# Learning-Based Bibliometric Research Performance Assessment of Degree Awarding Institutes using Stochastic Models

**Ph.D. Thesis**

Mamoona Anam

134-FBAS/PhD/CS



**Supervisor**

Dr. Jamal Abdul Nasir
Assistant Professor

**Co-supervisor**

Dr. Malik Sikander Hayat Khayal
Professor
Preston University, Kohat Campus Islamabad

**Department of Computer Science**

**Faculty of Computing and Information Technology**

**International Islamic University, Islamabad, Pakistan**

**2024**

*A dissertation submitted to the*

*Department of Computer Science*

*International Islamic University, Islamabad*

*as a partial fulfillment of the requirements*

*for the award of the degree of*

*Doctor of Philosophy in Computer Science.*

***DEPARTMENT OF COMPUTER SCIENCE***

# INTERNATIONAL ISLAMIC UNIVERSITY ISLAMABAD

**FINAL APPROVAL**

Dated: 30th Aug 2024

It is certified that we have read the thesis titled **"Learning Based Bibliometric Research Performance Assessment of Degree Awarding Institutes using Stochastic Models"** submitted by Ms. Mamoona Anam (134-FBAS/PhD (CS)/F15). It is our conclusion that this thesis is of sufficient standards to warrant its acceptance by the International Islamic University, Islamabad for the PhD Degree in Computer Science.

## COMMITTEE

### EXTERNAL EXAMINER

Dr. Waseem Shahzad
Director, Professor
FAST, NUCES

Dr. Arif Jamal Malik
Associate Professor,
Foundation University

### INTERNAL EXAMINER

Dr. Sadia Saleem
Assistant Professor,
Department of Computer Science
FC&IT, IIU, Islamabad

### DEAN

Prof.Dr Asmatullah khan
Professor,
Faculty of Computing & Information Technology
IIU, Islamabad

### CHAIRMAN

Dr Muhammad Nadeem
Assistant Professor,
Department of Computer Science
FC&IT, IIU, Islamabad

### SUPERVISOR & CO-SUPERVISOR

Dr Jamal Abdul Nasir
Assistant Professor
University of Galway Ireland

Dr Malik Sikander Hayyat Khiyal
Professor
Preston University, Islamabad

# Declaration

I, Mamoona Anam, hereby declare that my Ph.D. thesis titled ***"Learning Based Bibliometric Research Performance Assessment of Degree Awarding Institutes using Stochastic Models"*** is my work, neither as a whole nor as part of thereof has been copied out from any source except where due reference is made in the text. It is further declared that I have not previously submitted the work presented in the thesis report for partial or full credit for the award of degree at this or any other university.

**Mamoona Anam**

**134-FBAS/PhD/CS**

# Acknowledgement

I am deeply grateful to Almighty Allah for His favors and blessings, which encouraged me to work on writing this dissertation. I owe thanks to Allah SWT for giving me a life full of strength and inspiration to accomplish this task.

Moreover, I would like to express my gratitude to my supervisor, Dr. Jamal Abdul Nasir, who provided me with valuable advice and cooperation, which enabled me to carry out this entire work. It was an honor working under his supervision.

Furthermore, I also extended special thanks to Dr. Malik Sikander Hayat Khiyal, the co-supervisor. It has been a privilege to have his insight and academic support. I am humbly thankful for his availability, generosity, fruitful discussions, and remarks that helped me to improve the entire work.

**Mamoona Anam**

**134-FBAS/PhD/CS**

# Dedication

*To my beloved father Rana Asghar Ali Shahid, mother, Dr. Malik Sikander Hayat Khayal, Giovanni Abramo, My husband Muhammad Munawar and my beloved daughter Mahnoor, Ayesha, Hadia and Fajar*

## Abstract

In the rapidly evolving landscape of higher education, the performance assessment of degree awarding institutes has become increasingly important to various stakeholders, including students, faculty, funding agencies, and policymakers. Bibliometric indicators, such as publication output, citation impact, research collaboration, and funding, serve as essential tools for evaluating research performance. This study aims to develop a novel learning-based bibliometric research performance assessment model using stochastic models to offer a comprehensive understanding of institutional research performance, capturing the uncertainties and dynamics of bibliometric data. The proposed methodology begins with data collection and preprocessing, acquiring bibliometric data from established databases, and cleaning the data to ensure its accuracy and reliability. The develop stochastic models that consider various performance indicators to assess research performance, integrating the learning-based approach for model adaptation. Model calibration and validation are conducted by comparing the results with existing assessment methodologies and case studies. The application of proposed learning-based stochastic models to assess the performance of degree-awarding institutes reveals meaningful insights into their research performance. The comparison of the models with existing methodologies demonstrates the effectiveness and robustness of the proposed approach, accounting for uncertainties and dynamics in bibliometric data. Sensitivity analysis and case studies further validate the reliability and applicability of proposed model in different contexts and institutional types. The results indicate that the learning-based stochastic models provide a comprehensive and adaptable framework for evaluating research performance. The model's adaptability and robustness allow stakeholders to make informed decisions regarding institutional performance, resource allocation, and policy development. Furthermore, the model offers valuable insights into the strengths and weaknesses of institutions, promoting continuous improvement and growth in the higher education sector. While the study provides valuable insights into research performance assessment, some limitations exist, such as the reliance on bibliometric data and potential biases in citation and collaboration patterns. Future research could explore alternative data sources and the incorporation of additional

performance indicators to further enhance the model's accuracy and comprehensiveness. Moreover, the application of advanced machine learning techniques and the development of adaptive models that consider the evolving nature of research fields can further improve the proposed framework.

***Keywords:*** *Learning-based, Bibliometric, Research performance assessment, Degree-awarding institutes, Stochastic models, Higher education, Publication output, Citation impact, Research collaboration, Funding, Model, adaptation, Uncertainties, Dynamics, Sensitivity analysis, Model validation*

# Table of Contents

# List of Acronyms

| | |
|---|---|
| P | Publication |
| D | Documents |
| C | Citations |
| TC | Total Citations |
| CB | Cited By |
| F | Funding |
| PC | Publication Count |
| CC | Citation Count |
| CO | Co Author Network |
| H | h Index |
| I | i Index |
| G | g Index |
| ARWU | Academic Ranking of World Universities |
| Min | Minimum |
| Max | Maximum |
| IF | Impact Factor |
| FWIC | Field Weighted Citation Impact |
| CRI | Composite Researcher Index |
| JIF | Journal Impact Factor |
| HEI | Higher Education Institute |
| DAI | Degree Awarding Institute |
| IPI | Institutional Performance Index |
| FSR | Funding Success Rate |
| CI | Collaboration Index |
| CIA | Comparative Institutional Analysis |
| SSB | Subject Specific Benchmarking |

| | |
|---|---|
| DEA | Data Envelopment Analysis |
| DMU | Decision Making Unit |
| VRS | Variable Return to Scale |
| CRS | Constant Return to Scale |
| SBM | Stochastic Block Model |
| CD | Community Detection |
| NG | Network Generation |
| LP | Link Prediction |
| SCM | Stochastic Citations Model |
| SFA | Stochastic Frontier Analysis |
| SAOM | Stochastic Actor Oriented Model |
| PHEC | Punjab Higher Education Commission |
| MTDF | Medium Term Improvement System Sindh HEC |
| PMDC | Pakistan Medical and Dental Council |
| PEC | Pakistan Engineering Council |
| PCATP | Pakistan Architecture and Town Planner |
| WFME | World Federation and Medical Education |
| PMC | Pakistan Medical Commission |
| IPEA | International Professional Engineer Agreement |
| WPE | Weighted Power Edge |
| WPN | Weighted Power Node |
| RPN | Researcher Power Node |
| KPI | Key Performance Indicators |
| CDF | Cumulative Distributed Function |
| SD1 | Stochastic Dominance 1 |
| FSS | Fractional Scientific Search |
| CV | Coefficient of Variation |
| SD | Standard Deviation |
| FS | Feature Selection |

| | |
|---|---|
| PK | Plus K |
| SM | Simple Markov Model |
| LAS | Lasso Regression Model |
| RFM | Random Forest Model |
| GBRT | Gradient Boosted Regression Tree |
| SVM | Support Vector Machine |
| CART | Classification and Regression Tree |
| AI | Author Influence |

# List of Tables

# LIST OF FIGURES

# List of Acronyms

| | |
|---|---|
| P | Publication |
| D | Documents |
| C | Citations |
| TC | Total Citations |
| CB | Cited By |
| F | Funding |
| PC | Publication Count |
| CC | Citation Count |
| CO | Co Author Network |
| H | h Index |
| I | i Index |
| G | g Index |
| ARWU | Academic Ranking of World Universities |
| Min | Minimum |
| Max | Maximum |
| IF | Impact Factor |
| FWIC | Field Weighted Citation Impact |
| CRI | Composite Researcher Index |
| JIF | Journal Impact Factor |
| HEI | Higher Education Institute |
| DAI | Degree Awarding Institute |
| IPI | Institutional Performance Index |
| FSR | Funding Success Rate |
| CI | Collaboration Index |
| CIA | Comparative Institutional Analysis |
| SSB | Subject Specific Benchmarking |

| | |
|---|---|
| DEA | Data Envelopment Analysis |
| DMU | Decision Making Unit |
| VRS | Variable Return to Scale |
| CRS | Constant Return to Scale |
| SBM | Stochastic Block Model |
| CD | Community Detection |
| NG | Network Generation |
| LP | Link Prediction |
| SCM | Stochastic Citations Model |
| SFA | Stochastic Frontier Analysis |
| SAOM | Stochastic Actor Oriented Model |
| PHEC | Punjab Higher Education Commission |
| MTDF | Medium Term Improvement System Sindh HEC |
| PMDC | Pakistan Medical and Dental Council |
| PEC | Pakistan Engineering Council |
| PCATP | Pakistan Architecture and Town Planner |
| WFME | World Federation and Medical Education |
| PMC | Pakistan Medical Commission |
| IPEA | International Professional Engineer Agreement |
| WPE | Weighted Power Edge |
| WPN | Weighted Power Node |
| RPN | Researcher Power Node |
| KPI | Key Performance Indicators |
| CDF | Cumulative Distributed Function |
| SD1 | Stochastic Dominance 1 |
| FSS | Fractional Scientific Search |

# Chapter 1 Introduction

In the contemporary landscape of higher education, the evaluation and enhancement of research performance among Degree Awarding Institutes (DAIs) have transcended from being mere administrative exercises to critical endeavors that shape academic reputation, innovation, and societal progress. The role of these institutes extends beyond the traditional boundaries of knowledge dissemination; they serve as crucibles of innovation, catalysts for societal progress, and conduits for scholarly excellence. The assessment of their research output, impact, and contributions to diverse fields is not only a measure of academic achievement but also a litmus test of their effectiveness as knowledge hubs. In response to the increasing complexities of academic landscapes and the burgeoning diversity of research activities, the evaluation methodologies have evolved from traditional approaches to more advanced techniques, harnessing the power of bibliometric analysis and stochastic modeling[1][2].This paper embarks on a comprehensive exploration of research performance assessment for DAIs, with a specific emphasis on a learning-based framework, fortified by the capabilities of stochastic models. By amalgamating machine learning and stochastic methodologies, this study endeavors to transcend the limitations of conventional assessment methodologies and usher in a new era of nuanced and data-driven evaluations. As the contours of research endeavors continue to metamorphose, bolstering assessment techniques with stochastic models emerges as a pivotal stride towards holistic and contextual evaluation.

The importance of research performance assessment, particularly within the sphere of higher education, is irrefutable. In a knowledge-driven world, where innovation propels progress and inquiry fuels insights, the contributions of DAIs hold profound significance. The evaluation of their research activities serves as an intricate web that connects institutional objectives, policymaking, funding allocation, and the advancement of disciplines. It necessitates an expansive vantage point, one that transcends disciplinary

boundaries and delves into the multifaceted dimensions of scholarly activities. The conventional approach to research assessment, often reliant on quantitative metrics such as publication counts and citation indices, albeit valuable, paints a partial picture, insufficient to capture the intricate interplay of variables that encapsulate research performance. As a response to this challenge, researchers, evaluators, and institutions alike have begun to embrace more sophisticated and comprehensive approaches that combine quantitative rigor with stochastic models.

## 1.2 Bibliometric Analysis Indicators

Bibliometric analysis, a burgeoning field rooted in quantifiable indicators like publications and citations, has gained prominence as an effective tool for research performance assessment. Its quantitative nature has rendered it particularly suitable for generating insights into research productivity, impact, and collaborations [3]. Yet, the limitations of bibliometrics in encapsulating the diverse spectrum of research performance have prompted a quest for augmenting its capabilities with stochastic models. These models, encapsulating uncertainty and variability, provide a framework that is not only cognizant of the intricacies of research phenomena but also capable of encapsulating the uncertain trajectories that define scholarly activity. By integrating these models into the research performance assessment paradigm, institutions and stakeholders stand to gain a more profound understanding of the dynamics that underpin scholarly progress [4].The integration of stochastic models and bibliometric analysis is emblematic of the evolving nature of research assessment methodologies. Machine learning techniques, often leveraged in tandem with stochastic modeling, have the potential to unveil latent patterns, relationships, and trends within large and intricate datasets. This confluence of methodologies empowers the assessment process by enabling the extraction of insights that extend beyond mere quantitative measurements. Notably, such methodologies align with the increasing call for multidimensional evaluations that acknowledge the holistic nature of research contributions.

The Bibliometric analysis is attached model in given in figure 1.1. Models contain document type.

### 1.2.1 Bibliometric Indicators for Research Performance

Bibliometrics is a quantitative technique to studying and analyzing scientific literature that uses metrics such as publication counts, citation rates, and cooperation patterns. Bibliometrics is commonly used to assess the research performance of individuals, companies, or nations, as well as to identify trends.

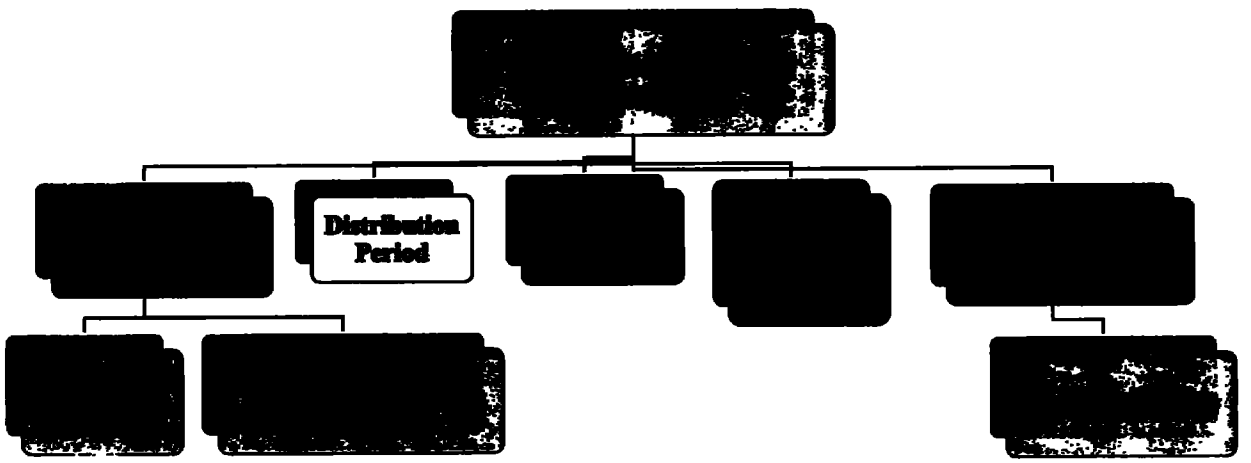



Figure 1.1: Bibliometric Analysis Indicators

1. **Document (Type)**

   In bibliometric analysis, the "document type" refers to the various forms of scholarly publications that are being examined. These can include journal articles, conference papers, books, book chapters, theses, patents, and more. Analyzing document types provides insights into the diverse ways research is communicated and contributes to the understanding of publishing trends within a specific field.

2. **Critical Papers**

Critical papers are seminal research publications that have had a significant impact on a particular field of study. These papers often introduce novel concepts, groundbreaking methodologies, or transformative findings that reshape the direction of research. Identifying critical papers through bibliometric analysis helps researchers recognize key contributions and understand the historical and intellectual development of the field.

3. **Evolutionary Years**

As a bibliometric indicator, "evolutionary years" is not a widely recognized term. However, if it refers to a specific concept introduced after my last update in September 2021, it might pertain to analyzing the progression of research over time, tracking how ideas and theories have evolved across different years or periods within a particular field.

4. **Distribution Period**

"Distribution period" is not a standard bibliometric term as of my last update. However, it could potentially refer to the temporal distribution of certain research elements, such as citations, keywords, or document types, over a specific time frame. Analyzing distribution periods could provide insights into the popularity and relevance of research concepts within different time intervals.

5. **Published Journal**

The "published journal" refers to the scholarly journals in which research articles are published. Analyzing published journals in bibliometric studies involves assessing factors such as journal impact factor, publication frequency, and citation patterns. This helps researchers understand the influence of specific journals within a field and their role in disseminating research findings.

6. **Influential Author**

An "influential author" in bibliometrics is a researcher whose work has garnered significant attention and citations within a field. Identifying influential authors helps researchers acknowledge thought leaders, collaborators, and experts who have made substantial contributions to advancing knowledge in a particular area of study.

7. **Institution Country**

The "institution country" indicator involves examining the geographic distribution of research institutions contributing to a field. This analysis provides insights into the global landscape of research, collaboration patterns between countries, and the concentration of expertise in specific regions.

8. **Keywords Frequency**

"Keywords frequency" refers to the analysis of the frequency of specific terms or phrases in research publications within a field. This indicator helps researchers identify prevalent research themes, emerging trends, and the terminology commonly used by scholars to discuss subjects.

### 1.2.2 Applications of Bibliometrics in Research Evaluation:

Bibliometrics is a quantitative assessment of academic publications and citations that may be used to evaluate research. Bibliometrics examines publishing patterns, citation counts, and other bibliographic data to gain insights into the impact, visibility, and productivity of research output. This data is essential for evaluating research at all levels, including individual researchers, research groups, institutions, and funding agencies. One significant use of bibliometrics is to evaluate the performance and productivity of individual scholars. Bibliometrics analyses metrics such as the number of publications,

citation counts, h-index, and co-authorship networks to offer quantitative estimates of an individual's scholarly

influence and contribution to the field. This information aids in recognizing notable scholars, supporting merit-based evaluations, and directing academic career development decisions. Bibliometrics, on a larger level, allows for the evaluation of research groups and organizations. Bibliometrics identifies research strengths, collaborations, and areas for development by collecting and analyzing publication and citation data from various researchers connected with a group or institution. It helps in benchmarking performance against peer institutions, funding choices, and promoting research collaborations. Bibliometrics is very significant in funding agency decision-making procedures. Funding agencies can examine bibliometric indicators such as the impact factor of journals where researchers publish, citation rates, and research collaboration networks to assess the potential impact of research projects. This data aids in the allocation of research money and resources to projects and scholars who are more likely to make important contributions to their respective disciplines. Furthermore, bibliometrics is useful for tracking research trends and social effect. By evaluating publishing trends and citation networks, bibliometrics may help uncover emerging research topics, interdisciplinary collaborations, and areas of societal value. This information supports policymakers, industry stakeholders, and funding agencies in making educated judgements on research areas that are aligned with national or global goals [5].

### 1.2.3 Research Performance Assessment

Bibliometric Analysis provides procedure to perform research performance assessment. The model is given in figure 1.2. It has two main parts Institutional researcher and evaluation of individual researcher.

#### 1.2.3.1 Individual Researcher Evaluation:

Bibliometrics is frequently used to assess the research performance of individual researchers. It helps in evaluating productivity, impact, and influence, which can be considered for funding decisions, promotions, or tenure evaluations.

### 1.2.3.2 Institutional Research Assessment:

Bibliometric analysis provides a quantitative means to evaluate the research output and impact of institutions. It aids in comparing research performance among institutions, identifying areas of strength, and informing resource allocation decisions.

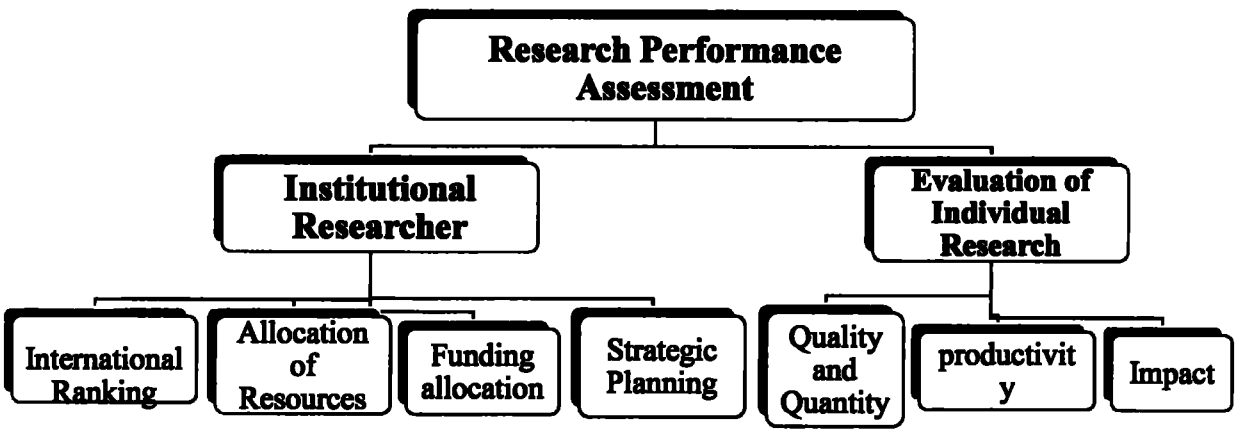



*Figure 1.2: Research Performance Assessment*

## 1.3 Background and Motivation

Higher education institutions have an essential role in fostering innovation, social and economic progress, and human capital generation. The increasing number of degree-granting institutes in Pakistan highlights the need for a rigorous and comprehensive framework for evaluating their research performance. This is critical for making educated decisions, allocating resources, developing policies, and continuously improving. A wide range of stakeholders, including students, professors, funding agencies, and

policymakers, are interested in Pakistani higher education institutions' research performance. Bibliometric variables such as publication production, citation impact, research cooperation, and financing are critical instruments for assessing research success.

Traditional ways to assessing research success frequently depend on deterministic models and aggregated indicators, which may fail to reflect the inherent uncertainties and volatility of bibliometric data. As a result, a robust and flexible framework for evaluating research performance that accounts for these limitations is essential, particularly in the context of Pakistani degree-awarding universities. Bibliometric research performance assessment, which incorporates quantitative examination of scholarly publications and citations, is critical in evaluating degree granting institutes' research production and effect. Traditional bibliometric techniques place emphasis on variables like publication numbers and citation measures. However, there has been a rising interest in merging learning-based methodologies and stochastic modelling for a more thorough assessment of research success in recent years.

The motivation behind learning-based bibliometric research performance assessment is twofold. Firstly, it addresses the limitations of traditional bibliometric approaches, which often overlook the contextual nuances and qualitative aspects of research. By integrating machine learning and statistical modeling, it becomes possible to capture more intricate patterns, correlations, and predictive relationships in research data [6]. Second, including stochastic models in bibliometric analysis gives a probabilistic framework for understanding research performance. By accounting for aspects such as random fluctuations, publication rate changes, and altering research trends, stochastic models enable researchers to evaluate research progress in a more dynamic and unpredictable setting. This stochastic viewpoint more properly shows the difficult nature of research performance. Furthermore, applying learning-based methodologies enables the

construction of predictive models capable of forecasting future research performance based on previous data. It is now feasible to uncover significant predictors of research quality and construct models that can help in strategic decision-making, resource allocation, and policy formation for degree granting institutes by employing machine learning algorithms.

The combination of learning-based techniques with stochastic modeling presents an attractive opportunity to improve the assessment of research performance in degree-granting institutions. It gives a more in-depth understanding of the underlying dynamics, trends, and potential research performance routes. Using this method, we may gain a deeper knowledge of the factors that contribute to research excellence and recommend methods to encourage continual progress in academic institutions. The implementation of learning-based bibliometric research performance evaluation for degree-granting institutes using stochastic models. The technique, difficulties, and potential advantages of this approach will be explained. By doing so, we seek to encourage research evaluation practices and aid in evidence-based decision-making at higher education institutions.

## 1.4 Research Problems

There is a noticeable gap in the literature regarding the application of learning-based stochastic models specifically for Pakistani degree-awarding institutes. The existing research performance assessment methodologies primarily rely on deterministic models and aggregated metrics, which may not fully capture the inherent uncertainties and dynamics in bibliometric data.

1. **Limited Integration of Qualitative Factors :** There is a research gap in incorporating qualitative factors, such as research collaboration networks, funding sources, and interdisciplinary research, into the learning-based stochastic models.

Incorporating such factors can provide a more comprehensive and holistic evaluation of research performance.

2. **Lack of Attention to Contextual Variations:** Existing approaches often neglect these contextual variations, leading to biased evaluations. Addressing this research gap involves developing stochastic models that can account for the contextual nuances and variations.

3. **Inadequate Consideration of Temporal Dynamics:** There is a research gap in developing stochastic models that can effectively capture and predict temporal dynamics in research performance.

4. **Limited focus on Uncertainty and Stochasticity:** There is a research gap in developing stochastic models that explicitly consider the probabilistic nature of research performance. This involves incorporating random variables, Monte Carlo simulations, or Bayesian approaches to capture the inherent uncertainties and stochastic elements of research performance.

5. **Insufficient Validation and Benchmarking:** This research gap needs comparison studies that compare the efficiency and accuracy of the proposed models to established bibliometric methodologies, proving their reliability and utility in assessing the research performance of degree-granting schools. Addressing these research gaps can help to develop the subject of bibliometric research performance assessment greatly. The proposed learning-based stochastic models can provide more accurate, nuanced, and informative evaluations of research performance in degree awarding institutes by incorporating qualitative

factors, accounting for contextual variations and temporal dynamics, considering uncertainty, and conducting rigorous validation.

## 1.5 Research Questions

1. Which learning based models are used Detecting Rising Stars in Dynamic Collaborative Networks for degree awarding institutes of Pakistan?

2. To what extent are the author-level bibliometric indicators, outlined in the exploratory study, appropriate in the evaluation of researchers from different disciplines and different academic seniorities?

3. Which models are used to Predict Citation-Based Impact Measures for learning based techniques?

## 1.6 Objectives of the Study:

The objectives of the study are as follows:

1. **To develop a learning-based approach for bibliometric research performance assessment of degree awarding institutes.**

a) To capture precise patterns, correlations, and prediction relationships in research data, machine learning techniques should be applied.

b) Incorporate qualitative and contextual aspects into the performance evaluation approach.

**2. To employ stochastic modeling to account for the dynamic and uncertain nature of research performance.**

a) Develop stochastic models that consider random fluctuations, variability in publication rates, and evolving research trends.

b) Provide a probabilistic framework to better understand and assess research performance.

