond to same paper ID. The weight of edge is increased by one if the same group of authors here, A_i and A_j has co-authored in another paper.

The construction of author-author citation network occur as, if a paper P_i cite another paper P_j, then all corresponding authors of paper P_i are connected with a directed edge

to all authors of P_j. Further, in consequent iteration if there already exist a link between a pair of authors then, the weight of edge is again increased by one. After construction of the network, iterative modified 'PageRank' algorithm is executed which is as discussed below. The individual score for each author from different layers are calculated.

4.4.3 Measuring C^3-index

The proposed $C^3 - index$ is computed using a set of iterative formulas. The $C^3 - index$ of the j^{th} author A_j at iteration level t, denoted by $C_j^{3(t)}$, is obtained as:

$$C_j^{3(t)} = (1-\theta) + \theta \times (ACI_j^{(t)} + AAI_j^{(t)} + PCI_j^{(t)})$$

In the above formula the terms $ACI_j^{(t)}$ and $AAI_j^{(t)}$, denote the scores of author A_j in author-author citation network and the author-author co-authorship network, respectively, that are obtained using the following iterative formulas:

$$ACI_j^{(t)} = (1-\theta) + \theta \times \sum_{A_k \in C(A_j)} \frac{ACI_k^{(t-1)}}{outdeg(A_k)}$$

$$AAI_j^{(t)} = \sum_{A_k \in CA(A_j)} \frac{AAI_k^{(t-1)}}{deg(A_k)}$$

where $C(A_j)$ denote the set of authors who cited at least one paper of author A_j, $CA(A_j)$ denote the set of authors who co-authored with author A_j in at least one paper, $outdeg(A_k)$ denotes the sum of the degrees of the outgoing edges from node A_k in the author-author citation layer of the network, $deg(A_k)$ denotes the sum of the degrees of the edges incident on node A_k in the author co-authorship layer, and θ is the *damping factor* for our strategy. Here we set it to 0.5 following the suggestion made by Chen et al. [24].

The third component in the formula, $PCI_j^{(t)}$ denotes the paper citation index score for author A_j at the iteration level t that are obtained from the paper citation layer of the network. It is the sum of the paper credits shared at that level for the publications made by author A_j distributed uniformly (or some other rule) among all the co-authors of the paper

using the formula:

$$PCI_j^{(t)} = \left(C_j^{3(t-1)}\right)^{\alpha} \times \sum_{P_k \in P(A_j)} \frac{PQI_k^{(t-1)}}{\sum_{A_l \in A(P_k)} \left(C_l^{3(t-1)}\right)^{\alpha}}$$

where $P(A_j)$ denote the set of papers published by the author A_j, $A(P_k)$ denote the set of authors for the paper P_k, and $PQI_k^{(t)}$ is a paper quality index score representing the credit of the paper that is obtained from the paper citation layer of the network using a PageRank based algorithm as follows:

$$PQI_i^{(t)} = (1-\theta) + \theta \times \sum_{P_k \in C(P_i)} \frac{PQI_k^{(t-1)}}{outdeg(P_k)}$$

where $C(P_i)$ denote the set of papers citing paper P_i, and $outdeg(P_k)$ denote the number of the outgoing edges from node P_k of the paper citation layer. We use the same damping factor θ for all the PageRank formulas mentioned here.

As a final note, we represent PCI_j as a generalized formula, where α is used as a *model parameter* to decide the way credit from an individual paper would be distributed among its authors. If it is set to 0, as is the case in the current experiments, then the credit will be distributed uniformly to all the co-authors. But for other values of α, the credit will be distributed on the basis of their current C^3-index. If α is a positive value, then authors having higher C^3-index would receive a larger share of the credit, whereas if α is negative, the authors with lower C^3-index will receive a larger share.

4.4.4 Convergence of the proposed algorithm

As we propose a recursive algorithm, we have to ensure that the algorithm would eventually converge. As we shall see later, for ACI, AAI, and PQI component scores, we use algorithm derived from Google PageRank. Thus we may claim that the above three component scores would definitely converge, as does Google PageRank. For PCI score, since it is derived from PQI score values, we may also claim through logical reasoning that PCI scores would converge. Finally, as $C^3 - index$ is the normalized sum of three convergent sequences, we may claim its convergence. As empirical support of the convergence of the algorithm, we provide Table 4.2 which shows the number of steps of termination over a few trial runs of the algorithm. These results indeed validate our claim.

Table 4.2: An illustration of the convergence for individual component scores of $C^3 - index$. *Precision* denotes the precision level chosen for the trial. The remaining three columns show the iteration step at which the changes of the corresponding individual component become less than the chosen Precision value for all the nodes in the corresponding layer of the multi-layer network. The highest among the three corresponds to the iteration step at which the $C^3 - index$ converges for the whole network. The values somehow validate our claim regarding the convergence of $C^3 - index$.

Trial #	Precision	PCI Convergence	ACI Convergence	Sum Convergence
1	1e-12	93	85	118
2	1e-14	122	113	146
3	1e-16	150	141	175

4.5 Results and Discussion

We begin with comparing $C^3 - index$ and h-index by calculating co-relation score between the two indices and then analyzing them on temporal scale. Finally observed relevant components. Statistically validate that $C^3 - index$ is more competent than h-index in predicting future prospect of an author.

4.5.1 $C^3 - index$ vs. H-index

After carrying out the study and proposing a completely new metric $C^3 - index$, an intuitive question that arises is how accurate are author ranking score of our proposed metric compared to other widely used indexes such as h-index. To measure the correlation we use Spearman Rank Correlation Coefficient between the pair-wise ranks of author (Table 4.3).

A strong correlation of h-index with ACI layer reflect from the coefficient values. In contrast, comparatively narrow correlation is seen with the other two. It also follows in line with our earlier observation as seen in Table 4.1. Another interesting observation that is reflected in Figure 4.4 is the author points which are closely present to diagonal of each subplot represent those authors for whom the $C^3 - index$ values are highly correlated with the other indices such as, h-index and g-index strategies. On the other hand, there exist few authors with low h-index but high $C^3 - index$ value (upper-left portion), and vice-versa (lower-right portion). We analyzed the citation profile for few of the authors selected among them in Table 4.1.

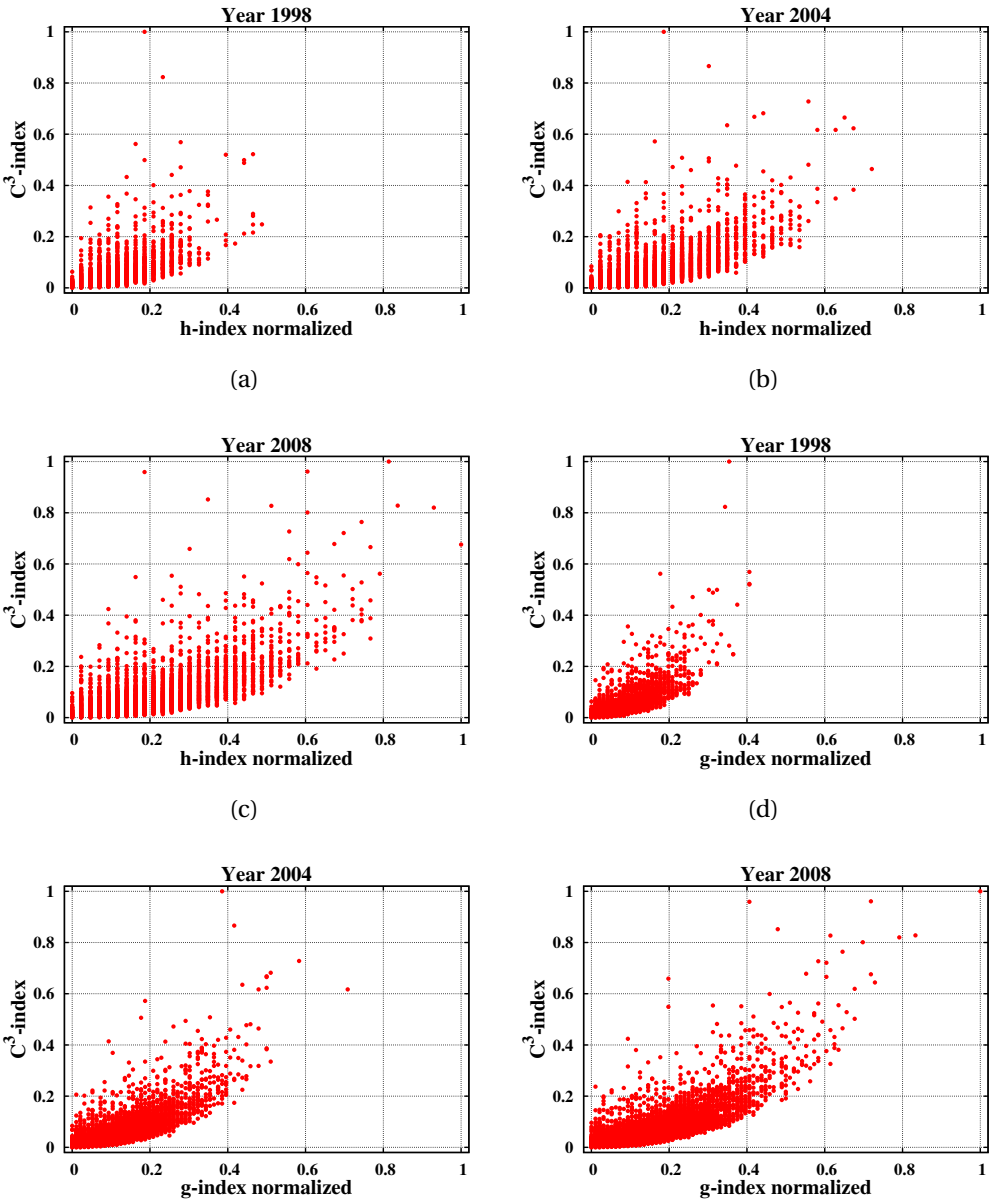


Figure 4.4: Comparison study of $C^3 - index$ against normalised (0 -1) values of h-index & g-index during the years 1998, 2004 and 2008. It is being observed that the value of $C^3 - index$ for majority of the authors remains almost consistent with their respective h-index as well as with their g-index. However, few inconsistent points mostly in upper-left portion of the plots, indicating those authors having low h-index (g-index), but high $C^3 - index$. This is possibly an indication of low citation but high co-authorship credit for the corresponding authors. In Table 4.1, we selected some of authors having such inconsistencies and analysed their behaviour over the years.

Table 4.3: The Spearman Rank Correlation Coefficient between h-index, $C^3 - index$ and its components. The values indicate that h-index is highly correlated to the ACI score, as compared to that for other two components, and hence with $C^3 - index$ as a whole. Thus we hypothesize that the information carried by $C^3 - index$ would be significantly different from that of h-index.

Year	H-index vs $C^3 - index$	H-index vs ACI	H-index vs PCI	H-index vs AAI
1998	0.577136	0.989151	0.467660	0.401122
2004	0.604968	0.988483	0.517128	0.426008
2008	0.613174	0.988427	0.539801	0.437871

A comparative study is done to see the correlation between $C^3 - index$ and h-index, g-index for all unique authors in the data set are give. In Figure 4.4, the sub-plots refer to author score calculated using $C^3 - index$ and h-index, g-index for a given year with a constraint that for a given author the publication entries are set up to that particular year (i.e., by removing from the data set the papers which are published after that year, the citations that are made after that year, and the authors who made their first publication after that year). Further, for all temporal studies the same procedure is followed. Eventually, heading towards recent time, the data volume of the complex network gradually increase in all aspect.

4.5.2 Temporal growth pattern

The performance variance of different indices (h-index as well as $C^3 - index$) on a time basis study is plotted in Figure 4.5. Four set of authors are selected from a pool of author from year 1998. Figure 4.5(a) corresponds to the set of authors who have relatively low ACI-score, but high AAI-score in 1998.

Over the years, temporal growth in respective author score using proposed $C^3 - index$ and h-index strategy is given for 31 selected author. In Figure 4.5(b) we plot the same results for set of 48 author who had low ACI-score and low AAI-score in 1998. While comparing the above two plots, we see that the indices for most of the authors in Figure 4.5(a) tend to end up with much high values as compared to that for authors in Figure 4.5(b), in both the cases the lines, start nearly from the same point. Consequently, it refers to the fact that the ACI component of $C^3 - index$ has strongly correlates with future performance behavior of a given author for whom $C^3 - index$ is calculated. In Figures 4.5(c) and 4.5(d) we plot similar growth curves for another two sets of authors, both having high ACI scores but different AAI

scores. Here also we observe that the major portion of authors from the author set having higher AAI scores ended up with higher performance indices.

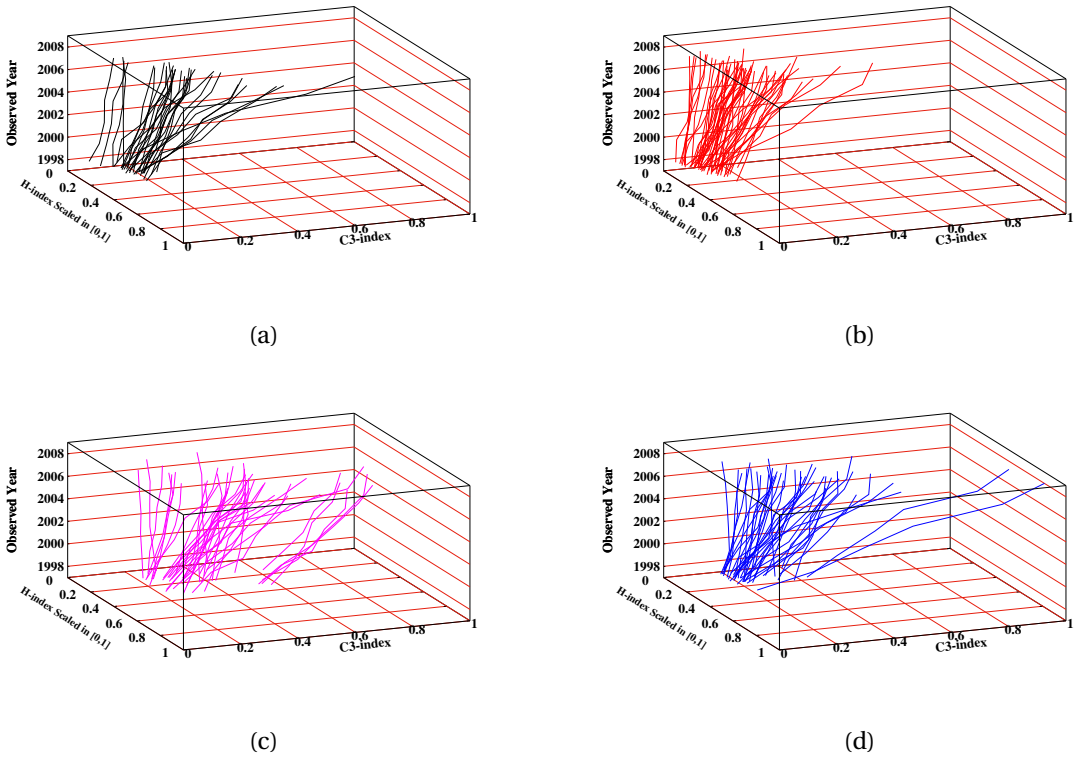


Figure 4.5: Proposed $C^3 - index$ has three components: ACI, PCI and AAI, respectively. We observed that h-index (and also g-index) has high correlation with ACI component, but has low correlation with the other two (Table 4.3). Here we select four sets of authors: **(a)** authors having ACI ≤ 20% of the ACI_{max}, AAI ≥ 80% of AAI_{max}, **(b)** authors having ACI ≤ 20% of the ACI_{max}, AAI ≤ 20% of AAI_{max}, **(c)** authors having ACI ≥ 80% of the ACI_{max}, AAI ≥ 80% of AAI_{max}, **(d)** authors having ACI ≥ 80% of the ACI_{max}, AAI ≤ 20% of AAI_{max}. The scores are selected on the basis of the year 1998. We plot 3D line curves for the corresponding authors in the respective sub-figures. In general, the figures suggest that the authors having high AAI-score improved more during the time period 1998-2008 than those having low AAI-scores. This suggests that the inclusion of AAI-score in the proposed $C^3 - index$ has brought future prediction capability in it.

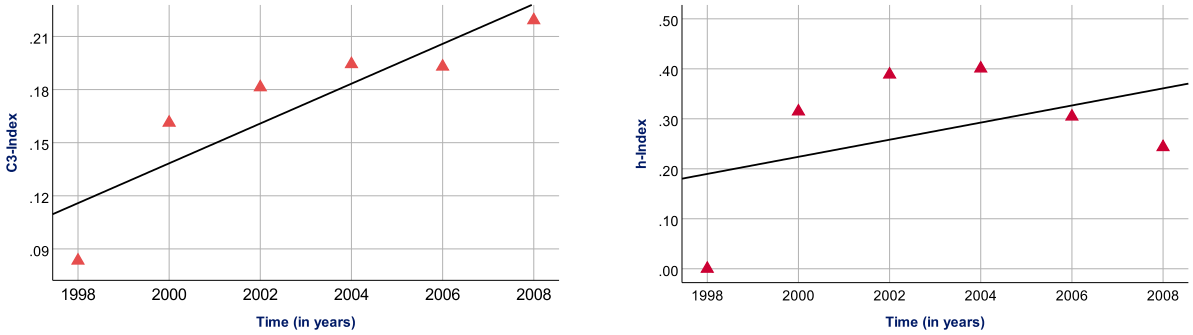
4.5.3 Predicting future prospect of authors using $C^3 - index$

As h-index is mainly calculated using citation and publication count of a given author; consequently, it takes more time for a change in its value. On the other hand, $C^3 - index$

reflects an instant change in the author's score while an author publishes a paper. As we are taking indexing score from three different layers which not only depends upon citations but also collaboration impact that is, the co-author's impact with whom an author is working. On comparison to h-index using regression analysis shows that $C^3 - index$ is statistically more acceptable.

4.5.3.1 Statistical regression analysis between *h-index* and $C^3 - index$

We have considered a sample of authors whose h-index scores are 0 in 1998 and which has increased to a range of 7 to 12 in 2008. Next, we calculate $C^3 - index$ for the same group of authors. Time series data is calculated in an interval of two years. We calculate the average for both indexes for each year taken into consideration and do a linear regression analysis.



(a) (b)

Figure 4.6: (a) Linear regression trend line for $C^3 - index$. R^2 value for $C^3 - index$ is 0.78 **(b)** Linear regression trend line for h-index. R^2 value for $h - index$ is 0.19

Consequently, to assess future predictive performance of h-index and $C^3 - index$; we find $C^3 - index$ is a better metric with R^2 value of 0.78 rather than h-index which has R^2 value of 0.19. This implies that for 78% of the times, variation in $C^3 - index$ could be accurately predicted. This is due to the consideration of academic social interactions of an author besides his publication and citation count.

Multiple layers taken into consideration for calculation are a paper citation, author citations and author collaboration. Equations for C3 index and h-index are $Y = 0.0112x - 22.35$ and $Y = 0.0171x - 34.007$ respectively. Significance value (p-value) in regression test for $C^3 - index$ and for h-index is 0.048. Overall, it refers that $C^3 - index$ is statistically more significant in predicting future performance of an author than h-index. In the next section,

we try to find which component of $C^3 - index$ is actually significant in predicting the future of author's prospect.

4.5.4 Inter-layer relationship analysis of $C^3 - index$

After thorough study, we see that the the ACI component of $C^3 - index$ is strongly correlated with h-index and a weak correlation is seen with the other two components.

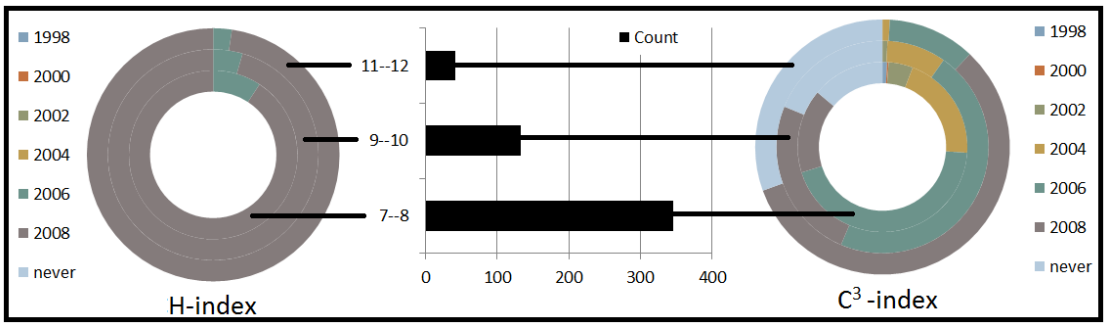
The Figure 4.5 is plotted to see if any meaning insight can be derived from $C^3 - index$. It is seen from the figure that authors having high AAI score show faster growth in their career trajectory than the authors with low AAI score. $C^3 - index$ can predict the future success of an author in initial days of publication of an author which its variants lack to do. We present multi-level pie-charts in Figure 4.7 to show whether $C^3 - index$ is capable of predicting the future success of authors for a selected set of authors to validate the above mentioned fact.

Authors who have obtained zero h-index in year 1998 and gradually attained h-index ranging between 7 to 12 are plotted in Figure 4.7(a). Bar plots in the middle show count of such set of authors in three equally-divided h-index bins. The multi-level pie-chart in the left shows the gradual improvement of h-index as observed over time for the authors in each bin during the time span observed in two-year separations. The multi-level pie-chart in the right points to the fraction of authors present in respective bins shown in the bar plot exceeding a chosen $C^3 - index$ bound in a given year.

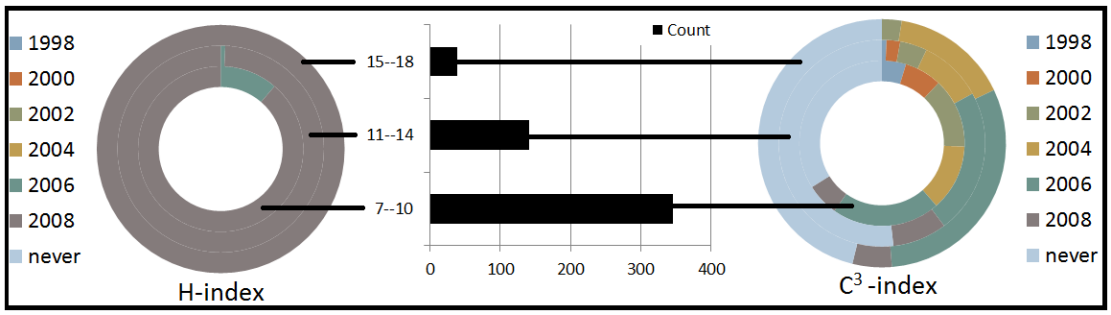
Three different bounds are chosen for three different bins, viz. 0.02 for 7-8 bin, 0.03 for 9-10 bin, and 0.04 for 11-12 bin. In left-hand pie-chart, we pinpoint the fraction of authors that reached the next h-index bin in the respective year. We observe from the left-hand pie chart that no fraction of authors reach the next h-index bin prior to 2006. On the other hand, it is apparent from the right-hand pie-chart that significant fraction of authors reached the next bin level much earlier than the above, which suggests that $C^3 - index$ is able to capture the change much ahead of time than h-index. This somehow establishes the predictive power of $C^3 - index$.

We are now interested to see whether above mentioned future-predictive behavior of $C^3 - index$ holds for authors present in other portion of the author spectrum. In Figure 4.7(a), a set of authors are selected whose h-index lay in the range of 4-7 in 1998. We may decently assume that such authors may be considered as medium-performers during the time when our observation begins. We observe that by 2008, the values of the selected authors' h-index lie in the range 7-18, which may indicate that some portion of the author gained high

visibility (i.e., gained high h-index) in 2008; whereas the rest fail to acquire enough visibility. The bar plot in the middle shows the number of those authors in three distinct bins similar to Figure 4.7(b).



(a)



(b)

Figure 4.7: Bar graph in the middle shows a total number of authors in equally divided h-index bins. Colour code in the multi-level pie chart on the left represents the year when a particular set of authors have acquired a corresponding range of h-index mapped to bar graph in the middle. Similarly, right side multi-level pie chart corresponds to equally divided C3 index bins. Colour code refers to the years when a set of authors have acquired the corresponding C3 index bins.
(a) Figure depicts the above description for a set of authors which is extracted from the data set having zero h-index in 1998 but acquired moderate h-index ranging between 7 to 12 in 2008.
(b) Figure depicts above description for a set of authors which is extracted from the data set having h-index ranging from 4 to 7 in 1998 but acquired moderate to high h-index ranging between 7 to 18 in 2008.

The multi-level pie-chart in the right pinpoints the fraction of authors lying in respective bins shown in the bar plot surpassing a chosen $C^3 - index$ bound in a given year. Three dif-

ferent bounds are chosen for three different bins, viz. 0.08 for 7-10 bin, 0.14 for 11-14 bin, and 0.17 for 15-18 bin. In left-hand pie-chart, we show the fraction of authors that reached the next h-index bin in the respective year. We observe from the diagram that a major fraction of authors reach the next level after 2006, and only a small fraction reaches this level during 2006, and none does the same before 2006.

On the other hand, for $C^3 - index$, future stars (those falling in 15-18 bin), are capable of surpassing the predefined boundary set during 2004. For the others, it has been much earlier – a fraction, although small, from 7-10 bin reaches this level even in 1998. This observation leads us to believe that proposed $C^3 - index$ has the capability of predicting future stars in advance.

4.6 Conclusion

We have proposed an efficient ranking strategy $C^3 - index$ in which prior deficits of its variants are overcome. On a temporal scale, we observe that $C^3 - index$ captures the growth of an author much ahead of time due to the presence of high co-authorship credit from the author-author collaboration layer. Performance of new researchers are evaluated efficiently, and ties are occurring between new and medium ranked researchers are overcome easily. $C^3 - index$ also, reveals the capability to predict the future success of an author.

4.7 Summary

∗ With the increasing rate of authors since the 1990ś, several author performance metrics have evolved which are mostly calculated based on citation and publication of an author. One such popularly used metric is h-index. The major deficit of h-index and other related citation-based ranking systems are that for a large number of authors h-index calculated is 0 due to which research beginners or medium ranked researchers; both cannot be measured properly. Also, it is difficult to break ties between new and medium ranked researchers.

∗ Our research begins with upgrading and devising a new metric, $C^3 - index$ for author performance measure using PageRank based algorithm. One advantage of using this algorithm is that we can add to multiple features. Here, we carefully select multiple features that strongly evaluate an author's performance and model it into a multi-layered network model.

∗ The three features modeled into a three-layered network model include author-author collaboration or co-authorship index (AAI), author-author citation index (ACI) and paper-paper

citation index (PCI).

∗ In comparison with the existing metrics that is, h-index, g-index it is found that $C^3 - index$ almost consistently grows with their respective h-index, g-index whereas, few inconsistent points are also observed between pair-wise ranks using Spearman Rank Correlation Coefficient which refers to authors having low h-index but high $C^3 - index$. It occurs due to low citation count collected by an author but high co-authorship credit from the AAI layer for the corresponding authors.

∗ $C^3 - index$ could be used to capture the future performance of an author. It is also validated from regression test that $C^3 - index$ is statistically more significant in determining future prospect of author than its other variants. For selected authors, $C^3 - index$ on a temporal scale during the year range 1998-2008 reveals that authors having higher $C^3 - index$ in 1998 shows steeper growth as they progressed over the years. Further, we find that authors having high AAI score (author-author collaboration index) during 1998-2008 improved more than those having low AAI score. Thus, the presence of AAI score component predicts future success for an author in advance.

∗ Hence, $C^3 - index$ is an efficient ranking strategy that helps to break ties between new and medium ranked researchers as $C^3 - index$ captures growth and visibility of an author much ahead of time than h-index.

Chapter 5

Citation count – a key attribute in the quality measure of research impact

5.1 Introduction

Though efficient dynamic models [116] are being implemented over the years to gauge the impact, productivity and contributions of a research outcome and scientific community as a whole, citation-based indicators stand out as being in the center of all models in terms of reliability, popularity, and frequency of usage. Significant earlier research performed by Eugene Garfield [117] claims that bibliographic analysis over growing serves as a constitutional quantifier for gauging the present day advancement in science [4]. It could lead the scientific community towards a more selective, developing and promising key areas of research. The elementary idea behind is that 'the more frequently a paper gets cited, the greater is its prominence and impact on a particular research ground' [118].

While studying citation dynamics of Computer Science articles over its effective lifetime, a generalized observation [1] reveals that, following publication there is a gradual expansion in citation sphere for a research work in the beginning that continues till two to three years (*evolving phase*), followed by a constant crust, i.e., the frequency of incoming citations becomes stagnant for the later two years (*perpetual phase*), and then, finally there is a gradual descent over the rest of the publication age for the article (*decline phase*) and gradually, at some point no further activity is observed (*obsolete phase*). An exhaustive empirical detection is done with 110 years of citation data from *physical review* journals [56][119] reveals that major segment in the citation distribution curve fits unusually well to a log-normal

form. In our present work, we conduct a thorough experiment on *Computer Science* domain using an extensively large bibliographic data set, and find that most part of the cumulative distribution of citations does fit to a *log-normal* form as well, and the extreme tail decays exponentially to a *power-law* form of curvature.

The motivation of our present paper stems from a fundamental question raised by Ruiz-Castillo [120] in connection with scientometrics that goes as follows: 'Are distributions in citations very identical for different fields of science or rather contrasting?' In [121], Radicchi et al. claimed that any science field could be characterized by the same citation distributions apart from the changing scaling factor. On the other hand, Waltman et al. [122] observed to have non-universal citation distributions in quite a number of fields. Research fields including Engineering Sciences, Social sciences and Material sciences with a comparatively low average citation count per publication are found to follow such trends. Such a conflicting observation about citation dynamics in case of engineering and scientific fields compel us to initiate a similar study for the field of *Computer Science*, which we believe has the flavor of both engineering and science. Another factor that distinguishes Computer Science domain from other fields is the dominance of conferences parallel to journals in its publication pool.

A common trend is observed in case of Computer Science field where papers published in conferences quickly start to acquire large volume of citations and thus, gain active popularity as well as visibility [57] within initial few years after publication; and likewise, such papers also result in experiencing an abrupt decay of citations and thus, popularity [123]. But in the case of journal papers, though they take a considerable amount of time to get published and become noticed in the scientific network; most journal papers consistently continue to collect citations for a long time. Also, we find in the Computer Science field that the total count of conference papers is more than papers published in journals. Several citation based statistical studies done on physical review journals [119] indicate that in theoretical fields of science such as Quantum theory, Nuclear Physics, etc.; authors tend to publish more in journals rather than conferences. We motivate our study in an applied domain of research such as Computer Science where the scenario follows a combination of distribution among journals and conferences though some significant contributions in Computer Science domain have been acclaimed by conference papers. Then it might be interesting to observe the fate of the findings of [56] or [119] when they are put against the applied domains like the present one. It might also be interesting to see a variation of citation dynamics of conference papers against journal papers in connection with the said application domains,

as the formers would contribute significantly higher in case of applied domains in contrast with the theoretical domains like Nuclear Physics, or Quantum theory.

Our objective throughout the present paper will be – (i) to study the impact of journals and conferences on the citation activity of well-cited papers. We analyze by profiling the citations of top journal & conference papers into 5 different citation trajectories [124]; (ii) taking into account both the volume of citations collected as well as regularity of intervals in which citations have been received, we redefine the 'hot publications' and study their citation profiles; and (iii) to characterize unique citation patterns of highly-cited papers. One such interesting line of papers that exhibit *'Delayed Recognition'* phenomenon can be classified under *'Sleeping Beauty'* or *'Revived Classics'*. The term 'Sleeping beauty' in reference to those papers whose citation trajectory depicts late peak over an effective lifetime was first coined by van Raan [125]. He suggested that there are three factors that result into revival of an old classic – (a) *'Hibernation period'* or the number of years for which the paper has remained completely unrecognized, (b) the average citation count received during its dormancy or hibernation period and (c) *'Awakening intensity'* or the volume of citations accumulated during 4 years after it has awakened from its deep sleep or from the year it experienced a sudden citation burst [58].

In our present study, we analyze a wide volume of Computer Science articles: over more than 2 million papers published in many eminent journals & conferences during the timespan of 1859 to 2012. We find that out of 1,088,452 papers which have received at least one citation, 598,203 papers are conference papers, and the remaining 490,249 papers have been published in journals. (i) We redefine the criteria for *'hot publications'* to be the papers that collects greater than 1500 citations and whose ratio of *average citation age* to *publication age* (defined in section 5.3) is greater than $\frac{2}{3}$. We get 23 hot papers in Computer Science domain and study their citation profiles. (ii) Considering the traditional preferential attachment model, we observe that the attachment rate has considerably increased when measured in the year range 2000-2009. This is due to increased visibility of research works in the Computer Science domain during the late 1990's and early 2000. (iii) For 'Sleeping Beauties', we have set the threshold to be more than 250 citations and the ratio of average citation age to publication age greater than 0.7. We get 8 Sleeping Beauties in our searched domain, and thus, study the behavior of their citation trajectories to identify peak citation values in net effective publication age. (iv) We find 41 *major breakthrough discovery papers* by setting the threshold to be more than 500 citations and the aforesaid ratio of less than 0.4. We observe

that, within a very small duration of average citation age, such papers manage to collect a huge volume of citations. Out of these 41 discovery papers, a maximum citation of 1020 with an average citation age of 3.43 years has been collected by a paper with the publication year 2001 titled "*A Transport Protocol for Real-Time Applications*". (v) While performing a comparative study on the influences of journals & conferences among the top 10 highly cited papers, 80% is found to be journal papers whereas 20% are conference papers. Journals have shown a major impact in the increasing rate of citations even in applied Computer Science domain. Profiling the citation activity of top journal or conference papers, a generalized observation reveals that journal papers usually depict late peaks and monotonic increase in their citation trajectory, whereas conference papers tend to experience multiple peaks and due to a rapid increase in popularity, the profile exhibits sharp increase and monotonic fall over its effective lifetime.