**3. To identify key predictors of research excellence and develop predictive models.**

a) Utilize machine learning algorithms to identify factors that contribute to research excellence.

b) Develop models that can anticipate future research performance based on historical data.

**4. To evaluate the effectiveness and potential benefits of the proposed learning-based and stochastic modeling approach.**

a) Compare the performance evaluation results of the proposed technique to those of known bibliometric approaches.

b) Evaluate the produced models' prediction powers and accuracy.

c) Examine the effect of combining qualitative elements and stochastic modeling on the evaluation of research performance.

By achieving these objectives, this research aims to contribute to the advancement of research performance assessment in degree awarding institutes needs to develop a more comprehensive and accurate evaluation framework. The results of this study have the potential to support evidence-based decision-making, resource allocation, and policy formulation in higher education institutions.

## 1.7 Significance and Contribution

The integration of learning-based approaches with stochastic models in research performance assessment holds immense potential for addressing the complexity and dynamism of scholarly activities. By bridging the gap between quantitative metrics and probabilistic modeling, this study offers a novel methodology that enriches the assessment process with multidimensional insights. The findings of this research can guide DAIs, policymakers, and stakeholders in developing more informed strategies, fostering interdisciplinary collaborations, and refining their contributions to knowledge advancement.

## 1.8 Overview of the Pakistani Higher Education Landscape

Pakistan's higher education business has expanded dramatically in recent years, with a growing number of degree-granting schools emerging around the country. The Higher Education Commission (HEC) of Pakistan is in charge of regulating and improving the quality of higher education in the nation. The HEC is in charge of establishing rules, regulations, and recommendations for degree-granting institutions, as well as providing financial aid for research and development endeavors. Despite improvements, Pakistan's higher education industry has a number of obstacles, including a lack of resources, poor facilities, and discrepancies in access to excellent education. Assessment of research performance is crucial for identifying areas of strength and weakness, facilitating cooperation, and cultivating an excellence culture in Pakistani degree-granting institutes. The HEC's initiatives have had a significant impact on research productivity and quality in Pakistan. The availability of research grants and funding programs has led to an increase in research output, as evidenced by the growing number of publications by Pakistani researchers in reputable national and international journals. The HEC's emphasis on research quality, integrity, and ethical practices has also contributed to enhancing the standard of research conducted within the country. These efforts have resulted in greater recognition and acceptance of Pakistani research findings both

domestically and internationally, thereby elevating the reputation of Pakistani researchers and institutions [84].

## 1.9 Thesis Organization

The remainder of this study is organized as follows: Chapter 2 presents a literature review of research performance assessment, bibliometrics, existing models and methodologies, and the application of stochastic models in research assessment. Chapter 3 introduces the theoretical framework for the study, including the integration of bibliometric indicators and stochastic models. Chapter 4 describes the methodology, including data collection, model development, and validation. Chapters 5 and 6 present the results, analysis, and discussion of the findings, followed by the conclusion and recommendations for future research in Chapter 7.

# Chapter 2

## 2.1 Overview of Research Performance Assessment

The relevance of research performance evaluation in higher education institutions has expanded since it provides important insights into the quality of research outputs, the efficacy of resource allocation, and the overall competitiveness of institutions. Methods for evaluating research performance often entail the investigation of numerous quantitative and qualitative variables that indicate an institution's research production, impact, and quality. The literature on research performance assessment covers a wide range of topics, including the development of assessment frameworks, the identification of appropriate performance indicators, the comparison of evaluation methodologies, and the exploration of challenges and opportunities in research performance evaluation [7] [8]. The assessment of research performance is critical in determining the scholarly output and influence of researchers, research groups, and academic institutions. It entails a systematic and quantitative examination of research activities, publications, and citations in order to assess the quality, productivity, and impact of research efforts. The significance of research performance evaluation stems from its potential to give significant insights and indicators that drive academic institutions' decision-making processes, resource allocation, and strategic planning. Institutions can identify areas of strength, recognize researchers' achievements, and provide resources to support continuing greatness by evaluating research performance Furthermore, research performance assessment may be used to facilitate partnerships, obtain funding possibilities, and improve institutional reputation. The primary goals of research performance evaluation are to uncover research strengths and shortcomings, to promote accountability, to assist career advancement and tenure choices, and to facilitate evidence-based decision-making in research policy and funding distribution. Institutions may efficiently track progress, set goals, and optimize research outputs by analyzing

research performance, resulting in increased competitiveness and excellence in the academic environment.

## 2.2 Bibliometrics in Research Performance Evaluation

Bibliometrics is the quantitative evaluation of scientific publications and their citation patterns to assess the research performance of individuals, organizations, or nations. Bibliometric indicators such as publication productivity, citation impact, research cooperation, and funding have been widely used in assessing research success [9][10]. Several bibliometric databases, including Web of Science, Scopus, and Google Scholar, give extensive data on publications and citations, allowing researchers and institutions to assess their research success across multiple dimensions. However, the use of bibliometric indicators in research performance evaluation has been criticized for potential biases and limitations, such as an emphasis on quantity over quality, underrepresentation of certain research fields, and the possibility of citation pattern manipulation [7][11]. Traditional approaches to research performance assessment involve the use of established bibliometric indicators and metrics to evaluate the productivity, impact, and influence of research activities. These approaches focus on quantitative measures and often rely on data derived from scholarly publications and citation databases. Here is a detailed overview of traditional approaches to research performance assessment:

### 2.2.1 Publication Count

One of the fundamental metrics used in research performance assessment is the number of publications. It provides an indication of the quantity of research output by individuals, research groups, or institutions. The total number of publications is often considered a simple measure of research productivity [39].

### 2.2.2 Citation Count

Citation counts are widely used as a measure of research impact and visibility. They reflect the number of times a research article has been cited by other researchers. High citation counts suggest that a publication has made a significant contribution to the field and has been influential in subsequent research [26][40].

### 2.2.3 h-index

The h-index is a composite measure that combines publication quantity and citation impact. It measures both the productivity and impact of a researcher's work. An individual's h-index is the highest number h for which they have published at least h papers, each of which has received at least h citations [24].

### 2.2.4 Journal Impact Factor

Journal Impact Factor (JIF) is a metric used to assess the prestige and influence of academic journals. It measures the average number of citations received by articles published in a specific journal over a certain period. Researchers often aim to publish in journals with higher impact factors to enhance the perceived quality and visibility of their research [27].

### 2.2.5 Field-Weighted Citation Impact

Field-Weighted Citation Impact (FWCI) adjusts citation counts based on the average citation rates within a specific research field. It allows for more accurate comparisons of research impact across different disciplines, accounting for variations in citation practices among fields.

### 2.2.6 Collaboration Metrics

Collaboration metrics evaluate the extent and impact of research collaborations. They include measures such as co-authorship networks, which assess the level of collaboration among researchers or institutions. Collaborative research is often considered a positive indicator of research quality and innovation.

### 2.2.7 Grant Funding and Awards

Research performance can also be assessed by considering the acquisition of external research grants, fellowships, and prestigious awards. The ability to secure competitive funding and receive recognition through awards highlights the perceived excellence and impact of research work.

### 2.2.8 Peer Review

Although not strictly a bibliometric approach, peer review is a traditional method used to evaluate the quality and significance of research. Peer-reviewed publications are often considered more reliable and of higher quality compared to non-peer-reviewed sources. These traditional approaches to research performance assessment provide quantitative measures that aid in evaluating research productivity, impact, and visibility. While they offer valuable insights, it is important to acknowledge their limitations, such as potential biases, disciplinary variations, and the focus on quantitative indicators rather than qualitative aspects of research.

## 2.3. Bibliometrics in Research Performance Evaluation of Pakistani Universities

There has been a rise in interest in the use of bibliometrics to measure research performance at Pakistani universities in recent years. The use of bibliometric measures, such as publication output, citation impact, research cooperation, and financing, gives a quantitative way to assessing Pakistani higher education institutions' (HEIs') research productivity and impact. This section provides an overview of the literature on the application of bibliometrics in Pakistani institutions.

### 2.3.1 Publication Output and Citation Impact:

Several studies have evaluated the publishing output and citation effect of Pakistani institutions, focusing on various academic domains and disciplines [12][13]. These studies have provided valuable insights on the research productivity of Pakistani HEIs,

emphasizing the need to increase the quality and quantity of research output in order to boost their global competitiveness [14]. In addition, citation analysis has been used to investigate the citation patterns and impact of research produced by Pakistani institutions, indicating the impact of foreign collaboration and financing on citation impact [14][15].

### 2.3.2 Research Collaboration

Research collaboration is a crucial aspect of research performance, as it contributes to the sharing of knowledge, resources, and expertise. Several studies have investigated the research collaboration patterns of Pakistani universities, both at the national and international levels [13][14]. These studies have found that international collaboration is positively correlated with research productivity and citation impact, emphasizing the importance of fostering collaborative research networks for Pakistani HEIs. Research funding is an important factor in influencing university research success. In Pakistan, the Higher Education Commission (HEC) is a significant source of financing for R&D endeavors. Several studies have been conducted to investigate the link between research financing and research performance at Pakistani institutions, with findings indicating a favorable relationship between funding and publication output, citation impact, and research collaboration.

## 2.3. Challenges and Limitations:

Despite the increasing use of bibliometrics in assessing research progress in Pakistani universities, there are several issues and constraints to employing bibliometric indicators. These include possible biases in citation and cooperation patterns, underrepresentation of certain research topics, and bibliometric database constraints in capturing the complete spectrum of research outputs [11][12]. Furthermore, the specific setting of Pakistani higher education, with limited resources, weak facilities, and inequality in access to excellent education, provides additional hurdles for the use of bibliometrics in research performance evaluation.

*Table 2.1: Bibliometrics in Research Performance*

| Author & Year | Focus | Methods | Findings |
|---|---|---|---|
| Kausar & Mahmood [18] | Publication output | Bibliometric analysis | Revealed the growth of research output in Pakistani universities, with variations across disciplines. |
| Meho & Yang [20] | Citation impact | Citation analysis | Analyzed citation patterns and impact of research conducted by Pakistani universities. |
| Arshad & Ameen [21] | Research productivity | Bibliometric analysis, interviews | Highlighted the need for improving the quality and quantity of research output in Pakistani universities. |
| Yousafzai et al. [22] | Citation impact and collaboration | Bibliometric analysis | Confirmed the influence of international collaboration and funding on citation impact in Pakistani universities. |
| Khan & Fatima [22] | Research collaboration | Bibliometric analysis | Found that international collaboration is positively correlated with research productivity and citation impact. |
| Hicks et al [7] | Challenges and limitations | Review | Criticized the use of bibliometric indicators in research performance evaluation for potential biases and limitations. |
| Rafols et al [11] | Challenges and limitations | Review | Discussed the challenges and limitations. associated with the application of bibliometrics in research performance evaluation. |
| Charnes et al. [23] | Data envelopment analysis (DEA) | Non-parametric technique | Introduced DEA as a method for evaluating the relative efficiency of decision-making units, such as higher education institutions. |

Contribution of different authors for bibliometric research performance is presented in Table 2.1 Highlighted the focus of the study, methods and findings their research. These studies have contributed to the understanding of research performance in the context of higher education institutions, including Pakistani universities. However, further research is needed to address the unique challenges and opportunities in the Pakistani higher education sector and develop novel approaches for research performance evaluation.

## 2.4. Existing Models

Several models and methodologies have been developed in the literature to assess research performance using bibliometric indicators. Some of the widely used approaches are given in table 2.4 covers the distinction of the model/methodology, key metrices/indicators and the applications of different authors.

***Table 2.2: Research Performance using Bibliometric Indicators.***

| Model/Methodology | Description | Key Metrics/Indicators | Applications |
|---|---|---|---|
| **Citation Metrics [26]** | Measures the influence and impact of scholarly publications based on the number of citations received. | Total Citations, Citations per Document, H-index | Individual researcher evaluation, comparison of research impact |
| **Journal Impact Metrics [27]** | Evaluates the influence and prestige of academic journals. | Journal Impact Factor (JIF), Eigen factor Score | Journal ranking and comparison, assessment of journal quality |
| **Co-authorship Networks [28]** | Analyzes co-authorship patterns and networks to identify collaborative research efforts. | Co-authorship network visualization, degree of collaboration | Collaboration analysis, identification of research networks |
| **Institutional Collaboration Index [29]** | Measures the degree of collaboration between institutions based on shared authorship. | Degree of collaboration between institutions | Assessment of collaborative research efforts |
| **Science Mapping and Visualization [30]** | Utilizes bibliometric techniques to map the intellectual structure of a research field. | Co-citation analysis, co-authorship analysis, visualization techniques | Identification of research clusters, interdisciplinary collaborations |

These models and methodologies provide a framework for assessing research output, impact, collaboration, and visibility. They serve various purposes in research evaluation, including individual researcher evaluation, institutional assessment, funding allocation, and policy development.

### 2.5 Ranking Systems

THE World University Rankings assess universities globally based on teaching, research, citations, industry income, and international outlook, providing insights into institutional excellence and reputation [32]. QS ranks universities worldwide using indicators like academic reputation, employer reputation, faculty/student ratio, citations per faculty, and international faculty and student ratios, offering a comprehensive view of university performance [31].

ARWU ranks universities primarily based on research output, quality, and impact, focusing on factors like Nobel laureates, highly cited researchers, and publications in high-impact journals [33]. News ranks universities globally based on research performance, global research reputation, publications, and collaboration, aiming to help students make informed decisions [34]. Leiden Ranking focuses on scientific impact, measuring universities' contributions to scientific publications and their influence relative to their size and subject area [35].

Nature Index tracks high-quality research output published in natural sciences journals, showcasing institutions' contributions to scientific advancements [36]. Webometrics ranks universities based on their web presence, visibility, and impact, offering insights into their digital engagement and online influence [37].

The Academic Ranking of World Universities (Shanghai Ranking) evaluates universities worldwide based on research performance, alumni and staff winning Nobel Prizes and Fields Medals, and articles published in high-impact journals [38].

These studies of the ranking system are presented in table 2.3. Highlighting the Description, criteria considered and key features.

***Table 2.3: Ranking Systems***

| Ranking System | Description | Criteria Considered | Key Features |
|---|---|---|---|
| **Times Higher Education World University Rankings[32]** | Global university rankings based on various indicators. | Teaching, Research, Citations, Industry Income, International Diversity | Broad range of indicators, international scope |
| **QS World University Rankings [31]** | Comprehensive rankings of global universities. | Academic Reputation, Employer Reputation, Faculty/Student Ratio, Citations per Faculty, International Faculty/Student Ratio | Reputation-based indicators, faculty-to-student ratio |
| **Academic Ranking of World Universities (ARWU) [33]** | Ranking system focusing on research performance and productivity. | Publications, Citations, Highly Cited Researchers, Nobel Prizes and Fields Medals | Emphasis on research output and recognition |
| **U.S. News & World Report Best Global Universities Rankings [ 31]** | Rankings of global universities based on academic research performance. | Global Research Reputation, Regional Research Reputation, Publications, Normalized Citation Impact | Reputation-based indicators, citation impact |
| **Leiden Ranking [35]** | Ranking system based on bibliometric indicators. | Publications, Impact, Collaboration | Emphasis on publication output and impact |
| **Nature Index [36]** | Ranks institutions based on their contribution to high-quality scientific research. | Article Count, Fractional Count, Share | Focus on high-quality scientific output |

## 2.6. Composite Indicators

Multiple bibliometric indices, such as the H-index or the g-index, are combined into a single composite index to provide a summary evaluation of research performance [24][25]. In these models, many bibliometric variables are combined into a single composite index to provide a summary assessment of research performance. Here are a few examples:

### 2.6.1. H-index

The h-index is a bibliometric metric that takes into account a researcher's research output (number of publications) as well as impact (number of citations). The h-index is the

highest value of h for which a researcher has at least h papers that have each been referenced at least h times [25]. To compute the h-index, perform the following steps: Compile a list of publications for a researcher, research group, institution, or nation, as well as their citation counts. This data comes from sources such as Web of Science, Scopus, and Google Scholar. Where "rank" refers to the position of a publication in the sorted list, and "citations" represents the number of citations received by the publication. It is important to note that the h-index has some limitations, such as its insensitivity to highly cited publications and its dependence on research field and career stage. Therefore, it is essential to use the h-index in conjunction with other bibliometric indicators for a more comprehensive assessment of research performance. The h-index is defined as the maximum value of h for which a researcher has at least h publications that have each been cited at least h times. The h-index can be represented as:

$$h - index = max(h) \text{ where } h \leq C(h) \quad (2.1)$$

Where: h-index is the value, we want to calculate. Ax(h) represents the maximum value of h that satisfies the condition(h) is the number of citations for the h-th publication when publications are sorted in descending order by the number of citations.

### 2.6.2. G-index

The g-index is a bibliometric measure that, like the h-index, combines an individual's research output (number of publications) with impact (number of citations). The g-index, on the other hand, gives greater weight to publications that have been widely referenced. The g-index is defined as the greatest value of g at which the top g articles have garnered at least g2 citations combined. This data comes from sources such as Web of Science, Scopus, and Google Scholar [25]. The g-index may be calculated mathematically as follows: Mathematically, the g-index can be defined as:

$$g - index = max(g) \; where \; g \leq Cumulative \; Citations(g) \quad (2.2)$$

Where g-index is the value, we want to calculate. max(g) represents the maximum value of g that satisfies the condition. Cumulative Citations(g) is the cumulative citation count for the g-th publication when publications are sorted in descending order by the number of citations. Findings and limitation of g-index are given in table 2.4.

**Table 2.4 Composite Indicator [25]**

| | Findings | Limitations |
|---|---|---|
| **Impact Factor** | Higher impact factor indicates higher perceived prestige and influence of a journal. | Focuses solely on journal-level impact, not individual articles. Can be influenced by self-citations and citation practices. |
| **h-index [24]** | Measures an author's productivity and impact based on their most highly cited papers. | Does not consider the distribution of citations across papers. Can be influenced by self-citations. |
| **Field-Weighted Citation Impact (FWCI)[ 41]** | Measures the citation impact of articles relative to the average impact in the field. | Limited to articles indexed in specific databases. May not capture the full impact of interdisciplinary research. |

These composite indicators offer valuable insights into research performance, impact, collaboration, and funding success. However, they also have limitations that should be considered when interpreting their findings. It is important to use these indicators as part of a comprehensive evaluation process, considering disciplinary variations, qualitative aspects of research, and contextual factors.

### 2.6.3. Benchmarking

Benchmarking approaches involve the comparison of an institution's research performance against a set of peers or best practices, using a range of bibliometric indicators [8]. These models involve comparing an institution's research performance against a set.

## 2.7 Benchmarking Approaches

Benchmarking provides valuable insights into research performance, identifies areas for improvement, and facilitates knowledge sharing. However, it also has certain limitations.

That should be considered when interpreting its findings. Additionally, benchmarking should be used as part of a comprehensive evaluation process that considers disciplinary variations, qualitative aspects of research, and specific institutional goals. The performance approaches are present in Table 2.5.

***Table 2.5 Benchmarking Approaches***

| Benchmarking | Findings | Limitations |
|---|---|---|
| **Comparative Institutional Analysis [42]** | Allows institutions to compare their research performance with peer institutions. | Difficulty in identifying appropriate peer institutions for comparison.<br><br>Differences in institutional size, focus, and resources can affect comparability. |
| **Subject-specific Benchmarking [43]** | Evaluates research performance within specific disciplines or subject areas. | Variations in publication and citation practices across disciplines. |

## 2.8 Data Envelopment Analysis (DEA):

DEA is a non-parametric technique that uses linear programming to evaluate the relative efficiency of decision-making units, such as higher education institutions, based on multiple input and output indicators [23]. This diagram is a simplified representation of the various bibliometric models, and there may be additional models and variations not included in this illustration.

Data Envelopment Analysis (DEA) is a flexible and widely used method for assessing the efficiency and productivity of decision-making units. It considers multiple inputs and outputs and provides efficiency scores, benchmarking information, and insights into resource allocation and performance improvement.

The choice between input and output orientation, as well as the consideration of returns to scale, enables a comprehensive analysis of efficiency. Additionally, DEA can be used to compare performance over time and identify sources of productivity change using the Malmquist Productivity Index.

## 2.9 Stochastic Models in Research Assessment

Data Envelopment Analysis (DEA) is a flexible and widely used method for assessing the efficiency and productivity of decision-making units [44]. It takes into account a variety of inputs and outputs and delivers efficiency ratings, benchmarking data, and insights into resource allocation and performance improvement. The ability to analyze efficiency in depth is enabled by the option between input and output direction, as well as the consideration of returns to scale. DEA may also be used to monitor performance over time and identify productivity change factors using the Malmquist Productivity Index [45].

The use of stochastic models in research performance evaluation has grown in popularity in recent years. By accounting for the inherent uncertainties and dynamics in bibliometric data, these models provide a more robust and flexible way to measuring research productivity and effect. Adoption of stochastic models in Pakistani higher education institutions (HEIs) has the potential to deliver more accurate and context-sensitive insights

regarding research success. The purpose of this literature review is to investigate the present level of knowledge on stochastic models in research evaluation, with a special emphasis on their use in Pakistani institutes. Some of the key stochastic models used in research assessment include:

### 2.9.1 Stochastic Block Models (SBMs)

Stochastic block models are a class of probabilistic models that partition a network into blocks, where the probability of connections between nodes (e.g., authors, institutions, or research articles) depends on their block memberships. SBMs have been used to analyze

collaboration networks and to detect communities or clusters of researchers with similar research interests [46]. Stochastic Block Models (SBMs) offer valuable insights into community detection, network generation, link prediction, and network comparison.

However, they also have certain limitations that should be considered when interpreting their findings. These limitations include challenges in handling complex networks with overlapping communities, sensitivity to parameter choices, difficulties in capturing dynamics beyond the block structure, and limitations in link prediction accuracy. Awareness of these limitations is crucial for the appropriate application and interpretation of SBMs in network analysis. The Stochastic Block Model (SBM) is a generative probabilistic model for representing the structure of networks, where the probability of an edge existing between two nodes depends on the blocks (groups) to which the nodes belong. Here is the mathematical representation of the SBM:

$$L\,(G|B,P) = \prod_{(u,v)\in E} P\,(u,v) * \prod_{(u,v)\notin E}(1 - P\,(u,v)) \qquad (2.3)$$

Let $G = (V,E)$ be an undirected graph, where V is the set of nodes, and E is the set of edges. Partition the nodes V into K disjoint blocks (groups) represented by $B_1, B_2, \ldots, B_k$. Each node belongs to one and only one block. Define a $KxK$ symmetric matrix P, where P (i, j) represents the probability of an edge existing between nodes belonging to block i and block j. The diagonal elements of the matrix, $P\,(i,i)$, represent the probability of an edge within the same block i. For each pair of nodes $u, v \in V$, an edge $(u,v) \in E$ exists with probability given by the corresponding entry in the connection. Probability matrix P, i.e., $P\,(u,v) = P\,(i,j)$ if $u \in B_i$ and $v \in B_j$. The goal of fitting the SBM to a given network is to find the optimal partitioning of nodes into blocks and the connection probability matrix P that best explains the observed network data. This can be achieved using various statistical inference techniques, such as maximum likelihood estimation, Bayesian inference, or spectral clustering. The maximum likelihood estimation involves finding the partitioning and connection probabilities that maximize the likelihood of the observed network data given the model.

The Stochastic Block Model provides a mathematical framework for representing and analyzing the structure of networks, such as co-authorship or citation networks, in the context of research performance evaluation. By modeling the probability of edge formation based on node group membership, the SBM enables the identification of research communities and the quantification of collaboration patterns among researchers or institutions.

***Table 1.6: Stochastic Block Models (SBMs)***

| Stochastic Block Models (SBMs)[46] | Findings | Limitations |
|---|---|---|
| **Community Detection [47]** | SBMs are effective in identifying communities or clusters within a network. | - Assumption of well-defined communities may not hold in complex networks with overlapping or ambiguous community structures.<br>- Difficulty in handling large-scale networks due to computational complexity. |
| **Network Generation [48]** | SBMs can generate synthetic networks that mimic the observed network's characteristics. | - Assumes that the network can be accurately represented by a block structure, which may not always be the case.<br>- Replication of observed characteristics does not guarantee an accurate representation of the underlying generative process. |



### 2.9.2 Stochastic Citation Models

Stochastic citation models are probabilistic models that describe the citation process as a random mechanism, with the likelihood of a paper receiving citations depending on various factors, such as its age, quality, and visibility [49]. Some notable stochastic citation models include the Price model [50], the Yule-Simon model [51], and the Barabási-Albert model [52]. Here is a mathematical description of Price's Model:

$$P(i) = \frac{c_i}{\sum_j c_i} \qquad (2.4)$$

Consider a set of papers, where each paper i has a citation count c_i. At each time step t, a new citation is added to the system. The new citation is assigned to a paper i with

probability: where the sum is taken over all papers j in the system. Repeat steps 2 and 3 to grow the citation counts over time. Price's Model results in a power-law distribution of citation counts:

$$P(c) \propto c^{-\alpha} \qquad (2.5)$$

where P(c) is the probability of a paper having c citations, and α is a constant exponent that typically falls between 2 and 3 for real-world citation networks.

***Table 2.7 Stochastic Citation Models (SCMs), Findings, Limitations***

| Stochastic Citation Models (SCMs) | Findings | Limitations |
|---|---|---|
| Citation Dynamics [26] | SCMs provide insights into the temporal evolution of citation networks, capturing the growth, decay, and patterns of citations over time. | Simplified assumptions about citation behavior may not fully capture the complexity of real-world citation dynamics. |
| Network Generation [48] | SCMs can generate synthetic citation networks that replicate observed statistical properties, such as degree distribution, clustering, and community structure. | Assumptions made in the model may not fully capture the underlying mechanisms that drive citation network formation. |
| Impact Prediction [53] | SCMs allow for the estimation of future citation impact based on past citation patterns and network structure. | Predictive accuracy is affected by the assumptions and limitations of the model used. |

Stochastic Citation Models (SCMs) offer valuable insights into citation dynamics, network generation, impact prediction, and policy evaluation. However, they also have certain limitations that should be considered when interpreting their findings. These limitations include simplifying assumptions, challenges in accurately capturing real-world citation dynamics, difficulties in reproducing all observed network properties, sensitivity to model parameters, and the inability to fully account for external factors and non-citation indicators. Awareness of these limitations is crucial for the appropriate application and interpretation of SCMs in citation analysis. Another important class of

stochastic citation models is the fitness-based models, which incorporate an intrinsic quality or fitness factor for each paper.

In these models, the probability of a paper receiving a new citation depends on both its current citation count and its fitness. A mathematical description of a fitness-based model is as follows: Consider a set of papers, where each paper i has a citation count c_i and a fitness value f_i. At each time step t, a new citation is added to the system. The new citation is assigned to a paper i with probability: where the sum is taken over all papers j in the system. Repeat steps 2 and 3 to grow the citation counts over time.

$$P(i) = \frac{f_i * c_i}{\Sigma_j (f_i * c_j)} \qquad (2.6)$$

In fitness-based models, papers with higher fitness values are more likely to receive citations, even if they have fewer citations than other papers. This allows for the possibility of papers with high intrinsic quality to accumulate citations faster and eventually surpass papers with lower quality, leading to a more diverse and dynamic citation distribution.