5.2 Data set

For analyzing citation data set, a treasure trove of openly accessible citation-related information of research publications is nowadays available. They are considered as relics of scholarly communication representing the trend of previously published works as to how they were referred to by the citing authors. As the Web is becoming a more and more innovative and influential medium for scientific communication, citation analysis and other bibliometric practices have found some major applications in studying this new phenomenon in scholarly communication.

We crawled a massive publication data set of Computer Science field from one of the largest archived, *Microsoft Academic Search (MAS)*[1]. Initially, the rank list given by the MAS was used to obtain the list of respective paper Ids. The paper Ids, thus obtained, were then used to fetch all the data and required credentials of the publications. In order to avoid overloading of a particular server with spurting traffic, we distributed the crawling to different systems using *Tor*[2]. We crawled information related to 6 million papers. For efficient browsing and extraction, an exponential back-off time was devised to send the request again in case of server or connection failure. From the perspective of both client and server, robot restrictions imposed by the servers were strictly followed for avoiding any disruption in efficient crawling of data.

[1] http://academic.research.microsoft.com/
[2] https://www.torproject.org/projects/torbrowser.html.en

Table 5.1: Details about the data set

Attribute	Count
Total number of papers	2101548
Paper count with at least one citation	1088452
Conference papers with at least one citation	598203
Journal papers with at least one citation	490249
Number of authors	2662300
Number of authors per paper	2.48
Number of papers per author	5.17

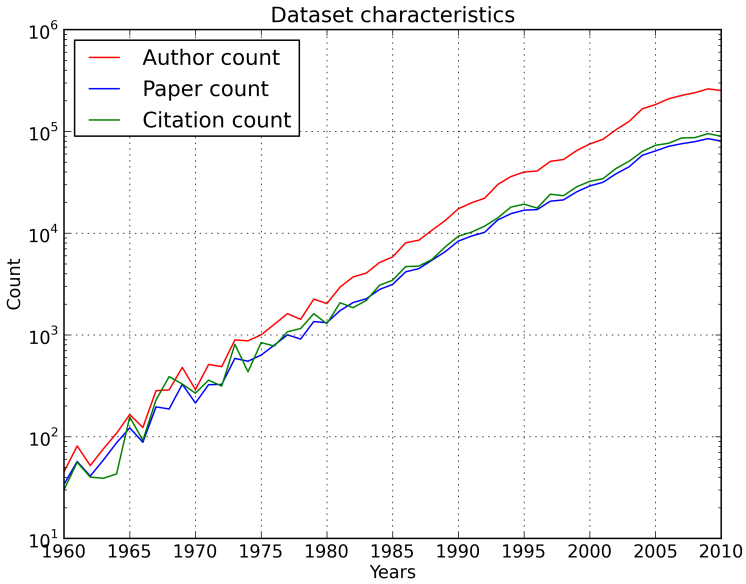


Figure 5.1: Data set characteristics- Growth of citation, author and paper count over time. x-axis denotes the years in which citations have been accounted whereas y-axis denotes count of authors, papers and citations per year.

The extracted data set had several inconsistencies that were removed by using a series of steps. We have filtered the data set as per requirement of our study by removing all such papers that had insufficient bibliographic attributes such as unique paper index, publication venue, list of authors, publication year, valid references, etc. Also, we have filtered out some

of the forward citations; i.e., citations to the papers which got published after the publication of the citing paper. We have considered papers published in the year ranging between 1859 to 2012 that have at least one incoming or outgoing citation, i.e., all the disconnected nodes with zero in-degree and zero out-degree are removed. The filtered data set now contains around 5 million papers. Some of the references that point to papers but not cited (swaying references) are also removed. Table 5.1 provides some of the salient features of the obtained data set.

For instance, searching with the paper title, we only found roughly 88.12% of the papers which can be identified with their respective research fields. Remaining (11.88%) of the papers were added using publication journal ID of the paper.

More than 19 million authors from diverse research domains and 38 million papers are weekly updated in MAS. In each field, on an average ten-year impact indicates the average number of citations that each individual paper within a field received from the papers of other fields during the years 1859 to 2012 are collected. Figure 5.1 highlights year-wise growth pattern of the various items in the data set. Overall, we analyze a large scale data set to strengthen the claims we have made in our paper.

5.3 Terminology and definitions

Publication age: Publication age refers to the difference between the year when a paper has collected its last citation to its original publication year.

Citation age: Citation age refers to the total count of productive years, i.e., we consider only those years when the paper has received at least one annual citation during its effective publication age.

Attachment rate: It gives a close probability that a new article is most likely to cite a paper which has already received k number of citations. Attachment rate acts as a significant parameter to model the growth of citations in different time windows leading to traditional cumulative advantage models.

Sleeping beauty: It refers to those papers which have remained unrecognized for a long period known as 'Hibernation period' and then, almost suddenly started to collect a huge volume of citations by another weighted article. The threshold is set as papers which were published before 1960, have received more than 250 citations and a ratio of average citation age to publication age of greater than 0.7.

Discovery papers: Papers that received more than 500 citations and a ratio of average

citation age to publication age of less than 0.4 are classified under Discovery publications. Such papers exhibit an increased citation count (k) in a very short average citation age.

Hot papers: We define hot papers as those papers which received greater than 1500 citations and whose ratio of average citation age to publication age is greater than $\frac{2}{3}$.

The heuristics that we follow for peak detection [1], we first try to find the potential peak (by normalizing each point by the maximum citation value) accounted for the papers. The two conditions that need to be satisfied are as:

(i) We consider the years, where citation count exceeds the max. Peak, i.e., 60% of the maximum citation gathered by that paper in any year.

(ii) Then, consider potential peaks, i.e., at least 2 years of gap between continuous peaks.

Limitations: In case, a situation is encountered where condition (i) is satisfied, but for condition (ii) the consecutive gap is less than 2 years; we then, treat it as a single peak and consider the last year as the time of peak.

Next, we categorize papers into either single peak or multiple peaks. We define the citation trajectories into five different categories:

i) Early peak (early): Papers whose trajectory depicts a single peak and which received its maximum citation peak within the initial five years excluding the first year after publication followed by a sharp fall. Moreover, the paper should receive at least one citation per year throughout its effective publication age. This condition is referred to as l-condition.

ii) Late peak (late): Papers that collect few or no citations in the initial years and then, it rises to a single peak in the duration of at least five years after publication (excluding the fifth year) are categorized into late peak rising paper. It should satisfy the l-condition also.

iii) Monotonically increasing (moninc): Papers which satisfy the l-condition and for which the citation count gets accumulated monotonically increasing starting from the year of publication until its effective publication age with sufficient volume of citations before the occurrence of a peak.

iv) Monotonically decreasing (mondec): Papers which got a single peak in the immediate next year after publication followed by a sharp decay of citations with sufficient volume of citations collected after the occurrence of the peak depicts monotonically decreasing curve. It further, should also satisfy the l-condition.

v) Multiple peaks (multi): Papers that experience multiple peaks at different time points, i.e., peaks occurring with a gap of more than two years and satisfying the l-condition fall under this category.

5.4 Results and Discussion

Raw citation count [126] is a determinant of a paper's acceptability and significance in a scientific community [78]. A comprehensive citation analysis of Computer Science domain covering 2,101,548 out of a total of 11,971,250 citations from the year 1859 to 2012 is performed. We bring out an extensive study on citation histories of highly-cited and significant papers in Computer Science domain which exhibits unique citation activity over its effective lifetime. Generalized work on identifying the hot publications and major breakthrough discovery papers considering 110 years of data of physical review journals are done earlier by Redner [56].

Here, we explore similar patterns in the Computer Science domain to expand such analysis. In addition, in each individual category, we study the impact of journals/conferences with the changing dynamics of citations. Further, this also helps us to examine how the Computer Science field has evolved in architecture and application over the years. We also try to extend our analysis by re-defining and profiling the activity of hot publications of Computer Science domain individually into five citation trajectories – a peak in the initial years after publication (early), a peak detected in the late years of its effective lifetime (late), multiple peaks (multi), multiple peaks monotonically increasing (moninc) and multiple peaks monotonically decreasing (mondec). Identifying the behavior exhibited by the 'Sleeping Beauties' of Computer Science, we find that such papers exhibit the highest single peak, but at much later years of its total publication age. Thus, precise predictability of citation-based measures is necessary to determine the quality of publications gauging impact factors of various scientific players.

5.4.1 Impact of venue on well-cited papers

The data set originally taken into account consists of 51.79% of papers with at least one citation. The same is analyzed to find the citation statistics as shown in Table 5.3. Further, we observe that out of those papers, 54.95% are conference papers, whereas 45.04% papers got published in journals. The initial observations that can be drawn from here are somewhat discouraging. The data reveals that publications received fewer than ten citations are the majority (nearly 80% of total papers). In contrast, among the small number of highly-cited papers, only 12 publications received more than 3,500 citations. In addition, further results clearly indicate that authors from Computer Science domain tend to publish more in conferences; possibly to gain popularity and visibility quickly in the research community, which

led to more number of conference papers in the field.

Table 5.2: Computer Science articles with more than 3500 citations till 2012

Title	Authors	Year	Citation count	Publication venue	Paper notation
Distinctive Image Features from Scale Invariant Key Points	David G. Lowe	2004	4794	Journal Id(373)	Lowe_IF
Digital communications	R Korn, P Wilmott	1985	5384	Journal Id(182)	DC
Classification and Regression Trees	Leo Breiman, J. H. Friedman, R. A. Olshen C. J. Stone	1984	3610	Conference Id(2614)	Breiman_RT
Numerical recipes in c: the art of scientific computing	William H. Press, S. A. Teukolsky, W. T. Vetterling & B. P. Flannery	1990	4598	Journal Id(902)	Press_NumC
Matrix Computations	David F. Mayers, Gene H. Golub, Charles F. van Loan	1986	4971	Journal Id(902)	Mayers_MC
Snakes: Active contour models	Michael Kass, Andrew P. Witkin, Demetri Terzopoulos	1988	3774	Journal Id(373)	Kass_ACM
Computers and Intractability: A Guide to the Theory of NP-Completeness	Michael Randolph Garey, David S. Johnson	1979	11788	Conference Id(277)	Garey_NPC
Graph-Based Algorithms for Boolean Function Manipulation	Randal E. Bryant	1986	3569	Journal Id(18)	Bryant_AlBfm
The mathematical theory of communication	W. Weaver, C. E. Shannon	1964	3734	Journal Id(909)	Weaver_MTC
Chord: A scalable peer-to-peer lookup service for internet applications	Ion Stoica, Robert Morris	2001	3559	Conference Id(1652)	Stoica_CIA
A tutorial on hidden Markov models and selected applications in speech recognition	L. R. Rabiner	1990	3844	Journal Id(416)	Rabiner_SR
Fuzzy Sets	Lotfi A. Zadeh	1965	6606	Journal Id(40)	Zadeh_FS

Comparatively, there exists a large number of conference papers (5,98,203) in the total pool of Computer Science papers. In contrast to it, it is found in Table 5.2 that journal papers tend to receive more number of citations as compared to conference papers. A large volume of journal papers exists among the highly cited articles with citation count (k) greater than 1500. This gives us a clear insight that journals continue to dominate and have a major impact even in the applied domain. Journal papers tend to collect consistent citations for a prolonged publication age. However, maximum citation count over the entire data set is found to be 11,788 which is received by a paper (`Garey_NPC`) of Garey and Johnson on Theory of NP-Completeness; and that is interestingly a conference paper.

Table 5.3: Citation count variation

Citation count (k)	No of publication
>3500	12
>1000	214
>500	882
>300	2199
>100	14133
>10	827262
>5	629411
>1	284252

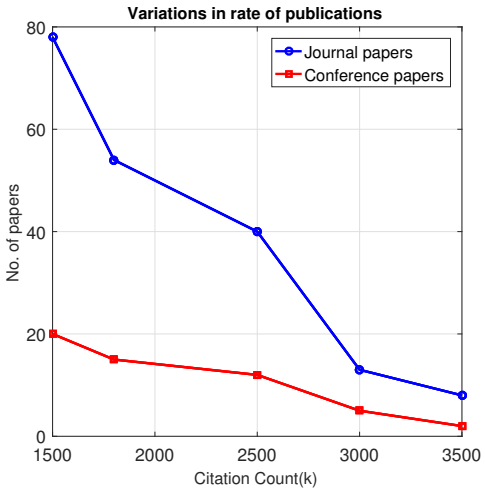


Figure 5.2: Variations in publication rate with citation count (k)- In the Table 5.3, we distinguish those papers which have collected at least one citation, i.e., 1,088,452 papers out of a total of 2,101,548 papers taken into consideration originally. Citation count (k) against the number of publications which has received at least (k) citations is recorded. In the figure alongside, the x-axis refers to citation count (k) whereas y-axis denotes the number of papers. For highly-cited papers with citation count (k) greater than 1500, it has been found that comparatively more number of papers are published in journal than conferences. The curve depicting journal papers leads all the way through the increasing citation count (k) along the x-axis. Also, with the increase in citation count (k) along the x-axis, there is a gradual decrease in the count of papers. This accounts for less number of highly cited publications.

We observe a line of such contrast patterns throughout our study. Though journal papers consistently continue to collect citations for a long time, some significant contributions are acclaimed due to conferences. Out of the top publications with more than 3500 citations, the ratio of papers being published in journals to conferences is $2:1$. Considering the top 20 publications, the ratio of journals to conferences is found to increase to $2.6:1$. Finally, among the top 100 papers, the proportion of journals to conferences is seen to further increase to $3.9:1$. The observations that we draw from varying statistics is that the proportion of journal papers is consistently growing. Journal papers collect large citations and have extended popularity even in applied fields such as Computer Science domain.

The journal topics that primarily influence large volumes of increasing citations among the top 12 publications include varied areas of research, such as Computer Networks, Com-

Table 5.4: Comparative study of the proportion of journals/conferences among the top 100 publications in Computer Science domain.

Publications	Citation count(k)	% of journals	% of conferences
Top 10	(>3500)	80 %	20 %
Top 20	(>3000)	77.77 %	22.22 %
Top 50	(>2500)	76.92 %	23.07 %
Top 70	(>1800)	78.26 %	21.73 %
Top 100	(>1500)	79.59 %	20.40 %

puter Vision, Information Control and Mathematics of Computation, Mobile Computing, etc. On the other hand, among the top four conference papers, the topics that influence more the citations in the field include Artificial intelligence, Communication, Parallel and distributed processing.

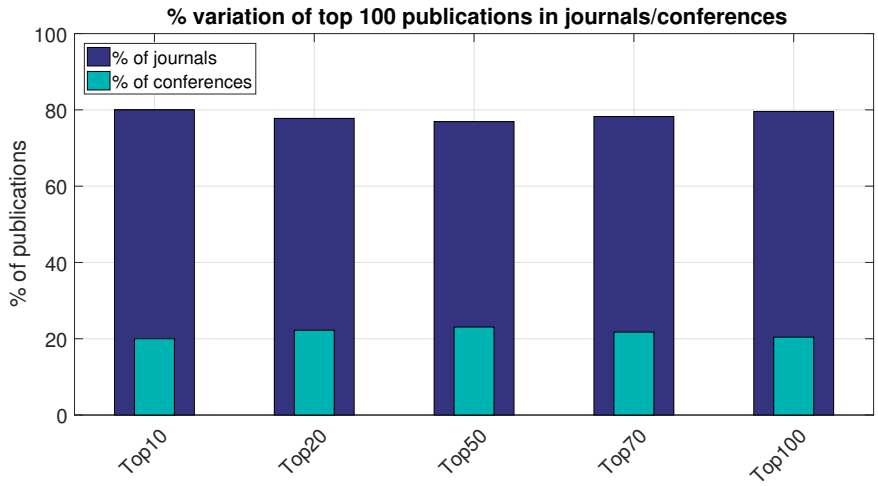


Figure 5.3: Percentage variation of top 100 publications in journals/conferences in the data set. Here, blue bars indicate percentage rise in journal papers whereas the sky blue bars depict percentage rise in conference papers. The y-axis depicts the percentage(%) of publications with ($k > 1500$). The graph clearly shows that, while considering the top 100 publications of the Computer Science field, the journal papers exhibit a consistently increasing rate of citations received as compared with the same for conference papers.

While studying the citation activity of top journal & conference papers, we observe that journals usually depict late peak in their citation trajectory; or mostly develop a monotoni-

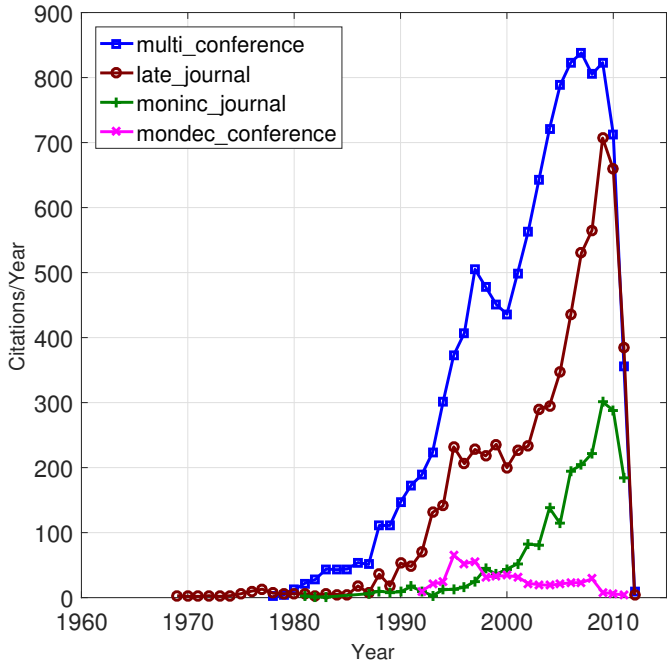


Figure 5.4: Citation profiles of top journal & conference papers - Profiling the citation activity of top journal & conference papers. A general trend as we find here is that; journals tend to depict monotonic increasing and late peaks, whereas conferences experience multiple peaks with monotonic decreasing range and early peaks.

cally increasing pattern of citations over time. This behavior indicates that through journal papers take sufficient time to get visible in the research community; they continue to receive citations throughout their effective publication age consistently.

Consistency is an important parameter while analyzing citation statistics. One of the journal paper (Zadeh_FS) has received the highest number (6606) of citations. It depicts a late peak, i.e., it rose to its maximum value at much later years of its lifetime. Also, late rising peak leads to a sharp denial of the 'first mover advantage or preferential attachment' models. Some of the top journal papers depict multiple peaks with a monotonic increase in citations over time. Similarly, a conference paper with maximum citation count over the entire data set (Garey_NPC) with 11788 citations shows multiple peaks in its entire citation age. A generalized observation is that conference papers usually depict an early peak or experience monotonically decreasing multiple peaks. Conference papers usually collect maximum

citation at an early age after publication. Multiple peaks depict an intermediary behavior between early rising and late rising papers.

5.4.2 Citation distribution

While performing analysis of citation distributions, an innate question that arises is how citation distributions are scaled over a given period; or rather, what is the probability $P(k)$ that a research article at least collects k number of citations? Further, re-scaling offers convincing evidence that these aggregated citation behaviors follow some generic scaling laws. Here, we could not plot a direct graph between probability distribution function $P(k)$ versus k, as it shows some significant statistical fluctuations. Instead, to get a uniform plot, we calculate cumulative citation distribution function as a summation of all probabilities within a given range as follows:

$$C(k) = \int_k^\infty P(k)\,dk$$

This demonstrates a probability of a paper receiving at least k incoming citations. Many of the existing literature focuses on citation distributions. Much debate has already been undergone on whether citation distribution follows an exponential decay or power-law characteristics or a log-normal form. Power-law scaling method is given by $P(k) \propto k^{v}$, where v is a positive integer, has many deficits and also, lacks the presence of a distinctive scale for appropriate citation analysis. To model correctly and avoid any misinterpretation of data, we plot the graph for $C(k)$ vs. k on a log-log scale. We observe that its negative curvature decays faster than the results given by small-scale studies of power-law form or an extended exponential form. Here,

$$C(k) \quad \propto \quad \exp -k^{\beta}$$

where β is less than 1. Over an extensive range of the graph, we find that the function exhibits a log-normal form. A log-normal process is the statistical realization of the multiplicative product of many independent random variables, each of which is positive. The best-fit log-normal curve is calculated as follows:

$$C(k) \quad = \quad a\exp\left[-b\ln k - c(\ln k)^2\right]$$

where, a=0.15, b=0.40, and c=0.16.

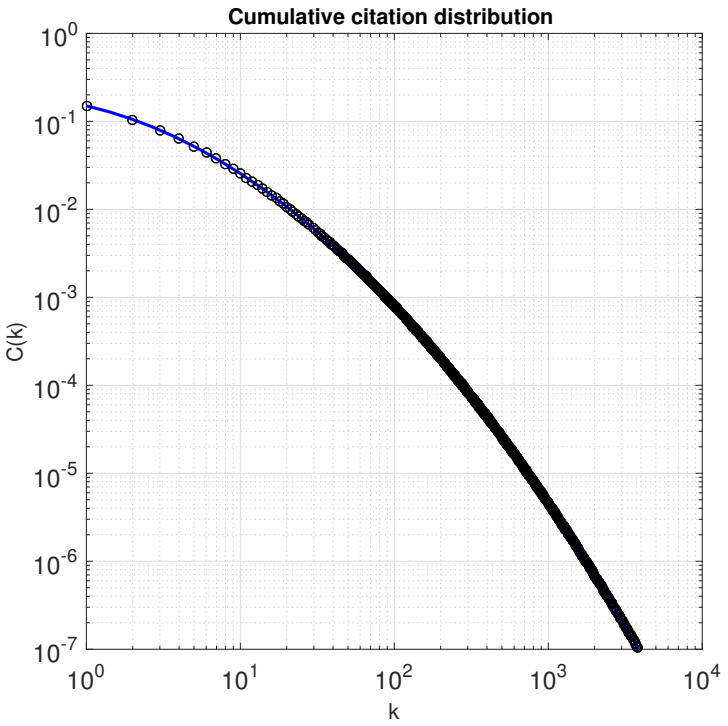


Figure 5.5: Cumulative citation distribution - Cumulative citation distribution function C(k) is plotted against the number of citations (k) in a log-log scale for all papers published between 1859 to 2012 in Computer Science journals/conferences. Circles reflect data points. The curve is a log-normal fit. Here, the graph reflects that, as the citation count k increases along the x-axis; the probability of getting atleast k citations decreases. For studying citation distributions, we consider here the cumulative function instead of taking probability P(k) as a metric to avoid statistical fluctuations. The values of constants are **a**=0.15, **b**=0.40, **c**=0.16.

The observations that we draw from cumulative citation distribution Figure 5.5 are: i) The graph clearly depicts that, as the citation count increases, the corresponding cumulative probability distribution decreases gradually. ii) The curve which scales to a log-normal form includes random multiplicative processes and shows a negative slope throughout. iii) We observe that the probability of a paper getting cited for more than 1000 times decreases to an order as low as 10^{-7}; however, the probability for papers which have fewer than ten citations is large as 0.1. iv) A separate study of the cumulative citation distribution function for journal & conference papers reveals a similar exponentially decaying curve without much deviation of the values a, b and c from the original values.

5.4.3 Preferential attachment model

Preferential attachment model highlights the well-established fact in the scientific community that papers which have high incoming citation history are the ones most likely to be cited again in the near future than the less-cited publications. As a result, we can conclude that the probability for an i^{th} paper to receive a citation again is directly proportional to the total citation count c_i that a paper has managed to collect in its effective past citation age. To measure the attachment rate, we consider the following: i) Citation count (k) for a paper in a given time frame. ii) Number of times each paper with a given citation count (k) was subsequently cited in another time window.

The observation reveals that the value of attachment rate (A_k) is a nearly-linear function for k less than 150. The log-normal behaviour of graph arises from a nearly linear preferential rate given by $A_k = k/(1 + alnk)$, where a is positive integer. The graph refers to the fact that A_k might be increasing slightly faster than linearly with respect to k. Computer Science domain has evolved in architecture and application from early 2000. Due to increased exposure to digital platforms and increased paper's visibility in applied research domains, the attachment rate for the year range 2000-2009 has shown a significant increase when compared with other attachment rates. However, linear preferential attachment rate is concluded with the following limitations – i) Linear rate usually implies the power-law form of growth, but we observe a log-normal form of distribution in our study. ii) It fails to predict the dynamics of citation bumps or sudden evolution of individual papers. iii) It does not take into account the fact that the probability of incoming citations varies as a function of time.

5.4.4 Citation age framework

With time, every research work's uniqueness fades away as new ideas and innovations are implemented in subsequent works. A key parameter that plays a significant role in citation analysis is the citation age structure. In order to investigate novel research contributions over a certain period of time; it is necessary to study citation dynamics as a function of age.

Whenever citation age equals the publication age, we can conclude that a paper has received citations uniformly during its effective lifetime and had consistent visibility throughout. Citation age distributions have two distinct underlying meanings: i) The total citation count received in a specific time period from a paper and vice versa (i.e., number of citations from a particular time frame to an article). ii) The total number of publications referred to in each citing year.

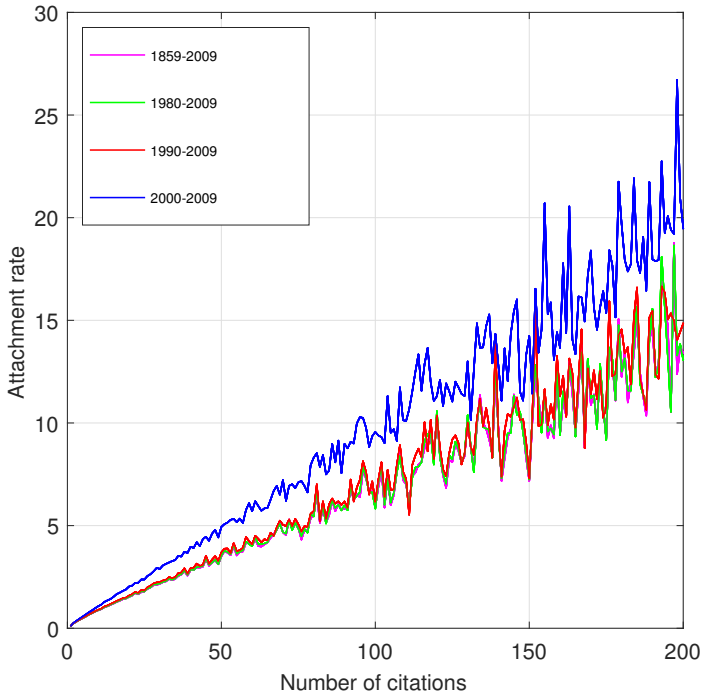


Figure 5.6: The attachment rate - The attachment rate A_k versus the number of citations k, especially for k less than 150 is being represented in a log-log scale. Different colors depict the change in k for different year ranges. Here A_k is a nearly-linear function, i.e., as the citation count increases along the x-axis, attachment rate rises almost linearly till $k < 150$. After comparison, it shows that the slope of line depicting the year range 2000-2009 is slightly steeper than the same for other year ranges. Also, we observe that the cumulative advantage models do not take into account the regularity of citations received as well as the fact that the probability of incoming citations also varies as a function of time. We find that the probability of citation decays after receiving a maximum citation in an initial growth phase, but here in this model, an older paper continues to collect citations irrespective of its age.

Over the entire data set, we find an average citation age (a_{ca}) of 7.04 years. It is also seen that the articles which were published before 2006 had (a_{ca}) less than two years and on an average received only 2.37 citations. The result indicates the fact that, in the case of less visible papers, citation count increases initially due to self-citations or due to the influence of publication venue; but soon after that, it becomes inconsiderate. In contrast, well-cited papers steadily continue to receive citations for a long period of time. Also in this process,

the reference of a highly cited paper gets an additional likelihood factor and chances of prolonged citation age. In our study, it comes out that papers with more than 100 citations have a_{ca} = 10.12 years. The 12 publications in the data set that received more than 3500 citations have a_{ca} of 19.68 years. The citation distributions varying with age decays exponentially with time. The average age with the number of citations roughly grows as $a_{ca} = k^{\alpha}$, with α almost equal to 0.3.

5.4.5 Reviewing unique citation profiles

For diverse citation analysis, we analyze highly cited papers from our data set and study their varying behavior over a span of citation age. To set accurate thresholds for 3 unique set of citation patterns, i.e., sleeping beauties, discovery papers and hot publications, we initially vary two parameters: citation count (k) and ratio of average citation age to the age of paper or publication age (r) and observe change in the count of papers Figure 5.7. As the ratio of average citation age to publication age (r) is varied from 0.1 to 1.0, publication count gradually increases. For citations greater than 250 and the average ratio (r) less than 1, we find a maximum publication count of 3019. This interesting observation reveals that when average citation age for a paper is maximum distributed over its effective publication age then the publication count increases. There is an increased count of such papers which have evenly maintained consistency in receiving citations from time to time and thus are useful for our study. Although, when citation counts k increases with varying ratio of r, paper count abruptly decreases. This leads to our previous observation of less number of highly-cited papers. The curve depicts that, as citations increase along the x-axis and the ratio r is varied from 0.1 to 1, the paper count shows an initial sharp peak followed by an abrupt decay across the surface.

5.4.5.1 Sleeping beauty or revived classics

The papers that exhibit sleeping beauty characteristics is due to factors such as the sudden expansion of research in a particular topic. The threshold that we set for sleeping beauty may differ in other domains. We get a total of eight papers in our data set that satisfies the criteria of sleeping beauties, i.e., the rebirth of an old classic. We draw a pattern Figure 5.9 for four revived classics and observe a striking similarity between them.

The statistics in Table 5.5 shows that a proportionate ratio of papers with such characteristics published in journals to conferences as 7:1. This clearly indicates that the impact of journals is always higher, and as a result, more weight may be given to journal papers. We

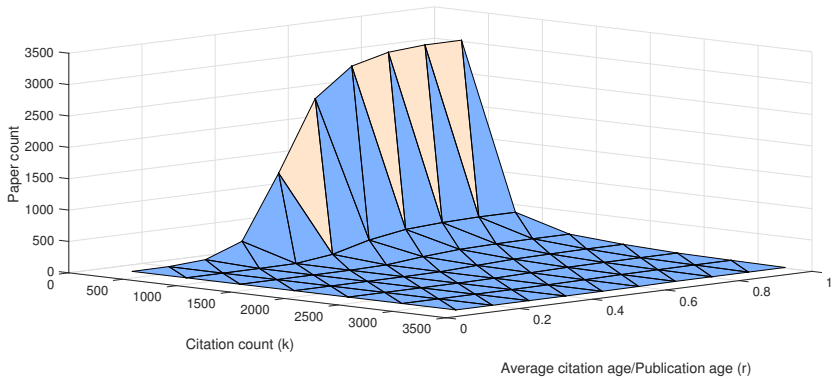


Figure 5.7: Variation in publication count with changing (k) and (r) - The x-axis denotes citation count k (i.e. greater than k citations), y-axis denotes ratio of average citation age to publication age (r) whereas z-axis denotes paper count. For finding the correct threshold in each category, i.e., for sleeping beauties, discovery papers, and hot papers, we vary x-axis and y-axis to observe the change in the count of papers. As the x-coordinate and y-coordinate values are changed, the count of papers depicts a sharp increase followed by a sharp decay from (>1000) citations throughout till citations greater than 3500.

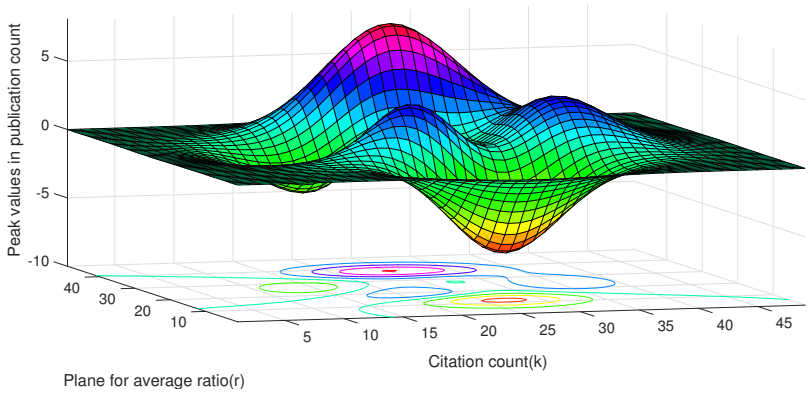


Figure 5.8: Peak points of publication count in 3D surface - Peak values obtained for count of papers where citation count (k) and ratio of average citation age to publication age (r) is varied across 49 × 49 matrix.

find one of the papers authored by Moran (Moran_SP) received in an a_{ca} of 48.77 years with 1235 received citations which are maximum among the eight sleeping beauties. The paper has an incoming citation burst after the year 2000 with an average of 80 citations/year after-

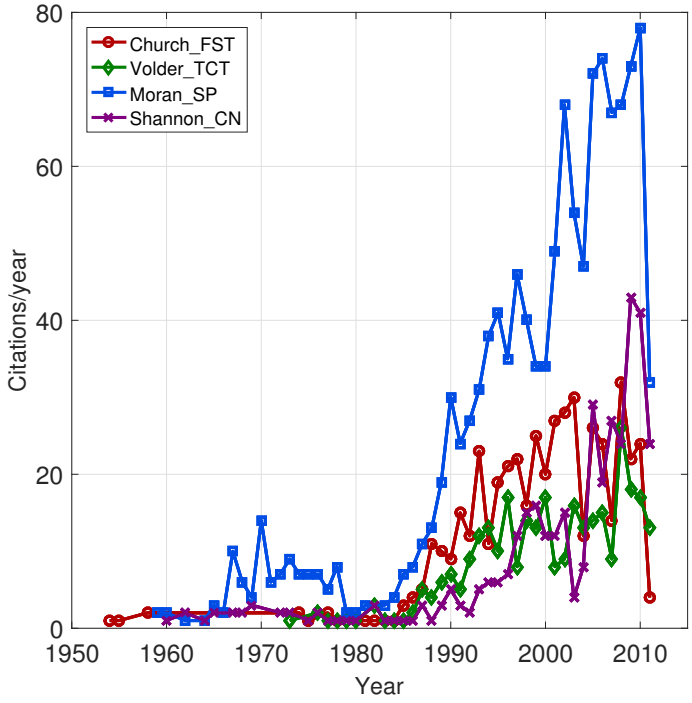


Figure 5.9: Four revived classics - The graph depicts the activity of four chosen sleeping beauties in Computer Science. The x-axis shows years after 1950 whereas annual citation count received is plotted along the y-axis. We study the activity of revived classics over its effective lifetime and find that such papers receive minimal citations and remain dormant for many years after publication, and suddenly, start to collect a huge volume of citations when lifted up by a prince (another weighted article). Though there are multiple distinct peaks encountered in its total effective lifetime, the maximum peak value is depicted in extreme later years of its citation age. For sleeping beauties, the average citation age comes out to be 49.75 years. Overall, such papers accumulate large volumes of citations with a considerably long average citation age.

ward. This draws the fact that sometimes a topic becomes central in a research network due to its diversified application in various domains.