### 2.9.3 Stochastic Frontier Analysis (SFA)

Finally, stochastic citation models provide a theoretical framework for investigating the growth and structure of citation networks, as well as the factors that influence citation patterns [54]. By modeling the citation accumulation process as a probabilistic mechanism, SCMs give insights into the dynamics of research performance, which can impact the creation of more robust and accurate citation-based assessments.

$$Y = f(X, \beta) * e^{(\varepsilon - u)} \qquad (2.7)$$

where: Y represents the observed output variable. X is a vector of input variables. β is a vector of unknown parameters $(X, \beta)$ is the deterministic production function, which represents the best-practice production frontier's is a random error term assumed to be normally distributed with mean zero represents the inefficiency component, which captures the deviations from the frontier due to factors such as managerial inefficiency or measurement error.

***Table 2.8 Stochastic Frontier Analysis***

| **Stochastic Frontier Analysis (SFA) [54]** | **Findings** | **Limitations** |
|---|---|---|
| **Efficiency Measurement [55]** | SFA allows for the estimation of technical efficiency by comparing observed production levels to the maximum achievable production levels given the inputs and technology. | Requires assumptions about the functional form of the production frontier, which may not accurately represent the underlying production process.<br><br>Sensitivity to model specification and choice of distributional assumptions.<br><br>Difficulty in disentangling inefficiency due to managerial factors from external factors beyond the control of the firm. |
| **Productivity Analysis [56]** | SFA provides insights into the sources of productivity differences across firms or units by decomposing total factor productivity into technical efficiency and technological change components. | Assumes a specific production technology and may not account for variations in technology across firms.<br><br>Reliance on data quality and accuracy for precise estimation of productivity components.<br><br>Limited ability to capture and measure non-technological factors influencing productivity, such as market conditions or strategic choices. |
| **Input Allocation and Resource Management [57]** | SFA helps identify optimal input allocation strategies by assessing the efficiency of resource utilization. It provides guidance on resource allocation decisions for improved performance. | Assumes that the input-output relationship is known and accurately represented by the production frontier.<br><br>Ignores variations in factor prices, market imperfections, and other external factors that influence resource allocation decisions.<br><br>Limited ability to account for qualitative aspects of resource management, such as innovation or strategic investments. |
| **Policy Evaluation[58]** | SFA can be used to evaluate the impact of policy interventions or managerial | Difficulty in establishing causal relationships between policies and |

| | | |
|---|---|---|
| | practices on firm efficiency and productivity. | efficiency improvements due to endogeneity and potential reverse causality.<br><br>Limited ability to capture the full range of policy effects beyond the efficiency dimension. |

SFA can help with efficiency measurement, productivity analysis, input allocation, and policy evaluation. It does, however, have certain limitations that should be considered before drawing conclusions from it. These constraints include assumptions about the production frontier, model sensitivity, difficulties disentangling various factors affecting efficiency, reliance on data quality and accuracy, and difficulties capturing non-technological factors and establishing causal relationships in policy evaluation. Understanding these constraints is critical for the proper implementation and interpretation of SFA in efficiency and productivity analyses.

### 2.9.4 Stochastic Actor-Oriented Models (SAOMs) [59]

Stochastic actor-oriented models are dynamic network models that describe the evolution of a network (e.g., a collaboration network) because of individual actors' decisions and interactions. SAOMs have been used to analyze the co-authorship networks of researchers and to study the factors influencing collaboration patterns and research productivity [59].

***Table2.9: Stochastic Actor-Oriented Models (SAOMs)***

| Stochastic Actor-Oriented Models (SAOMs) [59] | Findings | Limitations |
|---|---|---|
| **Network Dynamics [60]** | SAOMs provide insights into the dynamic processes of network formation, evolution, and adaptation. They capture how individual actors' behavior and network attributes interact over time. | Reliance on assumptions about individual decision-making processes, which may not fully capture real-world complexities.<br>Difficulty in accurately capturing and modeling the full range of mechanisms driving network dynamics.<br>- Limited ability to incorporate unobserved or latent variables that may influence actor behavior. |
| **Actor Heterogeneity** | SAOMs can account for actor heterogeneity by modeling individual-level characteristics | Challenges in accurately specifying and measuring individual-level characteristics |

| | | |
|---|---|---|
| [61] | and their influence on network dynamics and behavior. They capture variations in individual propensities and preferences. | and their impact on network dynamics.<br>Difficulty in capturing and modeling the interplay between individual attributes, behavior, and network structures.<br>- Potential for omitted variable bias and unobserved heterogeneity. |

## 2.10 Application of Stochastic Models in Pakistani HEIs

The application of stochastic models in research assessment of Pakistani institutes is an emerging area of research. Some notable studies that have employed stochastic models in the context of Pakistani HEIs include applied stochastic frontier analysis to assess the research efficiency of Pakistani universities. They found that larger universities with greater research funding and international collaboration tend to have higher research efficiency. The study also highlighted the importance of government support and research infrastructure in promoting research productivity and impact. Siddiqi et al. Stochastic block models used to analyze the research collaboration networks of Pakistani universities [62].

| Study | Stochastic Model | Application in Pakistani HEIs |
|---|---|---|
| Airoldi et al. [46] | Stochastic Block Models (SBMs) | Analyzing collaboration networks and detecting communities of researchers with similar research interests. |
| de Solla Price [50] | Price Model | Describing the citation process as a random mechanism, with likelihood of a paper receiving citations depending on various factors. |
| Simon [51] | Yule-Simon Model | Analyzing citation patterns and understanding the probability of citations for research publications. |
| Barabási & Albert [52] | Barabási-Albert Model | Investigating preferential attachment and network growth in citation networks. |
| Worthington & Lee [54] | Stochastic Frontier Analysis (SFA) | Assessing research efficiency of HEIs, considering research inputs, outputs, and environmental variables. |
| Snijders et al [59] | Stochastic Actor-Oriented Models (SAOMs) | Analyzing co-authorship networks and factors influencing collaboration patterns and research productivity. |
| Khattak et al. [54] | Stochastic Frontier Analysis (SFA) | Assessing research efficiency of Pakistani universities, accounting for size, funding, and international collaboration. |

**Table 2.10: Application of Stochastic Models**

The table presents application of stochastic models in research assessment, covering various models and their applications, such as analyzing collaboration networks, citation patterns, and research efficiency. While the application of stochastic models in Pakistani HEIs is an emerging area of research, existing studies have demonstrated the potential of these models to provide more accurate and context-sensitive insights into research performance in this context.

## 2.11. Gaps and Opportunities

The literature on research performance assessment using bibliometric indicators has extensively covered various models and methodologies. However, there is a noticeable gap in the application of learning-based stochastic models specifically tailored for Pakistani degree-awarding institutes. Most of the existing research performance assessment methodologies focus on deterministic models and aggregated metrics, which may not capture the inherent uncertainties and dynamics in bibliometric data. Moreover, these methodologies lack adaptability to the evolving nature of research fields and the unique challenges faced by Pakistani higher education institutions.

The opportunity lies in the development and application of a novel learning-based bibliometric research performance assessment model using stochastic models for Pakistani degree-awarding institutes. By addressing the limitations of existing methodologies, the proposed model can offer a more accurate, adaptable, and robust framework for evaluating research performance in the Pakistani context. The integration of learning-based techniques can enable the model to adapt to the evolving nature of research fields, as well as account for the unique challenges and opportunities faced by Pakistani higher education institutions. In addition, the proposed model can contribute to the literature on research performance assessment by providing valuable insights into the application of stochastic models in the context of Pakistani higher education. The

development of a tailored assessment model for Pakistan can also support stakeholders in the higher education sector, such as policymakers, funding agencies, and institutional administrators, in making informed decisions about resource allocation, policy development, and the promotion of research excellence.

***Table 2.11: Research Performance Assessment Gaps***

| **Research Performance Assessment using Bibliometric Indicators [63]** | **Gaps** | **Opportunities** |
|---|---|---|
| **Quantitative Assessment [64]** | Overreliance on traditional bibliometric indicators like citation counts, which may not capture the full impact or quality of research. | Development of new bibliometric indicators that go beyond citation counts and incorporate alternative metrics and qualitative assessments. |
| | Difficulty in accounting for variations in research disciplines and methodologies. | Incorporation of diverse research outputs and impacts, including non-traditional outputs and contributions to society. |
| | Limited consideration of non-traditional outputs and impacts, such as datasets, software, patents, and societal contributions. | Customization of indicators based on research disciplines and methodologies to ensure fair and meaningful comparisons. |
| **Bias and Limitations [65]** | Lack of standardization in bibliometric data collection and processing, leading to inconsistent and unreliable results. | Development of transparent and standardized data collection and processing protocols for bibliometric data. |
| | Inherent biases in citation practices, such as language bias, self-citation bias, and disciplinary differences. | Advances in methods to address biases in citation practices and account for interdisciplinary research. |
| | Inability to fully capture interdisciplinary research and collaborations. | Integration of diverse data sources, including non-traditional sources like social media and altimetric, to provide a more comprehensive assessment. |
| **Contextualization and Interpretation [66]** | Limited ability to account for research context, such as collaboration patterns, funding sources, and institutional differences. | Development of contextualized bibliometric indicators that consider collaboration patterns, funding sources, and institutional characteristics. |
| | Difficulty in distinguishing between high-impact research and research with high citation rates due to factors like self-citations or topic popularity. | Integration of qualitative assessments and peer review to complement bibliometric indicators and provide a more holistic evaluation. |
| | Challenges in interpreting bibliometric indicators in relation to other forms of | Increased emphasis on responsible use and interpretation of bibliometric |

| | |
|---|---|
| assessment, such as peer review or societal impact. | indicators, considering their limitations and biases. |

The assessment of research performance using bibliometric indicators has several limits and areas for improvement. Addressing quantitative assessment gaps, bias and limitations, contextualization and interpretation, and contextualization and interpretation can lead to more rigorous and meaningful evaluation practices. There are several possibilities for developing new indicators, standardizing data collection, addressing biases, adding non-traditional outputs, contextualizing assessments, and integrating qualitative evaluations. By taking use of these opportunities, research evaluation may become more complete, fair, and representational of research's numerous contributions and ramification.

**Summary:**

The challenges and research questions presented shed light on the intricate landscape of bibliometric analysis and its limitations when evaluating research impact and quality. Traditional bibliometric indicators, such as citation counts, often fall short in capturing the nuanced dimensions of scholarly contributions. They may not fully account for the quality and societal impact of research outputs. Additionally, these indicators may disregard interdisciplinary collaborations, hinder the assessment of diverse research methodologies, and overlook non-traditional outputs like datasets, software, and patents.

The first research question, which delves into the employment of learning-based models to identify Rising Stars in Dynamic Collaborative Networks among Pakistani degree awarding institutes, directly addresses the challenge of overreliance on conventional metrics. By seeking innovative methods, this question aligns with the broader challenge of capturing impactful research beyond traditional boundaries, while also acknowledging the variations in research practices.

The second research question explores the appropriateness of author-level bibliometric indicators across disciplines and academic seniorities. This question is particularly relevant in light of the challenge posed by variations in research disciplines and

methodologies. It underscores the complexities of evaluating researchers with diverse backgrounds and acknowledges the limitations of generic evaluation metrics.

Lastly, the third research question examines models for predicting Citation-Based Impact Measures using learning-based techniques. This aligns with the challenges related to the inherent biases in citation practices and the overemphasis on citation counts. The research question suggests a forward-looking approach that integrates machine learning to potentially mitigate biases and enhance the assessment of research impact.

In summary, the research questions mirror the recognized challenges of bibliometric analysis. They seek to explore alternatives, such as learning-based models, interdisciplinary evaluations, and context-aware assessments. By addressing these challenges head-on, the questions reflect a commitment to advancing research evaluation methodologies and embracing a more comprehensive understanding of scholarly contributions.

# Chapter 3 Proposed Methodology

1. **Existing Literature**

   The existing literature serves as the foundation for any scholarly endeavor. In this study, a comprehensive review of existing literature in the field of learning-based bibliometric research assessment will be conducted. This review will encompass studies that have investigated novel methodologies for assessing research impact, evaluating scholars, and understanding scholarly networks. By analyzing the existing body of work, the study will identify gaps, trends, and key insights that contribute to the advancement of learning-based bibliometric assessment.

2. **Start**

   The study begins with a clear objective to enhance the understanding of research impact assessment through learning-based bibliometric methodologies. By embracing innovative approaches, the study aims to overcome the limitations of traditional metrics and provide a more comprehensive view of scholarly contributions.

3. **Determining Keywords**

   To initiate the research, a strategic approach to determining keywords will be undertaken. This involves employing various search strategies, including author searches, subject categories, and affiliation-based searches. By identifying relevant keywords, the study ensures the inclusivity of pertinent publications within the research scope.

4. **Determine Data Source of Publications (Scopus)**

   The selection of an appropriate data source is crucial for reliable research outcomes. In this study, Scopus will be chosen as the primary data source for

publications. Scopus offers a comprehensive repository of scholarly articles, conference papers, and other academic materials, making it a suitable choice to capture a wide spectrum of learning-based bibliometric research.

5. **Examine Trends of Learning-Based Bibliometric Research Assessment Using Stochastic Model**

   The study aims to explore the trends in learning-based bibliometric research assessment by employing stochastic models. Stochastic models offer a probabilistic framework to understand uncertain and dynamic systems. This approach allows for a nuanced analysis of bibliometric trends, considering factors such as author collaboration patterns, citation dynamics, and the evolution of impact over time.

6. **Chapter 4 Network Analysis (Co-author, Power Graph, Impact Graph)**

   In Chapter 4, the study delves into network analysis techniques. Co-authorship networks provide insights into collaboration patterns among researchers. Power graphs reveal influential nodes in the network, and impact graphs unveil the relationships between scholarly impact and collaboration. These analyses contribute to a holistic understanding of the research ecosystem.

7. **Chapter 5 Multivariate Stochastic Models Using Dominance Orders**

   Chapter 5 focuses on multivariate stochastic models employing dominance orders. This approach enables the integration of multiple variables to assess research impact. By establishing dominance relationships, the study enriches the evaluation process, accommodating diverse factors that influence scholarly contributions.

8. **Chapter 6 Learning-Based Models (Author Age, Paper Age, Author Impact, Paper Impact)**

   Chapter 6 introduces learning-based models that consider author age, paper age, author impact, and paper impact. These models leverage machine learning techniques to predict research impact beyond traditional metrics. By incorporating these dynamic elements, the study enhances the accuracy of impact assessment.

9. **Overlay Visualization of Analysis for Results of Selected Publications**

   The study employs overlay visualization techniques to present the analysis results of selected publications. Overlay visualization provides a graphical representation that allows researchers to discern patterns, trends, and relationships in the data. This approach enhances the accessibility and interpretability of complex bibliometric analyses.

10. **Discussion**

    The discussion section critically engages with the findings of the study. It contextualizes the results within the existing literature, identifies implications for research assessment, and addresses any discrepancies or novel insights that emerged during the analysis. The discussion contributes to a deeper understanding of the study's contributions to the field.

11. **End**

    The conclusion marks the culmination of the study's journey. It succinctly summarizes the key findings, discusses their significance in the broader context of research assessment, and offers avenues for future research. The conclusion underscores the importance of learning-based bibliometric approaches in advancing the accuracy and comprehensiveness of research impact evaluation.

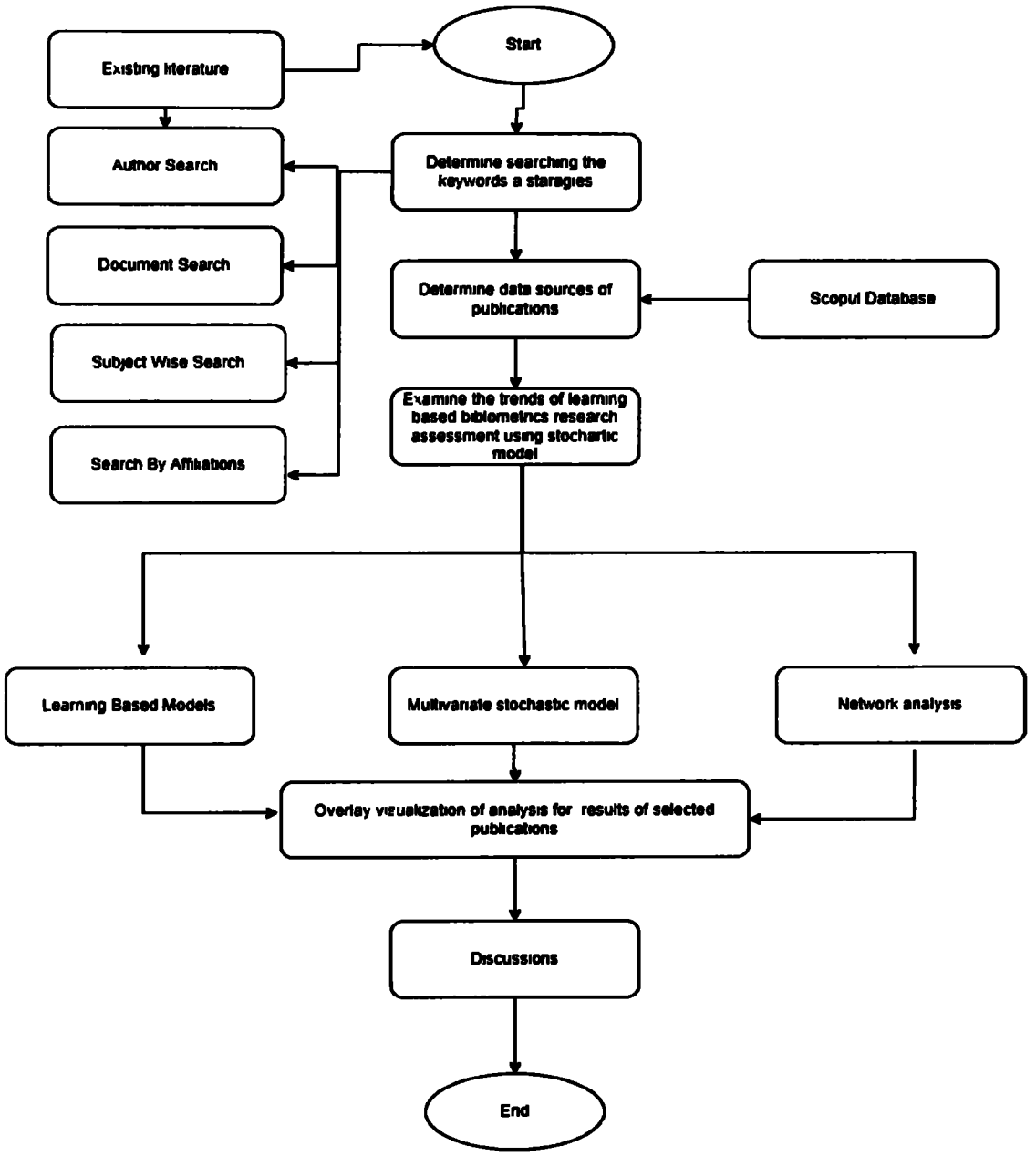



*Figure 3. 1: Proposed Methodology*

# Chapter 4 Detection of Rising Stars using Co-author, Power graph and Datamining Techniques

## 4.1 Introduction:

In recent years, the academic environment in Pakistan has seen tremendous expansion, with an expanding number of degree awarding institutes (DAIs) striving for excellence in research and innovation. Identifying and developing emerging stars, or prospective researchers with great potential for future success, is critical for these institutes to establish a flourishing research environment. By evaluating time-evolving co-authorship graphs and power graphs, this work proposes a thorough data mining technique for identifying rising stars in Pakistani DAIs.

Co-authorship graphs provide a valuable source of information on the patterns of collaboration among academics, reflecting the dynamics of research communities and the structure of academic networks [28]. In contrast, power graphs depict the hierarchical relationships between researchers based on their influence and production. We may acquire significant insights about academics' performance, cooperation habits, and network influence by examining these networks, which can then be utilized to identify prospective rising stars.

Our suggested methodology combines data mining techniques such as clustering algorithms, community identification methods [47], centrality measures, machine learning algorithms, and temporal analysis techniques [82]. These tools are used in a multi-stage procedure to analyze time-evolving co-authorship graphs and power graphs, which capture the dynamics of research cooperation and the development of emerging stars in Pakistan's DAIs. To begin, we gather publishing data from a variety of sources, including the Higher Education Commission (HEC) of Pakistan and foreign databases, in order to create time-evolving co-authorship graphs. We then use clustering algorithms

and community identification approaches to group academics that collaborate in similar ways, exposing the underlying structure of research communities.

Then, in the co-authorship and power graphs, we construct centrality measures for each researcher to identify prominent researchers based on their location and connectedness within the network. High centrality ratings suggest a major involvement in the network, which might be linked to emerging stars.

We use machine learning methods to forecast the potential of researchers based on previous performance data and network factors. Support vector machines, decision trees, and neural networks, for example, are trained on a labelled dataset of known rising stars and non-rising stars to categories researchers based on key performance indicators (KPIs) and network properties. The evolution of co-authorship and power graphs across time is studied using temporal analytic approaches. We may get insights into the dynamics of the research environment and the rise of emerging stars in Pakistan's DAIs by evaluating changes in cooperation patterns, research communities, and researchers' importance over time.

The validation and sensitivity studies indicate that our technique is reliable and robust, giving us confidence in our findings and their implications for spotting rising stars in Pakistani DAIs. The combination of these data mining approaches enables a full examination of researchers' performance, cooperation patterns, and network influence, providing significant insights for identifying and developing talented researchers in academic institutions. The study provides a thorough examination of relevant works concentrating on identifying emerging stars in academia, network analysis of co-authorship graphs, and the implementation of data mining techniques in the academic arena. This review is divided into the areas listed below:(1) Rising Stars in Academia, (2) Co-authorship Networks, (3) Power Graphs, (4) Data Mining Techniques in Academia, and (5) Identification of Rising Stars using Data Mining. Researchers with outstanding potential for future success and a substantial effect on their respective subjects are referred to be rising stars in academia. It is critical to identify and nurture these talented

individuals in order to stimulate research and innovation in academic institutions. This section examines rising stars' qualities, factors influencing their success, and approaches used to detect them in the academic arena. In conclusion, understanding the characteristics of rising stars, the factors affecting their performance, and the methodologies used to identify them is crucial for fostering research and innovation in academic institutions. By recognizing and supporting these talented individuals, institutions can cultivate a thriving research environment and contribute significantly to the advancement of knowledge in their respective fields.

### 4.2 Characteristics of Rising Stars

Rising stars often possess a unique combination of traits and abilities that set them apart from their peers. Some common characteristics of rising stars include [87]:

1. **High Productivity:** Rising stars tend to publish frequently, contributing significantly to the body of knowledge in their fields [88].

2. **High Impact:** Their publications often receive more citations than their peers, indicating their work is influential and recognized by other researchers [89].

3. **Collaboration:** Rising stars tend to have extensive collaboration networks, working with researchers from various institutions and countries [28].

4. **Interdisciplinarity:** They are often engaged in interdisciplinary research, bridging gaps between different fields and fostering innovation [84].

5. **Funding success:** Rising stars are more likely to secure research funding, reflecting their ability to develop compelling research proposals and their potential for significant contributions to their fields.

6. **Factors Affecting Rising Stars' Performance:** Various factors can influence the performance and success of rising stars in academia, including [87].

7. **Institutional Support:** Access to resources, mentorship, and a supportive research environment can significantly impact the success of rising stars [85].

8. **Networking:** Building strong professional networks can lead to increased collaboration opportunities, better access to resources, and higher visibility in the academic community [87][88].

9. **Time Management:** Balancing research, teaching, and administrative responsibilities can be challenging for rising stars, and effective time management is essential for maintaining productivity [87]. The following notation for the periods that we study since we process the cooperation graphs at different points in time to examine the evolution of an author through time: $t_0$ describes the period that the author published a research work for the first time. $t_n$ the period we examine, which must be after $t_0$. $t_{end}$ the most recent period in the publication database.

10. **Career Stage:** The career stage of a researcher can influence their potential as a rising star. Early-career researchers may have a higher potential for future success, while established researchers may already have a track record of high impact [89].

## 4.3. Methodologies for Identifying Rising Stars

Various methodologies have been proposed to identify rising stars in academia. Some of these methods include [87][88][89].

1. **Bibliometric Indicators:**

Researchers often use publication output, citation impact, and other bibliometric indicators to assess the performance and potential of rising stars [1][7][9].

### 2. Network Analysis:

Co-authorship networks and centrality measures can provide insights into researchers' collaboration patterns and their influence within the academic community, which can be used to identify potential rising stars [60].

### 3. Machine Learning:

Supervised machine learning algorithms can be trained on historical performance data and network properties to predict the future success of researchers, helping identify potential rising stars [90].

***Table 4.1: Methodologies for Identifying Rising Stars***

| # | Author & year | Key Finding | Method | Limitations |
|---|---|---|---|---|
| 1 | *Cameronet al.[91]* | Role of individual factors in predicting scientific impact | Statistical analysis | May not capture complex network dynamics |
| 2 | *Radicchi et al.[92]* | Universal impact factor for identifying rising stars across fields | Citation analysis | May not account for qualitative factors |
| 3 | *Liu et al.[93]* | Deep learning-based approach for predicting impact in computer science | Deep learning, network analysis | Limited to the field of computer science |
| 4 | *Ali Daud et al [94]* | Identifying rising stars using collaboration and citation networks | Network analysis, statistical analysis | May not capture discipline-specific dynamics |
| 5 | *Calra et al [95]* | Predictors of early-career research success in Italian academia | Regression models, bibliometric | Limited to Italian academia |

### i. Co-authorship Networks

Co-authorship networks are a useful technique for analyzing research cooperation trends and academic community structure. These networks depict researchers as nodes and their partnerships as edges, allowing them to provide insights into the dynamics of research communities through time. This section examines the features, analysis tools, and academic uses of co-authorship networks. The following is a list of research and their

significant conclusions about co-authorship networks. These studies investigate the structure, development, and function of individual and environmental elements in co-authorship networks.

*cit* ($i$, $t_k$): Citations that paper $i$ received at period $t_k$.

*aut*($i$) defines Number of authors of publication $i$.

While *per*($i$)=The period of publication for article $i$.

The generic co-authorship graph of a given period (e.g., year) tn is formally represented as: QGtn = (V, WE), where V is the collection of authors and a weighted edge, we = v1, 2, w1,2 WE signify that author 1 and 2 co-authored w1,2 works from t0 to tn. The weight includes the aggregated information from all articles published within that time period, whether it be just the number of publications or the cumulative number of citations, or any other time penalized cumulative score. More specifically, in the Quantity Graph, we define a quantity edge weight W Equan as the number of articles co-authored by authors x and y in a certain period 1. On the other side, in the influence Graph, we define an edge weight W Imp as the influence of two authors' collaboration 3. More specifically, when we generate the Quantity Graph for a given time period tn, we collect all articles written by the same author (or co-written by a pair of writers).

$$WE_{num}\,(a, b) = \sum_{\forall\, i \in Pa \cap Pb} \cdot\ 1 \qquad (4.1)$$

In general, a co-authorship graph built at a specific point in time can contain information from the past up to that point; however, traditional such graphs provide only a snapshot of publications while ignoring useful information about past changes in the graph (e.g., the point in time when an edge was first created, or when an author appeared in the graph).

To circumvent this limitation, we adopt the edge weighting technique to compensate for the otherwise lost information. The citations obtained by an article from the time it was published till time period tn are not all equal on the Impact Graph. The most recent are more significant since they demonstrate the paper's direct influence on current science. As a result, the edge weight in the Impact Graph is now denoted by.

$$WE_{impact}\,(a, b) = \sum_{\forall\, i \in Pa \cap Pb} a. \frac{\sum_{tk=to}^{tn}(w(tk).cit(i,tk))+\beta}{aut(i).(1+tn-per(i))} \qquad (4.2)$$

Where

$$W(per(i)) = \frac{1}{1+t_n-per(i)} \qquad (4.3)$$

Normalizing citations based on time is a typical bibliometrics principle [Bornmann and Marx, 2015]. It has also been used in connection with bibliometric networks as an assessment measure for research groups developed from the Louvain community discovery method [Blondel et al., 2008]. Normalization of citations by document type, publication channel type, or study field might be valuable and must be applied at this step, depending on the dataset being analyzed and the purpose. In this case, $WE_{imp}$ calculates the total effect of a set of publications co-authored by authors x and y. In essence, the effect of an article at a particular time $t_n$ is proportional to the number of citations it received up to that point, and inversely related to the number of writers who co-authored it and the number of years since its publication. The values and represents an author's interest in the influence of his/her work or the number of his/her publications. Several metrics and methodologies have already been presented for balancing impact and quantity, as well as for simultaneously analyzing the influence of authors, journals, and publications individually or all of them together. For the sake of simplicity, we use a weighting method with =0.7 and =0.3 to prioritize impact above quantity. Of course, more research and even a training program are required here. If we want to fine-tune these two weights or train a classification model that distinguishes between writers of different styles, we may do so. Furthermore, the clear difference in the complexity of the two models derives from the fact that we wanted to see if a classic and straightforward method could be exceeded by a more complicated one. The latter's superiority is far from unarguable, because, as Occam's razor suggests, simple models can certainly reflect reality far better than more complicated ones in many circumstances. Our approach to this hypothesis is dependent on which of the two graphs is more important in the clustering process.

*Table 4.2: Co-authorship Networks Studies*

| # | Study | Key Finding | Method | Limitations |
|---|---|---|---|---|
| 1 | *Maria et al[96]* | Structural cohesion and embeddedness in co-authorship networks | Network analysis, statistical analysis | Focused on sociology, may not generalize |
| 2 | *Guillermo A [97]* | Collaboration dynamics and the role of trust in co-authorship networks | Qualitative interviews, network analysis | Limited sample size |
| 3 | *Dario et al [98]* | Multi-level structure and the role of individual attributes in co-authorship networks | Multilevel modeling, network analysis | Focused on South Korean academics |
| 4 | *Golness [99]* | Social network analysis for identifying key players in co-authorship networks | Network analysis, centrality measures | Limited to specific networks |
| 5 | *Pengsheng [100]* | Weighted network analysis of co-authorship networks | Network analysis, weighted centrality measures | Limited to specific fields and networks |

### 4.4 Properties of Co-authorship

Co-authorship networks exhibit various properties that have been studied extensively in the literature. Some of these properties include [101].

1. **Small-world Phenomenon:** Co-authorship networks often exhibit small-world properties, characterized by short average path lengths between nodes and high clustering coefficients [95].

2. **Scale-free Topology:** The degree distribution of co-authorship networks often follows a power-law distribution, indicating the presence of a few highly connected nodes (hubs) and a large number of nodes with low connectivity [102].

3. **Community Structure:** Research communities in co-authorship networks are often characterized by dense connections within communities and sparse connections between communities, reflecting the nature of research collaboration in specialized fields [47].

4. **Evolution over Time:** Co-authorship networks evolve over time as new researchers enter the field, collaborations are established or dissolved, and research interests shift [103].

## 4.5 Analysis Techniques for Co-authorship Networks

Various techniques have been developed to analyze the properties and structure of co-authorship networks. Some of these techniques include [104].

### 4.5.1 Network Visualization:

Visual representations of co-authorship networks can provide intuitive insights into the structure and patterns of research collaboration [99].

### 4.5.2 Clustering Algorithms:

Clustering techniques can be used to identify groups of researchers with similar collaboration patterns, revealing the underlying structure of research communities [105].

### 4.5.3 Community Detection Methods:

Community detection algorithms can be applied to co-authorship networks to uncover the structure of research communities and the relationships between them [47].

### 4.5.4 Centrality Measures:

Centrality metrics, such as degree centrality, closeness centrality, and betweenness centrality, can be used to identify influential researchers in co-authorship networks based on their connectivity and position within the network [82]. The centrality of an author x

at period tn measures the sociability of x, which is directly linked with that of his/her co-authors. Thus, an author's weight is proportional to the sum of weights of his neighbors normalized by λ, which is the maximum weight in the graph.

## 4.6 Applications of Co-authorship Networks

Co-authorship networks have been widely used to study various aspects of academia, including [28].

### 4.6.1 Performance Assessment:

Researchers have used co-authorship networks to assess the performance and impact of researchers, institutions, and countries [106]. Each researcher has a set of features that depict his/her success as an individual, without taking into consideration the co-authors' network. Such features generally account for the productivity, or else the volume and frequency of publications, and the impact they have, which is calculated using citations. Both productivity and impact have a strong temporal dimension, which must be considered when building an author's profile.

1. Weighted Cumulative Productivity: The number of articles written by an author x from t0 until a given time $t_n$, weighted by oldness (i.e. by the periods that have passed from the publication period of each article until tn).

$$\text{WCP} = \sum_{\forall t \in t_o}^{t_n} \frac{1}{(tn-t)+1} . Pub\ (x,t) \qquad (4.4)$$

2. Weighted Cumulative Impact: The number of citations made to the articles of author x, weighted by oldness (i.e. by the periods that have passed from the period of each citation until $t_n$).

Insert formula here.....

### 4.6.2 Collaboration Patterns:

Co-authorship networks have been used to investigate the nature of research collaboration, such as the formation and dissolution of collaborations, interdisciplinary research, and the role of geographic proximity in collaboration [28].

### 4.6.3 Research Communities:

Researchers have applied community detection methods to co-authorship networks to identify research communities, study their evolution over time, and understand the dynamics of knowledge production in different fields [47]. Co-authorship networks provide a valuable tool for analyzing research collaboration patterns, the structure of academic communities, and the dynamics of research communities over time. By examining these networks, researchers can gain insights into various aspects of academia, such as performance assessment, collaboration patterns, research communities, and the identification of rising stars.

## 4.7 Power Graphs

Power graphs are a type of network representation that captures the hierarchical relationships between researchers based on their influence and productivity. In power graphs, nodes represent researchers, and edges represent the dominance relationships between them. This section discusses the properties, analysis techniques, and applications of power graphs in the academic domain. Power graph research is relatively less explored compared to co-authorship networks. However, below are three studies that discuss power graphs or power graph analysis, highlighting their key findings [107]. Power

graphs provide a compact and readable representation of graph structure, allowing for more efficient visualization and analysis of complex networks [108]. The weights in the influence Power Graph measure the cumulative influence of a group of writers or their collaborations, whereas the weights in the Quantity Power Graph measure one author's, a group of authors', or a specific collaboration's productivity. Consider the example of an author x, who belongs in Power Node, to better understand the concepts of the graphs and what they represented. In the Power Graph, a clique motif will connect Power Node with itself through a Power Edge. In the de notions of Power Graphs we use $\chi$ or $\psi$ to refer to a Power Node (i.e. an author or group of authors).

$P\,N\chi$ =The set of Power Nodes connected to Power Node $\chi$.

$P\,C\chi$ =The set of Power Nodes that contain Power Node $\chi$.

$W\,P\,N(\chi)$ =The weight of Power Node $\chi$.

$W\,P\,E\,(\chi,\psi)$ =Weight of Power Edge connecting Power Node $\chi$ and Power Node $\psi$.

The weight of the Power Edge represents the strength of the clique's collaboration among writers, whereas the weight of the Power Node represents the clique's size. Similarly, there will be Power Edges that link to other Power Nodes, for example, or. Given the Quantity Power Graph, we choose the following characteristics for an author x who belongs in Power Node:

$$Sociability\,(x) = WPN(x).WPE\,(x,x) \qquad (4.6)$$

Where Sociability considers the size of the collaboration group to which x belongs (the weight of the Power Node) as well as the author cooperation strength within the clique. The number of coauthored publications is represented by the cooperation strength in the Quantity Power Graph.

$$\text{Psoc}\,(x|x\,\varepsilon\,\chi) = \sum_{\forall\xi\in PNx} WPN(\xi).\,\text{WPE}\,(\chi,\xi) \qquad (4.7)$$

$P_{soc}$ stands for Potential Sociability or the expanded clique. This approach aggregates

the power of all possible collaborators, i.e., the authors contained in the Power Nodes that link to through a Power Edge, multiplied by the weight of that Power Edge, since the stronger the connection, the more likely the author would eventually collaborate with them.

Given the Impact Power Graph, we choose the following characteristics for an author y who belongs in Power Node: Power graph analysis can be extended to weighted networks, providing a more comprehensive understanding of the structure and organization of complex weighted networks [109]. As power graph research is an emerging field, there are fewer studies available than in the co-authorship network domain. Nevertheless, here are two more studies related to power graphs: Power graphs can be employed to analyze dynamic networks by considering the evolution of the network structure over time, providing insights into the temporal patterns and changes in complex networks [110].

Power graphs can be utilized to model and analyze the structure of power distribution networks, helping to improve the design and management of these systems. These studies, together with the previous three, illustrate the potential of power graph analysis in various domains [111]. They highlight the benefits of power graph representation and analysis for understanding complex networks, including their organization, structure, and temporal dynamics. Further research could explore the application of power graphs to other areas, such as social networks, communication networks, and even co-authorship networks, to better understand their structure and dynamics. Due to the relatively nascent field of power graph research, finding additional studies specifically focused on power graphs can be challenging [112].

Graph compression techniques can be employed to analyze and represent large social networks, improving the efficiency of network visualization and analysis while preserving important structural features [113]. While these studies do not focus specifically on power graphs, they demonstrate the broader relevance of graph compression techniques, which include power graphs, in analyzing and representing complex networks. The insights from these studies can be useful for further research

exploring the potential of power graphs in various domains, such as co-authorship networks, social networks, and communication networks, among others.

*Table 4.3: Power Graphs*

| # | Author & Year | Key Finding | Method | Limitations |
|---|---|---|---|---|
| S 1 | Ghoniam [114] | Compact and readable representation of graph structure | Power graph analysis | Limited to specific network types |
| 2 | X ma et al [115] | Insights into the organization of biological networks | Power graph analysis, application to biology | Focused on biological networks |
| 3 | M.E.J Newman [116] | Extension to weighted networks | Power graph analysis, weighted networks | Limited to specific weighted networks |
| 4 | T.O.F et al [117] | Analysis of dynamic networks | Power graph analysis, temporal networks | Limited to specific dynamic networks |
| 5 | K.R et al [118] | Efficient representation of large social networks | Graph compression techniques, including power graphs | Focused on social networks, not specifically power graphs |

The thesis contains exclusively focus on finite groups. The vertex set of the undirected power graph of a finite group $G$ consists of the elements of $G$. Two distinct elements in this graph are considered adjacent if one element is a power of the other. The concept of a power graph was introduced [119]. The concept of an undirected power graph was introduced by [120]. Numerous intriguing findings about power graphs have recently been acquired, as evidenced by the references.

The concept of the intersection power graph for a finite group [121]. The concept of the intersection power graph. $\Gamma_{IP}(G)$ for each finite group called $G$. This graph is constructed by considering the group elements as the vertices and two distinct vertices a and b are connected in $\Gamma_{IP}(G)$ if $\langle a\rangle \cap \langle b\rangle \neq \{e\}$ and $e$ is adjacent to all other vertices of $\Gamma_{IP}(G)$, where $e$ is the identity element of $G$. A graph containing an edge connecting every pair of distinct vertices is called a complete graph. For the entire graphical analysis with $n$

vertices, we use the notation. $K_n$. An independent set of a graph $\mathcal{G}$ is defined as a subset $A$ of its vertices such that the induced subgraph on A contains no edges.

The independence number of a graph $\mathcal{G}$, denoted as $\alpha_0(\mathcal{G})$., is defined as the maximum cardinality of an independent set in $\mathcal{G}$. The notation $\alpha_0(\mathcal{G})$ will be used to represent it. A graph that is connected and contains only a single cycle is referred to as a unicyclic graph. A graph that lacks any edges is referred to as a null graph. The diameter of a graph is defined as the maximum distance between any two vertices within the graph. A friendship graph, denoted as $F_n$, is an undirected planar graph with $2n + 1$ vertices and $3n$ edges. The friendship graph can be constructed by connecting $n$ instances of the cycle graph. $C_3$ at a shared vertex.

The properties of the independent set polytope, denoted as $\Gamma_{IP}(G)$, are being discussed. In this section, we present a comprehensive overview of the fundamental characteristics of the function. $\Gamma_{IP}(G)$.

**Theorem 4.1**: Consider a finite group $G$. The unicyclic property of $\Gamma_{IP}(G)$ holds if and only if $G$ is isomorphic to either $\mathbf{Z}_3$ or $S_3$, where $S_3$ represents the symmetric group on three letters.

**Proof.** It is evident that the group $\Gamma_{IP}(\mathbf{Z}_3)$ can be represented as a cycle with a length of 3. The group $\Gamma_{IP}(S_3)$ exhibits a single cycle of length 3 that is generated by the identity element, along with two elements of order 3. On the other hand, let us assume that the induced subgraph $\Gamma_{IP}(G)$ is unicyclic. Subsequently, we shall demonstrate the following:

1. The group$|G|$ does not possess any prime divisor p where p is greater than or equal to 5, as the graph $\Gamma_{IP}(G)$ exhibits unicyclic behaviour. Consequently, the cardinality of the group $|G| = 2^m 3^n$.
2. Let us consider a Sylow 3-subgroup, denoted as $M$, of the group $G$. Let us assume that the cardinality of set $|M|$ is greater than or equal to 9. If the group $G$ contains an element $x$ with an order of 9, then the subgroup generated by $x$, denoted as $\langle x \rangle$, will have at least two cycles in the permutation representation $\Gamma_{IP}(G)$. This leads to a contradiction. If the set $M$ does not contain any elements with an order

of 9, then every non-trivial element in $M$ must have an order of 3. This in turn implies that the group $\Gamma_{IP}(G)$ must have at least two cycles, which contradicts the previous statement.

3. According to condition 2), it can be deduced that the order of the group G, denoted as $||G| = 2^m 3$. Additionally, it is known that G possesses a solitary Sylow 3-subgroup.
4. In a similar vein, it can be inferred that the set $G$ does not contain any elements that have orders of both 4 and 6
5. Let $P = \langle x \rangle$ denote the distinct Sylow 3-subgroup of group $G$. The subgroup $P$ is considered to be normal in the group $G$. According to the "N/C' Theorem, it is possible to embed the group $G/C_G(P)$ into the automorphism group of Aut$(P)$. Let us assume the existence of an element $y$ in the set $G \setminus P$

such that the elements $x$ and $y$ commute, i.e., $xy = yx$. The order of $y$ is 2. Given that the orders of $x$ and $y$ are coprime, it follows that the order of $xy$ is 6, which is deemed unattainable. This implies that the composition of the function $C_G(P) = P$. Furthermore, it is widely recognized that the automorphism group of the set Aut$(P) \cong \mathbb{Z}_2$. Therefore, the cardinality of the quotient group $\left|\frac{G}{P}\right| = 1$ 2. In the scenario, one. The condition $G = P = Z_3$ is satisfied as intended. In the latter scenario, it can be observed that the order of the group $|G| = 6$, and it can be readily demonstrated that $G = S_3$

**Theorem 4.2** states that for a finite group $G$ with order $n = p_1^{\beta_1} p_2^{\beta_2} \cdots p_m^{\beta_m}$, where $p_1, p_2, \ldots, p_m$ and are distinct prime numbers and $\beta_1, \beta_2, \ldots, \beta_m$ are natural numbers, $G$ possesses unique subgroups with orders $p_1, p_2, \ldots, p_m, u$ and $v$ are non-adjacent in the graph $\Gamma_{IP}(G)$, where $u, v(\neq e) \in G \Leftrightarrow$ if and only if the greatest common divisor between the order of $u$ and the order of $v$ is one.

**Proof.** Assuming that the highest common factor between the order of element $u$ and the order of element $v$ is one, it follows that the intersection of the subgroups generated by $u$

and $v$, denoted as ⟨u⟩ and ⟨v⟩ respectively, consists solely of the identity element e. Hence, the vertices $u$ and $v$ are not adjacent.

On the contrary, let us assume that $u$ and $v$ are not adjacent. If the greatest common divisor between the order of $u$ and the order of $v$ is not equal to one, then there exists a prime number $p_i$ such that $p_i$ divides the order of $o(u)$ and $p_i$ divides the order of $o(v)$. This suggests that the elements $\langle u \rangle$ and $\langle v \rangle$ should possess a subgroup characterized by an order of $p_i$. Given that the group has a distinct subgroup with an order $p_i$, $|\langle u \rangle \cap \langle v \rangle| \geq p_i$, it follows that the intersection of the subgroups generated by elements $u$ and $v$, Hence, the adjacency of $u$ to $v$ leads to a contradiction. Therefore, the greatest common divisor between the order of $u$ and the order of $v$ is one.

**Theorem 4.3.** In the case of a finite group $G$ with an order expressed as $p_1^{\beta_1} p_2^{\beta_2} \cdots p_m^{\beta_m}$, where $p_1, p_2, \ldots, p_m$ are distinct prime numbers and $\beta_1, \beta_2, \ldots, \beta_m$ are natural numbers, it can be observed that the graph $\Gamma_{lP}(G)$ is connected and has a diameter that is less than or equal to 4. This holds true when $G$ contains only subgroups with orders corresponding to $p_1, p_2, \ldots, p_m$.

**Proof.** Given that $p_i$ is a divisor of the order of the group $|G|$, for $i = 1,2,\ldots,m$, it follows that there exists an element $x_i \in G$ such that the order of $x_i$ is precisely $p_i$, for $i = 1,2,\ldots,m$. Let $x_i$ and $x_j$ denote two distinct elements in the group $G$, each having an order of $p_i$ and $p_j$, where $i \neq j$ and $1 \leq i,j \leq m$. Let $N_i = \langle x_i \rangle$ and $N_j = \langle x_j \rangle$ denote the subgroups of $G$. Based on our presumptions, $N_i$ and $N_j$ denote distinct subgroups characterized by their respective orders, $p_i$ and $p_j$. It can be inferred that both $N_i$ and $N_j$ are normal subgroups of the group $G$. Furthermore, it can be observed that $N_i$ and $N_j$, being a normal subgroup of $G$, possesses the property that its order is equal to the product of $p_i p_j$. Given that $N_i$ and $N_j$are cyclic subgroups, it follows that their product, $N_i N_j$ is also a cyclic subgroup. Consequently, there exists an element $y$ in $N_i N_j$with an order equal to the product of the orders of $N_i$ and $N_j$, denoted as $p_i p_j$. Based on our assumptions, the intersection of the sequence $\langle x_i \rangle \cap \langle y \rangle = \langle x_i \rangle$ and $\langle x_j \rangle \cap \langle y \rangle = \langle x_j \rangle$.

This suggests that the product of $x_i y x_j$ forms a path in the graph $\Gamma_{IP}(G)$. Let $u, v$ denote two elements in the group $G$. Next, we consider the existence of $p_i$ and $p_j$, where i and j are specific integers within the range of 1 to m, inclusive. It is required that $p_i$ divides the order of $u$, while $p_j$ divides the order of $v$. It should be noted that the expression $ux_i y x_j v$ represents a path connecting vertices $u$ and $v$ within the graph $\Gamma_{IP}(G)$. Therefore, it follows that the induced subgraph $\Gamma_{IP}(G)$ is connected and has a diameter of at most 4 i.e, $(\Gamma_{IP}(G)) \leq 4)$.

**Theorem 4.4.**

The complete graphs $K_5$ and $K_{3,3}$ are examples of non-planar graphs. The non-planarity of $\Gamma_{IP}(G)$ can be observed in the case of an abelian group $G$ with an order of either 12 or 18.

**Proof. Case 1:** The group G is cyclic.

For $|G| = 12, G \cong \mathbf{Z}_{12}$. Because $G$ includes four elements of order 12 and a unique subgroup of order 2, $K_5$ is a subgraph of $\Gamma_{IP}(G)$. According to Theorem 2.4, 2.4, $\Gamma_{IP}(G)$ is non-planar.

For $|G| = 18, G \cong \mathbf{Z}_{18}$. Because $G$ includes six components of order 18, $K_6$ is a subgraph of $\Gamma_{IP}(G)$. According to Theorem 4.4, $\Gamma_{IP}(G)$ is non-planar.

Case 2: $G$ is not a cyclic group.

For $|G| = 12, G \cong \mathbf{Z}_2 \times \mathbf{Z}_2 \times \mathbf{Z}_3$. $G$ has six elements of order 6, hence $K_6$ is a subgraph of $\Gamma_{IP}(G)$. According to Theorem 2.4, $\Gamma_{IP}(G)$ is non-planar.

For $G = 18, G \cong \mathbf{Z}_2 \times \mathbf{Z}_3 \times \mathbf{Z}_3$. $G$ has eight elements of order 6, hence $K_8$ is a subgraph of $\Gamma_{IP}(G)$. According to Theorem 2.4, $\Gamma_{IP}(G)$ is non-planar.

**Lemma 4.1.** For any two finite groups $H_1$ and $H_2, \Gamma_{IP}(H_1) \cong \Gamma_{IP}(H_2)$ if $H_1 \cong H_2$

Proof. Assume that $f: H_1 \to H_2$ is a group isomorphism. Let $a, b \in H_1$ be such that $a$ is next to b in $\Gamma_{IP}(H_1)$. Since $\langle a \rangle \cong \langle f(a) \rangle$, for each $a \in H_1, |\langle a \rangle \cap \langle b \rangle| = |\langle f(a) \rangle \cap \langle f(b) \rangle| \geq 1$. In $H_2$, $f(a)$ is next to $f(b)$. As a result, $\Gamma_{IP}(H_1) \cong \Gamma_{IP}(H_2)$.

**Remark 4.1:** The converse of Lemma 4.1 is false. Consider $(\mathbf{Z}_8, +_8)$ and the quaternion group $Q_8$ with order 8. It should be noted that $\mathbf{Z}_8$ is not isomorphic to $Q_8$, but rather $\Gamma_{IP}(\mathbf{Z}_8) \cong K_8 \cong \Gamma_{IP}(Q_8)$.

**Theorem 4.5:** $G, \Gamma_{IP}(G)$ is a tree $\Leftrightarrow G \cong D_2$ or $D_4$ for any dihedral group $G$.

Proof. If $G$ is isomorphic to $D_2$ or $D_4$, then its itersection power graph is $K_2$ or $K_{1,3}$. As a result, $\Gamma_{IP}(G)$ is a tree.

Consider the case where $\Gamma_{IP}(G)$ is a tree. Assume there is a prime number $p \geq 5$ that is a divisor of $|G|$. Because $G$ contains an element of order $p, \Gamma_{IP}(G)$ contains $K_p (p \geq 5)$ as a subgraph. This $\Gamma_{IP}(G)$ is not a tree, which is a contradiction. As a result, $|G| = 2^n 3^m$, where $n \geq 1$ and $m \geq 0$ are two integers.

Assume$|G| = 2^n 3^m$, where $n \geq 3$ and $m = 0$. Then, as a subgraph, $\Gamma_{IP}(G)$ contains $K_{2^{n-1}}$. This $\Gamma_{IP}(G)$ is not a tree, which is a contradiction.

Assume $|G| = 2^n 3^m$, where $n \geq 1$ and $m \geq 1$ are integers. This means that $G$ has an element $w$ of order 3. The subgraph induced by $\langle w \rangle$ now contains $K_3$. This $\Gamma_{IP}(G)$ has $K_3$ as a subgraph, which is a contradiction. This means that $G$ can be isomorphic to either $D_2$ or $D_4$. In this section, several finite groups are categorized as as whose $\Gamma_{IP}(G)$ has a maximum book thickness of two.

**Problem 4.1.** $m \geq 4, bt(K_m) = \left\lceil \frac{m}{2} \right\rceil$.

**Problem 4.2.** If $G \cong D_{2n}$ where $n = 1,2,3,4$, $bt(\Gamma_{IP}(G))$ is at most two for a dihedral group $G$.

Proof. Given that the intersection power graph of $D_2, D_4$ and $D_6$ contains at least one edge, it can be concluded that the book thickness for these graphs is at least one. Each subgraph in this context refers to a subset of a larger graph, specifically the one-page

embeddable graph associated with a given integer n. Therefore, the thickness of the lines representing the graphs in the book is one. The inclusion of the subgraph $K_4$ within the intersection power graph of $D_8$ implies, according to Theorem 3.1, that the minimum book thickness for this graph is two.

**Problem 4.3**, if $G$ is a finite abelian group and isomorphic to the trivial group of order one, $\mathbf{Z}_2 \times \mathbf{Z}_2 \times \cdots \times \mathbf{Z}_2, \mathbf{Z}_3 \times \mathbf{Z}_3 \times \cdots \times \mathbf{Z}_3$ or $\mathbf{Z}_4$., then the number of $bt(\Gamma_{IP}(G))$ (G)) is at most two.

**Proof.** The book thickness of the intersection power graph of the trivial group of order one is zero, as the graph is a null graph. Given the intersection power graph of the Cartesian product $\mathbf{Z}_2 \times \mathbf{Z}_2 \times \cdots \times \mathbf{Z}_2$ and $\mathbf{Z}_3 \times \mathbf{Z}_3 \times \cdots \times \mathbf{Z}_3$, it can be observed that there exists at least one edge. Consequently, the book thickness for these graphs is determined to be no less than one. The isomorphism between the intersection power graph of $\mathbf{Z}_4$ and $K_4$ can be observed. According to Theorem 3.1, the graph in question has a book thickness of two. In the current section, the values of $\alpha_0(\Gamma_{IP}(G))$ are achieved.