The publication *Stochastic processes* (Moran_SP) became popular in the years between 2000-2010 due to its varied application in modelling many networks. Such citation burst after 50 years of publication is uncommon and exemplary. Another paper (Shannon_CN) published in 1949 by Shannon shows a sudden rise in citations with an average of 40 cita-

tions/year and a_{ca}=52.52 years. This is due to the fact that the topic of Communication Theory and issues in Communication Networking got popular in the early 21st century which is also clearly reflected in the data set.

Some other papers in Table 5.5 make strong affirmations in different contexts like Church's paper of 1940 regarding 'Formulation of simple theory of types' and Nash's paper of 1950 on 'Equilibrium points in n-person games' has the longest citation age of 59.34 and 56.52 years, respectively in the history of Sleeping Beauties of Computer Science. The contributions by the American mathematician John Nash forms one of the most powerful concepts of Game Theory, 'The Nash Equilibrium'. Another momentous contribution by Huffman in 1952 commonly used for lossless data compression 'Huffman Coding Algorithm' receives the second highest of 934 citations with a_{ca}=48.64 years among the eight SB's. Further, 'Lattice theoretical fixpoint theorem and its application' by Tarski is of paramount relevance in developing many further concepts on Boolean Algebra, Set Theory, and Topology, Real Functions, etc. It receives a total of 496 citations in a_{ca} of 45.8 years. Long-term relevance of a paper reveals the fact that the papers whose citation activity lasts for a longer duration in the research community broaden the scope of further researches in the field.

Table 5.5: Seven sleeping beauties (threshold ≥ 250 citations, $r > 0.7$)

Title	Authors	Year	Citation count	Average age	Publication venue	Paper notation
A formulation of the simple theory of types	Alonzo Church	1940	483	59.3436853	Journal Id(137)	`Church_FST`
A lattice-theoretical fix point theorem and its applications	Alfred Tarski	1955	496	45.80040323	Journal Id(7339)	`Tarski_FxT`
The CORDIC Trigonometric Comp. Technique	Jack E. Volder	1959	313	41.09584665	Journal Id(1043)	`Volder_TCT`
Stochastic Processes	P. A. Moran	1950	1235	48.77246964	Journal Id(912)	`Moran_SP`
A Method for the Construction of Minimum-Redundancy Codes	D. A. Huffman	1952	934	48.64025696	Journal Id(17162)	`Huffman_CMRC`
Equilibrium Points in n-person Games	John Nash	1950	456	56.52412281	Journal Id(919)	`Nash_EqGT`
The Automatic Creation of Literature Abstracts	H. P. Luhn	1958	402	45.29353234	Conference Id(943)	`Luhn_ALA`
Communication in the Presence of Noise	C. E. Shannon	1949	373	52.52546917	Journal Id(17162)	`Shannon_CN`

5.4.5.2 Discovery papers

For publications that are recognized to contribute to some breakthrough discoveries; our citation data set inquisition reveals that discovery papers experience an expeditious growth in citations in a very short average citation age. All 41 discovery papers detected from 1974 to 2002 in our data set are listed on Table B.1 of which 10 of them are on Table 5.6.

Table 5.6: Out of 41, top 10 discovery papers (threshold: ≥ 500 citations, $r < 0.4$)

Title	Authors	Year	Citation count	Average age	Publication venue	Paper notation
Myrinet: A Gigabit-per-Second Local Area Network	N. J. Boden , Daniel I. A. Cohen , Robert E. Felderman , Alan E. Kulawik , Charles L. Seitz , Jakov N. Seizovic , Wen-king Su	1995	1013	6.421520237	Journal Id(401)	Boden_LAN
The Java Virtual Machine Specification	Tim Lindholm , Frank Yellin	1997	723	5.822959889	Conference Id(1589)	Lindholm_JVMs
Low Power CMOS Digital Design	A. P. Chandrakasan , R. W. Brodersen	1996	714	6.774509804	Journal Id(5320)	Chandrakasan_CMOSd
Equation-based congestion control for unicast applications	Sally Floyd , Mark Handley , Jitendra Padhye , J"rg Widmer	2000	626	4.725239617	Journal Id(218212)	Floyd_UA
RTP: A Transport Protocol for Real-Time Applications	H. Schulzrinne , S. Casner , R. Frederick , V. Jacobson	2001	1020	3.431372549	Journal Id(991)	Schulzrinne _RTP
The Unified modelling Language User Guide	Grady Booch , James E. Rumbaugh , Ivar Jacobson	1999	1919	4.540385618	Journal Id(107)	Booch _UML
Span: an Energy-Efficient Coordination Algorithm for Topology Maintenance in Ad Hoc Wireless Networks	Benjie Chen , Kyle Jamieson , Hari Balakrishnan	2002	662	3.996978852	Journal Id(275)	Chen _CATM
Technical Reports	Douglas H. Adams , Robert H. McMichael , George E. Henderson	1992	839	5.972586412	Journal Id(459)	Adams_TR
Active messages: a mechanism for integrated communication and computation	Thorsten von Eicken , David E. Culler , Seth Copen Goldstein , Klaus Erik Schauser	1992	748	7.04144385	Conference Id (86)	Eicken_AM
Parallel discrete event simulation	Richard M. Fujimoto	1990	844	8.630331754	Journal Id(209)	Fujimoto_PDES

For analyzing the discovery papers, we have set a threshold criterion that the papers should have more than 500 citations (k) and a ratio of average citation age to publication age (r) of less than 0.4. In the time span of 1974 to 2002, a total of 41 discovery papers got published. Here, we also find a similar pattern that journal publications have contributed more than conferences with the proportion varying as 4.85 : 1.

A significant discovery paper (Schulzrinne_RTP) on RTP ('A transport protocol for real-time applications') received 1020 citations in an average citation age of 3.4 years, and it is a journal paper. The earlier years of research publications saw up spring of fields such as Database, Operating System, Modelling and Logic Programming, that included works like Database Abstraction, Relational Database Models, UNIX Time-Sharing Systems (a 1974 paper Ritchie_UTS (see appendix) that had the longest citation age of 14.09 years) and Logic Synthesis & Optimization. Gray's 1978 paper 'Notes on Data Base Operating System' with $a_c a$ of 13.45 years and 683 citation count reflects this. Another 1987 paper (Jaffart_CLP) on 'Constraint logic programming' received 752 citations with a total citation age of 9.74 years. In the late '90s, the research domain significantly shifted towards improving Communication & Networking Models and providing extensive researches in Wireless Networking, Memory Models, Parallel Architecture, Object-Oriented Programming concepts and understanding its methodology through UML, etc.

Past 20 years have seen a major upsurge in eliminating drawbacks caused by communication and complex computational problems. A major segment of the Computer Science field involves access to information, data processing and sending that information to far away distant places. Some prominent contributions include 'RTP' (Schulzrinne_RTP) with 1020 citations. A 2002 paper 'Span' (Chen_WN) that made major contributions in developing energy-efficient coordination algorithms in wireless networking.

5.4.5.3 Hot papers

Along with citation counts, regularity of intervals in which citations are received is also a major factor in measuring the consistency of performance in research works. For defining hot papers, we carefully measure the distribution of citation age over its effective publication age and thus, consider those well-cited publications whose citation age is maximum distributed over its publication age to inspect the year-wise distribution of incoming citations for highly-cited papers. We thus set the threshold ratio of average citation age to publication age greater than two-thirds of unity.

Redner's 2005 publication [56] on physical review journals defined hot publications as those papers whose citation count is greater than 350 and ratio (r) of average citation age to publication age is greater than or equal to 2/3. We objectively re-define the threshold for hot papers as those papers which receives greater than 1500 citations and whose ratio (r) is greater than 2/3. We got 23 publications as 'hot papers' (see in appendix), out of which ten

Table 5.7: Out of 23, top 10 hot papers (threshold: ≥ 1500 citations, $r > 2/3$)

Title	Authors	Year	Citation count	Average age	Publication venue	Paper notation
Rough sets	Zdzisław Pawlak	1982	1789	24.11514813	Journal Id(79)	Pawlak_RS
The mathematical theory of communication	C. E. Shannon, W. Weaver	1964	3727	38.88596727	Journal Id(909)	Shannon_TC
Indexing by Latent Semantic Analysis	Scott C. Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard A. Harshman	1990	2214	15.22312556	Journal Id(141)	Dumais_LSA
Abstract interpretation: a unified lattice model for static analysis of programs by construction or approximation of fixpoints	Patrick Cousot, Radhia Cousot	1977	1828	25.49452954	Conference Id(255)	Cousot_LM
R-trees: a dynamic index structure for spatial searching	Antonin Guttman	1984	2231	18.67144778	Conference Id(370)	Guttman _RT
Particle swarm optimization	James N. Kennedy, Russell C. Eberhart	1995	2573	13.04897007	Conference Id(2133)	Kennedy _PSO
Classification and Regression Trees	Leo Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone	1984	3610	18.82742382	Conference Id(2614)	Brieman _RT
Network information flow	Rudolf Ahlswede, Ning Cai, Shuo-yen Robert Li,Raymond W. Yeung	2000	1577	8.562460368	Journal Id(433)	Ahlswede_NIF
Fuzzy Sets	Lotfi A. Zadeh	1965	6606	38.48622464	Journal Id (40)	Zadeh_FS
Computers and Intractability: A Guide to the Theory of NP-Completeness	Michael Randolph Garey, David S. Johnson	1979	11788	23.08322022	Conference Id(277)	Garey_NP

papers have been listed in Table B.3.

The ratio of journals to conferences among the said 23 hot papers is 2.83:1. This clearly reflects that journals have consistently continued to impact a major collection of well-cited hot papers. The two papers with the most citations over the entire data set, a journal paper with 6606 citations Zadeh_FS and a conference paper with 11788 citations Garey_NP are interestingly hot papers. The newest hot article on a breeding topic 'Network Information Flow' has collected 1577 citations in an average age of 8.56 years where extensive research on Networking and Communications expanded in the late 1990ś.

The topics that influence hot papers throughout including Artificial Intelligence, Com-

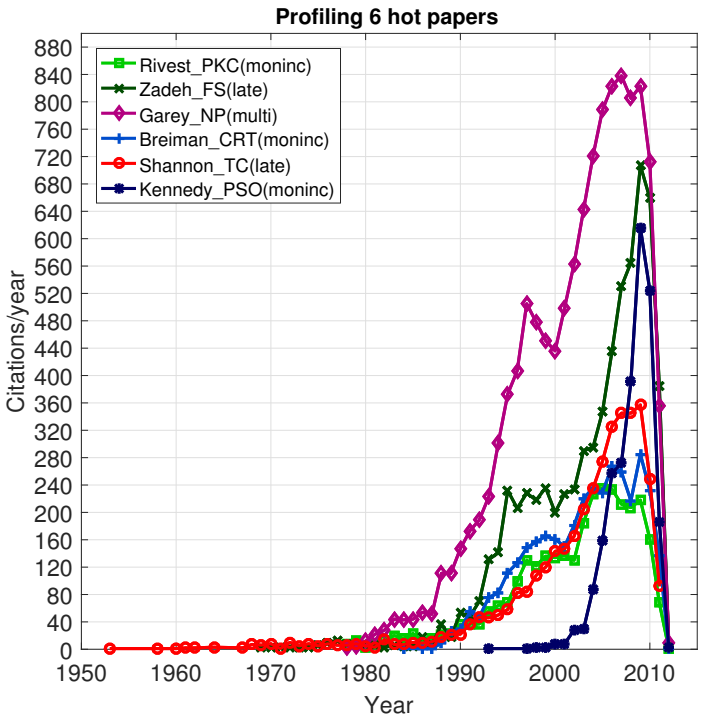


Figure 5.10: 6 Hot papers - In this figure, the x-axis denotes years after the publication of a paper, y-axis denotes annual citations (citations/year). We find that the citation curves mostly fall under three citation trajectories – late peaks, monotonically increasing curve or multiple peaks. The percentage of hot papers in monotonically increasing category (39.13%), multiple peaks (17.39%) and late peaks (43.47%) where hot papers refer to some significant research contributions whose citation activity is widely distributed for a very long citation age.

munication & Networking, and Algorithms. Another 1985 paper on growing topic Digital Communications receives 5376 citations in the average age of 18.44 years. An unusual 1963 paper `Gallager_PCC` which is 49 years old is also classified under hot papers.

As per Table B.3, we categorize the total number of net 23 hot papers into three citation trajectories. Studying the citation profiles of these 23 publications, we observe that the most of the hot papers exhibit either of the three categories of citation trajectories: *late peak*, *monotonic increasing curve* or *multiple peaks*, with maximum papers exhibiting *late peaks*. Such proportions clearly indicate that hot papers have maximum acquired citations in the year range after the 1990's due to increased visibility and gradual expansion of researches in

the Computer Science domain after those years. Though a majority of the hot papers took some initial years to attain sufficient visibility, such papers maintain growth in collecting citations even after several years of publication.

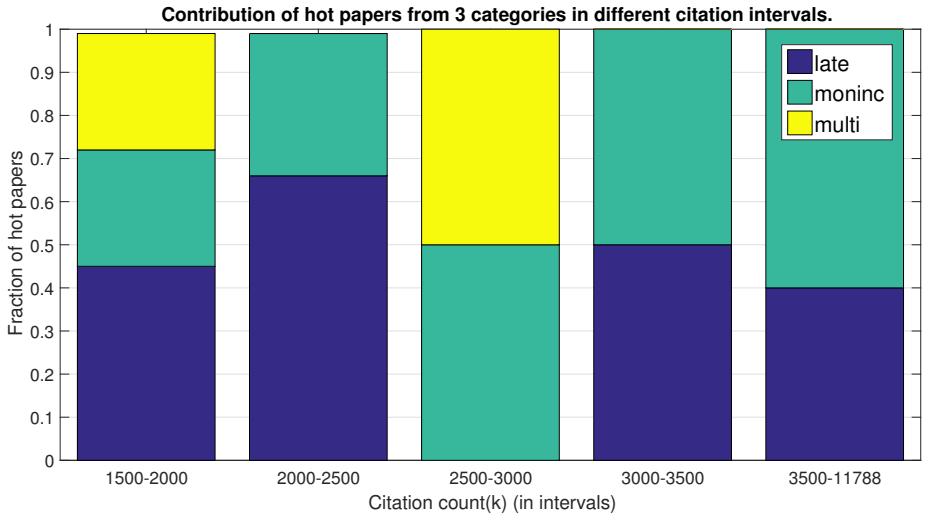


Figure 5.11: Fraction of hot papers in three different citation trajectories - x-axis denotes citation count (k), whereas y-axis denotes the fraction of hot papers in each of the three categories – late peak, monotonic increasing, and multiple peaks. The range of citation values is divided into five intervals starting from ($k > 1500$) in which for a specific category the share of hot papers is divided by the total publication count.

5.5 Conclusion

Performing statistical analysis on Computer Science citation data set, we retrieved a structured outlook in determining uniform patterns of citation distributions. We observed that citation distributions over time depict a log-normal behavior, and thus, it could be used to derive the probability of n different publications that are inclined towards getting cited in the near future. This chapter clearly provides strong evidences that the highly-cited publications get attached with a certain likelihood factor that helps them later to receive even more number of citations, and thus, to have increased in the net citation count (k). Moreover, characteristic behavior followed by major topics in Computer Science over the years has been studied, which give rise to visualization of the pattern of evolution of papers that have led to some ground-breaking researches and discoveries. We classified them under *Hot*

Papers and *Discovery Papers*. These papers tend to acquire a large number of citations within a very small average citation age. Analyzing citation count as a function of age, we find that citation-based comparisons favor papers that had been published in the past, and also, to established researchers. The unique *Sleeping Beauty* phenomenon has been widely investigated for papers which have remained dormant for long and has started receiving sudden citation bumps. Throughout the paper, we tried to establish fundamental metrics to gauge uniform predictability for the varying patterns of citation distributions.

5.6 Summary

∗ With increasing paper and author rate, size of bibliographic databases have grown ten-fold. We try to derive common citation patterns from paper-paper citation network and hence, scale citation distribution into generic models such as cumulative model and preferential attachment model. Cumulative citation distribution function depicts a log-normal form of a curvature with a negative slope throughout. Studying separately for journal & conference papers we could not observe any deviation in the values for a, b, c in the $C(k)$ function where $a = 0.15$, $b = 0.40$ and $c = 0.16$.

∗ Further, we observe an increase in attachment rate for time duration 2000-2009 due to a rapid stretch of visibility and an increase in research works in the Computer Science domain. Though there are certain drawbacks accounted in the generic models. Preferential attachment model continues to accumulate citation for papers even when a paper has reached its obsolete phase, i.e., when no further citation activity is accounted for. One major drawback here is that it does not take into account the probability of incoming citations as a function of publication age. It thus continues to credit papers with citations even after its effective publication age.

∗ A comparison is drawn, and impact of publication venue in common citation patterns are observed in the applied domain of research such as Computer Science. There is a number of conference papers in total; however, among the top 100 highly-cited papers, 79.59% are journal papers and only 20.40% are conference papers. Journals show consistent growth throughout effective publication age; on the other hand, conference papers gain quick popularity and therefore, play an important role in the applied research domain. In particular, the maximum cited paper present in our data set is a conference paper.

∗ We broadly categorize citation trajectories into five classes- Early peak, late peak, monotonically increasing peak, monotonically decreasing peak and multiple peaks. Citation activity of well-cited journal papers depicts monotonic increasing accumulation of citations and late peak; whereas, conference papers show multiple peaks at different time points with monotonic decreasing range.

∗ Citation is also studied as a function of age. Average citation age a_{ca} grows as $a_{ca} = k^{\alpha}$, where α almost equals 0.3. For papers which collect more than 3500 citations, a_{ca} comes out to be 19.68 years for them, that is, highly cited papers take a considerable amount of time to die down. The varying ratio of average citation age to publication age or age of paper r from 0.1 to 1.0, paper count also increases. This draws a conclusion that there is a significant collection of papers whose active citation age is maximum distributed over its effective publication age.

∗ We extract unique class of well-cited papers from our data set - Sleeping Beauties, Discovery papers, and Hot papers and hence, study their citation behavior. For 'Sleeping Beauties' in Computer Science domain, we set threshold as $k > 250$ and $r > 0.7$. We get eight sleeping beauties. Such papers exhibit late peak over effective publication age and become noticed much later after publication year. For 'Discovery Papers', we set threshold as $k > 500$ and $r < 0.4$. We get 41 discovery papers. Such papers experience an expeditious growth in citations in a very short average publication age. For 'Hot Papers', we set threshold as $k > 1500$ and $r > \frac{2}{3}$. We get 23 hot papers. Such papers have their average citation age maximum distributed over its effective publication age and hence, consistently collect citations.

Chapter 6

Exploring and reasoning citation patterns among journals

6.1 Introduction

In the research community, scientific journals are an important choice of publication venue. For most authors, publication in high impact factor journal plays a decisive role in hiring and promotion. For research groups, institutions and even nations, it determines ranking and funding decisions. In the pre-2000 era, a common trend depicts that fewer research articles were published along with a limited range of visibility and accessibility of journals. After the 2000-2005 period with wider usage of evaluation metrics such as impact factor for journal [127, 128] and h-index, g-index [23, 30] for an author, tremendous publication and citation pressure prevails in the academic community. On a macroscopic level, the impact of publication venue that is, journals are getting tossed up.

Related Work: Recently, there has been a volume of work that studies and report on issues like disrupting publication ethics [8, 13, 82, 83, 129, 130] and tampering with impact factor [17, 18, 20, 21, 127, 128] using unethical citations such as self-citation [81, 131–134], citation cartel [10, 11], citation stacking [9] and in general, adding anomalous citation [12, 16]. However, there are evidences that journal and their authors, editors and publishers attempt to artificially boost impact by manipulating citation and paper count [12, 13, 16, 20, 130].

Recent studies create awareness against unethical practices and addition of anomalous citations to bias impact factor. More commonly, it is reported for the journal where editors, publishers, and authors manipulate citation [8, 9, 20]. Wilhite and Fong in their paper

[8, 129] organized a survey for researchers in multiple business and sociology domain to understand the occurrence of coercive self-citation. Here, coercive self-citation is referred to the condition where an editor forces an author to add citation of editor's journal irrespective of content matching in lieu of publication. Analyzing 6,672 responses from survey and data of around 832 journals, this practice is uncommonly common in business disciplines more than sociology and psychology domain. Moreover, it is stated that 'Journal of Business Research' has the highest of 49 number of coercive observations. Other factors that affect it includes the author's academic rank, number of authors on a paper, the discipline of journal and type of publisher, etc.

Therefore, out of many anomalous citations reported in the literature, easiest trick is just by adding excessive 'self-citation' [9, 131–134]. Both authors and journals are practising it. In an article by Arnold [13], several cases of both author and journal misconduct have been reported. There are works which have started to question and redefine the metric 'impact factor' itself [21, 127, 128]. Chorus and Waltman [81] proposed a measure, called IFBSCP (Impact Factor Biased Self-Citation Practices), a ratio between the share of self-citation count to papers published particularly in last two years (considering impact factor time window) to the relative share of self-citations to papers published in that journal for preceding five years. Stephen M. Lawani [131] discusses different classes of self-citations, namely diachronous and synchronous self-citations.

One newly identified pattern is 'Citation Cartel' [83] where a group of journal disproportionately exchange citations mutually benefiting each other. Fister *et al.* [11] approached towards identifying cartels in authors by finding interlinked relationships in multi-layer networks, that is, paper-paper and author-author citation networks. Cui *et al.* [47] showed the relationship between researchers by combining multiple networks (paper citation, author citation, and author collaboration network) with an aim to improve link prediction in citation network.

Mongeon et al. [10] attempts to address issue of 'Citation Stacking' and 'Citation Cartels'. Another paper [9] identifies a case of citation stacking where, three Romanian physics journals of same publisher Editura Academiei Romane receive citations with unequal variance. Bai [12] discovered anomalous citations between articles using their collaborative time-factors. Fong [16] discussed problem of Coercive Citation (adding irrelevant citations) and Padded Citation (add unnecessary citations to manuscripts prior to submission).

Since 2009, Thomson Reuters (currently known as Clarivate Analytics) in annual jour-

nal citation report consisting of journals they index, started reporting and blacklisting those journals which were excessively self-cited and involved in other anomalous citation patterns such as “Citation Stacking" [17, 18, 81].

Motivation & Objective: Extensive study of anomalous citations has been done in domain of multi-disciplinary, biological, sociological, psychological and business sciences [8, 9, 129]. However, to the best of our knowledge extensive study in the domain of Computer Science journals is not available. Using similar techniques, it would be interesting to see the volume of such pieces of evidence in the Computer Science domain. Even though there are types of unethical citation practices reported in literature, however, general models or patterns of some anomalies are yet to be proposed. Along with identifying patterns, it is challenging to see if other microscopic features or parameters can be used to identify such anomalies.

Contribution: We propose some generalized citation patterns broadly classified into self-loops, pairwise mutual citations and group mutual citations which might be useful to detect citation anomalies. Based on these patterns, we collect evidence and conduct a thorough analysis. On a bibliographic data set in Computer Science domain consisting of 2,621 journals, we dig out factors that inherently give rise to such patterns. We also dig into other microscopic features such as time series change in impact factor, abrupt increase in average publication count of donor journal during that specific period, publisher’s network, author editorial board network, author co-author network, trending advertisement policies of some journals, narrow domain specialization which are inevitable and instinctively may give rise to such patterns. Our study helps to narrow down the sample size of possible data set of anomalies. It also helps to give a broader perspective on which features might or might not be competent to classify a journal being involved in the possible anomalous pattern? Because without solid ground proof, we should be very cautious at pointing out at a journal being involved in malicious citations.

This chapter is organized as, in Section 6.2, we summarize the data set used for our experiments and report a total count of several bibliographic attributes such as author, paper, journal, reference, etc. We also present journal centric characteristics of the data set. In Section 6.4, we describe methodology used. Section 6.4.1 categorizes self and mutual journal citation behavior into possible geometrical patterns. Section 6.5 reports a detailed analysis on time basis impact factor study for journals for all derived patterns and fundamental features behind such patterns. We study the general trend of impact factor curve along with the

visibility of sudden peaks. In Section 6.6 we conclude with a discussion on major findings, importance and future work in this domain of study.

6.2 Data set description

We crawl one of the largest available bibliographic data sets in Computer Science domain, “Microsoft Academic Search" (MAS) which includes as of 2012, fetched 2,281,307 papers with unique paper indices and corresponding bibliographic attributes such as publication year, venue, author, citation list, etc refer to Figure 6.1. We obtain a total of 11,86,413 authors and 1,64,63,489 citation. A total number of journal papers with at least one citation are 4,90,249. A total number of journals at least one citation are 2,621. We restrict our study to a publication period ranging from 1990-2012. In the last decade, there is 42% increase in journal count along with 60% inflation in publication count and 74% increase in citation count with respect to time. For journal centric data set summary, please refer to Table 6.1. On average, a journal gets citations from 62 journals and is effectively cited for seven years.

Table 6.1: Journal centric data set visualization

Entities	Minimum	Maximum	Mean	Standard Deviation	Inter Quartile Range
`Total incoming citations from all journals`	1	89,920	968	4448	147
`Total outgoing references for all journals`	1	96,003	968	3885	303
`Citation age of journals`	1	23	7.20	7.83	13
`Reference age of journals`	1	23	10.79	8.18	17
`Count of distinct journals referred`	1	970	61.53	100	75
`Distinct journals from which citations received`	1	1358	61.53	127.002	51

Further, dividing the research articles respective to their publication venue, we get 54.95% conference papers whereas 45.04% are journal papers. Since 2000, publication of research articles has seen 60% inflation with 74% growth in citations. In this chapter, we have considered only journal-journal citation graph. Almost a double increase in the number of articles per journal is observed where the ratio is seen to rise from 3,116 in 2000 to 6,332 in 2012.

6.3 Terminology and definitions

Citation link: A citation link is said to form between two journals J_i and J_k if there exists a directed edge from J_i to J_k or vice versa such that there is a cite/being cited by relationship.

Citation age (t_{ca})**:** Citation age refers to the count of years when a given journal (J_i) has received at least one citation.

Mean citation rate for a given journal (M_{cin})**:** For a given journal in the journal-journal citation network, M_{cin} refers to the total incoming citation collected in its effective citation

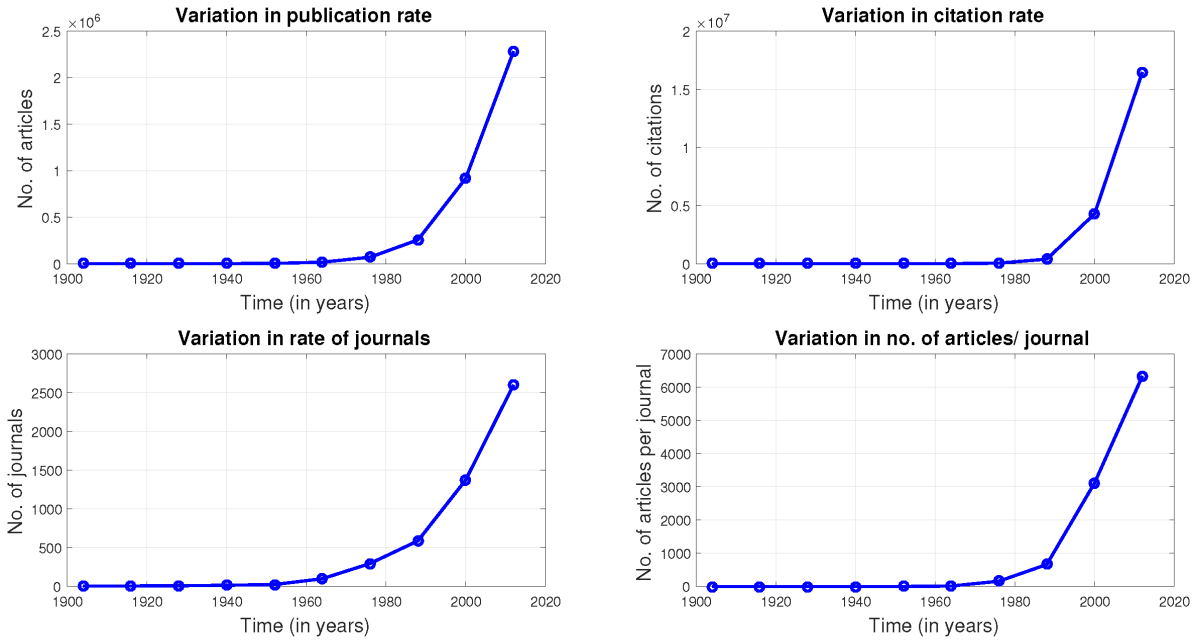


Figure 6.1: Growth in the rate of publication, citation, number of journals and number of articles published per journal on a time basis from 1900 till 2012. We observe that since a period from 1990 to 2000, there is rapid inflation in each of these figures. In the last decade, there is 42% increase in journal rate along with 60% inflation in publication rate and 74% increase in citation rate.

age as measured from year wise journal citation distribution.

Mean reference rate for a given journal (M_{rin})**:** For a given journal in the journal-journal reference network, M_{rin} refers to the total outgoing citation or references given by a journal in its effective citation age.

Overall mean of incoming citation for all journals (M_{acn})**:** It refers to mean of all incoming citations from remaining citation links in the incoming citation network where x_{ik}>=M_{cin}.

Overall mean of outgoing citation for all journals (M_{arn})**:** It refers to mean of all outgoing citations from the remaining citation links in the outgoing or reference network where x_{ik}>=M_{rin}.

Impact factor: Impact factor refers to the ratio between the total incoming citations received by the recently published articles in that journal for two preceding years to the total publication count considering the articles published in the same time window [117]. It mea-

sures the frequency with which on an average, an article in a journal has been cited during that particular time frame.

6.4 Methodology

Journals remain a predominant choice of publication even in the applied domain of research such as Computer Science due to its prolonged citation impact and visibility. Unlike conferences, journals exhibit a steady citation growth and a continuing long span of publication age. Therefore we conduct our experiments on detecting such citation patterns considering journals.

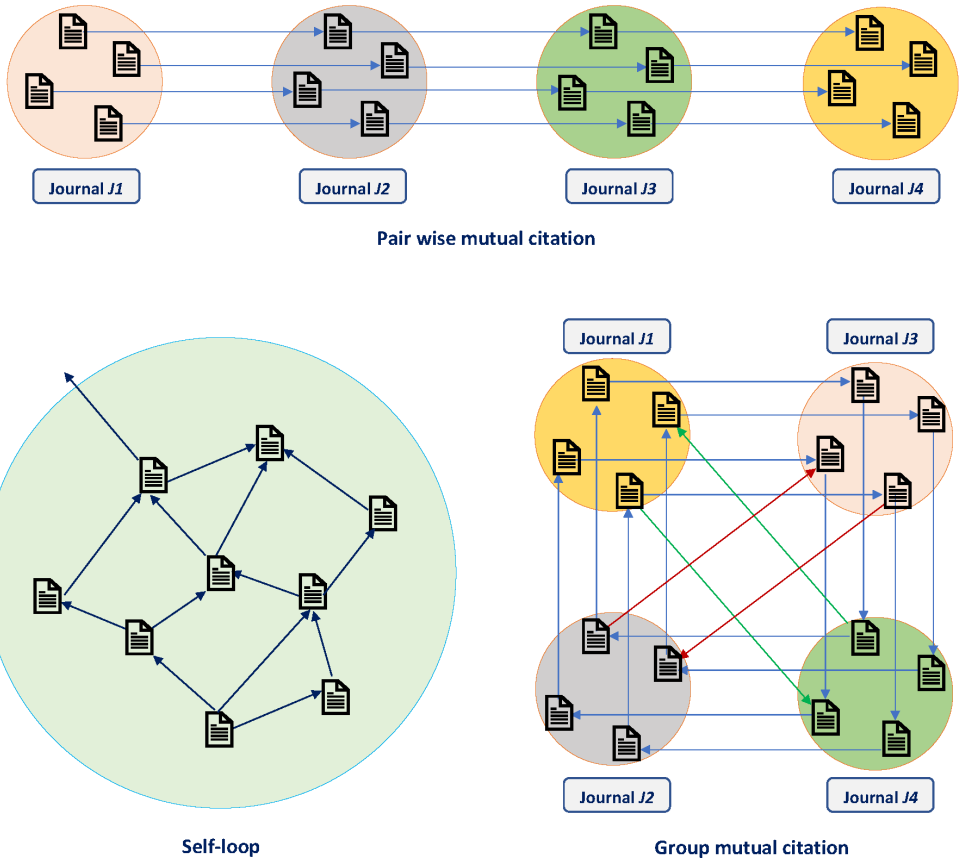


Figure 6.2: Citation patterns which are extracted from data set are shown. Circle depict journals. The figure shows how paper-paper citation network is grouped to form journal-journal citation network. Patterns are broadly classified into: *self-loop*, *pairwise mutual citation* and *group mutual citation*.

Initially, we simulate a paper-paper citation network with 490,249 nodes in the form of

a weighted directed graph for papers published in journals and are being cited/cite at least one corresponding journal publication. In the graph G(V,E); vertices (V) refer to journal papers and weighted directed edges (E) refer to incoming or outgoing citations between them. Next, we reduce the huge network formed with 490,249 vertices by grouping those papers based on the respective journals they are published in. A striking fall in vertex count is seen, and the resultant journal-journal citation network consists of only 2,621 nodes. To illustrate when papers P_1, P_2,.... P_n papers published in journal J_i cites Q_1, Q_2,....Q_m papers respectively published in journal J_k; a directed edge is drawn from J_i towards J_k with total outgoing citation (x_{ik}) from all papers of J_i pointed towards J_k as weight of the edge between them as in Figure 6.2. In form of adjacency matrix, it is represented as

$$A = [x_{ik}]_{m \times n}$$

where, no. of vertices in the graph is given by $m \times n$ and x_{ik} refers to citation count from vertex J_i towards vertex J_k.

Partitioning: From journal-journal citation graph, for a single vertex three types of directed edges, can exist- self-directed loop, incoming edge, and outgoing edge. Moreover, it is seen that reputed journals with large impact factor receive high incoming citation rate and in turn, refer to fewer journals in the same domain. In contrast, journals belonging to an evolving new inter-disciplinary research field are prone to increase outgoing citation rate and self-citations to enhance the range of visibility to a wider scope in the community. To take into account the contribution made by a vertex in all the directions equally, we partition the graph into two subgraphs-*incoming citation graph* and *outgoing reference graph*. The main objective to do a separate study is to individually consider the weight age of those journals which exhibit superfluous citation growth only in one direction. Common vertex and edge relationships are omitted from either of the networks.

Filtering out edges: Next, weak edges are filtered out. The data set reveals a soaring increase in citation and publication rate, i.e. (number of publications/journal) since 1990. Hence, we restrict our study to a publication period ranging from 1990-2012. For each vertex, we consider its citation distribution on a temporal scale and calculate mean citations collected (M_{cin}) or referred (M_{rin}) by a journal over its effective citation age t_{ca} (considering only those years when the journal has received or given at least one citation.) The filtered sub-network formed consist of all 2621 nodes but only those edges remain which has weight

Table 6.2: Citation graph (Incoming)

Mean citations	Nodes	Edges
>55 (80%)	520	1394
>110 (60%)	428	1047
>166 (40%)	353	845
>221 (20%)	319	717
Mean>=277	281	620
>332 (120%)	246	525
>388 (140%)	224	461
>443 (160%)	207	408
>498 (180%)	191	380
>554 (200%)	176	342

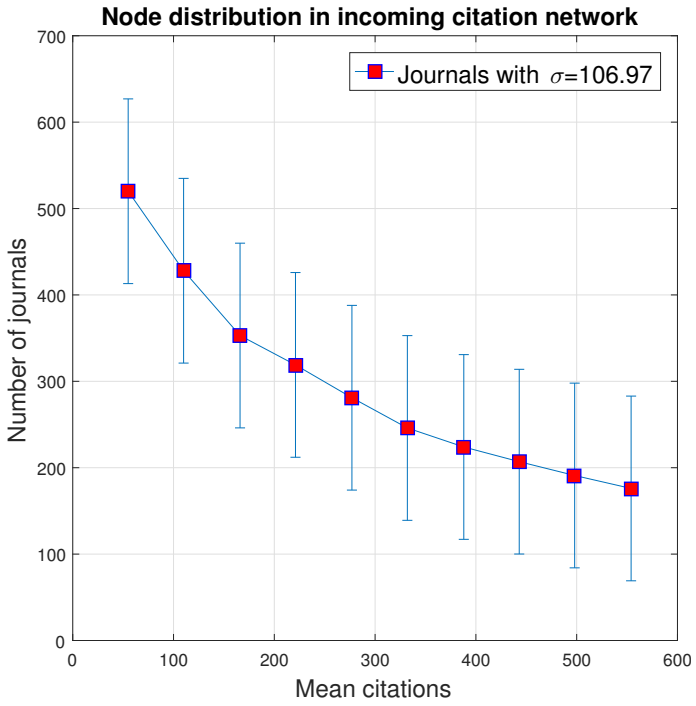


Figure 6.3: Mean variation in outgoing reference graph to filter out insignificant edges. In the resultant sub-graph, we consider 60% of mean (> 110 citations) 428 nodes, 1047 edges.

(x_{ik} greater than M_{cin} or M_{rin}. Thus, the remaining edges linked to a vertex has a weight greater than the mean citation. A shortcoming observed is when a vertex (journal) has a low mean citation rate over its effective citation age (i.e., M_{cin} and M_{rin} is very small), corresponding edges x_{ik} considered in the network are insignificant which do not make any contribution towards anomaly. Their effect on the network needs to be nullified. Hence, we filter out such vertices by recalculating mean citation rate (M_{acn} and M_{arn}) for the graph.

Filtering out vertices: M_{acn} and M_{arn} for incoming citation graph and outgoing reference graph is 277 and 182 respectively. Varying mean in both the directions, we see a change in a number of vertices and edges in the resultant network as shown in Table 6.2 and Table 6.3. The standard deviation (σ) for vertex count in incoming citation graph is 106.97 as shown in Figure 6.3, and outgoing reference graph is 128.75 6.4. For related reasons, we relax mean by σ taking 60% of M_{acn} (> 110 citations) in the incoming citation graph and 40% of the mean(> 109 citations) in the outgoing reference graph . Moreover, citation distribution on a temporal scale for impact factor years is feasible to study only if the weighted edges have values > 100. A drastic fall in vertex count from the resultant network with 428 vertices,

Table 6.3: Citation graph (Outgoing)

Mean Citations	Nodes	Edges
>36 (80%)	2111	688
>72 (60%)	1695	559
>109 (40%)	1407	492
>145 (20%)	1181	430
Mean>=182	1030	397
>218 (120%)	908	363
>255 (140%)	790	333
>291 (160%)	702	308
>328 (180%)	616	276
>364 (200%)	569	260

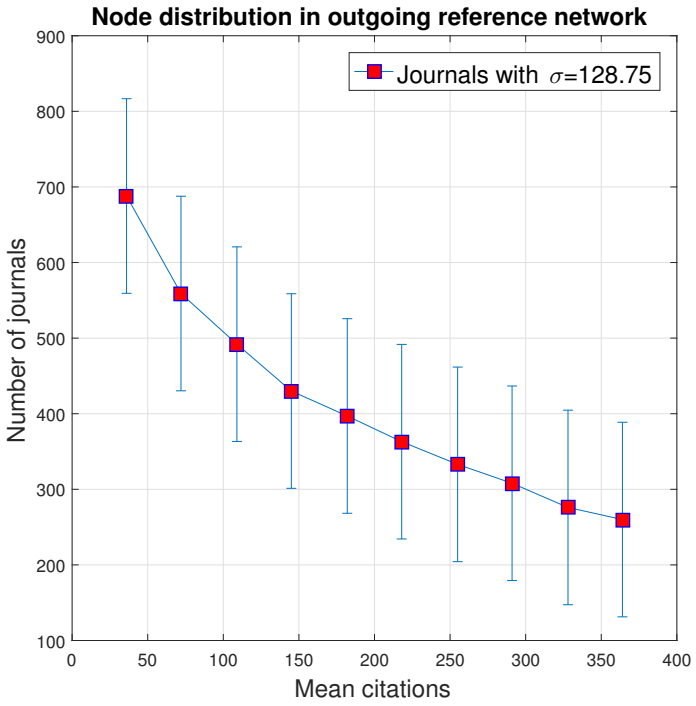


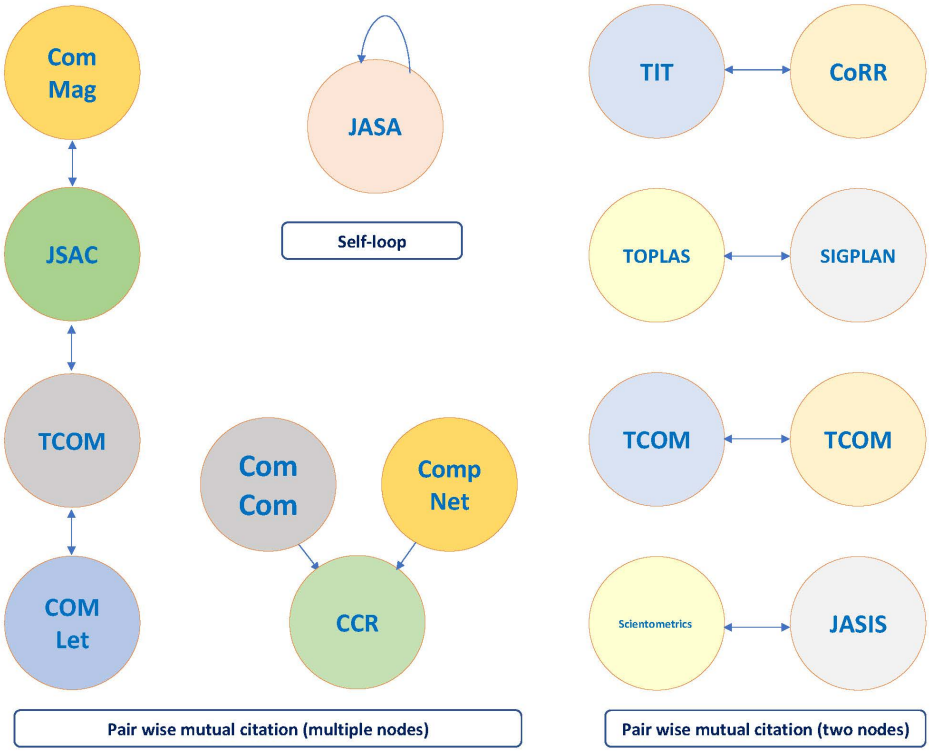
Figure 6.4: Mean variation in outgoing reference graph to filter out insignificant edges. In the resultant sub-graph, we consider 40% of mean (> 109 citations) 559 vertices, 1695 edges.

1047 edges, and 559 vertices, 1695 edges respectively in the resultant incoming citation and outgoing reference graph is obtained.

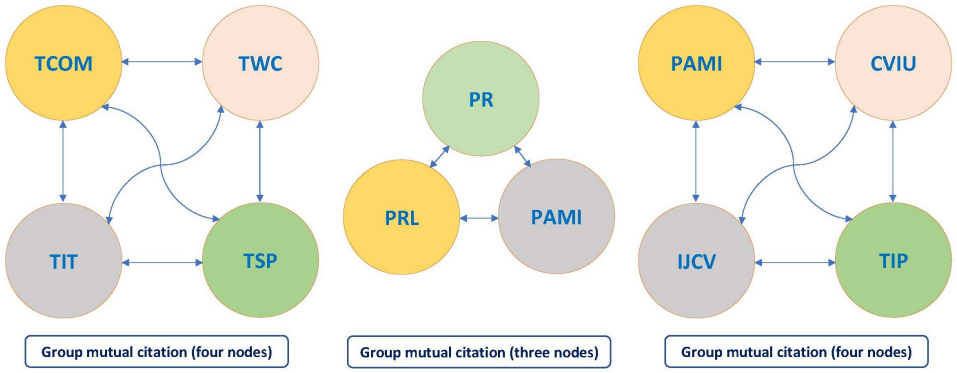
We further, do a microscopic study to measure impact factor biasness on a temporal scale where, for any given year (y), the ratio of number of citations collected by recent publications of journal for preceding two years (y-1), (y-2) and total number of publications in that time period is calculated. Next, on a temporal scale for each journal we find up to 3^{rd} highest citing contributors (journals) during the calculated impact factor years and remove incoming citations from them to re-calculate a revised impact factor (RIF).

6.4.1 Categorization of geometrical citation patterns among journals

From the resultant graph, we extract journals grouped together in the form of common geometrical patterns. These patterns include *self-directed loop*, *uni-directed citation* (a group of journals with large uni-directed citation flow towards any single journal), *pairwise mutual citations* (two journals connected by bi-directional weighted citation edge). Further, we extend *pairwise mutual citations* to other patterns such as *group mutual citation*.



(a)



(b)

Figure 6.5: **(a)** Different citation pattern extracted between journals are shown – *self loop*, *uni-directed citation*, *pairwise mutual citation*. In pairwise mutual citation, two journals are involved in mutual citation which is also extended in form of chain for more than two journals. **(b)** Pairwise mutual citation is extended to group mutual citation – *group mutual citation (3 nodes)* and *group mutual citation (4 nodes)*. In group mutual citation, three or four journals are all involved in mutual citation with each other.

Table 6.4: Number of cases identified in each of citation patterns

Name of pattern	Number of cases
Self loop	361
Pairwise mutual citation (only 2 nodes)	123
Pairwise mutual citation (more than 2 nodes)	56
Group mutual citation (3 nodes)	24
Group mutual citation (4 nodes)	11
Uni-directed citation	43

6.4.1.1 Self loop:

Self-loops are a common phenomenon. However, in recent years Thomson Reuter's indexing firm does not accept impact factor biased excessive self-cites in a journal. During 1990-2000 period, self-citation has abruptly increased due to the expansion of research in Computer Science domain. Few journals in our data set also receive as high as 80% to 90% self-cites. Excessive self-citation is a frequently observed property in the newly published journal. In contrast, we are more interested in finding why excessive self-cites consistently occur in some high impact factor old journals also? For instance, `Journal of The Acoustical Society of America` receive 85.71% self cites in its entire publication age of 82 years. We specifically study journals which receive greater than 55% of total citation from self-cite. We find 22 journals which are excessively self-cited. Top 5 excessively self-cited journals are `Journal of The Acoustical Society of America (JASA)`, `IEEE Transactions on Energy Conversion (ENCONV)`,`Environmental Modelling and Software (ENVSOFT)`,`ACM Sigcse Bulletin (ASB)` and `Computing Research Repository (CoRR)`.

6.4.1.2 Pairwise mutual citation (2 nodes):

A pair of journals connected by bi-directional weighted citation edge such that both of them are among highest citation contributing journals to each other present in the extracted list of 10 journals (see Figure 6.5). It might happen that both directed edges are not equally influential. Therefore, we assign a single $CouplingWeight$ (w). We calculate it as if there is a large difference between two citation weight such that $\sigma > \mu$ then larger weight among them becomes w; otherwise, we take mean of two weights. For an example, large mutual citation occur between two journal `IEEE Transactions on Information Theory (TIT)` and `Computing Research Repository (CoRR)` and `Bioinformatics` and `BMC Bioinformatics`.

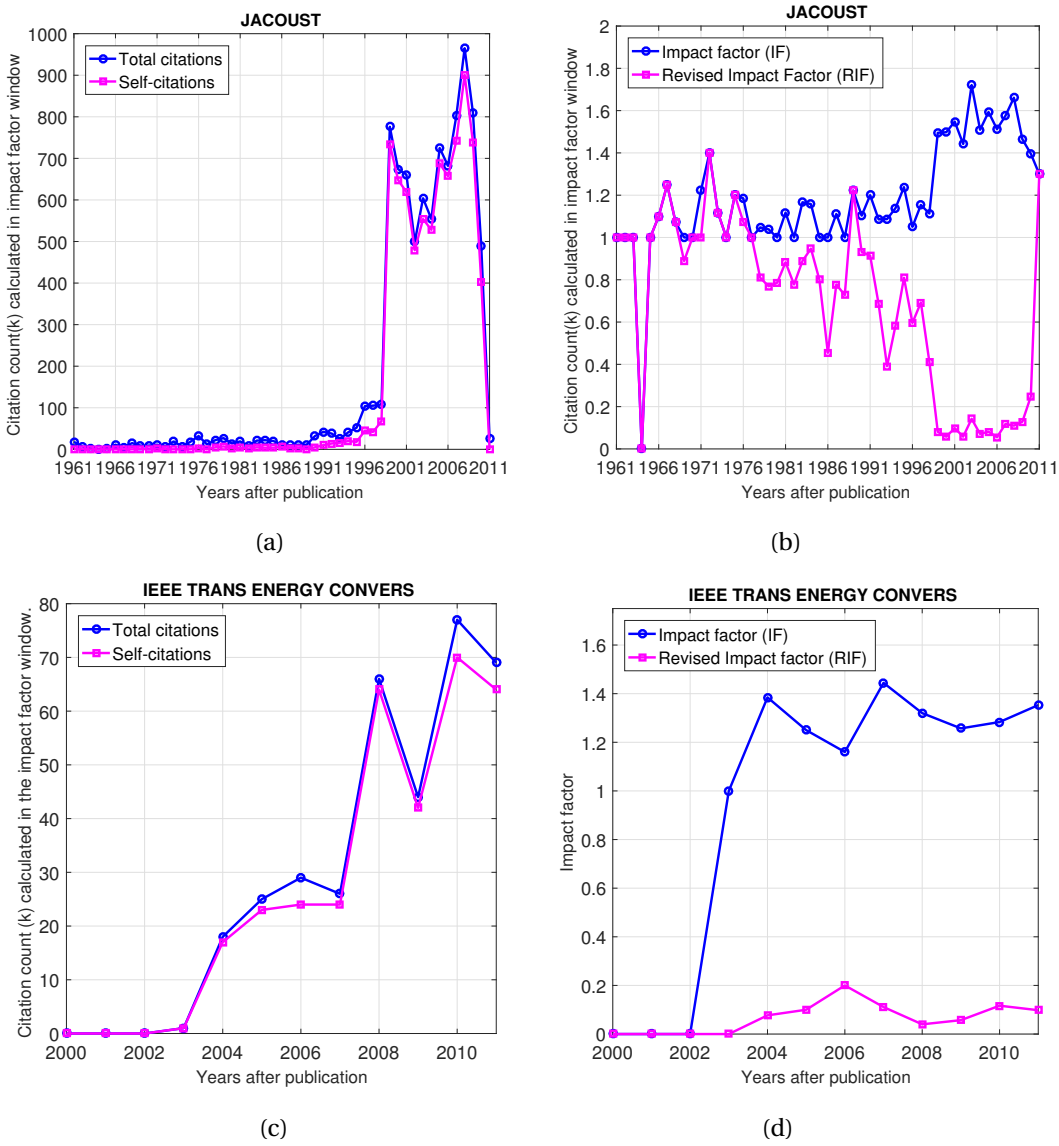
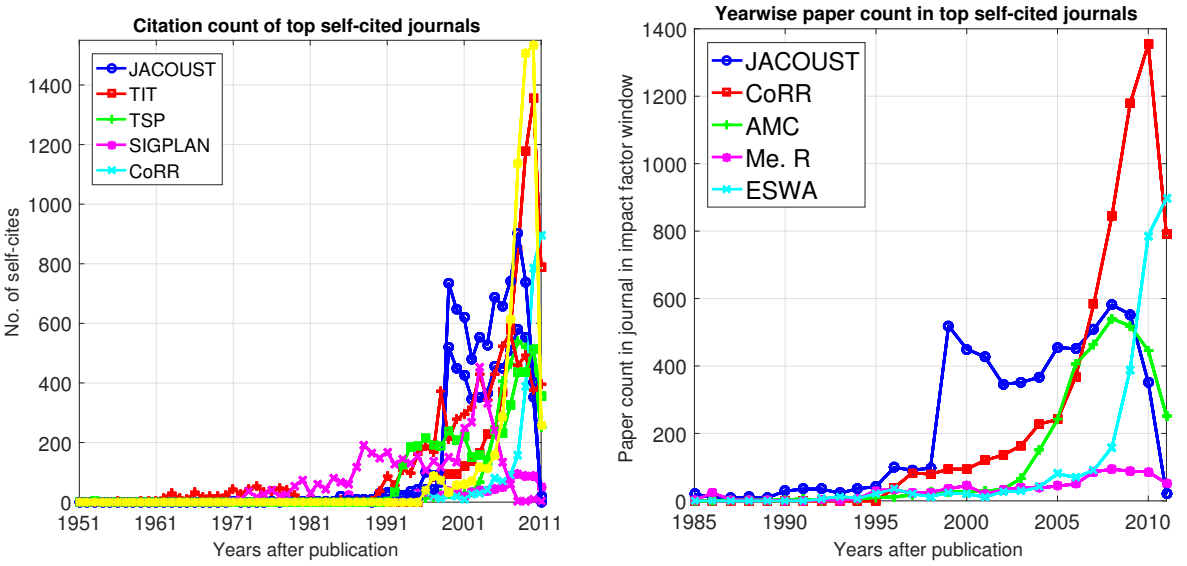


Figure 6.6: **(a, b)** Self-citation activity for JACOUST (Journal of the Acoustical Society of America) is depicted. Throughout lifetime, a total of 80% self-citations are collected. The two curve lines for total citations and self-citations almost overlap each other. A large variation in impact factor is observed especially since year 2000 when self-citations has gradually increased and the impact factor metric became more popular. **(c, d)** Self-citation activity for ENVCONV (IEEE Transactions on Energy Conversion) is depicted. A significant rise in self-citations to boost impact factor is observed since year 2004 with highest variation seen in year 2010, when the impact factor decreases from 1.28 to 0.11. In 2010, the publication count has also doubles along with increasing self-citations.



(a) (b)

Figure 6.7: (a) Citation activity of top 5 highly self-cited journals is depicted. JACOUST and SIGPLAN (Sigplan Notices) have experienced multiple peaks in their citation trajectory. CoRR has experienced a single late peak. Moreover, most journals (TIT (IEEE Transactions on Information Theory) and TSP (IEEE Transactions on Signal Processing)) are showing a monotonic increase in their self-citation rate in their citation profile. **(b)** The x-axis denotes time (in years) whereas y-axis denotes count of recently published articles calculated for the two preceding years, i.e., in the impact factor time window. For better visualization, we plot the graph along x-axis after 1985. Rapid inflation in the frequency of publications is observed for all the journals since 1995. The highest self-cited journal JACOUST has consistently maintained a high frequency of publication experiencing multiple peaks since 1999. CoRR experienced a sudden peak in 2009-2010 for publishing a huge volume of articles for those two years. Both CoRR and ESWA (Expert Systems With Applications) experience late peaks.

6.4.1.3 Pairwise mutual citation (more than 2 nodes):

Mutual citation exchange between a group of three or four journals such that all vertices do not have direct bi-directional weighted citation edge between them but, they are connected in the form of an open chain (see Figure 6.5). Although, we obtain 56 such relations of different lengths 2, 3 and 4 from resultant graph; the majority of connecting edges in this pattern belong to either medium or low pairwise mutual citation bucket with two nodes. For example, we find four pair of mutually cited journals in form of open-chain `IEEE Communications Letters (ComLet)` → `IEEE Transactions on Communications Home (TCOM)`

→IEEE Journal on Selected Areas in Communications (JSAC) → IEEE Communications Magazine(ComMag) where ComLet and ComMag are sister journals published on a timely manner to review research on all aspects of communications.

We try to find bi-directional pairs where any two consecutive nodes in a are involved in the mutual citation. We combine bi-directional weights into a single edge weight as per the previous algorithm. Here, length of the chain (l) = number of linked edges (e) and number of nodes in a chain (n) = $e+1$ (Figure [6.8]). We get a total of 56 chain relations with 22.19% vertices and 25.59% edges involved where $2 <= l <= 4$. We get 69%, 23% and 7% chain relations with length 2, 3 and 4 respectively.

Next, we find coupling strength of a chain depending on weights of pivot edges. For a chain of l=2, n=3; where, intermediate node is the pivot node, coupling strength of the chain (w) is given by if range (x,y) is significantly large such that (σ(x,y) > mean(x,y))

$$w = max(x, y)$$

else if σ(x,y) is small,

$$w = mean(x, y)$$

. For a chain of l = 3, n = 4; given, y is weight of the pivot edge and x and z are the weights of adjoining edges of pivot edge

$$w = mean(mean(x, z), y)$$

Similarly, for a chain of l= 4, n = 5; given, y, z as weight of pivot edges and x, w as weights of adjoining edges, coupling strength of the chain is given by

$$w = mean(mean(x, w), y, z)$$

.

6.4.1.4 Group mutual citation (3 nodes):

Mutual citations are occurring between three journals in a closed loop such that if journal $J1$ mutually cites $J2$ and $J3$, then there has to be a bi-directional citation edge between $J2$ and $J3$ to form a Group mutual citation with 3 nodes (see Figure 6.5). For example, three journals IEEE Transactions on Pattern Analysis and Machine Intelligence

(PAMI), Pattern Recognition (PR) and Pattern Recognition Letters (PRL) are seen to form a triangle.

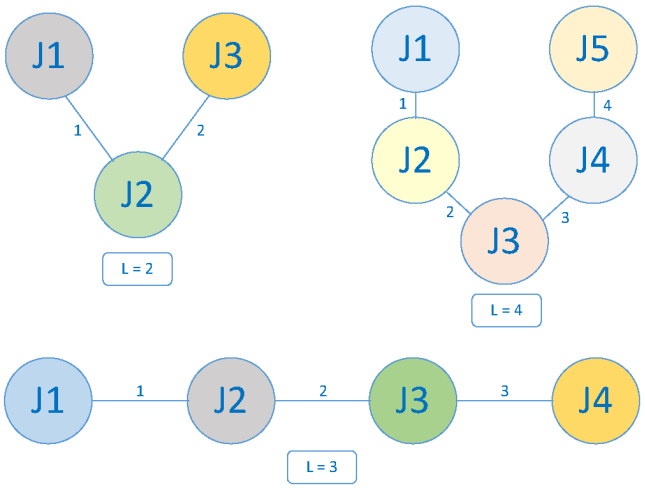


Figure 6.8: Chain relations with l = 2, 3 and 4 are depicted. Here, J1, J2, J3, J4 and J5 depict journals and 1, 2, 3, 4 depict dissolved bi-directional weight into a single undirected coupling weights.

6.4.1.5 Group mutual citation (4 nodes):

Mutual citations are occurring between a group of four or five journals. We can form a group of 4 mutually cited journals out of two groups of 3 mutually cited journals joined by a common edge (see Figure 6.5). If journals $J1$, $J2$, $J3$ form a *citation triangle* and journals $J4$, $J2$, $J3$ form another *citation triangle* and a bi-directional weighted citation edge exists between $J1$ and $J4$, then it forms a group of 4 mutually cited journals pattern. From mutual citations between two journals to group mutual citation between 4 journals, we find few journals which are co-incident and eventually their association increases. Such patterns reveal interesting observations on how journals mutually associate with each other in a citation relationship. For example, IEEE Transactions on Communications Home (TCOM), IEEE Transactions on Information Theory (TIT), IEEE Transactions on Wireless Communications (TWC) and IEEE Transactions on Signal Processing (TSP) belong to same publication house (IEEE). They form a citation mesh pattern. Such patterns are also inter-convertible into other mutual citation patterns.

6.4.1.6 Uni-directed citation:

When a journal receives large citations from a group of journals without citing them back at all then such pattern refers to uni-directional citations (see Figure 6.5). It is more likely to be seen in new and less visible journals which contribute excessive one-way citation to an older well-established journal in the field.

For example, `Computer Communications (COMCOM)` and `Computer Networks (CompNetw)` from same publisher (Elsevier) give excessive one-way citation to `Computer Communication Review (CCR)` published by `ACM` all belonging to same domain where, `CompNetw` is a new journal.

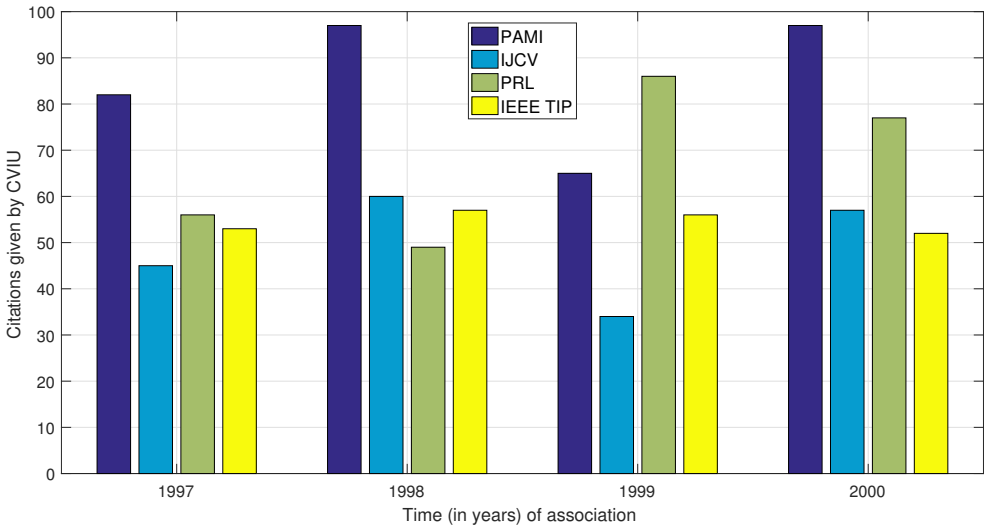


Figure 6.9: Journal Computer Vision and Image Understanding (CVIU) are seen to give citations during four year time period since 1997-2000 towards IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), International Journal of Computer Vision (IJCV), Pattern Recognition Letters (PRL) and IEEE Transactions on Image Processing (TIP) unidirectionally. y-axis denotes incoming citation count from CVIU towards these journals calculated in the impact factor window.

After deriving such patterns, we try to find whether such citation association leads to any impact factor change on a temporal scale. Impact factor (IF) for a journal in a year y is defined as the ratio of recently published articles in that journal during preceding two years $(y-1)$ and $(y-2)$ divided by total citation collected by those articles during the same time window. For a given journal, we calculate time basis impact factor and revised impact

factor (RIF) after removing influential citations contributed by either self-cites or highest citing journals to recently published articles. Next, we dig out possible features that lead to such grouping on a macroscopic scale such as the domain of journal, publication year, type of article published in a journal (review article or research article), publisher of a journal, etc. For such group of journals, we also calculate time basis author citations that are, citations collected by an author for publications published in that journal in the preceding two years.

Overall, the pattern study reflects that only patterns are not enough to detect the intent of a citation. We find that such patterns are uniformly distributed among good and reputed journals as between PR, PRL, and PAMI. Further, we study the citation activity of a journal on a temporal scale considering the impact factor time window. Our main aim is to see due to such association between journals how its impact factor changes over the entire publication age.

6.5 Detailed analysis

One of the prime intentions behind a journal getting its hands dirty could be to gain a high impact factor. We further try to understand growth in journal impact factor on a temporal scale so as to break down large mutual citations between a group of journals into a sliding window of 2 years.

6.5.1 Time series analysis of journal impact factor

Calculating cross citations over entire publication age, we get many large mutually cited group of journals; however in-depth study of time basis impact factor variation shows that their IF growth is consistent. Impact factor widely came in use since 2000. The general trend of IF curve is that it shows a monotonic increase. It is also evident from results that impact factor of a journal largely inflates in the first five years after publication due to self-citations and then, mostly for well-established journals it shows a monotonic increase. For increasing visibility, newly published journals are more prone to impact factor inclined self-citation. For instance, ENCONV after 2003, gives excessive self-cites as is evident from RIF curve which almost decreases to 0 (see Figure 6.10).

Contrastingly from patterns extracted above, we also get a group of journals whose temporal IF curve shows sudden spikes. The peaks are not seasonal. We further plot RIF curve for such mutually cited pair of journals. We find in few cases that removing citation contribution of donor journal, the RIF curve shows a sharp decrease. In order to detect outlier points and understand how much deviation occurs in such sudden peaks in IF temporal curve, we

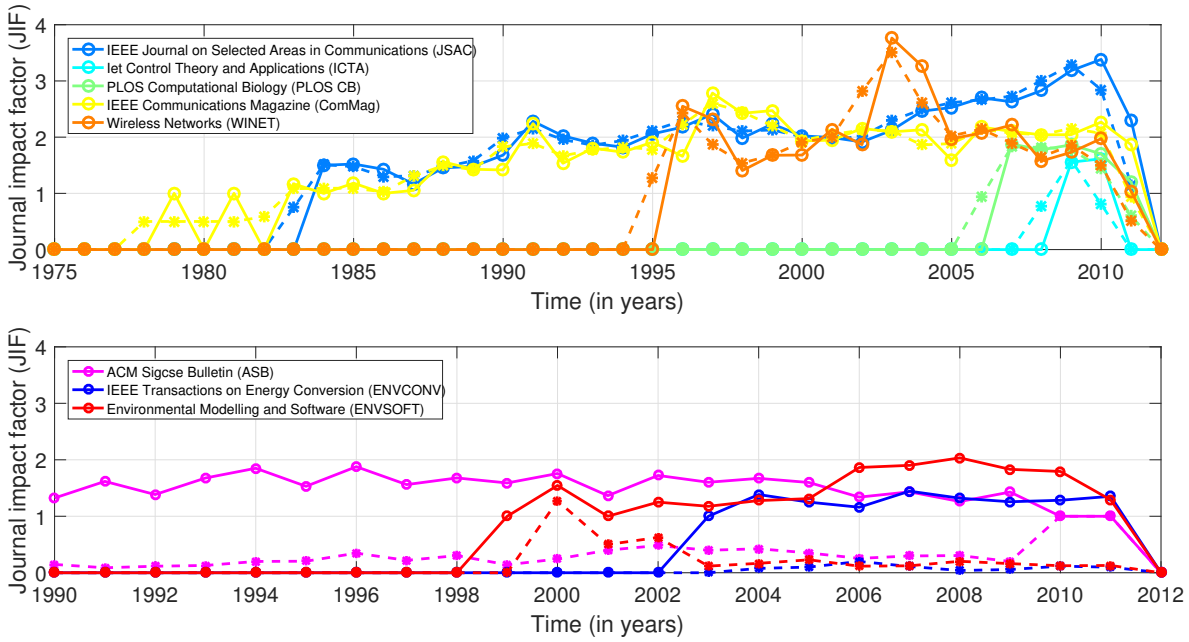


Figure 6.10: **(a)** Computed journal impact factor using simple moving average method is plotted in dashed line. Solid line represents actual impact factor curve. The curve is monotonic increase in nature for example, `ComMag`. Sudden peaks are visible during specific years in actual IF curve of journals `JSAC`, `WINET` and `lCTA`. **(b)** Variation in impact factor and revised impact factor on a time basis of three excessively self-cited journals – `ENVSOFT`, `ENVCONV` and `ASB`. Solid line represents impact factor and dashed line represents revised impact factor. Here, revised impact factor is calculated after removing self-citation. We see that revised impact factor curve almost decreases to 0 for all three journals. While `ENVSOFT` is a new journal which gets excessive self-cites after 5 years of publication. `ENCONV` after 2003, gets sudden peak in impact factor whereas revised impact factor curve for it shows a sharp decrease.

use a simple moving average method. It computes journal impact factor in a year $t+1$ taking an average of past three year observations. We find optimal window size = 3 that smoothens IF curve. All past observations are given an equal weight of $\frac{1}{3}$.

$$IF_{(t+1)} = \frac{1}{3} * \sum_{k=t+1-3}^{t} IF_k \tag{6.1}$$

Next, we plot calculated mean impact factor behavior sliding window of size 3 in dashed line, and against it, we also plot the actual impact factor curve of that journal. Comparing, we see large deviations at some specific years (see Figure 6.10). '`IEEE Journal on Selected`

Areas in Communications (JSAC)' is one of the journals among them. Along with 'JSAC' we get three journals in form of citation chain pattern (see section 6.4.1), ComLet → TCOM →JSAC → ComMag. All four journals are published by IEEE and belong to the same field of research "Communications". ComLet and ComMag are sister journals featured on a timely manner to publish up to date review research on all aspects of communications including technological and development advances, market trends, upgradation in services and systems, change in regulatory policies and issues whereas TCOM and JSAC mainly focus on telecommunications. Such an inherent grouping from same publisher and domain makes sudden spikes in IF curve of JSAC visible. It is a characteristic of review journals to publish up to date extended the content of parent journals which automatically adds references for them.