**Problem 4.4.** For a finite group $G$ with order$p_1^{\beta_1} p_2^{\beta_2} \cdots p_m^{\beta_m}$, where $p_1, p_2, \ldots, p_m$ are distinct prime numbers and $\beta_1, \beta_2, \ldots, \beta_m$ are positive integers, the inequality $\alpha_0(\Gamma_{IP}(G)) \geq m$ holds.

Proof. Given that each $p_i$ is a divisor of $|G|, G$, it follows that G possesses elements $a_i$ such that the order of each element $a_i$, denoted as $o(a_i) = p_i$, for $1 \leq i \leq m$, where $i$ ranges from 1 to $m$. It is important to highlight that the intersection of the sequence $\langle a_i \rangle \cap \langle a_j \rangle = \{e\}$, for each $i \neq j$, for every i not equal to j. The set $\{a_1, a_2, \ldots, a_m\}$ is an independent set of the induced subgraph $\Gamma_{IP}(G)$.. Therefore, the desired outcome can be inferred.

**Problem 4.5.** For a finite group with order $p_1^{\beta_1} p_2^{\beta_2} \cdots p_m^{\beta_m}$, where $p_1, p_2, \ldots, p_m$ are distinct prime numbers and $\beta_1, \beta_2, \ldots, \beta_m$ are positive integers, it is true that

$\alpha_0(\Gamma_{IP}(G)) = m \Leftrightarrow G$ if and only if. The group G possesses a distinct subgroup characterized by order of $p$, where $i$ ranges from 1 to $m$.

**Proof.** Let us consider a scenario in which the group $G$ possesses a solitary subgroup that has an order denoted as $p_i$, where $i = 1,2,\ldots,m$. If the independence number of the induced subgraph $\alpha_0(\Gamma_{IP}(G)) > m, G$, then the graph G contains an independent set $X$ with at least $m + 1$ elements. According to Theorem 2.2, the orders of elements in set $X$ are mutually prime. Given that the group $G$ has precisely m distinct prime divisors, it is not possible to identify m+1 elements within $G$ whose orders are mutually prime. The inequality $\alpha_0(\Gamma_{IP}(G)) \geq m$ holds. Furthermore, according to Theorem 4.1, it can be concluded that the lower bound of $\alpha_0(\Gamma_{IP}(G)) \geq m$ is greater than or equal to $m$. Therefore, the value of $\alpha_0(\Gamma_{IP}(G)) \geq m$ is equal to $m$.

On the other hand, let us assume that the value of $\alpha_0(\Gamma_{IP}(G)) = m$. According to Cauchy's Theorem, the set $G$ contains elements that have an order denoted by $p_i$, where $i = 1,2,\ldots,m$. Consider a set of elements $a_i \in G$, where $o(a_i) = p_i$, where $i = 1,2,\ldots,m$.

Let us consider a group $G$ that possesses two distinct subgroups, each having an order denoted by $p_i$, where i represents a specific index. Let the elements $b_i \in G$ be such that the order of each element, denoted as $o(b_i) = p_i$ . It is evident that the intersection of the sequences $\langle a_i \rangle \cap \langle b_i \rangle = \{e\}$ and so $\{a_1, a_2, \ldots, a_m, b_i\}$ is the independent set in in $\Gamma_{IP}(G)$ with $m + 1$ elements. This observation leads to a contradiction. Hence, the group $G$ possesses a distinct subgroup of size $p_i$, where $i = 1,2,\ldots,m$

### 4.7.1 Properties of Power Graphs

Power graphs exhibit several properties that distinguish them from other types of network representations [122].

1. **Directed Edges:** Unlike co-authorship networks, power graphs have directed edges, representing the dominance relationship between researchers [123]. A

directed edge from node A to node B indicates that researcher A is more influential or productive than researcher B.

2. **Hierarchical Structure:** Power graphs exhibit a hierarchical structure, with researchers ranked based on their influence and productivity. This structure can be used to identify key players and understand the distribution of power within research communities [124].
3. **Weighted Edges:** Edges in power graphs can be weighted to reflect the strength of the dominance relationship between researchers. For example, the weight of an edge can be determined by the difference in citation counts or publication output between the connected researchers [125].

### 4.7.2 Ranking Algorithms

Ranking algorithms can be applied to power graphs to determine the relative importance of researchers based on their influence and productivity. Examples of ranking algorithms include [127] and HITS [128]. A ranking system based on multiple KPIs to identify the rising stars in the degree awarding institutes (DAIs) of Pakistan. The ranking system integrates the clustering results, community detection results, and centrality measures, as well as other KPIs such as publication output and citation impact. In this Thesis a weighted average method to calculate the overall ranking score for each researcher, where the weights are determined based on the importance and relevance of each KPI to the research field and context. The rising stars are then identified based on their ranking scores and other criteria, such as their career stage and potential for future impact.

### 4.7.3 Community Detection Methods:

Community detection algorithms can be applied to power graphs to uncover the structure of research communities and the hierarchical relationships between them [47]. Community detection methods are used to identify the subgroups or communities of researchers within the co-authorship and power graphs. The Louvain algorithm, which is

a modularity-based method, to detect communities in the graphs. The Louvain algorithm optimizes a quality function based on the modularity score, which measures the degree of clustering or separation of the nodes in the graph. We evaluate the quality of the community detection results using the modularity score and the participation coefficient. Given the Impact Power Graph, we choose the following characteristics for an author y who belongs in Power Node:

$$\text{Impact (y)} = (\text{WPN } (\Psi).\ \text{WPE } (\Psi, \Psi) \qquad (4.8)$$

Where Impact takes into account the size of the clique that y belongs to (the weight of the Power Node) and the cumulative impact of all the collaborations between authors within the clique. The collaboration impact between two authors in the case of the Impact Power Graph is based on the respective edge, hence it is the impact of their paper.

$$Comimpact\ (y|y\ \varepsilon\ \Psi) = \sum_{\forall \xi \in PN\Psi} WPN(\xi).\text{WPE}\ (\Psi, \xi) \qquad (4.9)$$

Where Impact is Community Impact and acts in similar to equation 4.8.

### 4.7.4 Centrality Measures:

Centrality metrics, such as eigenvector centrality and Katz centrality, can be used to identify influential researchers in power graphs based on their position within the hierarchical structure [82]. Centrality measures are used to assess the influence and prominence of researchers in the co-authorship and power graphs. We calculate various centrality measures, such as degree centrality, betweenness centrality, and eigenvector centrality, to identify the most connected and influential researchers in the network. Degree centrality measures the number of direct connections a node has in the graph, while betweenness centrality measures the number of shortest paths that pass through a node. Eigenvector centrality measures the influence of a node based on the influence of its neighboring nodes. We evaluate the centrality measures using various network analysis techniques, such as node ranking and visualization.

The centrality of an author x at period tn measures the sociability of x, which is directly linked with that of his/her co-authors. Thus, an author's weight is proportional to the sum of weights of his neighbors normalized by λ, which is the maximum weight in the graph.

$$\text{Centrality} = \frac{1}{\lambda}\sum_{\forall x \in Cx}\left(Centerality\ (n)\right) \qquad (4.10)$$

Weighted Collaboration Impact: An author's x Weighted Collaboration Impact measures the impact of the author's collaborations at a given time. It is the sum of weights of all edges that contain x in the Impact graph.

$$\text{Impact} = \sum_{\forall x \in Cx}(WEqual(x, y)) \qquad (4.11)$$

### 4.7.5 Applications of Power Graphs

Power graphs can be used to study various aspects of academia, including:

1. **Distribution of Power and Influence:**

By analyzing power graphs, researchers can gain insights into the distribution of power and influence within research communities, which can help understand the dynamics of knowledge production and dissemination [111].

2. **Identification of key Players:**

Power graphs can be used to identify key players in academic networks, who may be influential in shaping research agendas and driving innovation [129].

### 4.8 Identification of Rising Stars:

Power graphs can be used to identify rising stars in academia by studying their position within the hierarchical structure and their dominance relationships with other researchers [87][88][94]. In conclusion, power graphs provide a valuable tool for analyzing the hierarchical relationships between researchers based on their influence and productivity.

By examining power graphs, researchers can gain insights into the distribution of power within research communities, identify key players in academic networks, and identify rising stars in academia. The collaboration graphs give us an idea on the sociability of an author and the impact of his/her direct collaboration. Power graphs generated from these collaboration graphs allow us to examine the potential of an author's collaborations, under the prism of the extended co-authorship society that the author belongs to. Thus, we assume an increased potential for authors who belong to an eminent co-authorship group or are part of a scientific (sub)network of high impact. Each of the following features are defined both for the impact and the quantity power graph. Author Power Node Weight: An author's Power Node weight AP Nw measures the volume or impact of papers published by author's x close community (authors that belong in the same power node $\chi$ with x.

$$APNw\,(x|x\,\varepsilon\,\chi) = WPN\,(\chi) \qquad (4.12)$$

Author Power Clique Weight: An author's Power Clique weight APCw measures the volume or impact of papers published by author's x wider community (including the co-authors of co-authors of x).

$$AP\,Cw\,(x|x \in \chi) = X\,\forall\mu \in P\,N\chi\,W\,P\,N(\mu) \cdot W\,P\,E(\mu,\psi) + X\,\forall\nu \in P\,C\chi\,W\,P\,N(\nu) \quad (4.13)$$

## 4.9 Data Mining Techniques in Academia

Data mining techniques have grown in popularity in academia for analyzing enormous amounts of bibliometric data and extracting significant patterns and insights. These strategies can assist in identifying notable scholars, research trends, and patterns of collaboration, among other areas of academia. This section examines several frequently used data mining techniques and their applications in academia. In the academic realm, many data mining approaches have been used to analyze bibliometric data and acquire insights into various areas of research [90]. Among the most prominent approaches are:

### 4.9.1 Clustering

By studying their position in the hierarchical structure and dominance connections with other academics, power graphs may be utilized to detect budding academic star [87] [94]. K-means, hierarchical clustering, and DBSCAN are examples of common clustering techniques. To analyze the collected publication data, we employ various data mining techniques, including clustering algorithms, community detection methods, and centrality measures. Clustering algorithms are used to group similar researchers based on their publication and citation data. We use the K-means clustering algorithm, which is a popular unsupervised learning method, to group researchers based on their citation counts, h-index, and collaboration indices. K-means clustering partitions the data into K clusters, where K is a predefined number chosen based on the data and research context. We use the elbow method and silhouette analysis to determine the optimal value of K and evaluate the quality of the clustering results using various metrics, such as the within-cluster sum of squares and the silhouette score.

Min Change: The minimum change of an author's feature value between two consecutive periods (minchangef).

$$\text{min change f} = min_{t\in\{1,n\}} (f(t_i) - f(t_i - 1) \quad (4.14)$$

Max Change: The maximum change of an author's feature value between two consecutive periods (max Changef ).

$$\text{Max change f} = max_{t\in\{1,n\}} (f(t_i) - f(t_i - 1) \quad (4.15)$$

Last Period Change: The change of a feature lastChange_f from the last period depicts the author's dynamics at time tn.

$$\text{last period change f} = last\ change_f (f(t_n) - f(t_n - 1) \quad (4.16)$$

Sum of Change: The sum of the features changes through the periods totalChange_f, portrays author's stability and shows the total change through time.

Total change f $=\sum_{v l \epsilon 1}(f(t_l) - f(t_l - 1))$ (4.17)

Overall, our data mining approach provides a comprehensive and systematic method to identify the rising stars in the DAIs of Pakistan, based on a range of publication and citation data, network analysis, and KPIs. However, it is important to note that our approach has certain limitations, such as the reliance on quantitative data and the potential biases and limitations associated with the use of data mining techniques. Further research is needed to validate and extend our approach to different contexts and research fields, and to address these limitations and challenges.

### 4.9.2 Classification

Classification algorithms can be used to categories researchers, organizations, or publications based on their characteristics. Researchers, for example, may be classified as emerging stars, established researchers, or inactive academics based on their publication output, citation impact, and collaboration patterns.[87][94][130]. Decision Trees, Support Vector Machines, and Nave Bayes are examples of common classification methods.

### 4.9.3 Association Rule Mining

In bibliometric data, association rule mining techniques can be used to find interesting links or trends. For example, association rule mining can be used by academics to uncover co-occurring terms, study subjects, or collaboration patterns [131].

### 4.9.4 Network Analysis

Network analysis techniques can be used to analyze the structure and dynamics of co-authorship networks, citation networks, and power graphs. These techniques can help

identify influential researchers, research communities, and collaboration patterns [87][130].

### 4.9.5 Text Mining

Text mining techniques can be employed to analyze the content of research publications, such as abstracts or full-text articles, to identify research trends, emerging topics, and interdisciplinary research patterns [94].

## 4.10 Data Collection

The collection of comprehensive and representative data for research performance assessment involves the integration of multiple features, each offering a distinct perspective on the scholarly landscape. Universities, as the focal points of knowledge creation, serve as essential data sources. Information about the universities' research output, faculty composition, collaborations, and institutional affiliations provides insights into their contributions and impact. Subject categories play a pivotal role in understanding the disciplinary distribution of research activities. By categorizing publications into distinct subject areas, the data collection process can capture the diversity of research endeavors and their alignment with various academic domains. Faculty data, encompassing academic expertise, research interests, and institutional affiliations, offers a granular view of individual researchers' contributions and their role within the university's research ecosystem. Authorship data, detailing the identities of authors, order of authorship, and collaboration patterns, contributes to understanding interdisciplinary collaborations, research networks, and the diffusion of knowledge. Finally, publication data, including publication titles, abstracts, publication venues, and citation counts, forms the core of bibliometric analysis. This data offers insights into the quality, impact, and dissemination of research outputs. Integrating these features in data collection ensures a comprehensive and multi-dimensional understanding of research performance, enabling more accurate and holistic assessments that encompass university dynamics, disciplinary strengths, collaboration patterns, individual contributions, and

scholarly impact. Objective of the work: identify Scopus publications authored by each Pakistan professors.

The Table contains the "unequivocal" identification of each university in the dataset (university names must be standardized). The first column must contain a primary key ("University") which is denoted by UID_1 of the record. The other fields can be the location, the nature (public/private), etc.

***Table 4.4 : Public and Private Universities of Pakistan***

| University | University Name | Sector | Chartered By | Discipline | Province | City |
|---|---|---|---|---|---|---|
| UID_1 | Abdul Wali Khan University | Public | Government of Khyber Pakhtunkhwa | General | Khyber Pakhtunkhwa | Mardan |
| UID_2 | Aga Khan University | Private | Government of Pakistan | General | Sindh | Karachi |

This table contains the "unequivocal" identification of fields used to classify professors in the dataset. The first column is a primary key ("SC_ID") of the record.

***Table 4.5 : Subject Categories of Public and Private Universities of Pakistan***

| SC_ID | Subject Category Name |
|---|---|
| SC_1 | Agriculture |
| SC_2 | Biochemistry |
| SC_3 | Chemistry |
| SC_4 | Computer Science |
| SC_5 | Economics |

Each professor must be identified by a unique identifier, so define a primary key (Professor) for professor (PD) and apply it to the first column of this table. Be sure that each professor is associated to a consistent university listed in Table 4.4. Be sure that each professor is associated to a consistent "subject category" listed in Table 4.5.

*Table 4.6: Faculty of Public and Private Universities of Pakistan*

| Professor | University_ID | NAME | GENDER | RANK | SC_ID |
|---|---|---|---|---|---|
| PD_1 | UID_1 | Mujahid Shah | Male | Chairman / Assistant Professor | SC_4 |
| PD_2 | UID_2 | Naveed ur Rehman | Male | Assistant Professor | SC_4 |
| PD_3 | UID_1 | Irfan Ullah | Male | Assistant Professor | SC_2 |
| PD_4 | UID_2 | Mahrukh Shakir | Female | Assistant Professor | SC_2 |
| PD_5 | UID_2 | Rafiq Nawab | Male | Assistant Professor | SC_4 |

For each professor indexed in the faculty data, table 4.6 lists the unique identifiers of SCOPUS publications (EID) associated to that professor.

*Table 4.7: Authorship ID of Public and Private Universities of Pakistan*

| Professor | EID |
|---|---|
| PD_1 | 2-s2.0-85030721271 |
| PD_1 | 2-s2.0-84891770309 |

For each EID listed in the Authorship table, this table contains the complete record of the corresponding SCOPUS publications. In the "Export document settings" flag both "Citation information" and "Bibliographical information". so to have a record with these fields: EID; Authors; Title; Year; Source title; Volume; Issue; Art. No.; Page start; Page end; Page count; Cited by; DOI; Link; Affiliations; Authors with affiliations; Correspondence Address; Editors; Publisher; ISSN; ISBN; CODEN; PubMed ID; Language of Original Document; Abbreviated Source Title; Document Type; Source.

*Table 4.8: Authorship ID of Public and Private Universities of Pakistan*

| EID | 2-s2.0-85015660640 | 2-s2.0-85021176570 |
|---|---|---|
| Authors | Aziz I., Siraj-ul-Islam, Asif M. | Ali U., Baig A.Q., Imran M., Abbas G., Asif M. |
| Title | Haar wavelet collocation method for three-dimensional elliptic partial differential equations | On topological properties of certain boron nanostructures |
| Year | 2017 | 2017 |
| Source title | Computers and Mathematics with Applications | Journal of Computational and Theoretical Nanoscience |
| Volume | 73 | 14 |
| Issue | 9 | 2 |
| Art. No. | | |
| Page start | 2023 | 887 |
| Page end | 2034 | 898 |
| Page count | | |
| Cited by | 3 | 8 |
| DOI | 10.1016/j.camwa.2017.02.034 | 10.1166/jctn.2017.6376 |
| Link | https://... | https://... |
| Affiliations | Department of ... | Department of ... |
| Authors with affiliations | Aziz, I., ... | Ali, U., ... |
| Correspondence Address | Aziz, I.; Department of Mathematics, University of Peshawar Pakistan; email: imran_aziz@upesh.edu.pk | Ali, U.; Centre for Advanced Studies in Pure and Applied Mathematics, Bahauddin Zakariya University Pakistan |
| Editors | | |
| Publisher | Elsevier Ltd | American Scientific Publishers |
| ISSN | 8981221 | 15461955 |
| ISBN | | |
| CODEN | CMAPD | |
| PubMed ID | | |
| Language of Original Document | English | English |
| Abbreviated Source Title | Comput Math Appl | J. Comput. Theor. Nanosci. |
| Document Type | Article | Article |
| Source | Scopus | Scopus |

## 4.11 Methodology to Detect Rising Stars for Specific DAI

Our study's approach entails gathering publishing data from a variety of sources, including academic databases and institutional repositories. We focused on scholars connected with degree-granting institutes in Pakistan and gathered their publication data from 2010 to 2020.Using publication data, we created time-evolving co-authorship graphs to reflect academics' cooperation patterns through time. The nodes in these graphs represent researchers, while the edges show their co-authorship ties. We utilized the Gephi program to create and show the co-authorship graphs, and we employed several

network analysis approaches to determine the network's essential characteristics and structures. To quantify the importance and prominence of researchers in the network, we estimated several network centrality metrics such as degree centrality, betweenness centrality, and eigenvector centrality. We created power graphs using co-authorship graphs to reflect the impact and prominence of researchers in the network. Power graphs are a sort of node-weighted graph in which each node has a weight given to it depending on its degree centrality in the original graph. Highly linked nodes, or "hubs," are allocated larger weights in power graphs, indicating their greater impact and importance in the network. We utilized the Net Miner programmed to generate and analyses the power graphs, which we then compared to the co-authorship graphs to discover the network's emerging stars.

We also estimated citation counts, h-indexes, and cooperation indices for each researcher. Citation counts are the number of times a researcher's articles have been referenced by other researchers, whereas the h-index is a measure of the researcher's productivity and influence based on the number of publications and their citation counts. cooperation indices quantify the breadth and variety of a researcher's cooperation network and may be estimated using the number of co-authors, the number of different universities or countries represented in the partnerships, and other parameters. We used a variety of data mining approaches, including clustering algorithms and machine learning models, to identify the network's emerging stars. We performed the data mining study with the Weka software and evaluated the performance and accuracy of the models with various validation and sensitivity analysis methodologies. Overall, our methodology employs time-evolving co-authorship graphs, power graphs, and other KPIs and data mining tools to present a complete and data-driven approach to identifying emerging stars in academia. It is crucial to highlight, however, that our study has several limitations, such as the focus on a specific geographical location and research subject, as well as the possible biases and restrictions connected with the use of bibliometric and network analysis approaches. More research is required to test and expand our methodologies to other settings and study disciplines, as well as to overcome these limits and problems.

## 4.12 Results and Discussions of Method apply to Different DAI

Data mining investigation uncovers numerous insights regarding cooperation trends and emerging stars in Pakistan's degree awarding institutes (DAIs). The time-evolving co-authorship graphs depict collaboration tendencies among DAI researchers from 2010 to 2020. The co-authorship network appears to be strongly linked and concentrated, showing significant collaborative relationships among academics inside and across universities. Based on their research interests and connections, we also identify different subgroups or communities of researchers inside the network. The Louvain algorithm discovers these communities and emphasizes the need of multidisciplinary connections for academic achievement. The power graphs created from the co-authorship graphs highlight the significant scholars even more and emphasize the necessity of cooperation for academic achievement. The power graph analysis indicates some highly linked and prominent scholars, or "hubs," who play an important role in linking diverse groups and boosting network collaboration. We also see that emerging stars are strongly linked and have a significant network presence, highlighting the importance of teamwork for their academic achievement.

The KPI-based rating methodology confirms the identified emerging stars by providing a quantifiable evaluation of their academic influence and future success potential. The ranking method incorporates numerous KPIs, such as citation counts, h-index, cooperation indices, and others, to analyze each researcher's overall performance and potential. We see that emerging stars have high rankings across numerous KPIs, suggesting persistent and diversified contributions to their profession.

$$\sum_{\forall t \in t_0}^{t_n} \frac{1}{(t_n - t)+1} \cdot \text{Pub}(x, t) \qquad (4.18)$$

The number of articles written by an author x from $t_o$ until a given time $t_n$, weighted by oldness (i.e., by the periods that have passed from the publication period of each article until tn).

$$\sum_{Vt\epsilon t_o}^{t_n} \sum_{AJ\epsilon Px} \frac{cit(t,j)}{t-per(j))+1} \qquad (4.19)$$

Overall, our findings shows that data mining approaches may be used to reveal collaborative trends and identify emerging stars in academia. Our findings can assist academics and educational institutions in making better decisions about research cooperation, funding distribution, and academic employment. It is crucial to emphasize, however, that our findings have certain limitations, such as the focus on a specific geographic location and study subject, as well as the inherent biases and restrictions connected with the use of bibliometric and network analysis approaches. More study is required to confirm and expand our findings to other contexts and research domains, as well as to overcome these limits and constraints. Our findings have important implications for the degree awarding institutes (DAIs) in Pakistan. The identification of rising stars can help these institutions allocate resources more effectively, promote interdisciplinary research, and foster a vibrant research culture.

First, identifying rising stars can help institutions allocate resources more effectively by targeting their support to the most promising and impactful researchers. Institutions can provide these rising stars with more funding, lab space, and mentoring, and create incentives for them to continue their research and collaboration efforts. This can help institutions build their research capacity and reputation and attract more talented researchers and students.

Second, the identification of rising stars can promote interdisciplinary research by highlighting the importance of collaboration and communication across different research fields and disciplines. Institutions can encourage these rising stars to form interdisciplinary teams and collaborate with researchers from different backgrounds and provide them with opportunities to participate in interdisciplinary research programs and projects. This can help institutions address complex and interdisciplinary research challenges and generate new insights and solutions.

Third, the identification of rising stars can foster a vibrant research culture by inspiring and motivating other researchers to pursue excellence and innovation. Institutions can celebrate and showcase the achievements of these rising stars and create a supportive and inclusive environment that encourages creativity, risk-taking, and continuous learning. This can help institutions attract and retain top talent and promote a culture of excellence and innovation.

However, it is important to note that the identification of rising stars is only one aspect of building a vibrant and successful research culture. Institutions need to create a supportive and inclusive environment that provides researchers with the resources, infrastructure, and mentorship they need to succeed. This includes investing in research facilities, equipment, and staff, providing access to funding and grant opportunities, and promoting collaboration and communication among researchers. Institutions also need to address the challenges and limitations associated with the use of bibliometric and network analysis methods and ensure that their evaluation and promotion policies are fair, transparent, and based on a range of criteria and indicators. In this section, the results and interoperations of the analysis are documented. Said differently, by considering the sample universities results and graphical analysis are documented separately for each university.

### 4.12 Air University

Air University is a degree-awarding institute in Pakistan in Network Visualization. There are 408 items,20 clusters, and links connected are 2255, and the total link strength are 5923. An analysis was conducted on several authors' publication records, considering the number of documents they authored and the "Total link strength" metric. Among the authors, I.M. Qureshi emerged with the most documents, having authored 131 publications. Additionally, Qureshi exhibited a significant total link strength of 311, indicating strong connections and collaborations with other researchers in the network. Following Qureshi, A. Jalal ranked second with 90 documents and a cumulative link strength of 227. A.R. Javed secured the third position with documents and a cumulative link strength 192. Similarly, M.S. Arif had 53 documents and a total cumulative link

strength of 150, while M.Y. Malik produced 58 documents with a total cumulative link strength 149. K.U. Rehman contributed 69 documents with a total link strength 147, while N. Naseer authored 58 documents with a total link strength of 133. Y.Y. Ghadi, with 27 documents, showed a high total link strength of 124, indicating strong collaborations despite a comparatively lower publication count. M.A. Khan and A. Raza published 46 and 35 documents, respectively, and shared a total link strength of 124. These findings shed light on these authors' research productivity and collaborative networks, emphasizing the significance of the number of documents authored and the strength of their connections with other researchers in the field.

**Table 4.9 Collaborative Cluster Based on Productivity of Air University**

| Author | Documents | Total Link Strength |
|---|---|---|
| Qureshi i.m | 131 | 311 |
| Jalal a.r | 90 | 227 |
| Javed a.r | 72 | 192 |
| Arif m.s | 53 | 150 |

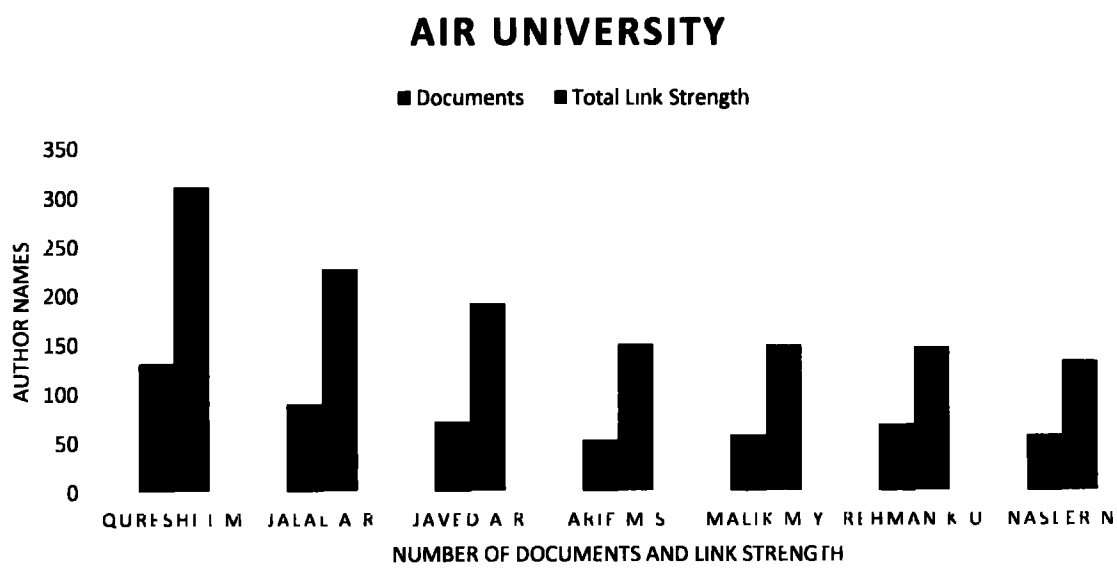



**Figure 4.2 Documents and total link strength of Air University**

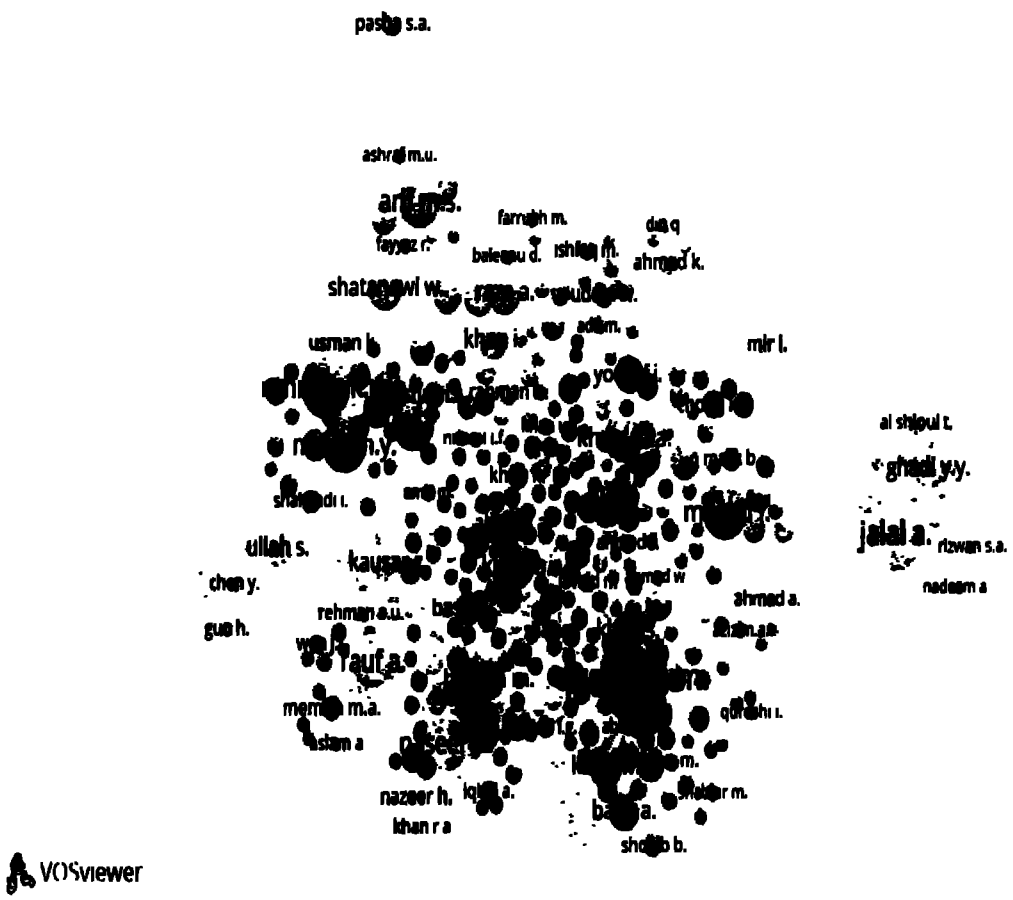



*Figure 4.1: Co-Authorship Graph of Air University*

*Table4.21: Collaborative Cluster Based on Productivity of Air University*

We can analyze the information based on the provided authors' data, documents, and total link strength to identify potential clusters or groups of authors. Clustering can help reveal patterns or associations among the authors based on their publication records and collaborative connections. Here is a clustering based on the provided data:

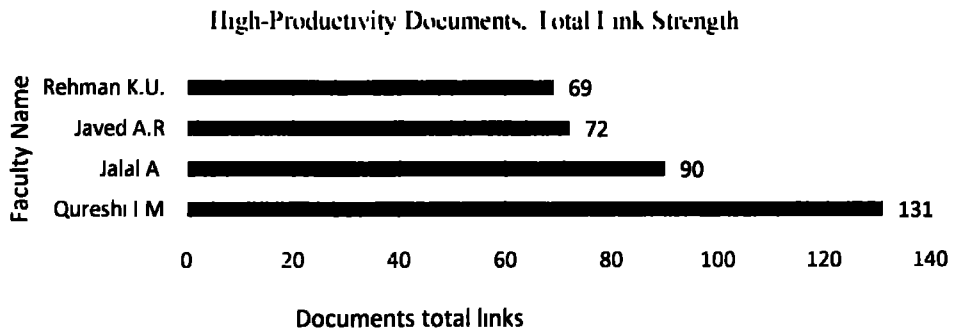



**Figure 4.3(a) Graphical Representation of High Productivity Documents of Air University**

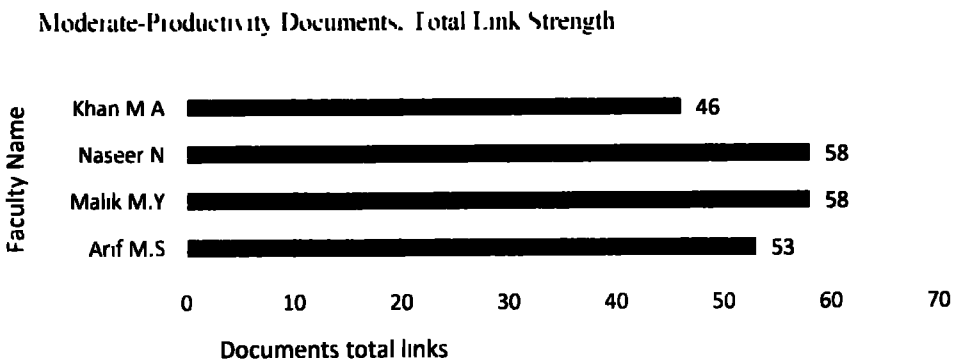



**Figure 4.3(b) Graphical Representation of Moderate Productivity Documents of Air University**

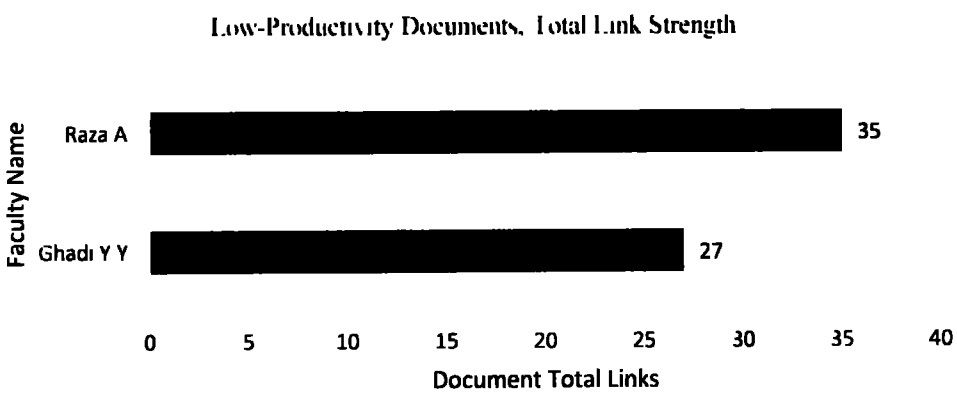



**Figure 4.3 (c) Graphical Representation of Low Productivity Documents of Air University**

**22 Bahria University**

Bahria University consists of 404 items which make 22 clusters of the researcher's links which connect them are 2191 and the total link strength is 5010. An analysis was conducted on the publication records of several authors, considering the quantity of documents they authored and the "Total link strength" metric. Among the authors, M. Ramzan stood out with the highest number of documents, having authored 115 publications. Furthermore, Ramzan displayed a notable total link strength of an unspecified value, indicating strong connections and collaborations with other researchers in the network.

Following Ramzan, K.N. Qureshi secured the second position with 73 documents and a total link strength of 169. Similarly, M. Usman ranked third with 52 documents and a total link strength of 167. G. Jeon contributed 58 documents and exhibited a substantial link strength of 158, while A. Ali authored 51 documents with a total cumulative link strength of 145. R.U. Haq produced 56 documents with a total link strength of 144, while A. Waqar had 42 documents and a total link strength of 122. S. Kadry authored 31 documents and exhibited a total link strength of 115.

M. Hamid contributed 36 documents with a total cumulative link strength of 113, while A. Ahmad had 42 documents and a cumulative link strength of 100.S. Khalid authored 52 documents with a total link strength of 97, and S. Iqbal produced 42 documents with a total link strength of 95. J.D. Chung displayed 26 documents and a 93 total link strength. A. Shafee and M. Hussain published 26 and 39 documents, respectively, with total link strengths of 90 and 82.

**Table 4.10 Collaborative Cluster Based on Productivity of Bahria University**

| High-Productivity | | Moderate-Productivity | | Low-Productivity | |
|---|---|---|---|---|---|
| **Authors** | Documents, Total Link Strength | Authors | Documents, Total Link Strength | Authors | Documents, Total Link Strength |
| **Ramzan M** | 115 | Waqar A | 42 | Chung J.D | 26 |
| **Qureshi K.N** | 73 | Kadry S | 31 | Shafee A | 26 |
| **Usman M** | 52 | Hamid M | 36 | Hussain M | 39 |
| **Jeon G** | 58 | Ahmad A | 42 | Ahmad S | 29 |
| **Ali A** | 51 | Khalid S | 52 | Chu Y.-M | 20 |

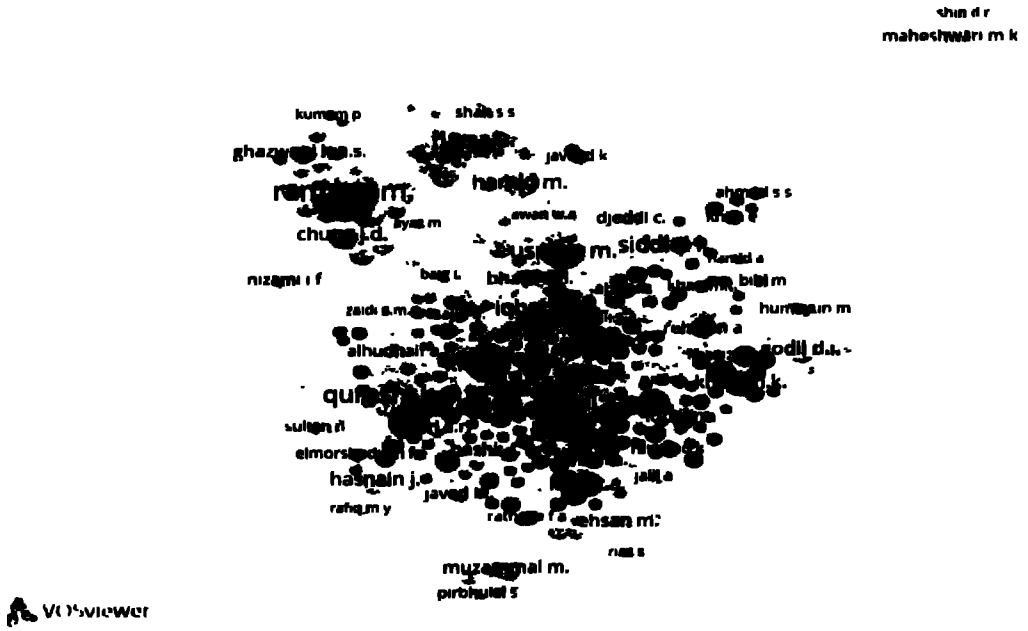



*Figure 4.5: Co-Authorship Graph of Bahria University*

Based on the given data of authors, documents, and total link strength, we can analyze the information to identify potential clusters or groups of authors. Clustering can help reveal patterns or associations among the authors based on their publication records and collaborative connections. Here is a clustering based on the provided data:

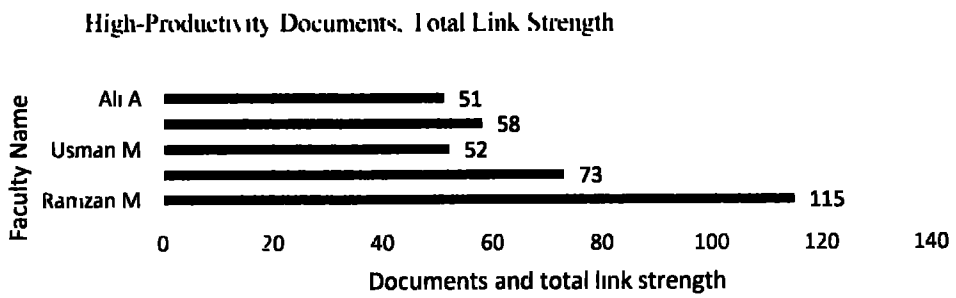



**Figure 4.5 (a) Graphical Representation of High Productivity Documents of Bahria University**

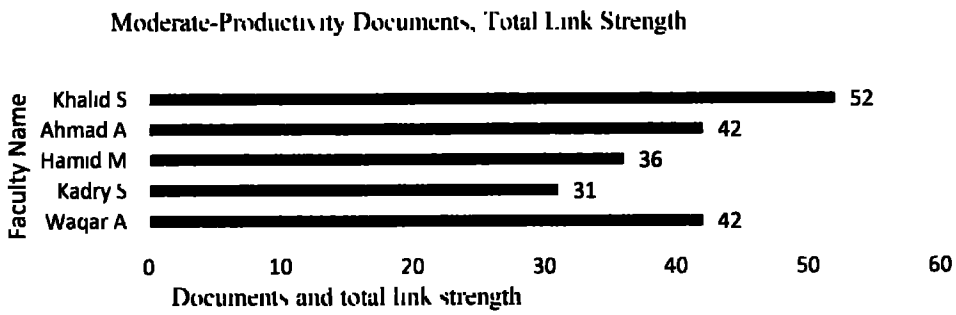



**Figure 4.5 (b) Graphical Representation of Moderate Productivity Documents of Bahria University**

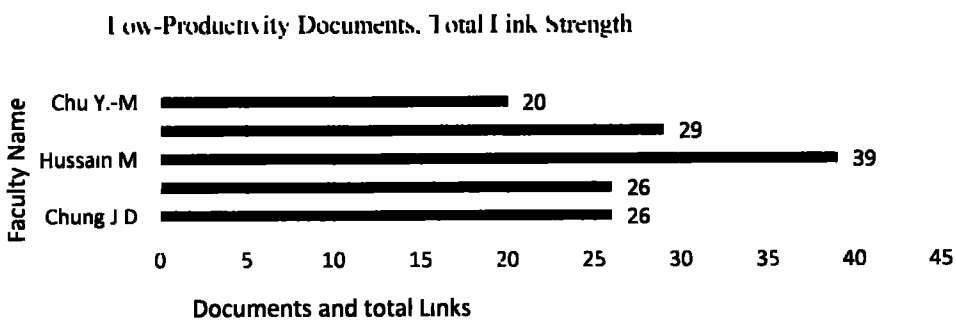



**Figure 4.5 (c) Graphical Representation of Low Productivity Documents of Bahria University**

## 4.13 Comsats University Islamabad

Comsats University consists of 567 items, making 27 clusters of the researcher's links, which connect them with 4191 and the total link strength is 1010. An analysis was conducted on the publication records of several authors, taking into account the amount of documents they authored and the "Total shared link" metric. Among the authors, K. Ayub emerged with the highest number of documents, having authored 59 publications. Additionally, Ayub displayed a total shared link value of 205, indicating strong connections and collaborations with other researchers in the network. Following Ayub, M.I. Khan ranked second with 57 documents and a total shared link value of 203. Similarly, S.U. Khan secured the third position with 55 documents and a total shared link value of 186. M. Imran contributed 48 documents with a total shared link value of 181, while A. Ali has authored 43 documents, with a collective shared link value of 152.A. Ahmad produced 52 documents with a total shared link value of 143, while T. Mahmood had 34 documents and a total shared link value of 131. A. Bokhari contributed 29 documents with a total shared link value of 120. J. Iqbal exhibited a total shared link value of 117, and S. Ahmad authored 41 documents with a total shared link value of 113.Furthermore, M.A. Khan and A. Hussain shared a total link value of 112 and 107, respectively. S. Khan authored 39 documents with a total shared link value of 104, while M.A. Gilani contributed 25 documents with a total of 103.

**Table 4.11: Collaborative Cluster Based on Productivity of Comsat University**

| High-Productivity | | Moderate-Productivity | | Low-Productivity | |
|---|---|---|---|---|---|
| **Authors** | Documents / Total Link Strength | Authors | Documents / Total Link Strength | Authors | Documents, Total Link Strength |
| **Ayub K** | 59 | Iqbal J | 49 | 0 | 0 |
| **Khan M.I.** | 57 | Ahmad S | 41 | 0 | 0 |
| **Khan S.U** | 55 | Khan M.A. | 39 | 0 | 0 |
| **Ali A.** | 43 | Khan S | 39 | 0 | 0 |

| | | | | | |
|---|---|---|---|---|---|
| **Ahmad A** | 52 | Gilani M.A. | 25 | 0 | 0 |

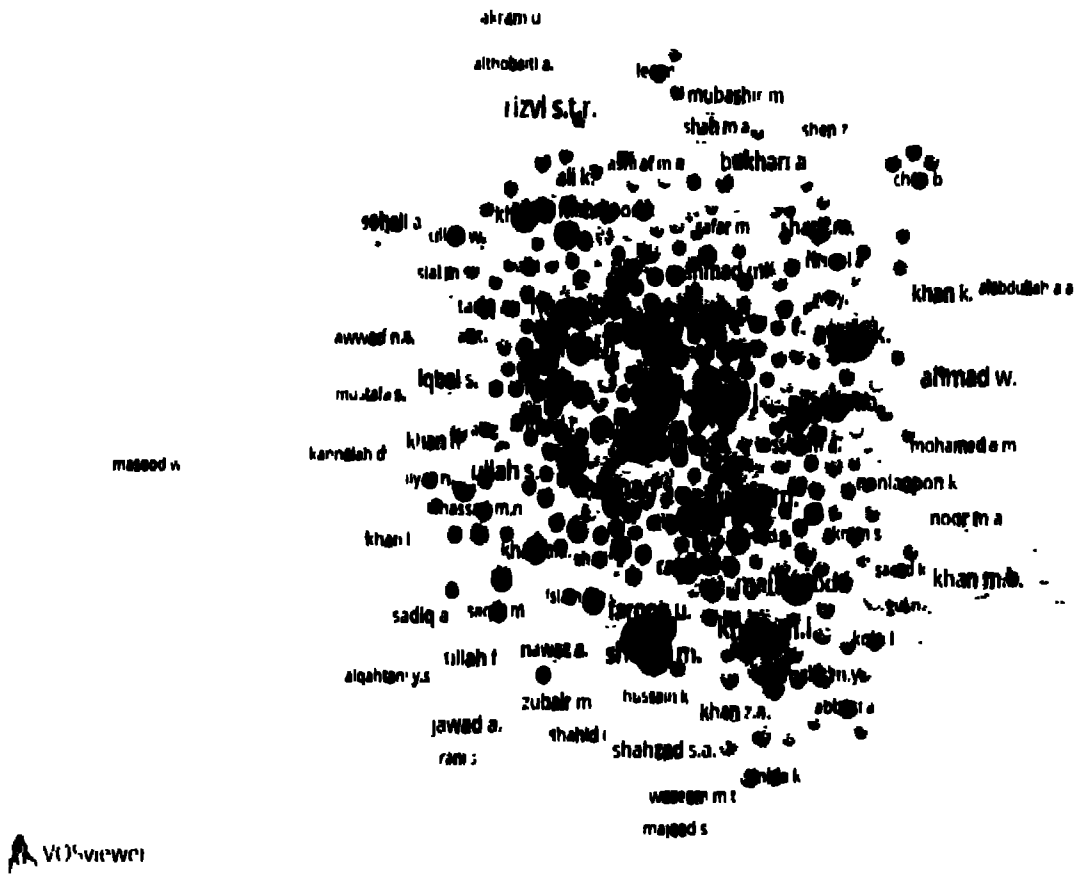



***Figure 4.6: Co-Authorship Graph of COMSATS University, Islamabad***

Based on the provided data of authors, documents, and total shared links, we can analyze the information to identify potential clusters or groups of authors. Clustering can help reveal patterns or associations among the authors based on their publication records and collaborative connections. Here is a clustering based on the provided data:

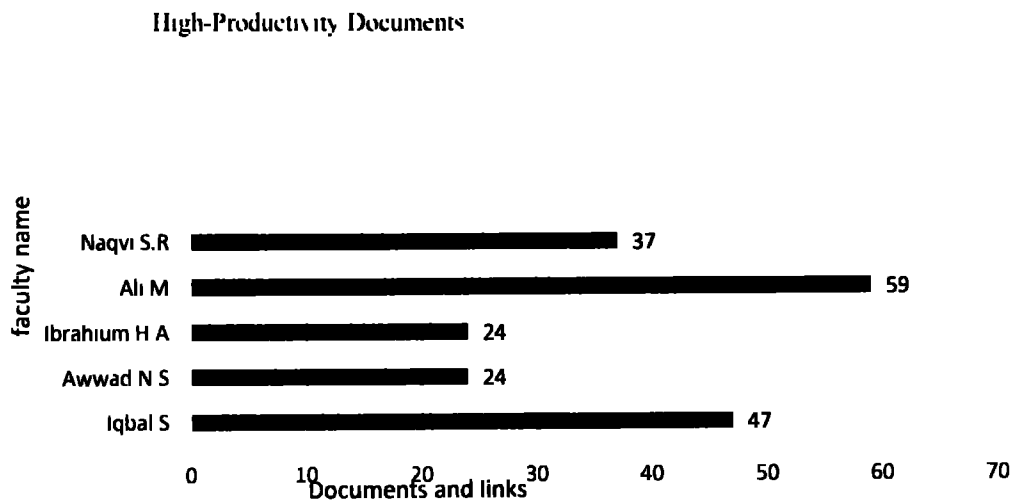



**Figure 4.7 (a) Graphical Representation of High Productivity Documents of Comsats University**

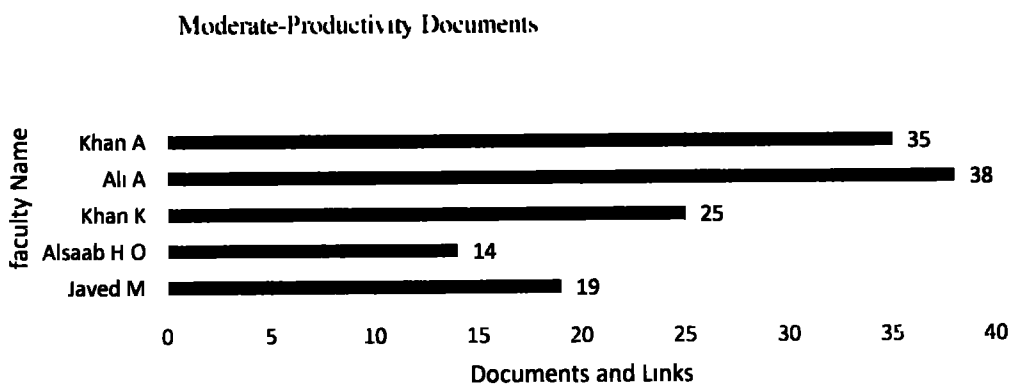



**Figure 4.7 (b) Graphical Representation of Moderate Productivity Documents of Comsats University**

## 4.14 Fast University, Islamabad

An analysis was conducted on several authors' publication records, considering the total number of documents they authored and the "Total link strength" metric. Among the authors, M.A. Jaffar stood out with the highest total number of documents, having authored 75 publications. Additionally, Jaffar exhibited an impressive total link strength

of 1800, indicating strong connections and collaborations with other researchers in the network. Following Jaffar, A.M. Mirza achieved the second position, having authored 59 documents and possessing a total link strength of 143. Similarly, A. Hussain secured the third position with 49 documents and a total link strength of 135.

M. Ahmad contributed 54 documents with a total link strength of 121, while M.A. Gondal authored 49 documents with a total link strength of 98. I. Hussain produced 42 documents with a total link strength of 96, while F.A. Khan and M.A. Khan both had 59 and 27 documents, respectively, with total link strengths of 93 and 89. M. Ali contributed 37 documents with a total link strength of 88, and M. Khan authored 57 documents with a total link strength of 81.

Additionally, T. Shah had 33 documents with a total link strength of 79, while A.R. Baig and I. Ullah both exhibited a total link strength of 75 with 52 and 45 documents, respectively. S. Anwar authored 18 documents with a total link strength of 73, while S. Ali contributed 33 documents with a total link strength of 72. A. Ali and M. Asim both published 46 and 34 documents, respectively, with total link strengths of 71.