We also present here two case studies on a mutually cited pair of journals. A pair of journals belonging in same domain – computational biology Bioinformatics and BMC Bioinformatics. BMC Bioinformatics is a new journal established in the year 2000. After the initial four years of publication, it suddenly starts giving a huge volume of citations to an older journal in the field Bioinformatics. Contrastingly, both journals are increasing their publication count two-fold each year in a period between 2005 to 2011. RIF calculated for Bioinformatics journal shows sharp decrease whereas for newer journal BMC Bioinformatics the curve monotonically decreases and it is less affected (see Figure 6.11). A similar case is observed between another pair of journals with the same publication house ACM – TOPLAS and SIGPLAN Notices which mutually cite each other throughout the entire publication period. Both journals belong to the same topic of interest "Advances in programming languages and systems". While TOPLAS is a premier journal, SIGPLAN Notices is an informal monthly publication journal giving review on several conference proceedings and SIGPLAN activities. RIF calculated for TOPLAS shows a sharp declining curve (see Figure 6.11). Thus, IF inflation is mostly domain and time specific. We find many journals in either pair or group showing a sudden peak in impact factor and then a constant rise. Removing influential contributions from donor journal, the revised impact factor calculated for recipient journal shows a sharp downfall.

6.5.2 Micro-level feature analysis

A high impact factor journal benefits an author also because it attracts citations for them anyway. Several instances in data set show the author adding self-referential and duplicate manuscripts and consequently, the journal is getting self-cited. But for journals belonging

to the narrow domain, these self-cited journals are the only ones which lead the field itself. One notable observation is that when a mutually cited group of journals experience sudden peaks in IF curve, there is always an abrupt increase in publication rate of donor journal thereby, attracting additional references. This trend is evident in Figure 6.11. Bar graphs in blue depict this trend for all four groups of mutually cited pair of journals.

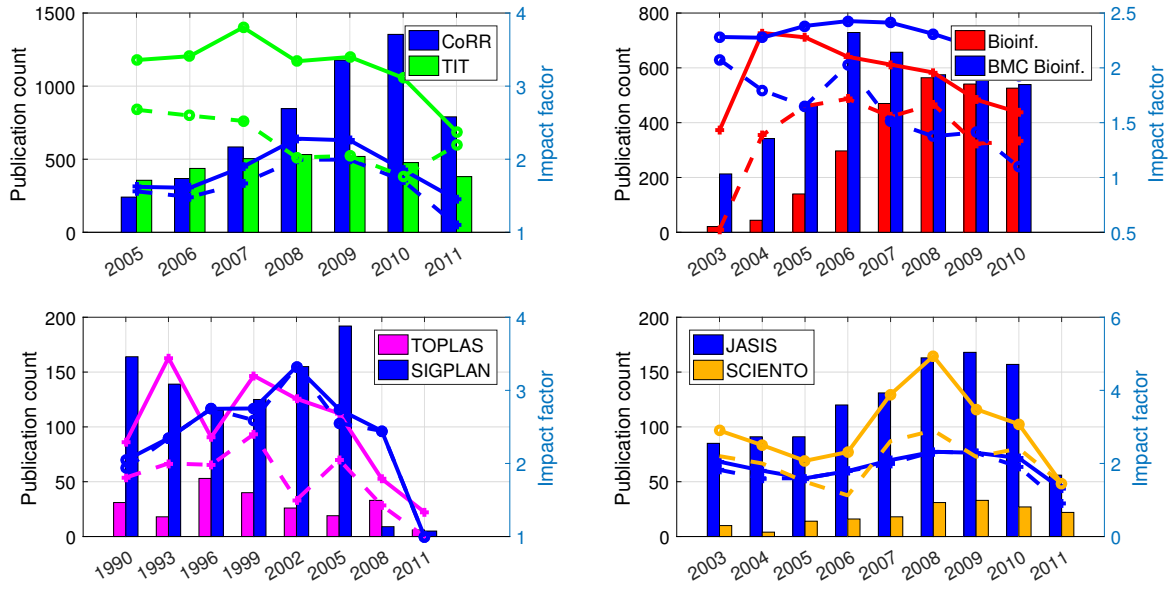


Figure 6.11: Time basis study of impact factor for four pair of mutually citing journals – `Computing Research Repository (CoRR)` and `IEEE Transactions on Information Theory (TIT)`, `Bioinformatics` and `BMC Bioinformatics` belong to *high weighted mutual citation bucket*, `ACM Transactions on Programming Languages and Systems (TOPLAS)` and `ACM Special Interest Group on Programming Languages (SIGPLAN)` belong to *medium weighted mutual citation bucket* and `Scientometrics (SCIENTO)` and `Journal of The American Society for Information Science and Technology (JASIS)` belong to *low weighted mutual citation bucket*. y-axis on left represents publication count and y-axis on right represents impact factor. Bar graph refers to the publication count of journals whereas, solid line graph and dashed line graph refer to variation in IF and RIF on time basis respectively. Since mutual citation patterns are time specific we only plot those data points where IF has largely been inflated due to mutual citations.

6.5.2.1 Narrow domain specialization of journals

Along with newly published journals, we find some old journals also which receive *excessive self cites* throughout the publication period. JASA is among few journals receiving 85.71% self cites and is also the highest self-cited journal 43,383 in our data set. It is so because it publishes articles in a specialized field of “Acoustics". Highly self-citing journals from small research fields do not do it for the purpose. It may simply happen that similar type of research is not published in other journals and one self-citing journal (or a small group of them) are the only ones that lead the field itself. From *mutual citation pattern* we get two journals Scientometrics (SCIENTO) and Journal of The American Society for Information Science and Technology (JASIS). Between 2006 to 2009, JASIS and SCIENTO exchange large mutual citations as is evident from the IF curve which shows a sudden peak during this specific time. Again, due to the presence of very few journals in the field of Scientometrics, it is likely that both journals will mutually cite each other.

Thus, almost every paper published in JASIS or SCIENTO cites an excessive number of papers from the other journals because simply most of the available information in this narrow research field was published in these journals. With the overlap of topics and a limited number of well-established journals in respective narrow field standing, such patterns are likely to be found.

6.5.2.2 Influence of publication houses

To some extent publisher's network influences the impact factor of journals. Six major publishing houses including *Elsevier, IEEE, ACM, Springer, Oxford Press, Wiley and others* involve in self-citation (see Figure 6.12) and mutual citation patterns of which *IEEE* and *Elsevier* make highest contributions. We divide mutual citation cases into three citation buckets with equal number of 41 cases each – *high weighted mutual citation bucket* ($w > 1200$), *medium weighted mutual citation bucket* ($w > 450$) and *low weighted mutual citation bucket* ($w < 450$) (w is the coupling weight). We find that 38%, 35%, and 21% cases are tagged with the same publication house in high, medium and low weighted mutual citation patterns respectively. Extending it to chain and triangle relations, we find 17% and 8% relations with same publication house. In mutual citation cases also, *IEEE* and *Elsevier* are sole contributors.

For instance, all four journals International Journal of Computer Vision (IJCV), IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Com-

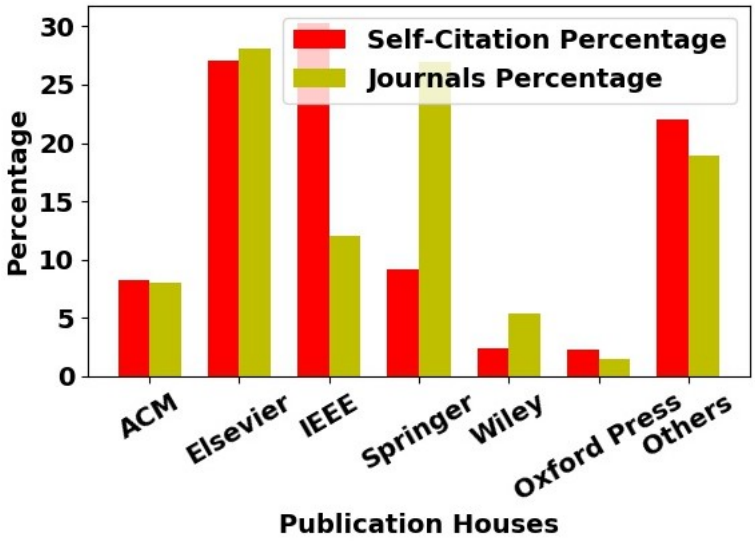
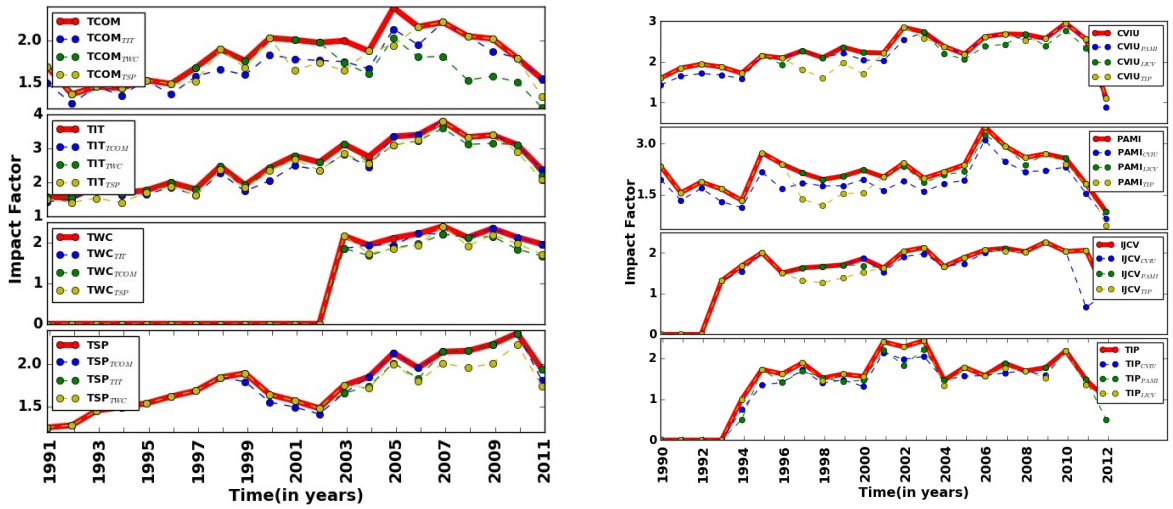


Figure 6.12: Six major publication houses such as *ACM, Elsevier, IEEE, Springer,Wiley, Oxford Press* and *Others* are represented along x-axis. Green bar graph depict percentage of journals published whereas red bar graph depict percentage of self-citation. Journals published by IEEE and Elsevier are highest self-cited.

puter Vision and Image Understanding (CVIU) and IEEE Transactions on Image Processing (TIP) belong to same domain and are published by different publication houses. TIP published by IEEE and CVIU published by Elsevier reciprocates large mutual citations between 1997 to 2000. Notable observations are *RIF* calculated for IJCV journal removing citations from CVIU in year 2011 shows a sharp decrease,(see Figure 6.13). Contrastingly, PAMI is one-way cited by CVIU for a long period between 1996 to 2010. Interestingly publication rate of CVIU and TIP journals abruptly increases during this specific duration. CVIU publishes maximum number of articles only during these 4 years. Another closed group of four journals TCOM, TIT, TWC and TSP (see section 6.4.1.5 for full journal titles) belonging to same publication house IEEE and same domain “Communications" mutually cite each other. Out of these four journals, TCOM's impact factor is inflated to 2.39 by other three journals where, citation contribution of donor journals 32.7% references from TWC and 20.8% from TSP are maximum. *RIF* calculated for TCOM after removing citations from TWC between 2007 to 2010 depicts a sharp monotonically decreasing curve (see Figure 6.13).

More such cases are observed in our data set. For instance, two sister journals from the same publisher Elsevier ‘PR and PRL’ moderately mutual cite each other. PRL which aims at



(a) (b)

Figure 6.13: Study of temporal impact factor of citation mesh pattern. Red line depicts variation in impact factor on time basis, whereas three dotted line refers to revised impact factor after removing citation from other three journals. Figure **(a)** refers to four journals, namely TCOM, TIT, TWC and TSP belonging to same publication house IEEE. All four journals are mutually citing each other in citation mesh pattern. Figure **(b)** refers to four journals, namely IJCV, PAMI, CVIU and TIP belonging to different publication house – IEEE and Elsevier.

the fast publication of concise review articles on 'Pattern Recognition' has fairly cited its sister journal. For many other cases of mutual citations in medium and low weighted buckets, although there is fair citation exchange over the entire publication period its influence on temporal impact factor study completely vanishes with negligible mutual cites. Combined with domain-specific nature, publisher's network may inherently give rise to such patterns and sudden peaks in IF curve. Publishers come up with occasional review journals. Also, sister journals give large citations to its parent journal which largely inflates the impact factor.

6.5.2.3 Author self-citation and author editorial relations

A high impact factor journal benefits an author also because it attracts citations for them anyway. Several instances in the literature show author adding self-referential manuscripts and consequently, a journal is excessively self-cited. Author self-citation is a major reason behind excessive self-cites for journals. For analyzing time specific visibility of sudden peaks in IF curve, we conduct author citation study on time basis to see which authors are responsible for large mutual cites in impact factor time window. We obtain a list of overlapping

authors among the highest contributors. The same author publishes in both journals and either cites its own papers or gives mutual citations to co-authorś paper. Hence, author self-citation and author co-author network play a vital role in such sudden inflation in IF.

TIT is an old journal published in 1953. A sudden citation exchange occurred between TIT and CoRR for a specific time period between 2005 to 2011 6.11. Although there are high cross citations, citations from CoRR to TIT is dominant. Donor journal is CoRR and recipient journal is TIT. Another notable indication is that publication count of CoRR journal rapidly increases from 2007 to 2011 with a maximum of 55.31% references to TIT in the year 2009. In return, CoRR receives 26.94% citations from TIT. When *RIF* for TIT is calculated by removing citations given by CoRR in impact factor window, it shows a sharp declining curve (see Figure 6.11). Also, CoRR is a newer journal than TIT. When we study author citations on time basis for these two journals in the duration of 6 years, we find an overlapping set of authors out of which, for recipient journal IF is largely increased by H. Vincent Poor the then editor in chief for TIT in a period between 2003 to 2007. Syed A. Jafar is found to be an associate editor for TIT from 2009 to 2011, and David Tse is a permanent member in fellowship for TIT journal [1]. While publishing in CoRR journal and simultaneously, positioned as an editorial board member in TIT journal authors either cite their own articles or papers of a co-author. Such underlying author co-author and author-editor relations influence largely IF inflations such as TIT in this case.

6.5.2.4 Well done marketing by newly published journals

Well-done marketing may lead to a distribution of information about papers published in the new journal to putative authors of an old journal such as marketing of PeerJ, PLoS ONE, some Chinese journals who send their TOCs to people who never signed for them. Elsevier, in general, has started now to send a handful of suggestions of papers based on the activity of respective researcher during the past week, etc. All these activities shape knowledge of the author, and the author cites not only the papers, which he/she already knows plus those, which she/he found when purposefully searching for some narrow topic.

6.6 Conclusion

Citation is a spontaneous process and it is tricky to understand possible intention behind a citation. From citation patterns extracted in form of common motif, we conclude

[1] http://www.itsoc.org/people/committees/publications/2005

that only pattern is not enough to detect anomalous nature of citation. However, it could reduce the sample size. These patterns are uniformly distributed among good and reputed journals. Therefore without understanding the dynamics of underlying features, it is not possible to identify them and model their impact. In general we study temporal changes in impact factor, impact of publishing houses etc. Other important features yet remain to be studied such as, author co-author network, editor-author network, author-author collaboration network etc. In future study, we aim to model these relations in a multi-layer approach. Machine learning algorithms could also help to detect outlier and their characteristic in such patterns.

General trend of IF curve is that it increases monotonically. Over entire publication age, we find a large mutually cited group of journals which shows consistent IF growth temporally. However, for some pair of journals, we find sudden spikes visible for specific duration and then a constant rise. An abrupt increase in the paper count of donor journal is a characteristic of this phase. Such inflation are time specific. From large scale data set analysis, we find narrow domain specialization, the influence of publication houses, author self-citation, author co-author relations, author editorial board relations that also add up to form such pattern. IEEE and Elsevier publishers contribute highest toward such a pattern. Newly published or less visible journals are more prone towards *self-citation* and *uni-directed citation*. Often, we find that review or sister journal from either the same or different publisher mutually inflate impact factor of its parent (older) journal in the same domain. Grouping between more than two journals in citation chain pattern is characteristic of such behavior. Overall we perform extensive exploratory feature analysis that affects mutual citation associations among journals. It could further help to classify a given citation pattern into an anomalous and non-anomalous category.

6.7 Summary

∗ Key motivation of this work was triggered after doing exhaustive literature survey [8],[10], [11], [12], [13], [16], [20], [81], [82], [83], [129], [130], [133], [135], [136] and some reports on blacklisting of journals [17, 18] by Thomson Reuters indexing firm. These works reflect on recent awareness created in the research community by citing evidences of unethical citation practices adopted by journals to mutually boost impact factor. From previous work, we find that the impact of publication venue influences an authorś citation. Therefore, it is crucial that journal assessment metric is also free from any error and manipulation. Some works

[21], [127],[128] also point towards inconsistency in Impact Factor metric for journals. The Impact Factor is calculated based on recently published articles and their citation in previous two-year window.

∗ For macroscopic study, we couple paper-paper citation network used in our previous work to a journal-journal citation network and hence, partition the network into two sub-networks- incoming citation network and outgoing reference network. Choosing proper thresholds, we filter out insignificant nodes and edges, and hence, incoming citation graph has resultant 428 nodes, and outgoing reference graph has 559 nodes.

∗ We derive common motifs such as self-loop, pairwise mutual citation (only 2 nodes), pairwise mutual citation (more than 2 nodes), group mutual citation (3 nodes), group mutual citation (4 nodes) and uni-directed citations. Pattern study is not enough to detect unusual citation nature in journals. Some good and reputed journals also exhibit similar patterns.

∗ We study change in impact factor for all such patterns and observe sudden spike in impact factor for some pair of journals. We observe that impact factor inflation is time specific. Although, large impact factor variation does not point towards an unusual case. From microscopic feature analysis we find that narrow domain specialization of journals sometimes lead to such patterns. Also, donor journal's recent publications during preceding two year window abruptly increases and consequently, impact factor largely inflates for recipient journal.

∗ Publication house also influences such patterns. Sister journals belonging to the same publishers and research domain and publishing updated content of its parent journal mutually cite each other; thus, a strong citation bond is seen to grow. Also, other microscopic entities such as authors publishing their work in both the journals inherently cite their own work. Author self-citation also inflates the journal impact factor. Further, we study few instances of author co-author network and author editorial network that also aids citation for a journal on macroscopic scale.

Chapter 7

Automated *CoI* management for reviewer assignment in conferences

7.1 Introduction

Peer review is the most commonly practiced approach for evaluating papers and research proposals in academic community [137]. Assignment of papers to expert reviewers is widely known as '*Reviewer Assignment Problem* (RAP)'. It is a critical challenge for journal editors, conference organizers and funding agencies as resources in the form of expert reviewers and research funds are limited, and demand is that research quality should be maximized. In this chapter, we mostly consider *Reviewer Assignment Problem* specifically in conferences where there is a need to review a large number of submitted papers within limited time frame by a handful of available reviewers. The common practice these days is to use some conference management systems (CMS) like 'EasyChair.com' where authors submit their papers along with keywords and list of co-authors. Reviewers as a part of the Technical Program Committee (TPC) member select the list of topics that match their expertise and also declare the conflict of interest. Finally, the Program Chair (PC) takes into consideration self-declaration of conflict of interest and topical matches and manually assigns each paper to a set of reviewers such that a load of all reviewers is more or less balanced. In contrast to EasyChair system which performs a manual assignment, the Toronto Paper Matching System (TPMS) uses a bayesian based scoring model similar to vector space model [138] and uses Latent Dirichlet Allocation (LDA) technique for topic modelling to perform automatic reviewer assignment to papers. TPMS has been integrated with Microsoft Conference Management Toolkit (CMT)

since 2012. Other matching systems include SubSift used in SIGKDD'09 and MLj for machine learning journals since 2010. In addition to existing systems there exist a large volume of research work dealing with automatic RAP. Since 1992, the research community studied the automatic RAP problem and modelled it, considering several factors to obtain an assignment. In the following, we review state of the art in this context.

Related work: Since 1992, the following approaches are being proposed to solve RAP. The first approach is *query-based information retrieval methods.* Here, a paper is used as a query and each reviewer in the database is represented by a text document with information about his field of expertise or publications. For a given paper, the problem is to retrieve the most relevant set of reviewers from the database. A major drawback is that retrieval process which makes an assignment between a reviewer and paper has to be independently done for each paper. The first automated solution to *Reviewer Assignment Problem (RAP)* was given by Dumais *et. al* [84] using information retrieval based methods. Due to common drawbacks such as uneven distribution of papers among reviewers where the workload is not balanced and the order in which papers are assigned, such methods became in-efficient in solving *RAP* [85].

The second approach is that *RAP is defined as a matching problem which is further solved using optimization techniques.* First step is to construct a weighted bipartite graph between paper and reviewer set where, the weight of an edge denotes relevance between a paper and a reviewer [139, 140]. The second step is to derive a matching in the form of final assignment such that a given objective function is satisfied considering several constraints such as each paper is reviewed by a certain number of reviewers and each reviewer's load is balanced. As a result of which recent studies have taken up matching based methods between paper and reviewer set [19, 80, 89, 141]. While calculating the relevance of an assignment, exhaustively studied factors include maximum field matching [79, 92, 93, 142]. Another paper by Devendra Kumar Tayal *et al.* captures the inherent imprecision of this NP-Hard problem by creating a type-2 fuzzy set and assigning relevant weights based on matching the expertise of reviewer [143]. Several other works have also proposed hybrid models using Genetic Algorithm, Ant Colony Optimisation and Tab Search [90, 144–147], optimal solution approach towards implementing convex cost flow problem in *RAP* [141].

Third approach is *feature-based machine learning models for automated paper to reviewer assignment* [94, 148]. In recent years, in order to improve the conventional peer review process several methods such as feature weighting, selection, and construction are

used to fine-tune the score matrix formed between a paper and reviewer set. Along with this score matrix, probabilistic models are used for making the decision on final assignments [138, 149].

The fourth approach is *use of recommender systems for an automated paper to reviewer assignment* [95, 150]. A recent work [149] has transformed the reviewer recommendation problem into a classification problem and used Word Mover's Distance Constructive Covering Algorithm (WMD-CCA). Here complex semantic relationships between submitted papers and reviewers are extracted from keywords using optimized WMD. Further CCA conducts a rigorous learning process for accurate predictions.

It may be noted, that the above reviewed works did not consider the biasness factor automatically derived from databases. In [19], the effect of CoI (conflict-of-interest) in *RAP* has been investigated after the assignment is performed which however did not consider biasness as an input factor. The paper mentioned four types of CoI: *collaborative relationship* where author and reviewer have collaborated in some paper, *colleague relationship* where author and reviewer have worked in the same affiliation in the past, *advisor-advisee relationship* where the author has been the Ph.D. advisor of the reviewer or vice versa. However, such estimating of biasness of different types is non-trivial, given that the standard data set [100] often lacks information like advisor-advisee.

Trend of co-authorship distance We crawl the list of TPC member and accepted paper for few conferences, like *ANTS* (2008-2011), *ASONAM* (2009-2011), *COMSNETS* (2010-2012), *HIPC* (1999-2011), *ICDCN* (2010-2012), *INFOCOM* (2000-2012), *MOBICOM* (1995-2012), *SIGCOMM* (2005-2011), *SIGIR* (2002-2012), *WALCOM* (2007-2012), *COMAD* (2005-2009) and *ICBIM* (2014, 2016) to see how co-authorship distance varies between TPC member and authors of accepted paper. All such information are crawled from their corresponding websites. However, the study is not based upon actual assignment mapping between assigned reviewer and accepted paper. For year 2005, the study reveals that co-authorship distance is 2 for maximum conference. On a time scale between 2002-2012, we see that average co-authorship distance has varied between 3 and 1 for most of the cases. We find that co-authorship relation exist between TPC member and author for maximum cases. Consequently, this relationship can be used to extract CoI without explicitly depending upon self-declaration from reviewer and author.

Motivation and objective: The main aim of this chapter is to explore if the biasness factor can be automatically derived from the recent academic databases (instead of relying on self-

declaration of CoIs) and the RAP be modelled using all three factors: topic similarity, load, and biasness. We further provide a greedy optimization algorithm to solve it.

Contribution: We redefine the objective function for RAP problem considering an additional factor 'biasness' (inversely proportional to collaborative distance) along with field similarity and reviewer load balancing. Further, we mathematically solve it using a greedy approach. The problem reduces to multi-criteria based multiple job, multiple worker assignment problem [151]. As a proof of concept, we validate our algorithm on a real conference data set of *ICBIM 2016* collected from EasyChair. First, we calculate the field similarity matrix using keyword matching technique. For reviewers, we mine data of their past publications from our hybrid data set described in chapter 3. A list of keywords and co-authors is formed. Similarly, for authors of submitted papers, we dig out the list of co-authors from our hybrid data set 3. Keywords are extracted from submitted papers itself. Second, we calculate collaborative distance and assign a value ranging between 0 and 3 from the co-author list of reviewer and authors. Our greedy algorithm calculates a score from two matrices and after each assignment calculates the load for reviewers. We find that assignment quality is better than that was achieved through the manual assignment as followed in practice.

This chapter is organized as follows. At first, we give a brief introduction. In section 7.2.1, we describe data set summary of real conference data *ICBIM 2016* collected from 'EasyChair' to validate our algorithm. In section 7.2.2, we present all mathematical notations and formulate a greedy algorithm to solve *RAP* considering an additional biasness factor. In section 7.5, we test run our algorithm on data set and measure assignment quality score as compared to manual assignments. Also, heuristics for running the algorithm are given.

7.2 Materials and methodology

In this section, we describe in brief the conference data set *ICBIM 2016* taken from EasyChair that is used later to validate the heuristics of our proposed algorithm. We describe how the factors which include *field similarity* and *collaborative distance* are calculated taking a real conference data set. We describe the methodology (extraction and calculation) required in each step before running our proposed algorithm.

7.2.1 Data set

We have collected the conference data set of *ICBIM 2016* [1] from 'EasyChair.com'. We consider three key factors which includes *field similarity percentage, collaborative distance*

[1] http://icbim2016.nitdgp.ac.in/

value whose inverse gives biasness measure and *reviewer load balancing*.

Table 7.1: ICBIM 2016 data set summary

Attribute	Count
Submitted papers	180
Accepted papers	59
TPC members	40
Unique author count in accepted papers	111
Average number of authors per paper	2.25
Average number of co-authors per paper	1.25

Before test running our algorithm on this data set for automated assignment, we have to extract field matching percentage for a set of all papers and corresponding TPC members. We also, need to calculate collaborative distance value. For calculating field matching percentage, keyword matching technique is used. For TPC members or reviewers, we dig out past publications of reviewers from 'hybrid data set' details of which are given in chapter 3. A dictionary containing a list of keywords and co-authors are extracted from their publications. Similarly, for authors of submitted papers, we use the same database for forming the co-author list. Further, keywords are extracted from submitted papers. Hence, keywords of submitted papers and keywords of reviewers are matched. Similarly, the two dictionaries of co-author list of reviewer and author are used to calculate the collaborative distance. In the next section, a detailed methodology is given.

7.2.2 Methodological overview

Review quality depends upon three indispensable factors. Assuming that the reviewer has bid for a paper and is willing to review, it is necessary that reviewer expertise and knowledge should match with the topics explored in the submitted paper. Next, it is important to take into consideration social academic relation between a reviewer candidate and authors of submitted paper. For fair and accurate review, it is expected that the reviewer assignment is done to a person with high co-authorship distance value and minimum CoI. Finally, the reviewer should not be overloaded as this might degrade review quality in some case (Figure 7.2).

Table 7.2: List of notations

Notation	Description
$R = \{r_i \mid 1 \leq i \leq m; where, m \in \mathbb{N}\}$	Set of reviewer
$P = \{p_j \mid 1 \leq j \leq n; where, n \in \mathbb{N}\}$	Set of paper
$T = \{t_k \mid 1 \leq k \leq u; where, u \in \mathbb{N}\}$	Conference topic set
$A(p_j) = \{a_x \mid 1 \leq x \leq p; where, p \in \mathbb{N}\}$	Author set for a paper p_j
$C(p_j) = \{c_y \mid 1 \leq y \leq q; where, q \in \mathbb{N}\}$	Co-author set of authors $A(p_j)$ for each paper p_j
$C'(r_i) = \{c'_z \mid 1 \leq z \leq r; where, r \in \mathbb{N}\}$	Co-author set of reviewers r_i
$W(p_j) = \{t_v \mid 1 \leq v \leq c; where, c \in \mathbb{N}\}$	Word set for each submitted paper p_j
$W(r_i) = \{t_w \mid 1 \leq w \leq d; where, d \in \mathbb{N}\}$	Word set representing each reviewer's archive r_i
$S(r_i, p_j)$	Topical similarity percentage between r_i and p_j
$D(r_i, p_j)$	Co-authorship distance between $A(p_j)$ and r_i
$CoI(r_i, p_j)$	Conflict of Interest (CoI) between $A(p_j)$ and r_i
$L(r_i)$	Work load of a reviewer r_i

7.2.2.1 Topic extraction and similarity measure

There are many topic modeling techniques such as *keyword matching* [152], *Latent Dirichlet Allocation (LDA)* [153], and, *Author Topic Model (ATM)* [79, 89, 91, 93]. In our work, we consider widely used topic modeling technique based upon LDA. Reviewer's expertise score is calculated in two step. First, the initial score is calculated by matching expertise of reviewer's (extracted from his/her past publication) and topic of submitted paper using a normalized word space model. Here, LDA algorithm is used to derive topic model. However, a comparative study shows that word space model performs better matching. Second, self-declared reviewer's expertise is considered as the ground truth value and combining with initial score, supervised prediction algorithm are run to get the final score.

For this experiment, we get reviewer's past publication from ArnetMiner data set [2] and then by using LDA algorithm extract topic set for each reviewer. The initial scores are obtained using the *word space model*, by matching this reviewer's archive with the submitted paper. Next, a second round of verification is done by running *Linear Regression Shared* algorithm to predict the final score by combining initial score and reviewer's self-declared

[2] https://aminer.org/citation

expertise. In the final step, a topic similarity matrix is formed. Working principle of the word space model is described as below:

1. Normalize word count: The submitted paper and archive of reviewer's publication are represented in form of vector. Next, the frequency of each word is normalized as given in Equation 7.1.

$$T(r_i) = \sum_{W \in R_p} \log f(W_{r_i}) \tag{7.1}$$

2. Smoothing: In order to better deal with the rare words, we Dirichlet smooth the reviewer's normalized word count. Here, μ is the smoothing parameter. W_{r_i} is the total number of word in reviewer's archive and, $|N_{w_k}|$ is the number of occurrences of word k in the word set of reviewer's archive. W is the total number of words in reviewer's corpus and $|N|$ is the number of occurrences of word k in the corpus.

$$f(W_{r_i}) = \left(\frac{|W_{r_i}|}{|W_{r_i}| + \mu}\right)\frac{|N_{w_k}|}{|W_{r_i}|} + \left(\frac{\mu}{W_{r_i} + \mu}\right)\frac{|N|}{W} \tag{7.2}$$

3. Normalize score: The individual score obtained for each reviewer and submission set are further, normalized by dividing by the length of paper (p_l).

$$T(r_i) = \frac{f(W_{r_i})}{p_l} \tag{7.3}$$

4. Initial expertise score: The initial expertise score is calculated as the dot product of reviewer's past publication and submitted paper such that,

$$S(r_i, p_j) = f(T_{r_i}) \bullet f(T_{p_j}) \tag{7.4}$$

7.2.2.2 Co-authorship distance measure

Second important factor is to consider the 'Conflict of Interest' (CoI) between reviewer and authors of submitted paper before assignment. We extract co-author list for each reviewer from ArnetMiner data set. Next, for each submitted paper we extract co-authors for distinct set of author in case, the paper is a multi-authored paper from same data set. Here, in this paper we consider the CoI factor from co-authorship graph as other types of conflict of interest mentioned in paper [19] can be easily derived from it. We cross verify it with

self-declared co-author list by reviewer and author of submitted paper. Next, we assign a co-authorship distance value which varies between 0 to 3. If an entity in the author set $A(p_j)$ of submitted paper p_j is directly matching with an entity in reviewer set R then, the co-authorship distance $D(r_i, p_j) = 0$ such that, $A(r_i) \cap p_j \neq \Phi$.