***Table4.12: Collaborative Cluster Based on Productivity of Fast University***

| **High-Productivity** | | **Moderate-Productivity** | | **Low-Productivity** | |
|---|---|---|---|---|---|
| **Authors** | Documents / Total Link Strength | Authors | Documents / Total Link Strength | Authors | Documents / Total Link Strength |
| **Jaffar M. A** | 75 | Hussain I | 42 | Shah T | 33 |
| **Mirza A.M** | 59 | Khan F. A | 59 | Baig A. R | 22 |
| **Hussain A** | 49 | Khan M. A | 57 | Ullah I | 25 |
| **Ahmad M** | 54 | Ali M | 37 | Anwar S | 18 |
| **Gondal M. A** | 49 | Khan M | 57 | Ali S | 23 |

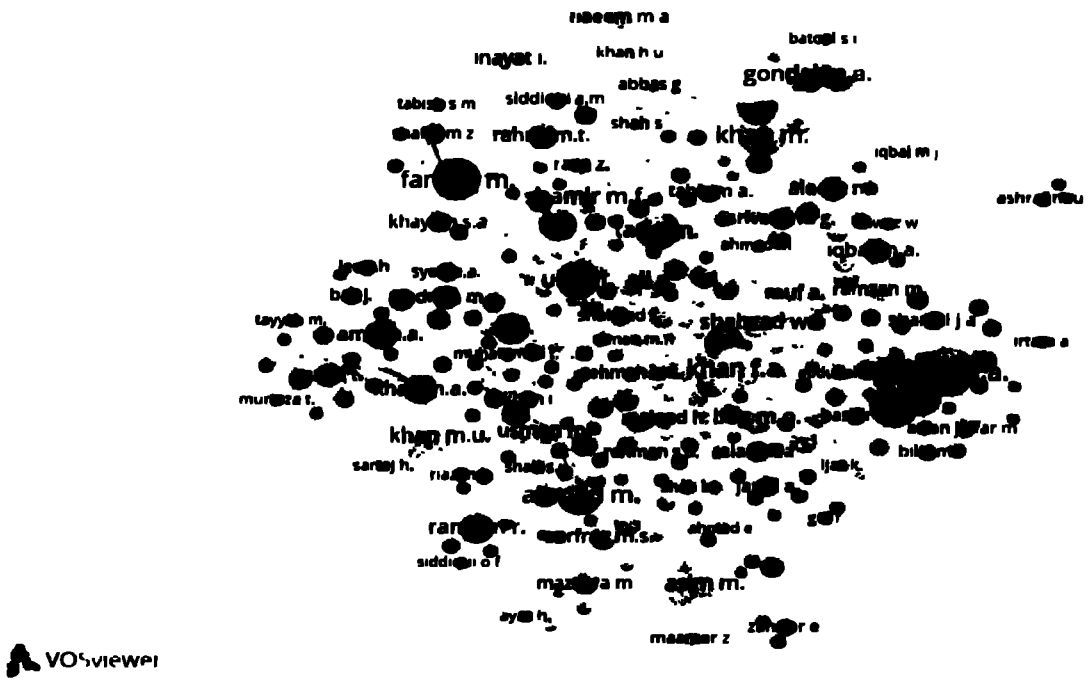



*Figure 4.8: Co-Authorship Graph of Fast University, Islamabad*

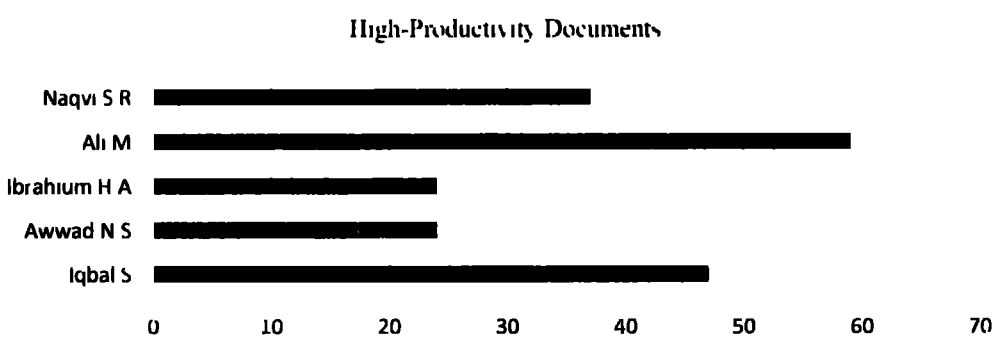



**Figure 4.9 (a) Graphical Representation of High Productivity Documents of Fast University**

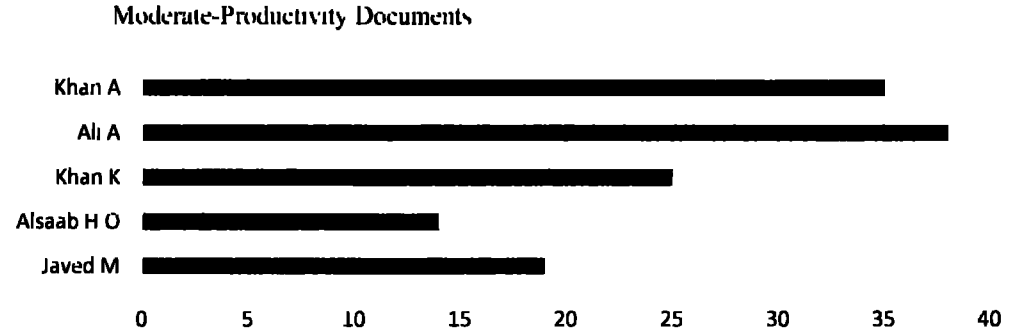



**Figure 4.9 (b) Graphical Representation of Moderate Productivity Documents of Fast University**

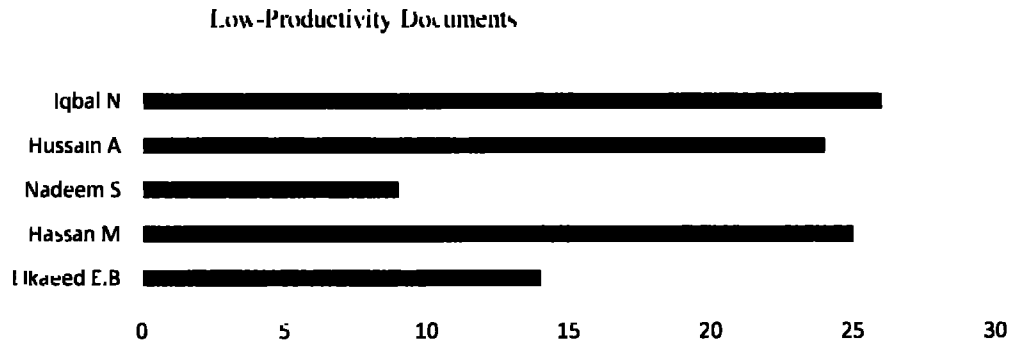



**Figure 4.9 (c) Graphical Representation of Low Productivity Documents of Fast University**

## 4.15 International Islamic University, Islamabad

An analysis was conducted on the publication records of several authors, considering both the documents they authored and the "Total link strength" metric. Among the authors, T. Mahmood stood out with the maximum number of documents, having authored 128 publications. Additionally, Mahmood exhibited an impressive total link strength of 290I, indicating strong connections and collaborations with other researchers in the network. Following Mahmood, Z. Ali secured the second position, having authored 98 documents and possessing a total link strength of 219. Similarly, M. Arshad secured the third position with 75 documents and cumulative link strength of 208. I. Ahmad contributed 91 documents with a total link strength of 195, while N. Ali authored 75 documents with a total link strength of 180. K. Ullah produced 57 documents with a

strength of 166, and A. Hussain had 64 documents with a total A. Ali contributed 42 documents with a total link strength of 131, while S. Khan exhibited a total link strength of 129. R. Ellahi authored 63 documents with a total link strength of 122, and A. Zeeshan contributed 55 documents with a total link strength of 113. Additionally, M. Ahmad had 40 documents with a total link strength of 111, while B. Uzair had 24 documents and a cumulative total link strength of 98. M.M. Bhatti contributed 35 documents with a total link strength of 90, and A. Ghani had 33 documents with a cumulative total link strength of 83. M. Ali and M.S. Khan shared total link strengths of 82, with 33 and 26 documents, respectively. Furthermore, A. Irshad authored 30 documents with a total link strength of 81, while A. Khan and A.A. Khan contributed 27 and 39 documents, respectively, with cumulative link strengths of 81.

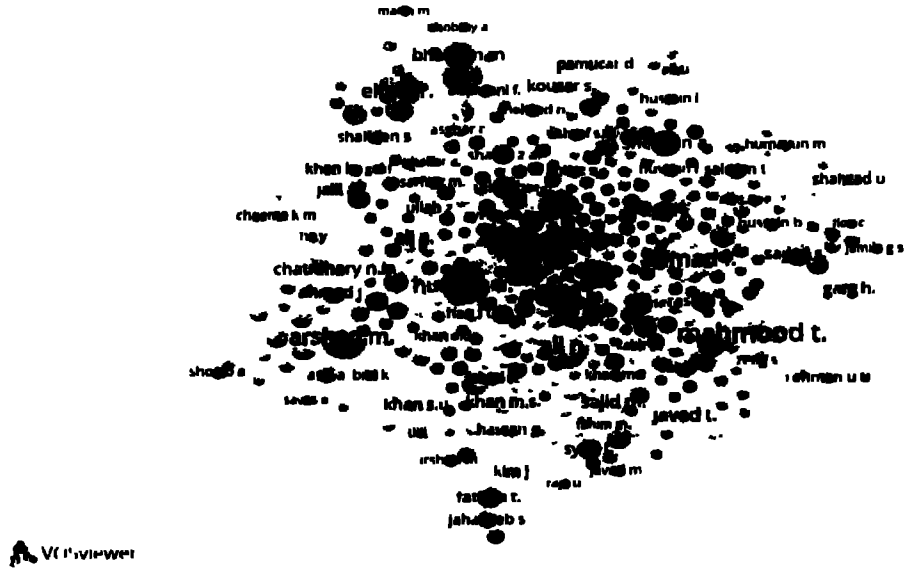



*Figure 4.10: Co-Authorship Graph of International Islamic University*

Based on the provided authors' data, documents, and total link strength, we can analyze the information to identify potential clusters or groups of authors. Clustering can help reveal patterns or associations among the authors based on their publication records and collaborative connections. Here is a clustering based on the provided data:

***Table 4.13 : Collaborative Cluster Based on Productivity of International Islamic University***

| High-Productivity | | Moderate-Productivity | | Low-Productivity | |
|---|---|---|---|---|---|
| **Authors** | **Documents / Total Link Strength** | **Authors** | **Documents / Total Link Strength** | **Authors** | **Documents/ Total Link Strength** |
| **Mahmood T** | 128 | Ali A. | 42 | Uzair B | 24 |
| **Ali Z** | 98 | Khan S | 40 | Bhatti M.M | 35 |
| **Arshad M** | 75 | Ellahi R | 63 | Ghani A | 33 |
| **Ahmad I** | 91 | Zeeshan A | 55 | Ali M | 33 |

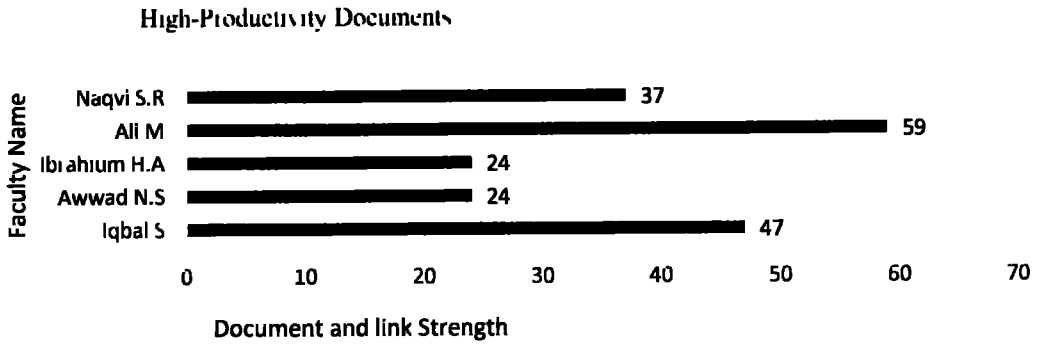



**Figure 4.11 (a) Graphical Representation of High Productivity Documents of IIUI University**

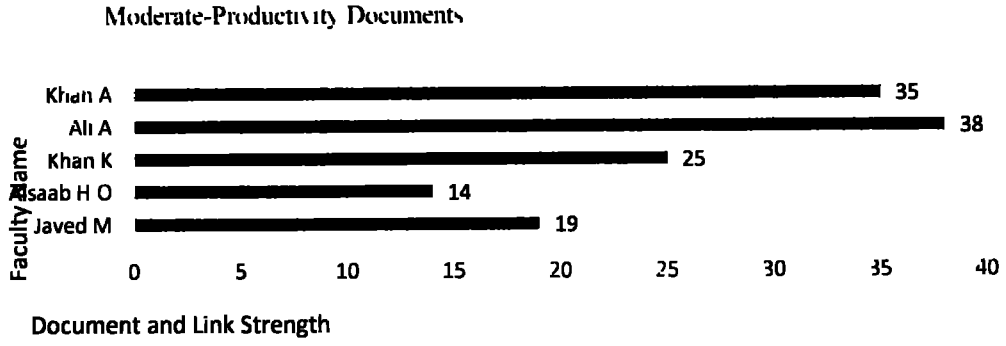



**Figure 4.11 (b) Graphical Representation of Moderate Productivity Documents of IIUI University**

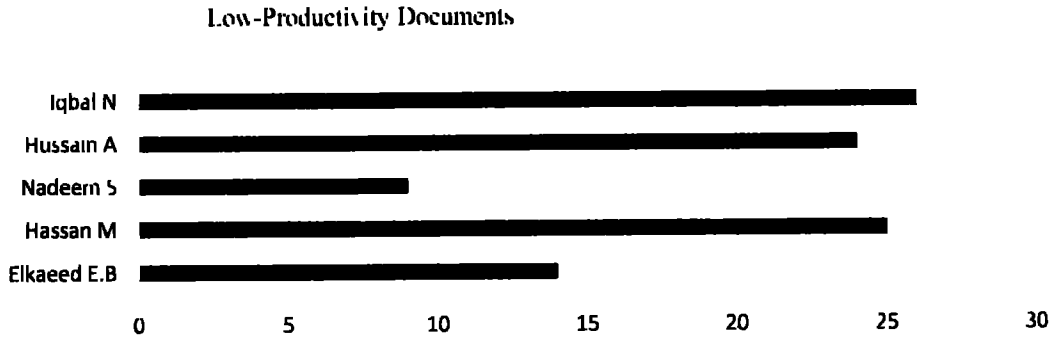



**Figure 4.11 (c) Graphical Representation of Low Productivity Documents of IIUI University**

## 4.16 National University of Science and Technology Islamabad

An analysis was conducted on the publication records of several authors, considering both the higher number of documents and a greater total link strength metric. Among the authors, S. Iqbal stood out with the most documents, having authored 47 publications. Additionally, Iqbal exhibited a substantial total link strength of 217, indicating strong connections and collaborations with other researchers in the network. Following Iqbal, N.S. Awwad and H.A. Ibrahim ranked second with 24 documents each and a total link strength of 162. Similarly, M. Javed contributed 19 documents with a total link strength of 131, while M. Ali authored 59 documents with a total link strength of 123. E.B. Elkaeed and H.O. Alsaab shared a total link strength of 105 and 97, respectively, with 14 documents each. S.R. Naqvi contributed 37 documents with a total link strength of 94, while Z. Said had 60 documents with a total link strength of 93. M.A. Khan and A.H. Khoja shared a complete link strength 90, with 33 and 27 documents, respectively. K. Khan authored 25 documents with a complete link strength of 86. Additionally, A. Ali contributed 38 documents with a link strength of 84, while A. Khan had 35 documents with a complete link strength of 80. M. Imran authored 40 documents with a complete link strength of 76, and M. Hassan contributed 25 documents with a total link strength of 73. M. Iqbal had 34 documents with a whole link strength of 72, while S. Nadeem authored 9 documents with a complete link strength of 69. Moreover, A. Hussain contributed 24 documents with a total link strength of 68, and N. Iqbal had 26 documents

with a total link strength of 67.

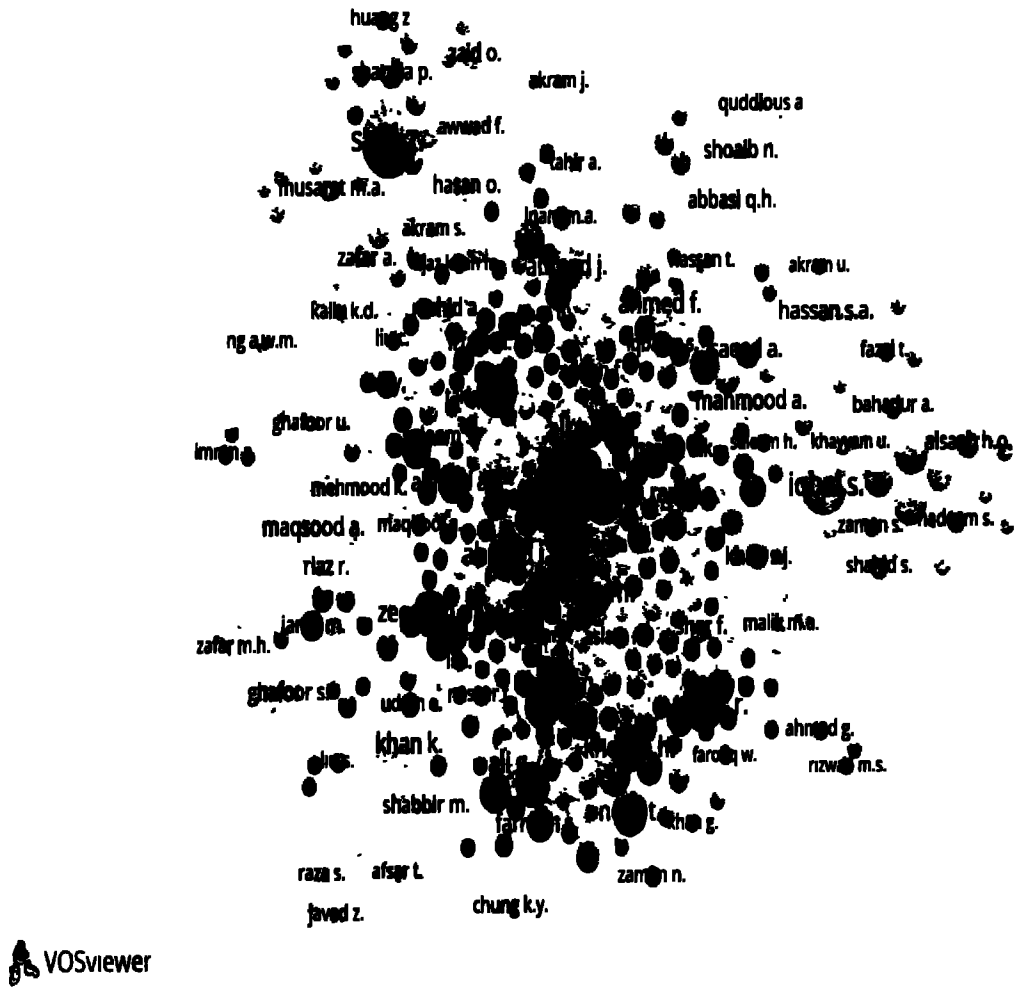



***Figure 4.12: Co-Authorship Graph of NUST***

Based on the provided data of authors, documents, and aggregate link strength, we can analyze the information to identify potential clusters or groups of authors. Clustering can help reveal patterns or associations among the authors based on their publication records and collaborative connections.

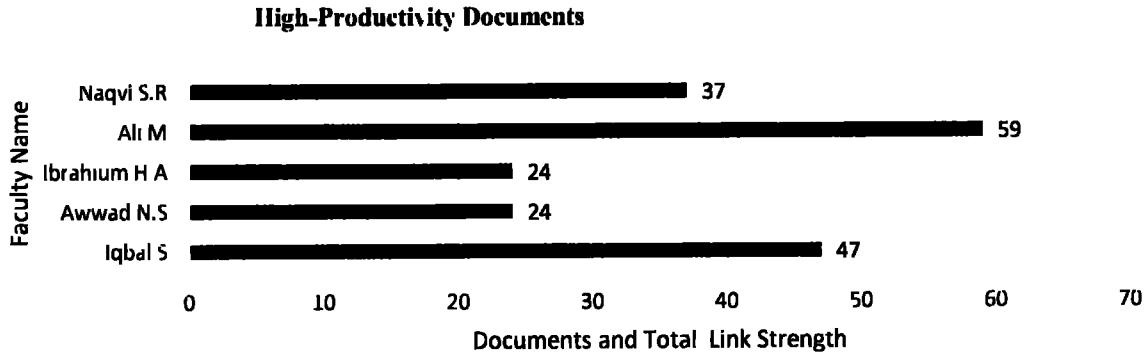



**Figure 4.13 (a) Graphical Representation of High Productivity Documents of NUST University**

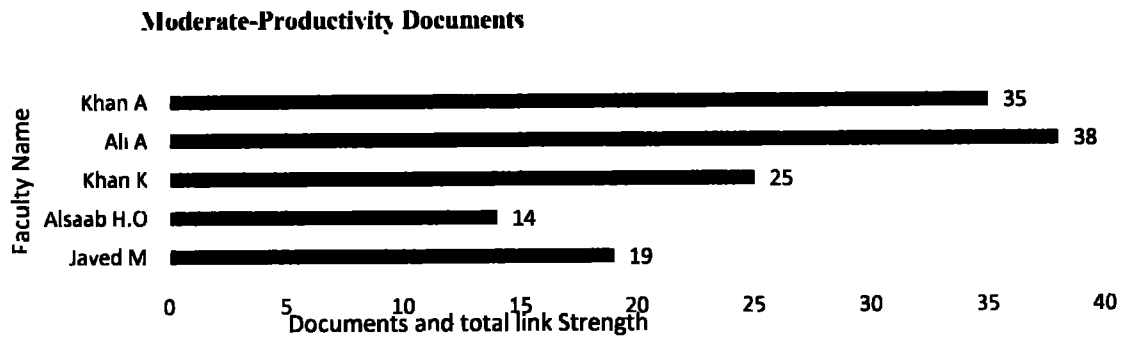



**Figure 4.13 (b) Graphical Representation of Moderate Productivity Documents of NUST University**

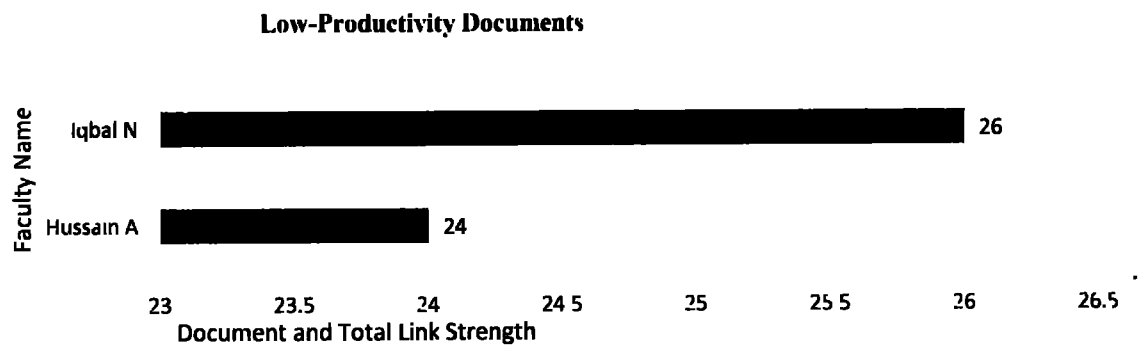



**Figure 4.13 (c) Graphical Representation of Low Productivity Documents of NUST University**

*Table 4.14: Collaborative Cluster Based on Productivity of NUST University*

| High-Productivity | | Moderate-Productivity | | Low-Productivity | |
|---|---|---|---|---|---|
| **Authors** | **Documents / Total Link Strength** | **Authors** | **Documents / Total Link Strength** | **Authors** | **Documents/ Total Link Strength** |
| Iqbal S | 47 | Javed M | 19 | Elkaced E. B | 14 |
| Awwad N. S | 24 | Alsaab H. O | 14 | Hassan M | 25 |
| Ibrahim H. A | 24 | Khan K | 25 | Nadeem S | 9 |
| Ali M | 59 | Ali A | 38 | Hussain A | 24 |
| Naqvi S. R | 37 | Khan A | 35 | Iqbal N | 26 |

## 4.17 Analysis of Universities their clusters, Links and Authors

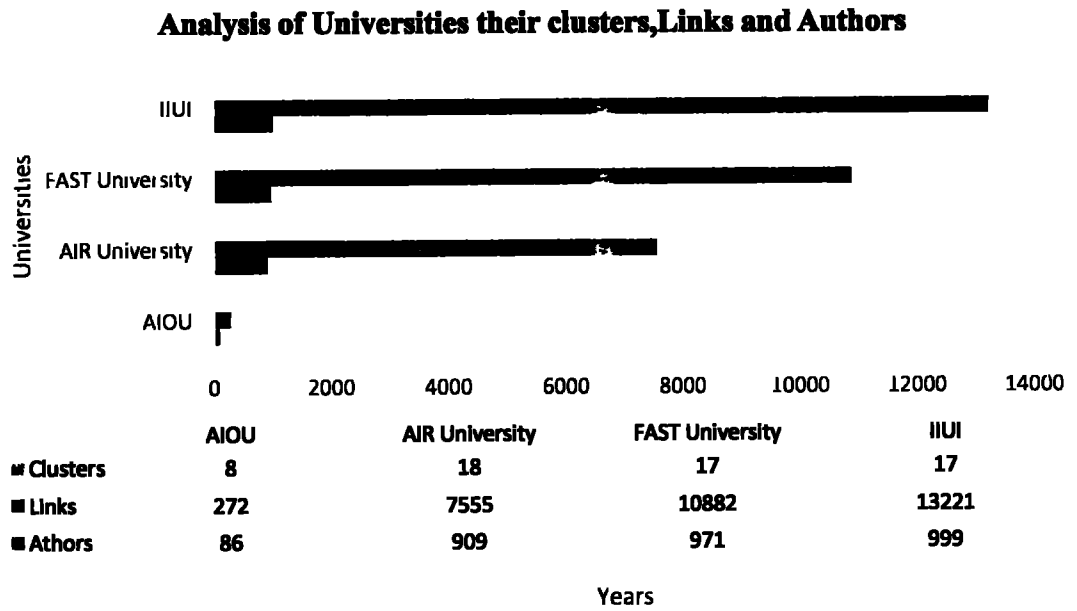



*Figure 4.14: Analysis of Universities their clusters, Links and Authors*

In the realm of academia, various educational institutions contribute significantly to the intellectual landscape through their research and scholarly endeavors. Cluster 8, comprising 272 links, represents one such distinctive node, bringing together authors who contribute to a diverse range of topics. Similarly, the AIR cluster, with 86 links, underscores the importance of collaboration and knowledge dissemination among professionals in the field. Moving ahead, the FAST cluster, encompassing 17 links and a remarkable 10,882 authors, signifies the extensive reach of research conducted at this institution. Equally noteworthy, the IIUI cluster, spanning 17 links and an impressive 13,221 authors, showcases the extensive scholarly network associated with the university. Overall, these clusters and their associated links and authors exemplify the profound impact of academic collaboration and research output within these esteemed institutions.