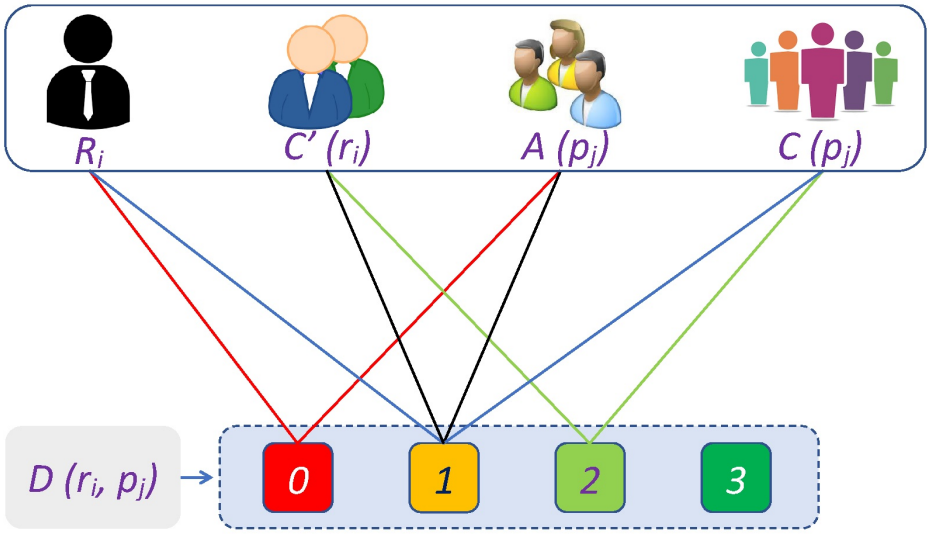


Figure 7.1: Co-authorship distance measure between reviewer (R_j) and author set (A_{p_j}) extracted using co-authorship graph.

If an entity in A_{p_j} belong to C'_{r_i} or conversely, an entity in R reviewer set is present in $C(p_j)$ such that either an author of submitted paper is co-author of reviewer or a reviewer is present in co-author set of submitted paper then, co-authorship distance $D = 1$. In set notation, it can be represented as, if $D(r_i, p_j) \neq 0$ and $C'(r_i) \cap A(p_j) \neq \Phi$ or $C(p_j) \cap (R_i) \neq \Phi$ then $D(p_i, r_j) = 1$. Finally, if co-authors of paper set C_{p_j} belong to the co-author set of reviewer C'_{r_j} then the co-authorship distance is given a value of 2 such that $D = 2$. In set notation, it can be represented as, if $D(p_i, r_j) \neq 1$ and $C(r_j) \cap B(p_i) \neq \Phi$ then $D(p_i, r_j) = 2$. For all other cases, $D = 3$ as no direct collaboration exist between authors and reviewer (refer to Figure 7.1). Mathematically it is represented as, if $D(p_i, r_j) \neq 2$ then $D(p_i, r_j) = 3$. 'Conflict of interest' is inversely proportional to co-authorship distance. We, filter out grouping of paper assignment set with selective available choice of reviewers and sort in a descending order such that paper with greater D value are mapped with reviewers at the beginning. $D(p_i, r_j) = 0$, is not considered for assignment.

7.2.2.3 Calculation of workload *(l)*

Third criteria before final assignment of papers to reviewer is to maintain a load counter for each entity in Technical Program Committee (TPC) or reviewer such that the overall work load of reviewers is equally balanced. This is also to ensure that a single reviewer is not overloaded due to his field of expertise. After every assignment, the load counter is incremented by 1 for a given reviewer, and if the load counter exceeds a certain value, such an assignment cannot be done. The load value is calculated as given in Equation 7.5. After making an assignment and incrementing the load counter, the next paper also goes through similar processing until all the papers submitted are assigned with best suitable reviewer option. For all cases, if none of any better match found then keeping that also for manual assignment.

$$l = \left(\left\lceil \frac{|P_j|}{|R_i|} \right\rceil \times m \right) \tag{7.5}$$

Algorithm 1: Weight matrix before assignment

$temp' = 0$
for $j \leftarrow 1 top_n$ **do**
  for $i \leftarrow 1 tor_m$ **do**
    $temp = s[j,i] * CoI[j,i]$
    $temp' = temp' + temp$
    $w[j,i] = \frac{s[j,i] * CoI[j,i]}{temp'}$

7.3 Problem formulation

In this problem, all graphs which are represented in the form of sets are complete bipartite graphs and directed. Here, graph G is given by $G = (R, P)$. Vertex set V is given by $V = R \cup P$ and edges set E is given by $E = (i, j) | i \in R, j \in P$. In terms of graph theory, the problem is to extract a minimum edge cover set for this complete bipartite graph G which satisfies maximal matching conditions for this problem.

Given a reviewer set as $R = \{R1, R2, R3, ..., R_m\}$ and paper set as $P = \{P1, P2, P3, ..., P_n\}$, after performing maximal matching we obtain a weight function w such that the assignment problem can be formulated as

$$\textbf{max} \sum_{i \in R} \sum_{j \in P} w(i,j) \times a_{ij} \tag{7.6}$$

Where a_{ij}=1, if reviewer R_i is assigned to paper P_j; a_{ij}=0, if reviewer R_i cannot be assigned to paper P_j. Here number of reviewers is denoted by m and number of papers is denoted by n.

Remark (Constraint 1)**.** *Each paper ($j \in P$) can be reviewed by at most* m *reviewers such that,*

$$\sum_{i \in R} a_{ij} = m | j \in P \tag{7.7}$$

Remark (Constraint 2)**.** *Each reviewer ($i \in R$) cannot be assigned more than l papers where $l = \left(\left\lceil \frac{|P_j|}{|R_i|} \right\rceil \times m \right)$ such that,*

$$\sum_{j \in P} a_{ij} <= n | i \in R \tag{7.8}$$

The problem is to find such an assignment that maximal matching occurs.

7.4 Solution approach

By maximal matching, the first condition that we refer is maximum topic similarity should occur between topic set of reviewer R (T_{r_i}) and topic set of paper P (T_{p_j}). The objective function for maximum topic matching can be formulated as

$$S_{max}(r_i, p_j) = \mathbf{max} \sum_{i=1}^{m} \sum_{j=1}^{n} x_{ik} \times x_{jk} | k \in T \tag{7.9}$$

The second condition that needs to be fulfilled is minimum CoI value should occur between reviewer set R and authors (A_{p_j}) of paper set P. CoI value is inversely proportional to co-authorship distance $D(r_i, p_j)$ which is calculated as in algorithm and Figure 7.1.

$$D_{max}(r_i, p_j) = \mathbf{max} \sum_{i=1}^{m} \sum_{j=1}^{n} x_{r_i a_{p_j}} + x_{r_i c_{p_j}} + x_{c'_{r_i} a_{p_j}} + x_{c'_{r_i} c_{p_j}} | i \in R, j \in P \tag{7.10}$$

$$CoI_{min}(r_i, p_j) = \mathbf{min} \left(\frac{1}{D_{max}(r_i, p_j)} \right) | i \in R, j \in P \tag{7.11}$$

Next, we calculate intermediate weight matrix by combining maximum topic similarity value and minimum CoI value following a greedy approach which can be formulated as

$$w'(i, j) = S_{max}(r_i, p_j) \times CoI_{min}(r_i, p_j) \tag{7.12}$$

Edge weight of final assignment which produces maximum matching are:

$$w_{max}(i,j) = \frac{w'(i,j)}{\sum_{i=1}^{m} w'(i,j)} | j \in P \tag{7.13}$$

Algorithm 2: Reviewer to paper assignment

for $j \leftarrow 1 top_n$ **do**
  for $i \leftarrow 1 tor_m$ **do**
    $w' = 0.5$
    $l[r_i] = 0$
    $assignment[p_j] = 0$
    while $w[j,i] \geq w'$ && $l[r_i] < \left(\frac{n}{m} \times q\right)$ **do**
      Assign $p_j \leftarrow r_i$
      $l[r_i] = l[r_i] + 1$
      $assignment[p_j] = assignment[p_j] + 1$
      Check **if** $assignment[p_j] > q$ **then**
        Remove p_j (as paper p_j is assigned to maximum required number of reviewers)
        $j = n - 1$
      Ensure : Until all assignments of $w[j,i] \geq w'$ are already made and still $l[r_i] < \left(\frac{n}{m} \times q\right)$
      $w' = w[j,i] - 0.05$
      Continue executing while loop
      if $l[r_i] > \left(\frac{n}{m} \times q\right)$ **then**
        Remove r_i (maximum load of reviewer exceeded)
        $i = m - 1$
      Break from while loop
    Go back to Algorithm 1 and repeat:

After each assignment of a reviewer to paper, we check the constraints such that each reviewer should not get more than l number of papers. As soon as a reviewer is assigned to l papers, we remove the corresponding reviewer, and as soon as a paper is assigned to at most m reviewers, we remove the corresponding paper from further assignments. Further, if a reviewer is removed from assignment, the complete step beginning from calculating topic

similarity matrix are followed again with remaining papers and list of reviewer.

7.4.1 Assignment quality score

The assignment quality score is a resultant score calculated based on three factors, *topic similarity percentage, co-authorship distance value and reviewer workload* to check how these factors are satisfied for a given assignment (Equation 7.14).

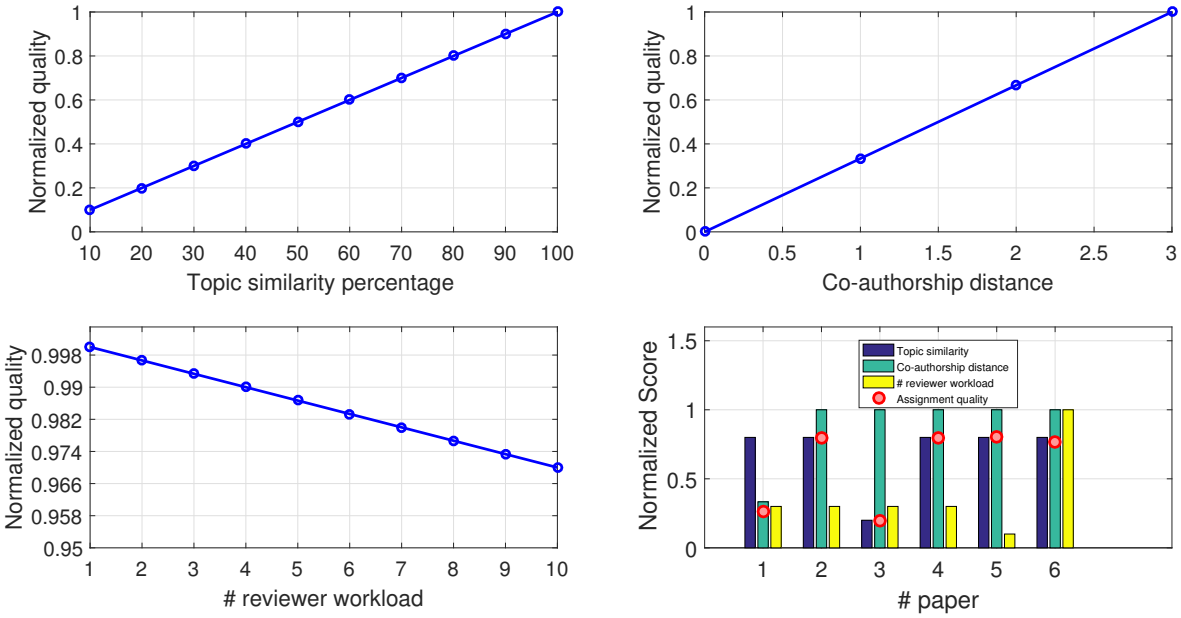


Figure 7.2: Variation in assignment quality is plotted in y-axis while varying topic similarity (in percentage), co-authorship distance value and reviewer work load along x-axis. In the bar graph, normalized scores of each factor is plotted and the calculated assignment quality is given in a red point. In paper 1 and 2, co-authorship distance is varied and other two factors are kept constant. Similarly, in paper 3 and 4 topic similarity percentage is varied and in paper 5, 6 reviewer work load is varied. Comparative study shows change in assignment quality while varying each factor and keeping the other two constant.

$$\frac{w'(i,j) - (max(l) - 1) + (max(l) - achieved(l(r_i))}{max(S(r_i, p_j)) * max(D(r_i, p_j))} \tag{7.14}$$

Here, w'(i, j) is as given in Equation 7.12. $max(S_{r_i,p_j})$ is the maximum topic similarity score that is, 100 and $max(D_{r_i,p_j})$ is the maximum co-authorship distance which a given

assignment could obtain that is, 3. l is the reviewer workload and $achieved(l(r_i))$ is achieved load of a given reviewer after assignment.

In order to check the relationship of three factors with assignment quality, we run three different experiments such that, one of the factor is varied between a range of different values and the other two are kept constant (Figure 7.2). The plots in Figure 7.2 show that both topic similarity and co-authorship distance factor are linearly proportional whereas reviewer workload is inversely proportional to assignment quality. The bar graph in Figure 7.2 depict same results. For papers 1 and 2, co-authorship distance is varied and the assignment quality is increased. For papers 3 and 4, topic similarity is varied and the assignment quality is increased. For papers 5 and 6, reviewer workload is varied and minimal change is seen in assignment quality.

As a sample test case, we demonstrate all the mathematical steps of our proposed algorithm in a toy example given in Appendix C.

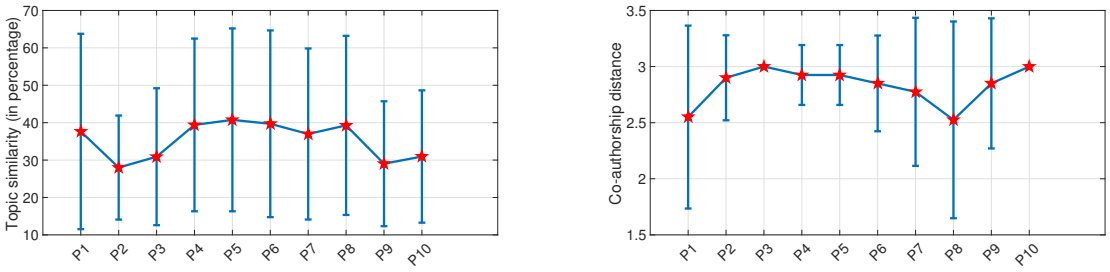
7.5 Experiment

As a proof of concept, we test run our proposed algorithm on a real conference data set *International Conference on Business & Information Management (ICBIM) 2016* [3] collected from 'EasyChair'. For comparison, we also run the automated *Toronto Paper Matching System* assignment strategy on this data set. Note that the collected data has assignment mapping between set of accepted paper and TPC member. Total number of accepted papers is 59 and number of TPC member is 40. Unique number of authors from accepted paper is 111. Average number of authors per paper is 2.25 and average number of co-authors is 1.25.

As mentioned in Section 7.2 for our proposed algorithm, we extract topic similarity and co-authorship distance matrix between set of all accepted paper and TPC member for *ICBIM 2016* data set. In Figure 7.3(a), we plot mean and variance of topic similarity and co-authorship distance in an error bar plot for a set of 10 randomly selected papers with all 40 available TPC members before assignment. Ideally, the mean topic similarity percentage and co-authorship distance should have a high value, and the variance should be less. When the mean topic similarity percentage and co-authorship distance value is small, and the variance is also comparatively small as seen for P2 in figure 7.3(b), such papers should be assigned to reviewers first due to scarce availability of expertise reviewer. The assignment made by 'EasyChair' for this data set considers two constraint. First, each paper needs to be assigned

[3] https://easychair.org/my/conference.cgi?welcome=1;conf=icbim2016

to two reviewers and, second, each reviewer can be assigned at most three papers. Hence, the reviewer workload for assignment is 3. Considering similar constraints, we assign papers using our proposed algorithm and existing benchmark RAP system 'TPMS'.



(a) (b)

Figure 7.3: (a) Mean topic similarity (in percentage) and its variance is plotted in an error bar graph for 10 randomly selected papers with available all 40 reviewers in *ICBIM 2016* data set is plotted. **(b)** Mean co-authorship distance value and its variance is plotted in an error bar graph for 10 randomly selected papers with available all 40 reviewers in *ICBIM 2016* data set is plotted.

7.5.1 Bench-marking methods

We briefly describe the existing benchmark automated RAP system which are being used in many of the conferences. It includes EasyChair and Microsoft Conference Management Toolkit (CMT) [4]. Since, the *ICBIM 2016* conference was conducted using 'EasyChair'; the data set can not be tested to obtain assignment using any other Conference Management System such as 'Microsoft CMT'. We implement TPMS which is the automated assignment strategy of 'Microsoft CMT'. CMT has been used in more than 3,500 conference. In EasyChair, author of submitted paper and reviewers explicitly declare all required data such as field of expertise, several conflicts of interest such as his/her co-author, collaborator from the same research group, peers from same affiliation, etc.

Conference program committee consider all such declarations before manually assigning a paper to reviewer. They can also opt for an automated assignment strategy. TPMS is another automated assignment strategy which consider only two factors topic similarity score and reviewer's workload. The reviewer profile is expressed as a set of topics taking a bayesian approach. Finally, it is used to calculate the score matrix before assignment.

[4] https://cmt3.research.microsoft.com/About

7.6 Result and discussion

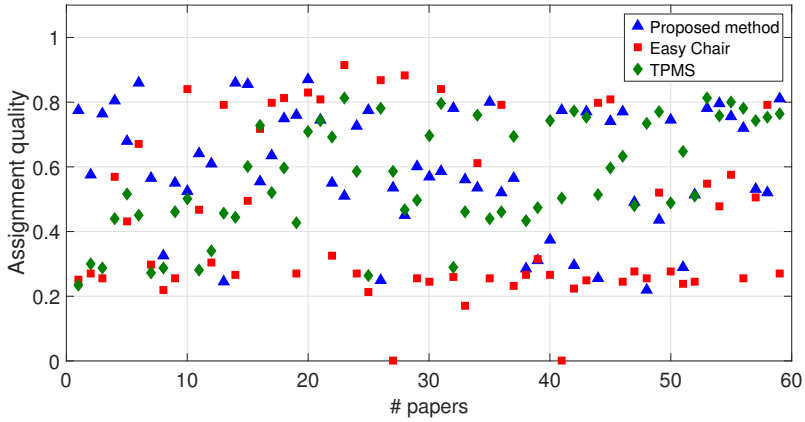
Our algorithm takes a time complexity of $O(m^2 * n^2)$ where m is the number of available reviewers and n is the number of submitted papers. Next, we compare the assignment performance of our algorithm with other bench-marking methods, EasyChair and TPMS. Actual 'EasyChair' assignment can be considered as a random test case assignment and TPMS considers only two factor topic similarity and reviewer's workload for doing the assignment. We aim to observe how inclusion of another factor that is, conflict of interest measured using co-authorship distance improves the assignment quality experimentally. As mentioned in Section 7.1, we collect entire conference details of ICBIM, 2016 along with actual assignment mapping between TPC member set and set of accepted papers. For each paper, we calculate the assignment quality using our proposed method, EasyChair and TPMS using Equation 7.14.

7.6.1 Performance comparison

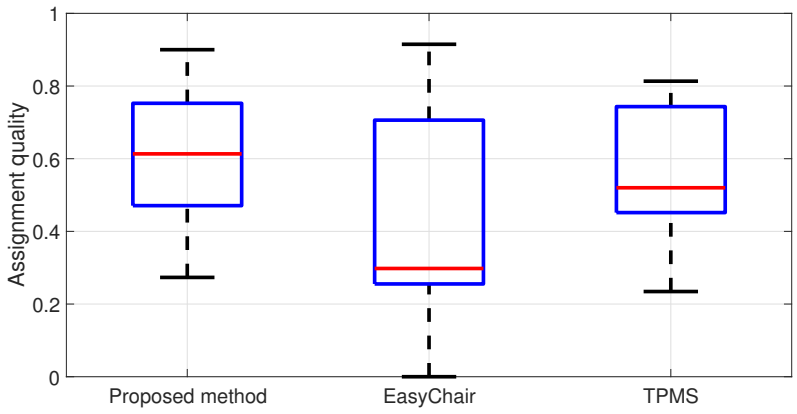
We perform a comparative study of assignment quality (as given in Equation 7.14) for three different assignment techniques, our proposed method, EasyChair and TPMS. Figure 7.4 (a), scatter plot shows that our proposed method and TPMS quality scores are consistently higher than actual EasyChair assignment quality. Moreover, Figure 7.4 (b), box plot shows a large deviation in actual EasyChair assignment quality which implies a significant difference than the other two groups. The median is lowest compared to all three that is, at 0.3 which implies that 50% of papers in lower quartile attain quality lesser than 0.3. The box plot for our proposed method is comparatively shorter than other two and the data distribution is symmetric which signifies that overall all papers get consistently higher score. The quality distribution for TPMS is left skewed. However, in order to understand whether there is statistical difference in quality between the three groups, we perform a paired sample t-test. We plot normalized topic similarity score and co-authorship distance value in a bar graph (Figure 7.5) after assignments by randomly selecting 17 paper from our data set. It is seen that assignments using our proposed method consistently maintain higher topic similarity score and co-authorship distance value. In contrast, TPMS maintains higher topic similarity score but in few papers attain low co-authorship distance value such as in P2 and P8. 'EasyChair' on the other hand, attain comparatively lower topic similarity score. 'EasyChair' assignment are also seen to maintain higher co-authorship distance values for most of the papers.

7.6.2 Hypothesis testing and statistical validation of quality score

We perform a hypothesis test using paired sample t-test and compare mean quality of assignment of our proposed method with that of EasyChair and TPMS. This test is done to statistically prove the importance of all three factors to attain optimal assignment quality.

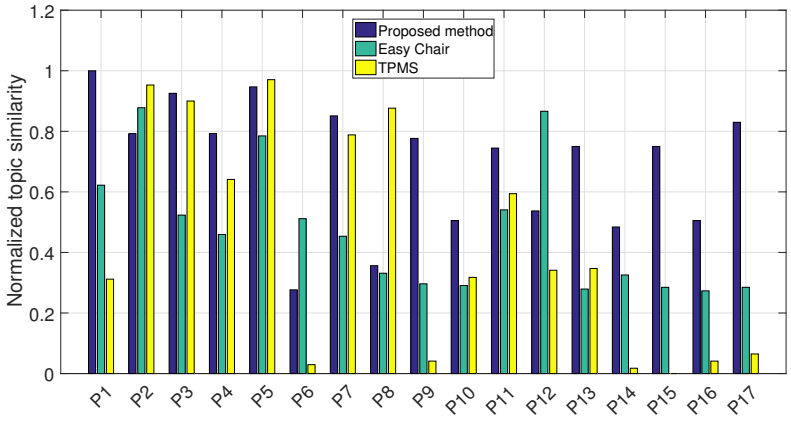


(a)

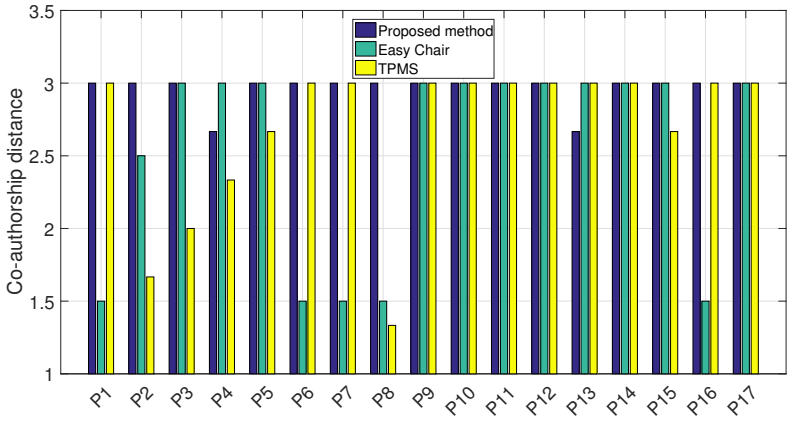


(b)

Figure 7.4: **(a)** Normalized assignment quality of our proposed method, EasyChair and TPMS for 59 accepted set of papers is plotted in scatter plot in blue, red and green respectively. **(b)** The assignment quality distribution for three different techniques considering all 59 accepted set of papers is plotted in a box plot.



(a)



(b)

Figure 7.5: **(a)** Normalized topic similarity score is plotted in a bar graph for 17 randomly selected set of papers for comparison between three different assignment techniques; our proposed method, EasyChair and TPMS after assignment is done. **(b)** Normalized co-authorship distance value is plotted in a bar graph for 17 randomly selected set of papers for comparison between three different assignment techniques after assignment is done.

Table 7.3: Comparative t-test of our proposed method (group 3, mean = 0.6103, variance = 0.0281) with EasyChair (group 1) and TPMS (group 2).

Bench-marking method	Mean quality	Variance	Sample size	t-stat	t-critical value	p-value
EasyChair (Group 1)	0.4435	0.0651	59	3.9900	1.6715	0.0000938
TPMS (Group 2)	0.5650	0.0300	59	1.6957	1.6715	0.0476

We have assignments obtained for 59 paper which are done using *group1 (EasyChair)*, *group2 (TPMS)* and *group3 (our proposed method.)* The same sample of 59 accepted papers is tested with inclusion and exclusion of different factor. For 'EasyChair' assignments, we consider it as a random assignment case in which three factors topic similarity, conflict of interest and reviewer's workload are randomly taken into consideration. Next, TPMS implicitly takes into consideration only two factor; topic similarity and reviewer's workload. Finally, we run our proposed method which takes into consideration all three factors.

7.6.2.1 Comparison of proposed method with EasyChair

Null hypothesis: There is no difference in mean assignment quality between two groups, group1 and group3.

$$\mu 1 - \mu 3 = 0$$

Alternate hypothesis: There is statistical difference in mean assignment quality between two groups, group1 and group3 that is,

$$\mu 1 - \mu 3 > 0$$

We calculate the t-statistic value as given in Equation (7.15).

$$t = \frac{(\bar{X1} - \bar{X3}) - (\mu 1 - \mu 3)}{\sqrt{S_p^2 * \left(\frac{1}{n1} + \frac{1}{n3}\right)}} \tag{7.15}$$

.

Here, $\bar{X1}$ and $\bar{X3}$ represent mean assignment quality of sample group1 and group3 respectively. $\mu 1$ and $\mu 3$ represent mean assignment quality of population group1 and group3 respectively which we consider as 0 in our null hypothesis. n1 and n3 is the sample size which is 59 for both the groups. S_p is pooled variance between two groups.

Since the t-statistic value for group1 assignment is 3.99 which is greater than t-critical value 1.6715 and the p-value at 95% confidence interval is 0.000093 which is less than 0.05. This implies that we can reject the null hypothesis and it is proven that the mean assignment quality of group1 and group3 is statistically different. The assignment performance of our proposed method is of superior quality than 'EasyChair' assignments.

7.6.2.2 Comparison of proposed method with TPMS

Similar comparative study of mean assignment quality is done for 59 accepted papers and 40 reviewers. The assignment quality is derived once using our proposed method and

the again using procedure used in TPMS.

Null hypothesis: There is no difference in mean assignment quality between two groups, group2 and group3.

$$\mu 1 - \mu 2 = 0$$

Alternate hypothesis: There is statistical difference between the mean assignment quality between two groups, group2 and group3 that is,

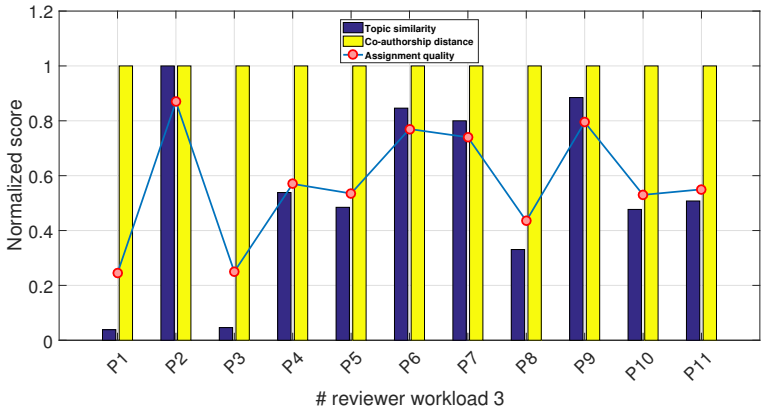
$$\mu 1 - \mu 2 > 0$$

We calculate the t-statistic value as given in equation 7.15. Since the t-statistic value 1.6957 is greater than t-critical value 1.6715 and the p-value at 95% confidence interval is 0.0476 which is less than 0.05. This implies that we can reject the null hypothesis. This implies that the mean assignment quality of group2 and group3 is statistically different with inclusion of a factor 'Conflict of Interest (CoI)' measured using co-authorship distance. The variance of our proposed method with actual *EasyChair* assignment, 0.0651 is more as compared to TPMS. As a result of which our proposed method gives an assignment quality more closer to TPMS.

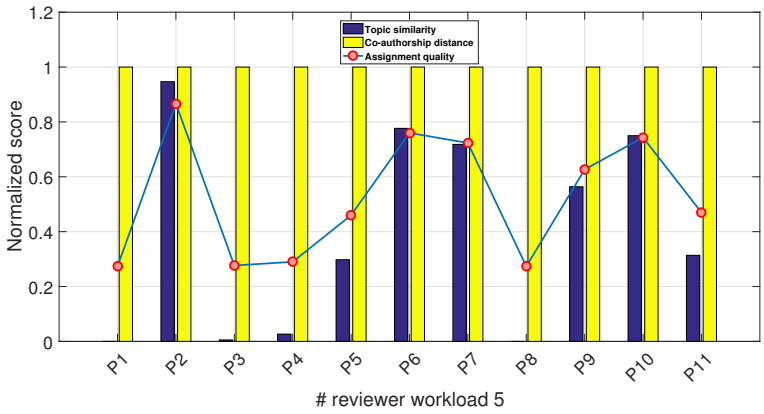
7.6.3 Comparative study on varying reviewer workload

Varying the constraints of our proposed method, we see change in assignment quality for a set of 11 randomly selected paper from our data set. The constraints in two different experiment are defined as each reviewer can be assigned at most 3 and 5 paper respectively. As seen in Figure 7.2, the assignment quality is inversely proportional to reviewer workload.

The same is validated while experimentally testing our proposed algorithm on a real conference data set. As reviewer work load increase from 3 to 5, the assignment quality decrease for few papers such as P4, P5, P8 etc. Next, we plot normalized topic similarity and co-authorship distance score considering reviewer workload 3 and 5 in Figure 7.6 for same set of papers after assignment. It is seen that the assignment quality mainly degrade due to comparatively lesser topic similarity score obtained by assignment considering *reviewer workload 5.*



(a)



(b)

Figure 7.6: Comparative study of assignment quality is done (line plot) when the reviewer workload constraint is 3 and 5 respectively. **(a)** Normalized topic similarity score and co-authorship distance is plotted in a bar graph for 11 randomly selected papers after assignment is done by our proposed algorithm considering constraint, reviewer work load = 3. **(b)** Normalized topic similarity score and co-authorship distance is plotted in a bar graph for same set of 11 randomly selected papers after assignment is done using our proposed algorithm considering constraint, reviewer work load = 5.

7.7 Conclusion

In our work, *CoI* issue is dealt using co-authorship graph instead of relying on self declaration by reviewer and authors. In this way, time required to collect such preliminary information from each reviewer and author can be avoided. Multiple constraint for optimizing multiple variables to obtain multiple matching sub-set between reviewer and paper makes

it a NP-Hard problem. The problem is solved using a weighted average based greedy algorithm by maximizing topic similarity, minimizing CoI and balancing reviewer workload. A new metric is proposed for evaluating performance of each assignment in terms of how the input factor are satisfied. Using this metric, a comparison of proposed method with existing CMS like *EasyChair* and *TPMS* assignment shows that the mean assignment quality by our proposed method is of superior quality. There is significant difference seen in mean quality with inclusion of an additional factor 'CoI'. Further, the proposed method can be incorporated in existing CMS for fair and competent review process. In future work, we aim to use different conference data set to test run our proposed method and measure the accuracy of our model.

7.8 Summary

∗ We re-formulate the reviewer assignment problem as an optimization problem by considering an additional factor that is, biasness measure into it. Unlike existing automated system which rely on self-declaration of conflict of interest, it is directly extracted from co-authorship relations in bibliographic data set.

∗ We use a greedy approach, and the heuristics are as $O(m^2 * n^2)$ where m is the number of available reviewers and n is the number of papers. Optimum assignment score is calculated from the algorithm such that maximum field similarity, minimum biasness and balanced load for all reviewers could be obtained.

∗ As a proof of concept, we test run our algorithm on a real conference data set *ICBIM 2016* collected from 'EasyChair.com' It is observed that the assignment quality score is consistently better than manual assignments as followed in many of the conferences.

∗ In future study, we aim to model the same problem using other optimization and machine learning techniques. Also, we would like to check how the assignment quality is improved in comparison to other automated conference management systems such as Microsoft Conference Tool Kit, etc.

Chapter 8

Conclusion

8.1 Summary of contributions

Throughout our research study, we have analyzed factors that lead to possible biasness in the evaluation of publication process encompassing key entities (paper, author and venue) and thus, re-define existing performance metrics. Besides, most of the widely used author-centric, venue centric or paper-centric metrics consider citation and publication as a fundamental quantifier of assessment which is highly susceptible to manipulation. To understand true credential of an author as well as a paper, we need to set ethical and strict standards of publication for the academic community otherwise; it can jeopardize the entire community.

In our present work, we propose an author ranking metric which is PageRank based multi-feature divided along a multi-layer network model; C^3-*index*. It takes into account three key features in determining performance of an author: the impact of citation received by papers published by an author, the impact of collaboration with his / her co-authors and the impact of other authors who cite his / her paper. Further, with available huge bibliographic data in Microsoft Academic Graph (MAG); through data processing and performing statistical analysis for Computer Science domain, we retrieve a structured outlook in getting a finite pattern of citation distributions that scale to generic models and measuring the impact of publication venue in influencing a paper's citation. Authors take a vital interest in journals/conferences they publish their work. Doing a citation profile study for well-cited papers, several factors such as self-citation, venue, evolving core or inter-disciplinary field of study becomes significant reasons in affecting a paper's citation over time.

The venue plays an elemental role in binding the entire academic community together. In order to identify nature of anomalies occurring in journal-journal citation network, we

filter and represent the network in the form of weighted directed graphs and thus, extract common motifs such as self-loop, bi-directional mutual citations, chains, triangles, mesh, etc. Further, we study the influence of several microscopic entities that lead to sudden superfluous inflation in impact factors for journals such as publication house, editorial board, author co-author relations, etc. It could accentuate in changing journal publication policies, selection of editorial board members, the board of governors, etc. Also, prior to publication, a fair reviewer assignment strategy is proposed to ensure maximum field matching with an expert reviewer, minimum biasness with a reviewer and balanced load on each reviewer as conferences attract huge manuscript submissions to be reviewed in a limited time frame.

8.1.1 Author-centric evaluation of performance metrics

Earlier for ranking an author's performance, there are two strategies existing in literature-indexing and scoring strategy. Author indexes which are widely in use since the 2000-2005 period for measuring performance use number of publications and citations for an author. In such indexing practices, it takes a long time for new authors who have fewer publications and corresponding very fewer citations to get a score of greater than 0 or more. While devising a scoring strategy, several related works had proposed PageRank based approach [44], as it promotes multiple features to be added but multiple features are required to be selectively chosen which directly impacts an author's performance.

In our work considering multiple features, we use a PageRank based methodology and propose a new author ranking quantifier C^3-*index*. The metric is designed in such a way that several inconsistencies of past metric such as h-index and its variant g-index, $\hbar$-index, p-index etc are removed. A significant issue of evaluating low rank authors who mostly attain 0 h-index is also removed when author performance is measured through C^3-*index*. C^3-*index* scores are continuous variables thus, dissolving ambiguity for low and medium ranked authors who mostly get discrete score values from h-index scoring strategy and its like variants.

Consequently, understanding the practical scenario we select the best features that could evaluate an author's performance better than other metrics. Author collaboration and co-citation are best promising features. Three features are carefully selected: first, citations received by papers of a given author, citation received by a given author from his/her peers and citation received by co-authors of the concerned author. Engraved in form of a multi-layer network, ranking from three separate layer is calculated that is, ACI-score, PCI-score and AAI-score etc. The ACI-score component of the multi-layered network alone behaves

similar to h-index and its variants. Our proposed metric can give more better insights than others.

C^3-index is also capable to understand the interconnection between the three layers. C^3-index calculated on a time series scale shows that authors obtaining a high AAI-score component eventually attain high h-index value at some point of time in future. Thus, a possible predictor variable of future performance of author is also considered to calculate his present performance.

8.1.2 Scientific article citation patterns

With huge and complete bibliographic records found in Microsoft Academic Graph (MAG) data set which is almost 100 GB in size, doing a statistical analysis of key entity paper and its references we retrieve finite patterns from citation structure of scientific publications over time. We find that citation distributions on a temporal scale depict lognormal behavior in citation graph and thus, it could be used to calculate the probability of n different papers that are likely to get cited in the near future. Our findings strongly conclude that well-cited publications get attached with a certain likelihood factor that helps them later to collect even more number of citations and thus, increase net citation count (k). We categorize citation profile study into five major peaks- early peak, multiple peaks, late peak, monotonically increasing and monotonically decreasing peak. Overall, citation peaks have mostly grown since the late 1990's due to increased visibility and growth in citation rate. Although more number of conference papers with multiple and monotonically increasing peaks exist in this applied research domain, journals also continue to be a predominant choice of publication exhibiting mostly late peak. Among top 100 well-cited publications, journal papers are more in number.

Characteristic citation profile following growth of major motor topics in Computer Science field over the years had been studied that gives rise to visualizing the pattern of evolution of papers that have led to some groundbreaking researches and discoveries. We classify such unique trajectories under- Sleeping beauty, hot papers and discovery papers. These papers tend to acquire a large number of citations within a very small average citation age $< a >$. Evaluating citation count as a function of age we find citation based comparisons favor papers that had been published in the past and also, established researchers. Also, the unique SB phenomenon had been widely investigated for papers which had remained dormant for a long age and had started receiving sudden citation bumps. Throughout the pa-

per, we tried to establish fundamental metrics to gauge uniform predictability for the varying patterns of citation distributions.

8.1.3 Reasoning different citation patterns among journals

The aim of this chapter is to explore various citation patterns that exist between journals. From our previous work in studying the impact of a publication venue, we find that journals continue to be a predominant choice of publication for authors. Using citation data from Microsoft Academic Search consisting of 2,621 journals, we extract pairwise mutual citation pattern and group mutual citation pattern along with self-loop and uni-directed citation. It is found that mostly such patterns are field-specific and highlight the intrinsic property of a publication house. We individually tag six major publishers: IEEE, Elsevier, Springer, ACM, Wiley, Oxford Press to each of such cases and find that IEEE and Elsevier, contribute highest in any of such patterns. A rapid increase in publication count inflating impact factor of its associate journal could inherently give rise to such patterns. Author self-citation is also a significant reason that uplifts impact factor of a journal on a macroscopic scale. Moreover, such grouping in journal-journal citation graph is influenced by underlying author co-author, editor co-author, author-author collaboration networks. An immediate future study would be to design an automated system that can take citation pattern early of a journal's career and predict if the journal experiences any anomalous citation pattern. One would further be interested in incorporating this anomalous citation pattern into the calculation of the impact factor and refine it.

8.1.4 Automated conflict management in reviewer assignment process

We take into account three factors affecting review process which include maximum field matching of submitted paper to reviewer's topical profile, maximum collaborative distance between author of submitted paper and reviewer and balanced workload for each reviewer. Here, we calculate the biasness measure by finding the collaborative distance from co-author relation. However, other types of conflict of interest might also exist [19] which can be extended as an addition to this work. We re-define the objective function for *RAP* and propose a greedy algorithm to optimize all three parameters. As a proof of concept, we have validated our algorithm on a real conference data set *ICBIM 2016*. It is observed that assignment quality is improved than manual assignment as practised in 'EasyChair' system. We automatically input all the required factors to our algorithm such as field and collaborative distance by extracting from bibliographic data set. No dependency on self-declaration from reviewer

and author set exist. Although reviewer and authors strictly follow code of ethics, automated approach could be beneficial to cross verify and speed up the assignment process. As future work other advanced optimization techniques such as genetic algorithm, machine learning algorithm, and recommender system can be used.

8.2 Future direction

Author is key entity of a research community, and several anomalies can exist in reference to measuring an author's performance such as excessive author's self-citation while calculating collaboration impact in a multi-authored paper, author's name ambiguity leading to irrelevant references, biasness occurring in scientific relations such as authors or reviewers belonging to the same affiliation where senior researchers or PhD. a guide is likely to be cited by junior researchers. Moreover, developing a future predictive model for measuring an author's performance by evaluating the author's work on a temporal scale in current as well as the previous field of studies could be done. Also, author and paper citation profiling [154] in terms of evolving topical influence is directly correlated with the performance of an author to extend further the scope of future author performance prediction through the proposed strategy.

Another future work follows where, tagging large bibliographic data sets such as Microsoft Academic Graph (MAG) based on continental or regional topical research growth, collaboration impact and factors affecting clustering propensity of researchers could be done. A comparative study in determining cross domain citation graphs is another research direction which needs to be explored. In multi-layer network models as used in C^3 index ranking strategy, removing author self-citations from author co-authorship relations in case of multi-authored paper forms another interesting topic of research As observed in Figure 4.5, C^3-index reveals the future outreach of a good fraction of the selected authors much earlier than the actual time they reached that milestone. This may indicate an additional scope of application for the proposed strategy than a mere ranking of authors based on their present performance. We shall also check the results on other data sets from different domains such as Physics, Biology to strengthen our claims.

Early detection of anomalous patterns in the venue by developing a framework on the basis of studying reasons behind the occurrence of anomalies such as editor-author network, publisher-author network, topical dependencies, collaboration or co-authorship relations leading to possible biasness. A multi-layer approach to take into account all such possible

reasons could be formulated. For fair reviewer assignment strategy in conferences, other advanced optimization techniques could be used by automatically taking into account all three factors discussed in our work. It includes maximum topical similarity with a reviewer, minimum biasness with the corresponding author of the assigned paper and balanced load on a particular reviewer.

Appendix A

$C^3 - index$: A toy example

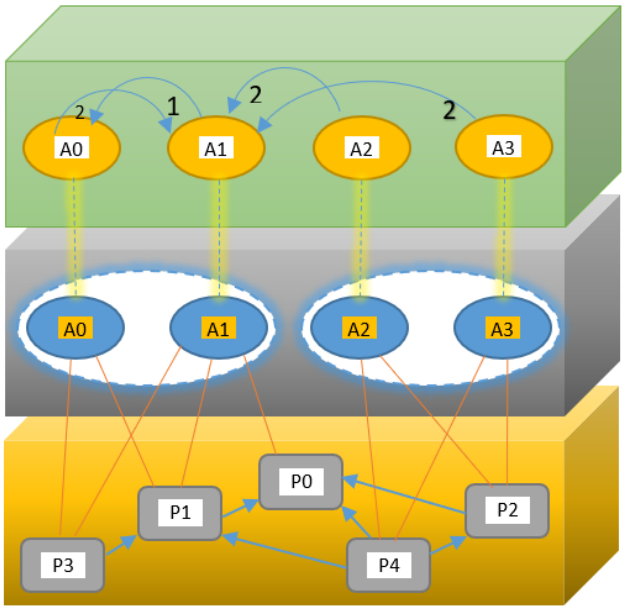


Figure A.1: Citation Network (Toy Example)

In the above figure, we take a small example of citation network where P0, P1, P2, P3 and P4 are papers published by authors {A1,A2}, {A0,A1}, {A2,A3}, {A0,A1} and {A2} respectively. In the bottom paper-paper citation layer, paper P0 gets citation from {P1, P2, P4} and paper p1 gets citation from paper {P3}. Correspondingly, author A0 gets citation from {A1}, author A1 gets citation from {A2, A0}, author A2 gets citation from {A1, A3}, author A3 did not get any

citation from any other authors.

For the above multi-layer model, we calculate each author's C3-index score and make a comparison Table A.1 with h-index as below. We find that author A1 with highest publication count 3 and citation count 4 gets h-index score 1 same as other mediocre authors. Besides, a tie occurs between author A0, A1 and A2. Also, author A3 even after having 1 publication gets no credit for his work due to 0 citation count whereas gets some credit in C3 index score calculated (0.058124). In C3 index author's score, the author's credit is more efficiently measured and it is easily distinguishable to break ties between authors.

AID	$PublicationCount$	$CitationCount$	$H-index$	$C3-Index$
$A0$	2	1	1	0.089603
$A1$	3	4	1	0.150423
$A2$	3	3	1	0.101850
$A3$	1	0	0	0.058124

Table A.1: Comparison scores of authors

Now, In the above network when we run our modified PageRank algorithm and we find author's score for different layers as described in Table A.2. For individual authors, combining scores obtained from the three layers chosen author-collaboration, author-citation and paper-citation, we add the three scores to calculate final score of the author. C3 index is calculated iteration wise till values obtained from three layers are almost normalised and there is minimal change.

AID	$AuthorCollaboration$	$AuthorCitation$	$PaperCitation$	$TotalScore$
$A0$	0.230747	0.098658	0.074576	0.089603
$A1$	0.336012	0.137919	0.145763	0.150423
$A2$	0.290619	0.113423	0.152542	0.101850
$A3$	0.142621	0.050000	0.027119	0.058124

Table A.2: Convergence values (Final scores)

Using modified page rank algorithm, C3 index is calculated and iteration wise values are shown in Table A.3 and Table A.4. It takes 3 iterations for paper-paper citation layer, 10 iterations for author-author citation layer and 11 iterations for author-author collaboration layer to converged.

AID	*AuthorCollaboration*	*AuthorCitation*	*PaperCitation*	*TotalScore*
Iteration:1				
*A*0	0.209722	0.087500	0.077778	0.066667
*A*1	0.348611	0.145833	0.155556	0.150000
*A*2	0.305556	0.116667	0.144444	0.133333
*A*3	0.136111	0.050000	0.022222	0.050000
*Iteration*2				
*A*0	0.233007	0.101042	0.070968	0.092963
*A*1	0.332110	0.135417	0.148387	0.145000
*A*2	0.291424	0.113542	0.154839	0.100926
*A*3	0.143459	0.050000	0.025806	0.061111
Iteration:3				
*A*0	0.229616	0.098047	0.074576	0.088563
*A*1	0.337384	0.138759	0.145763	0.151488
*A*2	0.290298	0.113194	0.152542	0.101665
*A*3	0.142702	0.050000	0.027119	0.058285
Iteration:4				
*A*0	0.231112	0.098839	0.074576	0.089969
*A*1	0.335465	0.137630	0.145763	0.149906
*A*2	0.290835	0.113531	0.152542	0.102065
*A*3	0.142589	0.050000	0.027119	0.058060
Iteration:5				
*A*0	0.230616	0.098599	0.074576	0.089457
*A*1	0.336208	0.138020	0.145763	0.150612
*A*2	0.290534	0.113381	0.152542	0.101764
*A*3	0.142643	0.050000	0.027119	0.058167
Iteration:6				
*A*0	0.230793	0.098678	0.074576	0.089655
*A*1	0.335942	0.137884	0.145763	0.150353
*A*2	0.290652	0.113438	0.152542	0.101885
*A*3	0.142613	0.050000	0.027119	0.058107
Iteration:7				
*A*0	0.230731	0.098651	0.074576	0.089585
*A*1	0.336037	0.137932	0.145763	0.150448
*A*2	0.290607	0.113417	0.152542	0.101837
*A*3	0.142625	0.050000	0.027119	0.058130

Table A.3: 1st to 7th iteration values

AID	*AuthorCollaboration*	*AuthorCitation*	*PaperCitation*	*TotalScore*
Iteration: 8				
*A*0	0.230753	0.098660	0.074576	0.089610
*A*1	0.336003	0.137915	0.145763	0.150414
*A*2	0.290623	0.113425	0.152542	0.101855
*A*3	0.142620	0.050000	0.027119	0.058121
Iteration: 9				
*A*0	0.230745	0.098657	0.074576	0.089601
*A*1	0.336015	0.137921	0.145763	0.150426
*A*2	0.290618	0.113422	0.152542	0.101848
*A*3	0.142622	0.050000	0.027119	0.058125
Iteration: 10				
*A*0	0.230747	0.098658	0.074576	0.089603
*A*1	0.336013	0.137920	0.145763	0.150423
*A*2	0.290619	0.113423	0.152542	0.101850
*A*3	0.142621	0.050000	0.027119	0.058124
Iteration: 11				
*A*0	0.230747	0.098658	0.074576	0.089603
*A*1	0.336013	0.137920	0.145763	0.150423
*A*2	0.290619	0.113423	0.152542	0.101850
*A*3	0.142621	0.050000	0.027119	0.058124

Table A.4: 8th to 11th iteration values

Appendix B

Unique citation profiles

Table B.1: Discovery papers (threshold: ≥ 500 citations, $r < 0.4$)

Title	Authors	Year	Citation count	Average age	Publication venue	Paper notation
Implementing data cubes efficiently	Venky Harinarayan, Anand Rajaraman, Jeffrey D. Ullman	1996	611	6.335515548	Journal Id(253)	Harinarayan_DC
STATEMATE: A Working Environment for the Development of Complex Reactive Systems	David Harel, Hagi Lachover, Amnon Naamad, Amir Pnueli, Michal Politi, Rivi Sherman, Aharon Shtull-trauring, Mark B. Trakhtenbrot	1990	520	8.767307692	Journal Id(27)	Harel_CRS
Memory consistency and event ordering in scalable shared-memory multiprocessors	Kourosh Gharachorloo, Daniel Lenoski , James Laudon , Phillip Gibbons , A. Gupta , J. Hennessy	1990	547	8.530164534	Journal Id(454)	Gharachorloo_MCMP
Equation-based congestion control for unicast applications	Sally Floyd , Mark Handley , Jitendra Padhye , J”rg Widmer	2000	626	4.725239617	Journal Id(218212)	Floyd_UA
Database abstractions: Aggregation and generalization	John Miles Smith , Diane C. Pirog Smith	1977	623	12.22792937	Journal Id(222)	Smith_DA
Plenoptic Modeling: An Image-Based Rendering System	Leonard McMillan , Gary Bishop	1995	572	6.611888112	Journal Id(804)	McMillan_PM

Continued on next table...

Table B.2: Discovery papers

Title	Authors	Year	Citation count	Average age	Publication venue	Paper notation
MIS: A Multiple-Level Logic Optimization System	Robert K. Brayton , Richard L. Rudell , Alberto L. Sangiovanni-vincentelli , Albert R. Wang	1987	507	8.037475345	Journal Id(667)	`Brayton_MIS`
Receiver-Driven Layered Multicast	Steven McCanne , Van Jacobson , Martin Vetterli	1996	627	6.256778309	Journal Id(294)	`McCanne_RLM`
Low Power CMOS Digital Design	A. P. Chandrakasan , R. W. Brodersen	1996	714	6.774509804	Journal Id(5320)	`Chandrakasan_CMOSD`
The cosmic cube	Charles L. Seitz	1985	555	7.432432432	Journal Id(209)	`Seitz_CC`
The Java Virtual Machine Specification	Tim Lindholm , Frank Yellin	1997	723	5.822959889	Conference Id(1589)	`Lindholm_JVMS`
Capacity of multi-antenna Gaussian channels	E. Telatar	2004	640	2.196875	Journal Id(1062)	`Telatar_CGC`
The UNIX Time Sharing System	Dennis M. Ritchie , Ken Thompson	1974	531	14.09416196	Journal Id(209)	`Ritchie_TSS`
A survey of active network research	D. L. Tennenhouse , J. M. Smith , W. D. Sincoskie , D. J. Wetherall , G. J. Minden	1997	510	4.807843137	Journal Id(1041)	`Tennenhouse_ANR`

Continued on next page

Table B.2 – *Continued from previous page*

Title	Authors	Year	Citation count	Average age	Publication venue	Paper notation
Semantic Database Modeling: Survey , Applications , and Research Issues	Richard Hull , Roger L. King	1987	527	8.557874763	Journal Id(210)	`Hull_SDM`
Extending the database relational model to capture more meaning	Edgar Frank Codd	1979	678	12.46312684	Journal Id(222)	`Codd_DRD`
Can programming be liberated from the von Neumann style?: A functional style and its algebra of programs	John W. Backus	1978	538	12.88475836	Journal Id(1779)	`Backus_PVNS`
Grid Information Services for Distributed Resource Sharing	Karl Czajkowski , Carl Kesselman , Steven Fitzgerald , Ian T. Foster	2001	603	4.253731343	Journal Id(800)	`Czajkowski_GIDR`
PVM: A Framework for Parallel Distributed Computing	Vaidy S. Sunderam	1990	516	8.337209302	Journal Id(305)	`Sunderam_PVM`
Myrinet: A Gigabit-per-Second Local Area Network	N. J. Boden , Daniel I. A. Cohen , Robert E. Felderman , Alan E. Kulawik , Charles L. Seitz , Jakov N. Seizovic , Wen-king Su	1995	1013	6.421520237	Journal Id(401)	`Boden_MLAN`

Continued on next page

Table B.2 – *Continued from previous page*

Title	Authors	Year	Citation count	Average age	Publication venue	Paper notation
Hypertext: An Introduction and Survey	Jeff Conklin	1987	635	9.792125984	Journal Id(11)	`Conklin_HT`
Efficient Implementation of a BDD Package	Karl S. Brace, Richard L. Rudell, Randal E. Bryant	1990	597	8.194304858	Conference Id(1099)	`Brace_BDDP`
A Survey of Wormhole Routing Techniques in Direct Networks	Lionel M. Ni , Philip K. Mckinley	1993	630	6.985714286	Journal Id(11)	`Ni_RTDN`
Three-Dimensional Computer Vision: A Geometric Viewpoint	O. Faugeras	1996	638	6.222570533	Journal Id(242)	`Faugeras_CV`
A case for end system multicast	Yang-hua Chu , Sanjay G. Rao , Hui Zhang	2002	686	3.883381924	Journal Id(16)	`Chu_ESM`
Technical Reports	Douglas H. Adams , Robert H. McMichael , George E. Henderson	1992	839	5.972586412	Journal Id(459)	`Adams_TR`
The Lorel Query Language for Semistructured Data	Serge Abiteboul , Dallan Quass , Jason Mchugh , Jennifer Widom , Janet L. Wiener	1997	584	4.686643836	Journal Id(64)	`Abiteboul_LQSD`

Continued on next page

Table B.2 – *Continued from previous page*

Title	Authors	Year	Citation count	Average age	Publication venue	Paper notation
Bandera: Extracting Finite-State Models from Java Source Code	James C. Corbett , Matthew B. Dwyer , John Hatcliff , Robby , Hongjun Zheng	2000	523	4.760994264	Conference Id(40)	`Corbett_FSMJ`
A reliable multicast framework for light-weight sessions and application level framing	Sally Floyd , Van Jacobson , Ching-Gung Liu , Steven McCanne , Lixia Zhang	1997	671	5.002980626	Journal Id(7)	`Floyd_LF`
Active messages: a mechanism for integrated communication and computation	Thorsten von Eicken , David E. Culler , Seth Copen Goldstein , Klaus Erik Schauser	1992	748	7.04144385	Conference Id(86)	`Eicken_ ICC`
Constraint logic programming	Joxan Jaffart , Jean-Louis Lassez	1987	752	9.742021277	Conference Id(225)	`Jaffart_CLP`
Memory Coherence in Shared Virtual Memory Systems	Kai Li , Paul Hudak	1989	669	8.361733931	Journal Id(221)	`Li_MCSM`
Querying Heterogeneous Information Sources Using Source Descriptions	Alon Y. Levy , Anand Rajaraman , Joann J. Ordille	1996	641	6.326053042	Conference Id(458)	`Levy_SD`
Notes on Data Base Operating Systems	Jim Gray	1978	683	13.45973646	Conference Id(2058)	`Gray_OS`

Continued on next page

Table B.2 – *Continued from previous page*

Title	Authors	Year	Citation count	Average age	Publication venue	Paper notation
Lightweight causal and atomic group multicast	Andre Schiper , Kenneth Birman , Pat Stephenson	1991	522	7.745210728	Journal Id(221)	`Schiper_AGM`
Parallel discrete event simulation	Richard M. Fujimoto	1990	844	8.630331754	Journal Id(209)	`Fujimoto_ES`
RTP: A Transport Protocol for Real-Time Applications	H. Schulzrinne , S. Casner , R. Frederick , V. Jacobson	2001	1020	3.431372549	Journal Id(991)	`Schulzrinne_RTP`
A Scheme for Real-Time Channel Establishment in Wide-Area Networks	Domenico Ferrari , Dinesh C. Verma	1990	627	8.051036683	Journal Id(16)	`Ferrari_CEWAN`
MPEG: a video compression standard for multimedia applications	Didier Le Gall	1991	554	7.557761733	Journal Id(209)	`Gall_MPEG`
The Unified Modeling Language User Guide	Grady Booch , James E. Rumbaugh , Ivar Jacobson	1999	1919	4.540385618	Journal Id(107)	`Booch_UML`
Span: an Energy-Efficient Coordination Algorithm for Topology Maintenance in Ad Hoc Wireless Networks	Benjie Chen , Kyle Jamieson , Hari Balakrishnan	2002	662	3.996978852	Journal Id(275)	`Chen_WN`

Table B.3: Hot papers (threshold ≥ 350 citations, $r \geq 2/3$)

Title	Authors	Year	Citation count	Average age	Publication venue	Paper notation
Indexing by Latent Semantic Analysis	Scott C. Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard A. Harshman	1990	2214	15.22312556	Journal Id(141)	`Dumais_LSA`
Abstract interpretation: a unified lattice model for static analysis of programs by construction or approximation of fixpoints	Patrick Cousot, Radhia Cousot	1977	1828	25.49452954	Conference Id(255)	`Cousot_LM`
R-trees: a dynamic index structure for spatial searching	Antonin Guttman	1984	2231	18.67144778	Conference Id(370)	`Guttman _RT`
Particle swarm optimization	James N. Kennedy, Russell C. Eberhart	1995	2573	13.04897007	Conference Id(2133)	`Kennedy _PSO`
Classification and Regression Trees	Leo Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone	1984	3610	18.82742382	Conference Id(2614)	`Brieman _RT`
Network information flow	Rudolf Ahlswede, Ning Cai, Shuo-yen Robert Li,Raymond W. Yeung	2000	1577	8.562460368	Journal Id(433)	`Ahlswede_NIF`

Continued on next page

Table B.3 – *Continued from previous page*

Title	Authors	Year	Citation count	Average age	Publication venue	Paper notation
Fuzzy Sets	Lotfi A. Zadeh	1965	6606	38.48622464	Journal Id (40)	Zadeh_FS
Computers and Intractability: A Guide to the Theory of NP-Completeness	Michael Randolph Garey, David S. Johnson	1979	11788	23.08322022	Conference Id(277)	Garey_NP
Information Retrieval	C. J. Van Rijsbergen	1979	1817	24.16895982	Journal Id(250)	Rijsbergen_IR
Digital Image Processing	B.R. Hunt	1981	1926	22.3219107	Journal Id(8536)	Hunt_DIP
How to Share a Secret	Adi Shamir	1979	1976	25.42459514	Journal Id(209)	Shamir_SS
New directions in cryptography	Whitefield Diffie, Martin E. Hellman	1976	2876	26.14638387	Journal Id (433)	Diffie_CYT
A Threshold Selection Method from Gray-Level Histograms	N. Otsu	1979	1760	26.5875	Journal Id(28)	Otsu_TSH
Low-Density Parity-Check Codes	R. G. Gallager	1963	1938	43.35603715	Journal Id(433)	Gallager_LPC
Digital communications	R Korn, P Wilmott	1985	5374	18.44436174	Journal Id(182)	author_DC
Data network		1982	1546	21.13195343	Journal Id(295)	author_DN
A method for obtaining digital signatures and public-key cryptosystems	Ronald L. Rivest, Adi Shamir, Leonard M. Adleman	1978	3108	23.91956242	Journal Id(209)	Rivest_DSC
R-trees: a dynamic index structure for spatial searching	Antonin Guttman	1984	2231	18.67144778	Conference Id(370)	Guttman_RT

Continued on next page

Table B.3 – *Continued from previous page*

Title	Authors	Year	Citation count	Average age	Publication venue	Paper notation
Particle swarm optimization	James N. Kennedy, Russell C. Eberhart	1995	2573	13.04897007	Conference Id(2133)	`Kennedy_PSO`
An Iterative Image Registration Technique with an Application to Stereo Vision	BD Lucas, T Kanade	1981	1896	24.47310127	Conference Id(64)	`Lucas_IRT`
Term-weighting approaches in automatic text retrieval	Gerard Salton, Christopher Buckley	1988	1676	16.82637232	Journal Id(45)	`Salton_ATR`
Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography	Martin A. Fischler, Robert C. Bolles	1981	2120	24.88396226	Journal Id(209)	`Fischler_RSC`
Scheduling Algorithms for Multiprogramming in a Hard-Real-Time Environment	Chung Laung Liu, James W. Layland	1973	3097	28.5586051	Journal Id(89)	`Liu_SA`
Rough sets	Zdzisław Pawlak	1982	1789	24.11514813	Journal Id(79)	`Pawlak_RS`
The mathematical theory of communication	C. E. Shannon, W. Weaver	1964	3727	38.88596727	Journal Id(909)	`Shannon_TC`

Appendix C

Reviewer assignment : A toy example

In this section, we demonstrate the assignment procedure followed by our proposed algorithm using dummy example data. In this example, we consider 12 papers to be assigned among 3 reviewers. Constraints for this example are defined as – (i) Each paper should be assigned to at least 2 reviewer. Hence, the value of $q = 2$. (ii) Each reviewer is assigned at most 8 papers $l = (\frac{12\times2}{3}) = 8$. Given a reviewer set $R = \{R1, R2, R3\}$ and paper set $P = \{P1, P2, P3, P4, P5, P6, P7, P8, P9, P10, P11, P12\}$; the problem is to find weight of an assignment edge $w(i, j)$ between R and P set such that maximal matching occurs between them. Maximal matching occurs if percentage of topic similarity that is, value of $S(r_i, p_j)$ between R and P set is maximum, minimum Conflict of Interest or inversely, maximum collaborative distance $D(r_i, p_j) >= 3$ exist between author set of paper p_j and reviewer set r_i.

Step 1: Firstly, we create a topic similarity matrix $S(r_i, p_j)$ by assigning each paper and reviewer a topic similarity percentage based on presence of distinct common keywords in topic set of paper T_p and topic set of reviewer T_r.

Table C.1: Topic similarity $S(r_i, p_j)$ matrix

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
R1	50	30	80	60	50	65	70	50	80	40	30	80
R2	70	40	30	75	35	75	80	30	60	50	40	90
R3	90	50	20	40	80	85	85	10	70	55	45	75

Step 2: Next, we calculate collaborative distance matrix $D(r_i, p_j)$ from author set of submitted paper P given by $A(p_j)$ and reviewer set R.

Table C.2: Conflict of Interest ($CoI(r_i, p_j)$) matrix

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
R1	3	0	0	3	2	3	3	3	3	3	1	2
R2	2	3	3	3	3	1	3	3	3	2	3	3
R3	3	3	3	3	3	3	2	1	3	3	3	3

Step 3: For calculation of weight matrix, we take product of topic similarity matrix $S(r_i, p_j)$ and collaborative distance matrix $D(r_i, p_j)$. For each paper, we calculate sum of weights column wise considering assignment of a single paper to multiple reviewers.

Table C.3: Weight matrix ($w'(i, j)$)

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
R1	150	0	0	180	100	195	210	150	240	120	30	160
R2	140	120	90	225	105	75	240	90	180	100	120	270
R3	270	150	60	120	240	255	170	10	210	165	135	225
SUM	560	270	150	525	445	525	620	250	630	385	285	655

Step 4: Finally we calculate a assignment matrix by dividing each element from the sum calculated in previous step. Next we continue iterating and make assignments of reviewer to papers by changing the weights as described below until any reviewer gets the maximum load l = 8.

1. After iteration 1 of assignment matrix, we assign reviewers to papers whose assignment weight is greater than 0.5. The assignments made in this iteration are $R1 \rightarrow \{P8\}$, $R2 \rightarrow \{P3\}$, $R3 \rightarrow \{P2, P5\}$.

2. After iteration 2 of assignment matrix, we assign reviewers to papers whose assignment weight is greater than 0.45. The assignments made in this iteration are $R3 \rightarrow \{P1, P6, P11\}$. The final set of assignments are $R1 \rightarrow \{P8\}$, $R2 \rightarrow \{P3\}$, $R3 \rightarrow$ *{P2, P5, P1, P6, P11}*.

3. After iteration 3 of assignment matrix, we assign reviewers to papers whose assignment weight is greater than 0.40. The assignments made in this iteration are $R2 \rightarrow \{P2, P4, P11, P12\}$, $R3 \rightarrow \{P3, P10\}$. The final set of assignments are $R1 \rightarrow \{P8\}$, $R2 \rightarrow \{P3, P2, P4, P11, P12\}$, $R3 \rightarrow \{P2, P5, P1, P6, P11, P3, P10\}$. *Note here that papers {P2, P3, P11} are assigned maximum of 2 reviewers so they are marked in red.*

4. After iteration 4 of assignment matrix, we assign reviewers to papers whose assignment weight is greater than 0.35. The assignments made in this iteration are $R1 \rightarrow \{P6, P9\}$, $R2 \rightarrow \{P7, P8\}$. The final set of assignments are $R1 \rightarrow \{P8, P6, P9\}$, $R2 \rightarrow \{P3, P2, P4, P11, P12, P7, P8\}$, $R3 \rightarrow \{P2, P5, P1, P6, P11, P3, P10\}$. *Note here that papers* $\{P6, P8\}$ *are assigned maximum of 2 reviewers so they are marked in red.*

5. After iteration 5 of assignment matrix, we assign reviewers to papers whose assignment weight is greater than 0.30. The assignments made in this iteration are $R1 \rightarrow \{P4, P7, P10\}$, $R3 \rightarrow \{P9\}$. The final set of assignments are $R1 \rightarrow \{P8, P6, P9, P4, P7, P10\}$, $R2 \rightarrow \{P3, P2, P4, P11, P12, P7, P8\}$, $R3 \rightarrow \{P2, P5, P1, P6, P11, P3, P10, P9\}$. *Note here that papers* $\{P4, P7, P9, P10\}$ *are assigned maximum of 2 reviewers so they are marked in red. Also note that maximum load of reviewer* $R3$ *has reached a maximum of 8 papers. For remaining paper set, we remove* $R3$ *and re-calculate.*

Here, we have used different colour codes which refer to corresponding paper and reviewer sets being assigned after each iteration. We use color BLUE when a paper is assigned to a reviewer and RED when a paper satisfies its minimum requirement.

Table C.4: Assignment matrix (Iteration 1)

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
R1	0.26	NA	NA	0.34	0.22	0.37	0.338	0.6	0.38	0.311	0.105	0.244
R2	0.25	0.44	0.6	0.42	0.24	0.14	0.38	0.36	0.28	0.2597	0.42	0.417
R3	0.48	0.55	0.4	0.22	0.53	0.48	0.27	0.04	0.33	0.42	0.47	0.343

Table C.5: Assignment matrix (Iteration 2)

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
R1	0.26	NA	NA	0.34	0.22	0.37	0.338	0	0.38	0.311	0.105	0.244
R2	0.25	0.44	0	0.42	0.24	0.14	0.38	0.36	0.28	0.2597	0.42	0.417
R3	0.48	0	0.4	0.22	0	0.48	0.27	0.04	0.33	0.42	0.47	0.343

Table C.6: Assignment matrix (Iteration 3)

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
R1	0.26	NA	NA	0.34	0.22	0.37	0.338	0	0.38	0.311	0.105	0.244
R2	0.25	0.44	0	0.42	0.24	0.14	0.38	0.36	0.28	0.2597	0.42	0.417
R3	0	0	0.4	0.22	0	0	0.27	0.04	0.33	0.42	0	0.343

Table C.7: Assignment matrix (Iteration 4)

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
R1	0.26	NA	NA	0.34	0.22	0.37	0.338	0	0.38	0.311	0.105	0.244
R2	0.25	0	0	0	0.24	0.14	0.38	0.36	0.28	0.2597	0	0
R3	0	0	0	0.22	0	0	0.27	0.04	0.33	0	0	0.343

Table C.8: Assignment matrix (Iteration 5)

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
R1	0.26	NA	NA	0.34	0.22	0	0.338	0	0	0.311	0.105	0.244
R2	0.25	0	0	0	0.24	0.14	0	0	0.28	0.2597	0	0
R3	0	0	0	0.22	0	0	0.27	0.04	0.33	0	0	0.343

Table C.9: Final assignment matrix for step 1

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
R1	0.26	NA	NA	0	0.22	0	0	0	0	0	0.105	0.244
R2	0.25	0	0	0	0.24	0.14	0	0	0.28	0.2597	0	0
R3	0	0	0	0.22	0	0	0.27	0.04	0	0	0	0.343

Step 5: In step 5 remaining paper set to be assigned include $\{P1, P5, P12\}$. We remove reviewer $R3$ from calculation and as in step 1, re-write topic similarity matrix and collaborative distance matrix for remaining paper sets. We re-calculate weights considering only 2 reviewer, $R1$ and $R2$.

Table A.10: $S(r_i, p_j)$

	P1	P5	P12
R1	50	50	80
R2	70	35	NA

Table A.11: $CoI(r_i, p_j)$

	P1	P5	P12
R1	3	2	2
R2	2	3	NA

Table A.12: Weight matrix ($w'(i, j)$)

	P1	P5	P12
R1	150	100	160
R2	140	105	NA
SUM	290	205	160

Table A.13: Assignment (Iteration 1)

	P1	P5	P12
R1	0.51	0.48	1
R2	0.48	0.51	NA

Table A.14: Final assignment (step 5)

	P1	P5	P12
R1	0.51	0.48	1
R2	0.48	0.51	NA

Table A.15: Optimal assignment

Paper	Reviewer 1	Reviewer 2
P1	R1	R3
P2	R2	R3
P3	R2	R3
P4	R1	R2
P5	R2	R3
P6	R1	R3
P7	R1	R2
P8	R1	R2
P9	R1	R3
P10	R1	R3
P11	R2	R3
P12	R2	R3

1. After iteration 1 of assignment matrix, we assign reviewers to papers whose assignment weight is greater than 0.5. Assignments made in this iteration are $R1 \rightarrow \{P1, P12\}$, $R2 \rightarrow \{P5\}$.

2. After final assignment matrix for step 2, the reviewer set of assignments (refer to table C.10) are $R1 \rightarrow \{P8, P6, P9, P4, P7, P10, P1, P12\}$, $R2 \rightarrow \{P3, P2, P4, P11, P12, P7, P8, P5\}$, $R3 \rightarrow \{P2, P5, P1, P6, P11, P3, P10, P9\}$.

3. After final assignment matrix, the paper set of assignments (refer to table C.10) are $P1 \rightarrow \{R1, R3\}$, $P2 \rightarrow \{R2, R3\}$, $P3 \rightarrow \{R2, R3\}$, $P4 \rightarrow \{R1, R2\}$, $P5 \rightarrow \{R2, R3\}$, $P6 \rightarrow \{R1, R3\}$, $P7 \rightarrow \{R1, R2\}$, $P8 \rightarrow \{R1, R2\}$, $P9 \rightarrow \{R1, R3\}$, $P10 \rightarrow \{R1, R3\}$, $P11 \rightarrow \{R2, R3\}$, $P12 \rightarrow \{R1, R2\}$

Table C.10: Final assignment of reviewer to paper

R1	P8	P6	P9	P4	P7	P10	P1	P12
R2	P3	P2	P4	P11	P12	P7	P8	P5
R3	P2	P5	P1	P6	P11	P3	P10	P9

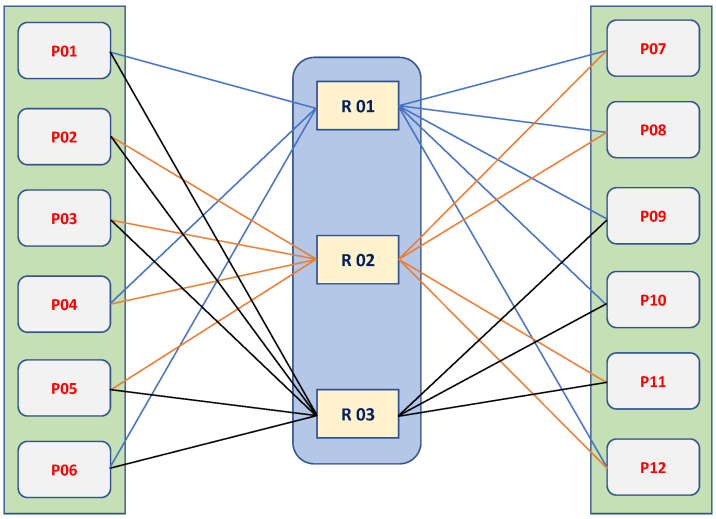


Figure C.1: Paper to reviewer final assignment

The final assignment is obtained as given in Table C.10 and is illustrated in figure C.1. The assignments are consistent to maximize topic similarity, minimize CoI value with consideration of reviewer's assignment load.

Bibliography

[1] Tanmoy Chakraborty, Suhansanu Kumar, Pawan Goyal, Niloy Ganguly, and Animesh Mukherjee. On the categorization of scientific citation profiles in computer science. *Communications of the ACM*, 58(9):82–90, 2015.

[2] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1067–1077. International World Wide Web Conferences Steering Committee, 2015.

[3] Ying Ding, Guo Zhang, Tamy Chambers, Min Song, Xiaolong Wang, and Chengxiang Zhai. Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information Science and Technology*, 65(9):1820–1833, 2014.

[4] Eugene Garfield and Robert King Merton. *Citation indexing: Its theory and application in science, technology, and humanities*, volume 8. Wiley New York, 1979.

[5] Jasleen Kaur, Emilio Ferrara, Filippo Menczer, Alessandro Flammini, and Filippo Radicchi. Quality versus quantity in scientific impact. *Journal of Informetrics*, 9(4):800–808, 2015.

[6] Lutz Bornmann and Hans-Dieter Daniel. The state of h index research. *EMBO reports*, 10(1):2–6, 2009.

[7] M. Kosmulski. A new Hirsch-type index saves time and works equally well as the original h-index. *ISSI Newsletter*, pages 4–6, 2006.

[8] Allen W Wilhite and Eric A Fong. Coercive citation in academic publishing. *Science*, 335(6068):542–543, 2012.

[9] Petr Heneberg. From excessive journal self-cites to citation stacking: analysis of journal self-citation kinetics in search for journals, which boost their scientometric indicators. *PloS one*, 11(4):e0153730, 2016.

[10] P. Mongeon, L. Waltman, and S. Rijcke. What do we know about journal citation cartels? A call for information. https://www.cwts.nl/blog?article=n-q2w2b4, 2016.

[11] Iztok Fister Jr, Iztok Fister, and Matjaž Perc. Toward the discovery of citation cartels in citation networks. *Frontiers in Physics*, 4:49, 2016.

[12] Xiaomei Bai, Feng Xia, Ivan Lee, Jun Zhang, and Zhaolong Ning. Identifying anomalous citations for objective evaluation of scholarly article impact. *Plos One*, 11(9):e0162364, 2016.

[13] Douglas N Arnold. Integrity under attack: the state of scholarly publishing. *SIAM news*, 42(10):2–3, 2009.

[14] John T Cacioppo. *Social neuroscience*. MIT Press, 2016.

[15] Filippo Radicchi and Claudio Castellano. Understanding the scientific enterprise: citation analysis, data and modeling. In *Social Phenomena*, pages 135–151. Springer, 2015.

[16] Eric A. Fong and Allen W. Wilhite. Authorship and citation manipulation in academic research. *Plos One*, 12(12):e0187394, 2017.

[17] Thomas Reuters Journal Citation Report. https://clarivate.com/products/journal-citation-reports/, 2009.

[18] Clarivate Analytics. Title suppressions from journal citation reports. *Erişim adresi: http://wokinfo. com/media/pdf/jcr-suppression. pdf*, 2017.

[19] Cheng Long, Raymond Chi-Wing Wong, Yu Peng, and Liangliang Ye. On good and fair paper-reviewer assignment. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 1145–1150. IEEE, 2013.

[20] Frank-Thorsten Krell. Should editors influence journal impact factors? *Learned Publishing*, 23(1):59–62, 2010.

[21] Robert M Carey. Quantifying scientific merit: Is it time to transform the impact factor? *Circulation research*, 119(12):1273–1275, 2016.

[22] Albert-László Barabási, Hawoong Jeong, Zoltan Néda, Erzsebet Ravasz, Andras Schubert, and Tamas Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications*, 311(3):590–614, 2002.

[23] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16569–16572, 2005.

[24] Peng Chen, Huafeng Xie, Sergei Maslov, and Sidney Redner. Finding scientific gems with google's pagerank algorithm. *Journal of Informetrics*, 1(1):8–15, 2007.

[25] S. Boccaletti, G. Bianconi, R. Criado, C.I. del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendiña-Nadal, Z. Wang, and M. Zanin. The structure and dynamics of multilayer networks. *Physics Reports*, 544(1):1 – 122, 2014.

[26] Rodrigo Costas and María Bordons. The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. *Journal of Informetrics*, 1(3):193 – 203, 2007. The Hirsch Index.

[27] Ludo Waltman and Nees Jan van Eck. The inconsistency of the h-index. *Journal of the American Society for Information Science and Technology*, 63(2):406–415, February 2012.

[28] Ludo Waltman, Rodrigo Costas, and Nees Jan van Eck. Some limitations of the h index: A commentary on ruscio and colleagues' analysis of bibliometric indices. *Measurement: Interdisciplinary Research and Perspectives*, 10(3):172–175, 2012.

[29] J. E. Hirsch. An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. *Scientometrics*, 85(3):741–754, December 2010.

[30] Leo Egghe. Theory and practise of the g-index. *Scientometrics*, 69(1):131–152, 2006.

[31] BiHui Jin, LiMing Liang, Ronald Rousseau, and Leo Egghe. The r- and ar-indices: Complementing the h-index. *Chinese Science Bulletin*, 52(6):855–863, 2007.

[32] Sidney Redner. On the meaning of the h-index. *Journal of Statistical Mechanics: Theory and Experiment*, 2010(03):L03005, 2010.

[33] Giovanni Abramo, Ciriaco Andrea D'Angelo, and Marco Solazzi. Are researchers that collaborate more at the international level top performers? an investigation on the italian university system. *Journal of Informetrics*, 5(1):204 – 213, 2011.

[34] Frank J. Trueba and Héctor Guerrero. A robust formula to credit authors for their publications. *Scientometrics*, 60(2):181–204, 2004.

[35] Jian Xu, Ying Ding, Min Song, and Tamy Chambers. Author credit-assignment schemas: A comparison and analysis. *Journal of the Association for Information Science and Technology*, 67(8):1973–1989, 2016.

[36] Teja Tscharntke, Michael E Hochberg, Tatyana A Rand, Vincent H Resh, and Jochen Krauss. Author sequence and credit for contributions in multiauthored publications. *PLOS Biology*, 5(1):1–2, Jan 2007.

[37] Nan Ma, Jiancheng Guan, and Yi Zhao. Bringing pagerank to the citation analysis. *Information Processing & Management*, 44(2):800–810, 2008.

[38] Ying Ding and Blaise Cronin. Popular and/or prestigious? measures of scholarly esteem. *Information Processing and Management*, 47(1):80–96, January 2011.

[39] Filippo Radicchi, Santo Fortunato, Benjamin Markines, and Alessandro Vespignani. Diffusion of scientific credits and the ranking of scientists. *Physical Review E*, 80(5):056103, 2009.

[40] José Luis Ortega. Influence of co-authorship networks in the research impact: Ego network analyses from microsoft academic search. *Journal of Informetrics*, 8(3):728 – 737, 2014.

[41] Xiaoming Liu, Johan Bollen, Michael L. Nelson, and Herbert Van de Sompel. Co-authorship networks in the digital library research community. *Information Processing & Management*, 41(6):1462–1480, December 2005.

[42] Ying Ding, Erjia Yan, Arthur Frazho, and James Caverlee. Pagerank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology*, 60(11):2229–2243, 2009.

[43] Michal Nykl, Karel Ježek, Dalibor Fiala, and Martin Dostal. Pagerank variants in the evaluation of citation networks. *Journal of Informetrics*, 8(3):683 – 692, 2014.

[44] Upul Senanayake, Mahendra Piraveenan, and Albert Zomaya. The pagerank-index: Going beyond citation counts in quantifying scientific impact of researchers. *PLoS ONE*, 10(8):1–34, 08 2015.

[45] U. Senanayake, M. Piraveenan, and A. Zomaya. Ranking scientists from the field of quantum game theory using p-index. In *Foundations of Computational Intelligence (FOCI), 2014 IEEE Symposium on*, pages 9–16, Dec 2014.

[46] U. Senanayake, M. Piraveenan, and A.Y. Zomaya. The p-index: Ranking scientists using network dynamics. *Procedia Computer Science*, 29:465 – 477, 2014.

[47] Jingyu Cui, Fan Wang, and Jinjian Zhai. Citation networks as a multi-layer graph: Link prediction and importance ranking. *CS224W Project Report*, 2010.

[48] Arda Halu, Raúl J. Mondragón, Pietro Panzarasa, and Ginestra Bianconi. Multiplex pagerank. *PLOS ONE*, 8(10):1–10, 10 2013.

[49] Manlio De Domenico, Albert Sole-Ribalta, Elisa Omodei, Sergio Gomez, and Alex Arenas. Ranking in interconnected multilayer networks reveals versatile nodes. *Nature Communications*, 6, April 2015.

[50] Ding Zhou, Sergey A Orshanskiy, Hongyuan Zha, and C Lee Giles. Co-ranking authors and documents in a heterogeneous network. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 739–744. IEEE, 2007.

[51] Hongbo Deng, Jiawei Han, Michael R. Lyu, and Irwin King. Modeling and exploiting heterogeneous bibliographic networks for expertise ranking. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '12, pages 71–80, New York, NY, USA, 2012. ACM.

[52] Erjia Yan, Ying Ding, and Cassidy R. Sugimoto. P-rank: An indicator measuring prestige in heterogeneous scholarly networks. *Journal of the American Society for Information Science and Technology*, 62(3):467–477, 2011.

[53] Dalibor Fiala, Lovro Subelj, Slavko Zitnik, and Marko Bajec. Do pagerank-based author rankings outperform simple citation counts? *Journal of Informetrics*, 9(2):334–348, 2015.

[54] Derek J de Solla Price. Networks of scientific papers. *Science*, 149(3683):510–515, 1965.

[55] Dashun Wang, Chaoming Song, and Albert-László Barabási. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, 2013.

[56] Sidney Redner. Citation statistics from more than a century of physical review. *arXiv preprint physics/0407137*, 2004.

[57] Sidney Redner. How popular is your paper? an empirical study of the citation distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, 4(2):131–134, 1998.

[58] Qing Ke, Emilio Ferrara, Filippo Radicchi, and Alessandro Flammini. Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences*, 112(24):7426–7431, 2015.

[59] Leo Egghe and Ronald Rousseau. Co-citation, bibliographic coupling and a characterization of lattice citation networks. *Scientometrics*, 55(3):349–361, 2002.

[60] Michal Brzezinski. Power laws in citation distributions: Evidence from scopus. *Scientometrics*, 103(1):213–228, 2015.

[61] Pedro Albarrán and Javier Ruiz-Castillo. References made and citations received by scientific articles. *Journal of the American Society for Information Science and Technology*, 62(1):40–49, 2011.

[62] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[63] William Shockley. On the statistics of individual variations of productivity in research laboratories. *Proceedings of the IRE*, 45(3):279–290, 1957.

[64] Young-Ho Eom and Santo Fortunato. Characterizing and modeling citation dynamics. *PloS one*, 6(9):e24926, 2011.

[65] Derek de Solla Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American society for Information science*, 27(5):292–306, 1976.

[66] Mingyang Wang, Guang Yu, and Daren Yu. Measuring the preferential attachment mechanism in citation networks. *Physica A: Statistical Mechanics and its Applications*, 387(18):4692–4698, 2008.

[67] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[68] Hua-Wei Shen, Dashun Wang, Chaoming Song, and Albert-László Barabási. Modeling and predicting popularity dynamics via reinforced poisson processes. *arXiv preprint arXiv:1401.0778*, 2014.

[69] Dinesh Pradhan, Partha Sarathi Paul, Umesh Maheswari, Subrata Nandi, and Tanmoy Chakraborty. C 3-index: revisiting author's performance measure. In *Proceedings of the 8th ACM Conference on Web Science*, pages 318–319. ACM, 2016.

[70] Paul L Krapivsky and Sidney Redner. Organization of growing random networks. *Physical Review E*, 63(6):066123, 2001.

[71] Michael H MacRoberts and Barbara R MacRoberts. Problems of citation analysis: A critical review. *Journal of the American Society for information Science*, 40(5):342, 1989.

[72] Eugene Garfield. Introducing citation classics-human side of scientific reports, 1977.

[73] Arnab Chatterjee, Asim Ghosh, and Bikas K Chakrabarti. Universality of citation distributions for academic institutions and journals. *PloS one*, 11(1):e0146762, 2016.

[74] Mayank Singh, Vikas Patidar, Suhansanu Kumar, Tanmoy Chakraborty, Animesh Mukherjee, and Pawan Goyal. The role of citation context in predicting long-term citation profiles: An experimental study based on a massive bibliographic text dataset. In

Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pages 1271–1280. ACM, 2015.

[75] Han Zhu, Xinran Wang, and Jian-Yang Zhu. Effect of aging on network structure. *Physical Review E*, 68(5):056121, 2003.

[76] Mike Thelwall and Paul Wilson. Regression for citation data: An evaluation of different methods. *Journal of Informetrics*, 8(4):963–971, 2014.

[77] Henk F Moed and Martijn S Visser. Developing bibliometric indicators of research performance in computer science: An exploratory study. *CWTS report*, 1, 2007.

[78] Mike Thelwall. The precision of the arithmetic mean, geometric mean and percentiles for citation data: An experimental simulation modelling approach. *Journal of Informetrics*, 10(1):110–123, 2016.

[79] Ngai Meng Kou, Nikos Mamoulis, Yuhong Li, Ye Li, Zhiguo Gong, et al. A topic-based reviewer assignment system. *Proceedings of the VLDB Endowment*, 8(12):1852–1855, 2015.

[80] Ngai Meng Kou, U Leong Hou, Nikos Mamoulis, and Zhiguo Gong. Weighted coverage based reviewer assignment. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 2031–2046. ACM, 2015.

[81] Caspar Chorus and Ludo Waltman. A large-scale analysis of impact factor biased journal self-citations. *Plos One*, 11(8):e0161021, 2016.

[82] Diana Hicks, Paul Wouters, Ludo Waltman, Sarah De Rijcke, and Ismael Rafols. The leiden manifesto for research metrics. *Nature*, 520(7548):429, 2015.

[83] Georg Franck. Scientific communication–a vanity fair? *Science*, 286(5437):53–55, 1999.

[84] Susan T Dumais and Jakob Nielsen. Automating the assignment of submitted manuscripts to reviewers. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 233–244. ACM, 1992.

[85] Xing Wei and W Bruce Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185. ACM, 2006.

[86] Seth Hettich and Michael J Pazzani. Mining for proposal reviewers: lessons learned at the national science foundation. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 862–871. ACM, 2006.

[87] J Scott Long. Measures of sex differences in scientific productivity. *Social Forces*, 71(1):159–178, 1992.

[88] Maryam Karimzadehgan, ChengXiang Zhai, and Geneva Belford. Multi-aspect expertise matching for review assignment. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 1113–1122. ACM, 2008.

[89] Wenbin Tang, Jie Tang, Tao Lei, Chenhao Tan, Bo Gao, and Tian Li. On optimization of expertise matching with various constraints. *Neurocomputing*, 76(1):71–83, 2012.

[90] Yong-Hong Sun, Jian Ma, Zhi-Ping Fan, and Jun Wang. A hybrid knowledge and model approach for reviewer assignment. In *System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on*, pages 47–47. IEEE, 2007.

[91] Maryam Karimzadehgan and ChengXiang Zhai. Constrained multi-aspect expertise matching for committee review assignment. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1697–1700. ACM, 2009.

[92] Xinlian Li and Toyohide Watanabe. Automatic paper-to-reviewer assignment, based on the matching degree of the reviewers. *Procedia Computer Science*, 22:633–642, 2013.

[93] Xiang Liu, Torsten Suel, and Nasir Memon. A robust model for paper reviewer assignment. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 25–32. ACM, 2014.

[94] Laurent Charlin and Richard Zemel. The toronto paper matching system: an automated paper-reviewer assignment system. *openreview.net*, 2013.

[95] Paul Resnick and Hal R Varian. Recommender systems. *Communications of the ACM*, 40(3):56–58, 1997.

[96] Tanmoy Chakraborty, Sandipan Sikdar, Vihar Tammana, Niloy Ganguly, and Animesh Mukherjee. Computer science fields as ground-truth communities: their impact, rise and fall. In *ASONAM*, pages 426–433, Niagara Falls, Canada, 2013.

[97] Tanmoy Chakraborty, Sandipan Sikdar, Niloy Ganguly, and Animesh Mukherjee. Citation interactions among computer science fields: a quantitative route to the rise and fall of scientific research. *Social Network Analysis and Mining*, 4(1):187, 2014.

[98] Jie Tang et al. Arnetminer: Extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 990–998, New York, NY, USA, 2008. ACM.

[99] Microsoft Academic Graph bibliographic dataset. https://academicgraph.blob.core.windows.net/graph-2016-02-05/MicrosoftAcademicGraph.zip. Accessed: 2016-05-27.

[100] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-june Paul Hsu, and Kuansan Wang. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246. ACM, 2015.

[101] Jonathan Stallings, Eric Vance, Jiansheng Yang, Michael W Vannier, Jimin Liang, Liaojun Pang, Liang Dai, Ivan Ye, and Ge Wang. Determining scientific impact using a collaboration index. *Proceedings of the National Academy of Sciences*, 110(24):9680–9685, 2013.

[102] James P Byrnes. Publishing trends of psychology faculty during their pretenure years. *Psychological Science*, 18(4):283–286, 2007.

[103] John T. Cacioppo. Metrics of science. *Association for Psychological Science*, 21(1), 2008.

[104] Dalibor Fiala, François Rousselot, and Karel Ježek. Pagerank for bibliographic networks. *Scientometrics*, 76(1):135–158, 2008.

[105] Ying Ding. Applying weighted pagerank to author citation networks. *Journal of the American Society for Information Science and Technology*, 62(2):236–245, February 2011.

[106] Karol Życzkowski. Citation graph, weighted impact factors and performance indices. *Scientometrics*, 85(1):301–315, 2010.

[107] Göran Melin. Pragmatism and self-organization: Research collaboration on the individual level. *Research policy*, 29(1):31–40, 2000.

[108] Gabriel Pinski and Francis Narin. Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing & Management*, 12(5):297–312, 1976.

[109] Johan Bollen, A. Marko Rodriquez, and Herbert Van de Sompel. Journal status. *Scientometrics*, 69(3):669–687, 2006.

[110] Matthew E Falagas, Vasilios D Kouranos, Ricardo Arencibia-Jorge, and Drosos E Karageorgopoulos. Comparison of scimago journal rank indicator with journal impact factor. *The FASEB journal*, 22(8):2623–2628, 2008.

[111] Dhiraj Murthy and Jeremiah P. Lewis. Social media, collaboration, and scientific organizations. *American Behavioral Scientist*, 59(1):149–171, 2015.

[112] Erjia Yan and Ying Ding. Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the American Society for Information Science and Technology*, 60(10):2107–2118, 2009.

[113] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[114] T. Chakraborty, S. Sikdar, V. Tammana, N. Ganguly, and A. Mukherjee. Computer science fields as ground-truth communities: Their impact, rise and fall. In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*, pages 426–433, Aug 2013.

[115] Jialu Liu, Kin Hou Lei, Jeffery Yufei Liu, Chi Wang, and Jiawei Han. Ranking-based name matching for author disambiguation in bibliographic data. In *Proceedings of the 2013 KDD Cup 2013 Workshop*, KDD Cup '13, pages 8:1–8:8, Chicago, Illinois, 2013. ACM.

[116] Matthew L Wallace, Vincent Larivière, and Yves Gingras. Modeling a century of citation distributions. *Journal of Informetrics*, 3(4):296–303, 2009.

[117] Eugene Garfield. The impact factor and using it correctly. *Der Unfallchirurg*, 48(2):413, 1998.

[118] Eugene Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178(4060):471–479, 1972.

[119] Filippo Radicchi and Claudio Castellano. Rescaling citations of publications in physics. *Physical Review E*, 83(4):046116, 2011.

[120] Javier Ruiz-Castillo. The role of statistics in establishing the similarity of citation distributions in a static and a dynamic context. *Scientometrics*, 96(1):173–181, 2013.

[121] Filippo Radicchi, Santo Fortunato, and Claudio Castellano. Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, 105(45):17268–17272, 2008.

[122] Ludo Waltman, Nees Jan van Eck, and Anthony FJ van Raan. Universality of citation distributions revisited. *Journal of the American Society for Information Science and Technology*, 63(1):72–77, 2012.

[123] Daniel Mcnamara, Paul Wong, Peter Christen, and Kee Siong Ng. Predicting high impact academic papers using citation network features. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 14–25. Springer, 2013.

[124] Dinesh Pradhan, Tanmoy Chakraborty, Saswata Pandit, and Subrata Nandi. On the discovery of success trajectories of authors. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 91–92. International World Wide Web Conferences Steering Committee, 2016.

[125] Anthony FJ Van Raan. Sleeping beauties in science. *Scientometrics*, 59(3):467–472, 2004.

[126] Frank Havemann and Birger Larsen. Bibliometric indicators of young authors in astrophysics: Can later stars be predicted? *Scientometrics*, 102(2):1413–1434, 2015.

[127] P.O. Seglen. *Why the impact factor of journals should not be used for evaluating research.*, volume 314(7079). British Medical Journal, 1997.

[128] E. Garfield. *Why the impact factor of journals should not be used for evaluating research.*, volume 161. Canadian Medical Association Journal„ 1999.

[129] John G Lynch. Business journals combat coercive citation. *Science*, 335(6073):1169–1169, 2012.

[130] Armen Yuri Gasparyan, Marlen Yessirkepov, Alexander A Voronov, Sergey V Gorin, Anna M Koroleva, and George D Kitas. Statement on publication ethics for editors and publishers. *Journal of Korean medical science*, 31(9):1351–1354, 2016.

[131] Stephen M Lawani. On the heterogeneity and classification of author self-citations. *Journal of the Association for Information Science and Technology*, 33(5):281–284, 1982.

[132] Dag W. Aksnes. A macro study of self-citation. *Scientometrics*, 56(2):235–246, 2003.

[133] Glänzel Wolfgang, Thijs Bart, and Schlemmer Balázs. A bibliometric approach to the role of author self-citations in scientific communication. *Scientometrics*, 59(1):63–77, 2004.

[134] James H Fowler and Dag W Aksnes. Does self-citation pay? *Scientometrics*, 72(3):427–437, 2007.

[135] Richard Van Noorden. Brazilian citation scheme outed. *Nature*, 500(7464):510–1, 2013.

[136] Petr Heneberg. From excessive journal self-cites to citation stacking: analysis of journal self-citation kinetics in search for journals, which boost their scientometric indicators. *PLoS One*, 11(4):e0153730, 2016.

[137] Fan Wang, Ben Chen, and Zhaowei Miao. A survey on reviewer assignment problem. *New frontiers in applied artificial intelligence*, pages 718–727, 2008.

[138] Simon Price and Peter A Flach. Computational support for academic peer review: A perspective from artificial intelligence. *Communications of the ACM*, 60(3):70–79, 2017.

[139] Laurent Charlin, Richard Zemel, and Craig Boutilier. A framework for optimizing paper matching. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 86–95. AUAI Press, 2011.

[140] Yordan Kalmukov. Architecture of a conference management system providing advanced paper assignment features. *arXiv preprint arXiv:1111.6934*, 2011.

[141] Wenbin Tang, Jie Tang, and Chenhao Tan. Expertise matching via constraint-based optimization. In *Web intelligence and intelligent agent technology (wi-iat), 2010 ieee/wic/acm international conference on*, volume 1, pages 34–41. IEEE, 2010.

[142] David Mimno and Andrew McCallum. Expertise modeling for matching papers with reviewers. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 500–509. ACM, 2007.

[143] Devendra Kumar Tayal, PC Saxena, Ankita Sharma, Garima Khanna, and Shubhangi Gupta. New method for solving reviewer assignment problem using type-2 fuzzy sets and fuzzy functions. *Applied intelligence*, 40(1):54–73, 2014.

[144] Tomasz Kolasa and Dariusz Król. Aco-ga approach to paper-reviewer assignment problem in cms. In *KES International Symposium on Agent and Multi-Agent Systems: Technologies and Applications*, pages 360–369. Springer, 2010.

[145] Tomasz Kolasa and Dariusz Krol. A survey of algorithms for paper-reviewer assignment problem. *IETE Technical Review*, 28(2):123–134, 2011.

[146] Fan Wang, Shaorui Zhou, and Ning Shi. Group-to-group reviewer assignment problem. *Computers & Operations Research*, 40(5):1351–1362, 2013.

[147] Yanqing Wang, Bingyu Liu, Kun Zhang, Yushan Jiang, and Fuquan Sun. Reviewer assignment strategy of peer assessment: Towards managing collusion in self-assignment. In *2nd International Conference on Social Science, Public Health and Education (SSPHE 2018)*. Atlantis Press, 2019.

[148] Judy Goldsmith and Robert H Sloan. The ai conference paper assignment problem. In *Proc. AAAI Workshop on Preference Handling for Artificial Intelligence, Vancouver*, pages 53–57, 2007.

[149] Shu Zhao, Dong Zhang, Zhen Duan, Jie Chen, Yan-ping Zhang, and Jie Tang. A novel classification method for paper-reviewer recommendation. *Scientometrics*, 115(3):1293–1313, 2018.

[150] Nicola Di Mauro, Teresa MA Basile, and Stefano Ferilli. Grape: An expert review assignment component for scientific conference management systems. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 789–798. Springer, 2005.

[151] Linzhong Liu and Xin Gao. Fuzzy weighted equilibrium multi-job assignment problem and genetic algorithm. *Applied Mathematical Modelling*, 33(10):3926–3935, 2009.

[152] Guo Chen and Lu Xiao. Selecting publication keywords for domain analysis in bibliometrics: a comparison of three methods. *Journal of Informetrics*, 10(1):212–223, 2016.

[153] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[154] Dinesh K. Pradhan, Joyita Chakraborty, and Subrata Nandi. Applications of machine learning in analysis of citation network. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, CoDS-COMAD '19, pages 330–333, New York, NY, USA, 2019. ACM.

Thesis related publications by the author

Journal

[1] Pradhan, D., Paul, P. S., Maheswari, U., Nandi, S., and Chakraborty, T. (2017). C^3-index: a PageRank based multi-faceted metric for authors' performance measurement. Scientometrics, 110(1), 253-273.

Conference

[1] Dinesh K. Pradhan, Joyita Chakraborty, and Subrata Nandi. 2019. Applications of Machine Learning in Analysis of Citation Network. In 6th ACMIKDD CoDS and 24th COMAD (CoDS-COMAD '19), January 3–5, 2019, Kolkata, India. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3297001. 3297053

[2] Pradhan, D., Paul, P. S., Maheswari, U., Nandi, S., and Chakraborty, T. (2016, May). C^3-index: revisiting author's performance measure. In Proceedings of the 8th ACM Conference on Web Science (pp. 318-319). ACM.

[3] Pradhan, D., Chakraborty, T., Pandit, S., and Nandi, S. (2016, April). On the discovery of success trajectories of authors. In Proceedings of the 25th International Conference Companion on World Wide Web (pp. 91-92). International World Wide Web Conferences Steering Committee.

Communicated

[1] Dinesh K. Pradhan, Joyita Chakraborty, Prasenjit Choudhary, and Subrata Nandi. An automated conflict of interest based greedy approach for conference paper assignment system. (Journal of Informetrics)

Other publications by the author

Journal

[1] Roy, Provas Kumar, Aditi Sur, and Dinesh Kumar Pradhan (2013). Optimal short-term hydro-thermal scheduling using quasi-oppositional teaching learning-based optimization. Engineering Applications of Artificial Intelligence 26.10 (2013): 2516-2524.

Conference

[1] Banerjee T., Pradhan D. and Choudhury P. (2016). Efficient Combinatorial Auction Mechanisms in Electronic Commerce. In Proceedings of the 18th International Conference on Enterprise Information Systems ISBN 978-989-758-187-8, pages 290-297.

[2] Pradhan, D. K., Nandi, S., and Choudhury, P. (2014). Implementing encryption with Enhanced Mobility Aware Routing Protocol for Bluetooth network. In Business and Information Management (ICBIM), 2014 2nd International Conference on (pp. 171-175). IEEE.

[3] Debnath, R., Kundu, B. C., Pradhan, M., and Pradhan, D. (2011). Lossless secure transmission in bluetooth scatternet, considering device mobility. In Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology (Wireless VITAE), 2011 2nd International Conference on (pp. 1-5). IEEE.

[4] Pradhan, Moumita, Dinesh Pradhan, and G. Bandyopadhyay (2010). Implementation of Fuzzy Approach to Improve Time Estimation [Case Study of a Thermal Power Plant Is Considered.], International Conference on Modeling, Optimization, and Computing (ICMOS 20110). Vol. 1298. No. 1. AIP Publishing, 2010.