# Chapter 5

# A Multivariate Stochastic Model to Assess Research Performance of Degree Awarding Institutes in Pakistan

## 5.1. Introduction:

Research is critical to the growth of a knowledge-based economy, since it drives innovation and boosts global competitiveness. Pakistan's higher education industry has grown significantly in recent years, with an increasing number of degree awarding institutes (DAIs) contributing to the research environment. Evaluating these institutions' research success is critical for politicians, educators, and stakeholders to efficiently allocate resources, create strategic goals, and track progress. In recent years, the higher education environment in Pakistan has seen substantial modifications, with an increasing emphasis on research and innovation. Degree awarding institutes (DAIs) play an important role in developing a knowledge-driven economy through advancing research, technology, and human capital. These institutes' research output is critical not just for the country's economic progress, but also for the country's capacity to compete worldwide and address significant societal concerns. However, the higher education system in Pakistan faces various problems that have an influence on research performance. These difficulties include a lack of financial resources, insufficient research infrastructure, a paucity of competent academic members, and a lack of incentives for research participation. Despite these limits, numerous DAIs in Pakistan have produced outstanding research and made significant contributions to their respective disciplines. Evaluating the research performance of Pakistani DAIs is critical in this setting for a variety of reasons. For starters, it assists policymakers, educators, and other stakeholders in understanding the present condition of the country's research environment. This knowledge may be used to guide strategic planning and resource allocation choices, allowing for the creation of tailored interventions to improve research performance. Second, evaluating research performance can aid in the identification of areas of strength and weakness in the higher

education sector. This data may be utilized to create customized capacity-building programs and boost national and international relationships across institutes. Finally, assessing research performance can help to foster a competitive research culture among Dais. Institutes can be pushed to consistently enhance their research production and impact by creating benchmarks and promoting a competitive spirit. Given the significance of evaluating research performance, a strong and dependable assessment system that can account for the inherent uncertainty and unpredictability associated with research activities is required. This work tries to fill that void by building a multivariate stochastic model that provides a full and nuanced view of research performance in Pakistan's degree-granting institutes.

## 5.2. Key objectives of the Multivariate Stochastic Model

The fundamental goal of this project is to create a multivariate stochastic model to evaluate the research performance of degree-granting institutes in Pakistan. The model is intended to provide a complete and rigorous evaluation framework by including a variety of quantitative and qualitative indicators that represent the many elements of research success across disciplines.

1. To account for the inherent uncertainty and randomness in research activities, offering a more reliable and accurate assessment of research performance compared to traditional deterministic approaches.
2. To facilitate a comparative analysis of research performance among Pakistani DAIs, enabling stakeholders to identify areas of strength and weakness within the higher education sector, and foster a competitive research culture.
3. To provide insights into the research landscape in Pakistan, helping policymakers, educators, and stakeholders make informed decisions regarding resource allocation, strategic planning, and capacity-building initiatives.
4. To evaluate the research performance of Pakistani DAIs across various disciplines, including natural sciences, social sciences, engineering, and humanities, highlighting discipline-specific trends and challenges.

5. To benchmark the research performance of Pakistani DAIs against international standards, offering a global context for the assessment and identifying opportunities for collaboration and knowledge exchange with international partners.

## 5.3. Scope of the Multivariate Stochastic Model

This study encompasses the application of the multivariate stochastic model to evaluate the research performance of degree awarding institutes in Pakistan. The study focuses on public and private universities, institutes, and colleges offering undergraduate and graduate degrees across the country. The key areas covered within the scope of the study are as follows:

### 1. Disciplinary Focus

The model will be applied to assess research performance across a wide range of disciplines, including natural sciences, social sciences, engineering, and humanities. This will enable a comprehensive understanding of the research landscape in Pakistan and help identify discipline-specific trends, challenges, and opportunities.

### 2. Quantitative and Qualitative Indicators

The study will incorporate various quantitative and qualitative indicators that capture diverse aspects of research performance. These may include research output (e.g., publications, patents), research quality (e.g., citation metrics, journal impact factors), research funding, collaborations, and research impact on society, among others.

## 5.4 Data Collection and Sources

The study will utilize data from multiple sources, such as higher education institutions, government databases, research funding agencies, and global databases like Scopus and Web of Science. This will help ensure the accuracy and reliability of the model and its results.

**1.National and International Comparison**

The research performance of Pakistani DAIs will be analyzed within the national context, enabling comparisons among different institutes and disciplines. Additionally, the study will benchmark the research performance of Pakistani DAIs against international standards, providing a global context for the assessment and identifying potential opportunities for collaboration and knowledge exchange with international partners.

**2.Policy Implications and Recommendations:**

The findings of the study will offer valuable insights for policymakers, educators, and stakeholders, informing strategic planning, resource allocation, and capacity-building initiatives aimed at enhancing the research performance of degree awarding institutes in Pakistan.

## 5.5 Multivariate Stochastic Models

Multivariate stochastic models have emerged as a promising approach for research performance assessment, as they can account for the inherent uncertainty and randomness in research activities, offering a more accurate and reliable assessment of research performance [44]. These models can incorporate multiple indicators and consider their interdependencies, providing a comprehensive evaluation of research performance across different dimensions. This allows for a more nuanced understanding of research performance, capturing the complex interactions between various aspects of research activities. In the context of degree awarding institutes in Pakistan, the development of a multivariate stochastic model addresses the need for a robust and reliable evaluation framework that captures the complexity and multidimensionality of research performance. By incorporating various quantitative and qualitative indicators, the model provides a comprehensive assessment of research performance, taking into account the uncertainty and randomness inherent in research activities. This approach can help inform policy decisions, allocate resources effectively, and foster a competitive research culture among degree awarding institutes in Pakistan.

### 5.5.1 Types of Stochastic Models

Stochastic models are mathematical models that incorporate randomness or uncertainty to describe the behavior of a system over time. Here are some common types of stochastic models:

1. **Markov Chains:** A Markov chain is a sequence of events where the future state depends only on the current state, and not on the sequence of events that preceded it. It's widely used in various applications, including finance, physics, and biology [132].

2. **Stochastic Processes:** Stochastic processes are mathematical models that describe the evolution of random variables over time. Examples include Brownian motion, Poisson processes, and birth-death processes [133].

3. **Monte Carlo Simulation:** Monte Carlo simulation involves generating a large number of random samples to estimate complex systems or perform numerical integration. It's widely used in finance, physics, and engineering [134].

4. **Queueing Theory:** Queueing theory deals with the study of waiting lines, such as those found in telecommunications, transportation, and customer service systems. Models often involve randomness in arrival times and service times [135].

5.**Stochastic Differential Equations:** Stochastic differential equations incorporate random noise into differential equations to model systems affected by uncertainty. They're commonly used in financial mathematics and physics [136].

6.**Discrete Event Simulation:** This involves modeling systems where events occur at distinct points in time. It's used to analyze complex systems with random inputs, such as manufacturing processes and computer networks [137].

**7.Hidden Markov Models:** Hidden Markov models involve observing a sequence of outputs that are influenced by an underlying sequence of hidden states. They're used in speech recognition, bioinformatics, and natural language processing [138].

**8.Random Walk Models:** A random walk is a mathematical formalization of a path that consists of a succession of random steps. It's often used to model stock prices, particle motion, and diffusion processes [139].

**9.Gaussian Processes:** Gaussian processes are used in machine learning and statistics to model functions where observations are subject to uncertainty. They're used in regression, interpolation, and optimization problems [140].

**10.Agent-Based Models:** Agent-based models simulate the actions and interactions of individual agents within a system. These models are used to study complex systems in economics, social sciences, and ecology [141].

**11.Spatial Stochastic Models:** These models incorporate randomness into spatial systems, such as population dynamics in ecology or the spread of diseases [142].

**12.Renewal Theory:** Renewal theory studies the times between consecutive events in a stochastic process, such as arrivals at a queue or failures in a system [143].

## 5.6. Multivariate Stochastic Models in Higher Education

Multivariate stochastic models have gained prominence in higher education research performance assessment due to their ability to account for the inherent uncertainty and randomness associated with research activities. These models provide a more accurate and reliable assessment of research performance compared to traditional deterministic approaches [44].

## 5.7. Advantages of Multivariate Stochastic Models

**1. Flexibility:**

Multivariate stochastic models can incorporate a wide range of quantitative and qualitative indicators, allowing for the assessment of research performance across multiple dimensions. This flexibility enables a comprehensive understanding of research performance, capturing the complex interactions between various aspects of research activities [144].

**2.Uncertainty and randomness:**

By accounting for the inherent uncertainty and randomness in research activities, multivariate stochastic models offer a more reliable and accurate assessment of research performance. This feature allows stakeholders to make informed decisions based on realistic expectations and accurate evaluations [145].

**3. Interdependencies:**

Multivariate stochastic models can consider the interdependencies between different indicators, providing a more nuanced understanding of research performance. This approach allows researchers to explore the relationships between various aspects of research activities and their impacts on overall performance [146].

**4.Comparability:**

Multivariate stochastic models facilitate the comparison of research performance across institutions, disciplines, and time periods, enabling stakeholders to benchmark progress, identify trends, and recognize areas of strength and weakness [147].

## 5.8. Applications in Higher Education Institutes

Multivariate stochastic models have been applied to various problems in higher education, including:

**1. Performance evaluation:**

Assessing research performance of higher education institutions, departments, and individual researchers, considering factors such as research output, quality, impact, collaboration, and funding [148].

**2. Resource allocation:**

Guiding the allocation of resources, such as funding and infrastructure, based on performance assessments and expected outcomes [149].

**3. Forecasting:**

Predicting future research performance and trends, informing strategic planning and decision-making in higher education institutions [150].

**4.Ranking:**

Developing ranking systems for higher education institutions based on comprehensive and robust assessments of research performance [127].By applying multivariate stochastic models to assess research performance in higher education, stakeholders can gain a comprehensive and nuanced understanding of the research landscape, informing policy decisions, resource allocation, and capacity-building initiatives. In the context of degree awarding institutes in Pakistan, the development of a multivariate stochastic model offers a robust and reliable evaluation framework, capturing the complexity and multidimensionality of research performance and addressing the unique challenges and opportunities faced by these institutions.

## 5.9. Research Performance Evaluation in Pakistan

Research performance evaluation has been a growing area of interest in Pakistan, with several studies focusing on assessing the performance of higher education institutions within the country. These studies have evolved over time, moving from simple metrics to

more comprehensive evaluation frameworks, highlighting the unique challenges and opportunities faced by Pakistani institutions.

### 5.9.1. Early Approaches: Simple Metrics

In the initial phase, studies primarily relied on basic metrics such as publication counts and citation analyses to evaluate the research performance of Pakistani higher education institutions (Mehmood & Shafique, 2010; Hameed et al., 2012). While these metrics provided a starting point for understanding the research landscape in Pakistan, they often failed to capture the nuanced aspects of research performance and could not account for the inherent uncertainty and randomness in research activities.

### 5.9.2. Bibliometric and Scientometric Indicators

To address the limitations of simple metrics, researchers began to employ bibliometric and scientometric indicators in the evaluation of research performance in Pakistan. These indicators provided more detailed insights into research output, quality, and impact, allowing for a more comprehensive understanding of the research landscape in the country [44]. However, these approaches still predominantly focused on quantitative aspects, potentially overlooking qualitative factors such as research relevance, collaboration, and societal impact.

### 5.9.3. Comprehensive Evaluation Frameworks

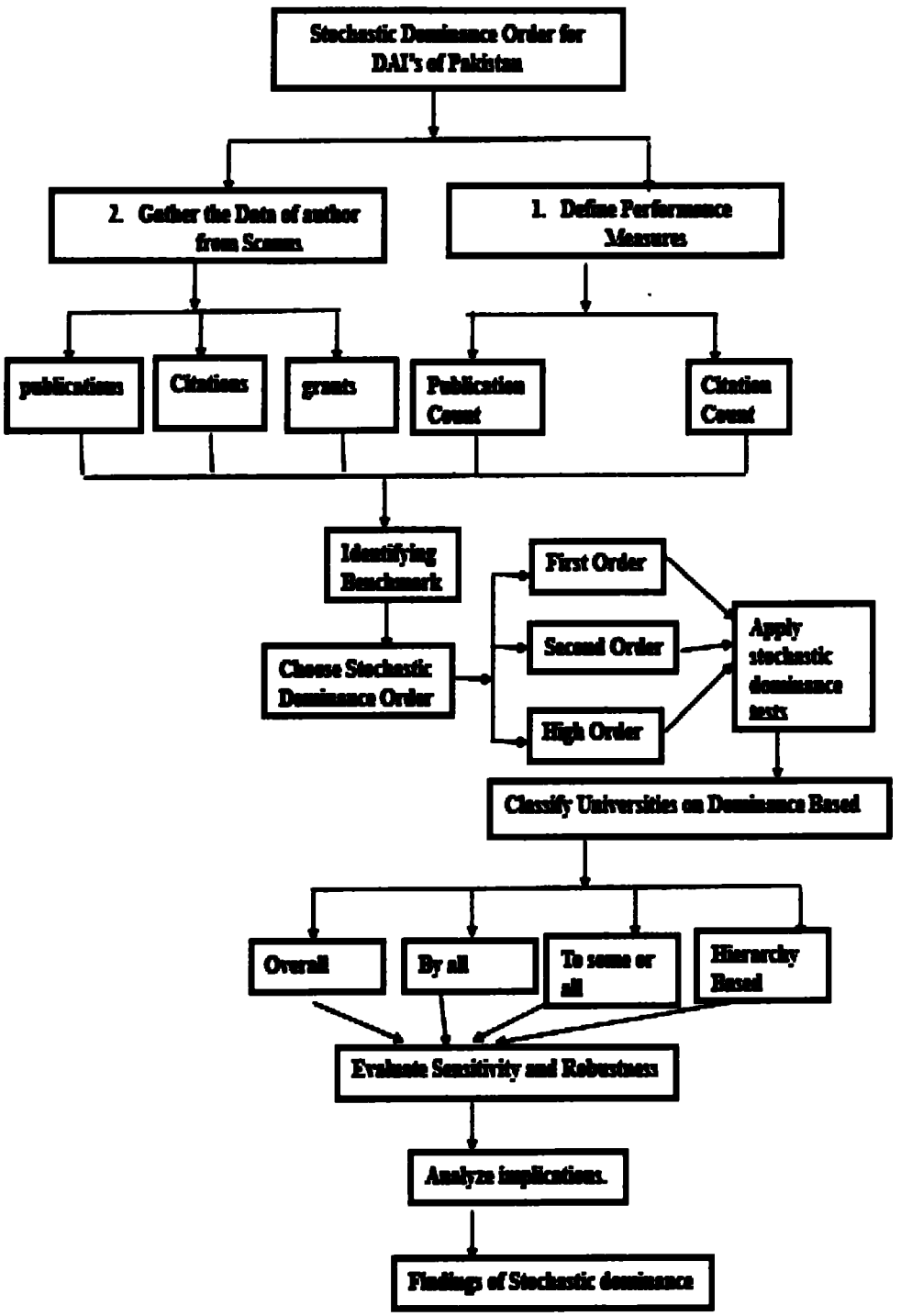
Recognizing the need for a more holistic assessment of research performance, more recent studies have attempted to develop comprehensive evaluation frameworks that incorporate both quantitative and qualitative indicators [44][45]. These frameworks provide a more balanced view of research performance in Pakistan, considering factors such as research output, quality, impact, funding, and collaboration. However, there remains a need for a robust and reliable evaluation model that can account for the uncertainty and randomness inherent in research activities.

### 5.9.4. Multivariate Stochastic Models for Pakistan

The development of a multivariate stochastic model for assessing research performance in Pakistan addresses the need for a comprehensive and reliable evaluation framework. By incorporating various quantitative and qualitative indicators, the model provides a more accurate assessment of research performance, accounting for the inherent uncertainty and randomness in research activities. This approach can help inform policy decisions, allocate resources effectively, and foster a competitive research culture among degree awarding institutes in Pakistan.

## 5.10 Methodology for Developing a Multivariate Stochastic Model

The methodology for developing a multivariate stochastic model to assess research performance in degree awarding institutes of Pakistan involves several key steps, ranging from selecting appropriate indicators to validating the model's results. This section provides an extended overview of the methodology and its various components. The first step in developing the multivariate stochastic model involves selecting a comprehensive set of quantitative and qualitative indicators that accurately capture various aspects of research performance. These indicators may include Publication counts, conference presentations, patents, and books. Citation metrics, journal impact factors, and normalized citation impact. External and internal grants, research budget, and funding per researcher. National and international research partnerships, co-authorship networks, and interdisciplinary collaborations. Research addressing societal needs, technology transfer, and policy influence. Number of PhD students, graduation rates, and faculty qualifications.




**Flow chart of Stochastic Dominance Approach**

## 5.11 Data Collection

Data for the selected indicators will be collected from various sources, including:

1. **Degree awarding institutes:** Institutional databases, annual reports, and research profiles.
2. **Government databases:** Higher Education Commission (HEC) of Pakistan and Pakistan Research Repository (PRR).
3. **Research funding agencies:** National and international funding organizations.
4. **Global databases:** Scopus, Web of Science, and Google Scholar.

## 5.12 Stochastic Model Development

Stochastic dominance models are constructed to compare the research performance of Pakistani universities based on the calculated performance measures. These models allow for a systematic and quantitative assessment of the dominance relationships among the universities. The construction of stochastic dominance models involves the following steps: Explanation of First-Order, Second Order, and Higher-Order Stochastic Dominance Models:

### 5.12.1 First-Order Stochastic Dominance:

In first-order stochastic dominance, one university is considered to dominate another if it has a higher value in at least one performance measure and a lower value in none. It means that for every possible performance level, the dominant university performs at least as well as the other university and strictly better for at least one performance level. First-order stochastic dominance can be represented mathematically using the cumulative distribution functions (CDFs) of two random variables. Let's consider two random variables X and Y, where X represents the performance measure of one university and Y represents the performance measure of another university. The mathematical equation for first-order stochastic dominance is as follows:

$$CDF = X(t) \leq CDF\ Y(t) \text{ for all } t \qquad (5.1)$$

This equation states that the cumulative probability of observing a value less than or equal to t for variable X (CDF_X(t)) should be less than or equal to the cumulative probability of observing a value less than or equal to t for variable Y (CDF_Y(t)) for all possible values of t. In simpler terms, first-order stochastic dominance implies that the CDF of X lies below or is equal to the CDF of Y for all values of t. This indicates that the probability of X being less than or equal to any specific value is always less than or equal to the corresponding probability for Y. In other words, one university's performance measure (X) dominates the other university's performance measure (Y) in terms of the cumulative distribution. It's important to note that the specific form of the CDFs may vary depending on the distribution of the random variables being considered. The equation for first-order stochastic dominance holds regardless of the specific functional form of the CDFs, as long as the inequality $CDF\_X(t) \leq CDF\_Y(t)$ holds for all t.

### 5.12.2 Second-Order Stochastic Dominance:

Second-order stochastic dominance requires that one university dominates another in terms of all performance measures simultaneously. It means that the dominant university performs better in every performance measure or has a superior distribution of outcomes across all measures. Second-order stochastic dominance provides a more stringent criterion for one university to be considered dominant over another. Second-order stochastic dominance provides a more stringent criterion for dominance than first-order stochastic dominance. It requires one random variable to dominate another in terms of all possible linear combinations of cumulative distribution functions (CDFs). Let's consider two random variables X and Y, representing the performance measures of two universities. The mathematical proof for second-order stochastic dominance is as follows:

*For all $\alpha > 0$, the equation*

$$\int [F\_X(t) - F\_Y(t)]\, d\alpha \leq 0 \qquad (5.2)$$

*where F_X(t) and F_Y(t) are the CDFs of X and Y, respectively, and the integration is performed over all α > 0.*

In simpler terms, this equation states that the integral of the difference between the CDFs of X and Y, weighted by α, must be less than or equal to zero for all positive values of α. The proof involves comparing the expected values of linear combinations of random variables X and Y. For any fixed α > 0, consider the random variable Z = αX + (1-α) Y, representing a linear combination of X and Y.

Using the properties of CDFs, we have:

$$\int [F\_X(t) - F\_Y(t)]\, d\alpha = \int [P(X \le t) - P(Y \le t)]\, d\alpha = \int P(X \le t, Z \le t) - P(Y \le t, Z \le t)]\, d\alpha = \int P(Z \le t)\, [P(X \le t \mid Z \le t) - P(Y \le t \mid Z \le t)]\, d\alpha \le \int P(Z \le t)\, [F\_X(t) - F\_Y(t)]\, d\alpha$$

Since P (Z ≤ t) ≥ 0 for all t, and [F_X(t) - F_Y(t)] ≤ 0 (by assumption of first-order stochastic dominance), it follows that.

$$\int P\,(Z \le t)\, [F\_X(t) - F\_Y(t)]\, d\alpha \le 0 \qquad (5.3)$$

Therefore, for all α > 0, the inequality ∫ [FX(t) - FY(t)] dα ≤ 0 holds, establishing second-order stochastic dominance. This mathematical proof demonstrates that second-order stochastic dominance is a stricter criterion for dominance than first-order stochastic dominance. It considers all possible linear combinations of the CDFs and requires the integral of their differences to be non-positive for all positive weights. By applying this proof, researchers can determine if one university's performance measure dominates another's across all possible linear combinations of performance measures.

### 5.12.3 Higher-Order Stochastic Dominance:

Higher-order stochastic dominance extends the concept to higher levels of comparison. Higher-order stochastic dominance models allow for comparisons that consider multiple

levels of performance measures and their distributions. They assess dominance relationships across multiple dimensions simultaneously, providing a more comprehensive understanding of the relative research performance of universities. The mathematical proof for higher-order stochastic dominance is as follows: For all possible linear combinations of weights $\alpha\_i \geq 0$, where i ranges from 1 to n, the following inequality holds:

$$\sum [\alpha i * FXi(t) - \alpha i * FYi(t)] \leq 0 \qquad (5.4)$$

for all t, where F_Xi(t) and F_Yi(t) are the CDFs of the ith performance measure for X and Y, respectively. This inequality states that the sum of the differences between the weighted CDFs of X and Y must be less than or equal to zero for all t and all possible weight combinations αi. The proof involves considering the joint distribution functions (JDFs) of the random variables X and Y for each performance measure. By comparing the JDFs, we can establish higher-order stochastic dominance.

*Let J_X and J_Y represent the JDFs of X and Y, respectively. The joint distribution can be represented as* $J\_X(t\_1, t\_2, ..., t\_n) = P(X\_1 \leq t\_1, X\_2 \leq t\_2, ..., X\_n \leq t\_n)$, *where X_i represents the ith performance measure of X.*

Using the properties of JDFs, we have:

$\sum [\alpha\_i * F\_Xi(t) - \alpha\_i * F\_Yi(t)] = \sum \alpha\_i * [P(X\_i \leq t) - P(Y\_i \leq t)] = \sum \alpha\_i * P(X\_i \leq t, Y\_i > t)$

*Since* $\alpha\_i \geq 0$ *and* $P(X\_i \leq t, Y\_i > t) \geq 0$ *for all i and t, it follows that the sum* $\sum \alpha\_i * P(X\_i \leq t, Y\_i > t) \leq 0$.

Therefore, the inequality $\sum [\alpha\_i * F\_Xi(t) - \alpha\_i * F\_Yi(t)] \leq 0$ holds for all t and all possible weight combinations α_i, establishing higher-order stochastic dominance. This mathematical proof demonstrates that higher-order stochastic dominance provides a more comprehensive comparison of the joint distributions of multiple performance measures. It

ensures that the sum of the weighted differences between the CDFs is non-positive across all possible weight combinations and for all values of t. By applying this proof, researchers can determine if one university's performance dominates another's across multiple dimensions of research performance simultaneously.

It is inappropriate to recommend a single performance metric due to the diversity of objectives among research organizations and overtime. However, this does not justify the widespread use of multiple indicators. The present research also presents a collection of five bibliometric indicators. Two of them are in areas where the co-authors' diverse contributions are acknowledged by their respective places in the paper's byline. Some many assumptions and simplifications must be used to evaluate research performance using bibliometric measurement.

As in other studies, a significant difficulty in the current study is the fair accessibility of resources across scientists within a certain profession. Research productivity is commonly regarded as the primary and essential performance indicator in most evaluation exercises. The current study examines this phenomenon by utilizing the FSS metric, which evaluates published works' quantity and quality. Following is the formula of FSS:

$$FSS_R = \frac{1}{t}\sum_{i=1}^{N} \frac{c_i}{\bar{c}} f_i \qquad (5.5)$$

The professor's work history is indicated in the observed period by the letters $t$, $N$, and $c_i$, respectively, while N denotes the number of publications and $c_i$ denotes the number of citations each article has received. The citation distribution means for all referenced works from the same year and subject area as the publication in question is represented by the variable "c." The variable "$f_i$" denotes the fractional contribution of the researcher to the publication labelled as "i". As opposed to the method of "full counting," which is used in this case, the methodology of "fractional counting" of research contributions is used. It is believed to be more in line with the microeconomic theory of production. The

methodology employed in this study allows for the systematic assessment of individual author contributions, considering their respective positions in the byline.

In certain academic disciplines, the fractional contribution is determined by arranging authors in alphabetical order and is mathematically defined as the reciprocal of the total number of authors involved. However, the fractional contribution is assigned different weights in various other disciplines. In life sciences, it is customary, both within Italy and internationally, for authors to indicate their contributions to published research by arranging their names in a specific order within the byline. Within these academic disciplines, the relative importance assigned to individual co-authors is contingent upon their position within the byline and the type of co-authorship involved, namely intra-mural or extra-mural. A publication's first and last authors affiliated with the same university receive 40% of the citations, with the remaining 20% going to the other authors.

When the first two and last two authors of a publication are affiliated with different universities, the citation distribution can be described as follows: According to the data, it is observed that 30% of citations are attributed to the authors listed first and lasted in the publication. Additionally, 15% of citations are ascribed to the authors listed second and second-to-last. The remaining 10% of citations are distributed among all other authors involved in the publication. Abramo et al. (2015) suggest that neglecting to consider the number and arrangement of authors in the paper's byline may result in notable distortions in individual rankings.

Productivity is an essential indicator for assessing the efficiency of production systems. Nevertheless, measuring research excellence is a crucial metric in evaluating performance, as it relates to the ability to generate pioneering findings. As a result, we quantify the number of articles attributed to the top 1% ($HCA_{1\%}$) and 5% ($HCA_{5\%}$) of global publications based on their citation count, concerning each professor. Within the realm of life sciences, it is customary for the initial author's byline entry in a publication to indicate the individual who originated the primary idea and the researcher who made the most substantial contributions to the study and composition. The final author's

position typically denotes the role of team leader in the research project in a corresponding manner. The assignment of primary or final authorship is considered a mark of prestige and is widely acknowledged in the scientific community. Following this, the number of scholarly articles in which a professor assumes the position of either the primary author or the last author is denoted as $First_A$ or $Last_A$, respectively. In general, it can be deduced that the Pakistan context is improbable to yield distorted performance measures due to variable returns to scale, which could be attributed to disparities in university sizes.

## 5.13 Application of Dominance Relationships to Compare Research Performance:

Once the stochastic dominance models are constructed, they are applied to compare the research performance of Pakistani universities. The models identify the dominance relationships between universities based on the calculated performance measures. By employing the dominance relationships, researchers and policymakers can evaluate the relative performance of universities and rank them accordingly.

## 5.14 Formulation of Dominance Rules Based on Calculated Performance Measures:

Dominance rules are formulated based on the performance measures to determine the dominance relationships. These rules define the criteria for one university to dominate another within the stochastic dominance models. For example, in first-order dominance, the rule states that a university A dominates university B if university A has a higher value in at least one performance measure and a lower value in none.

## 5.15 Consideration of Different Levels of Comparison and Data Availability:

The construction of stochastic dominance models considers different levels of comparison based on the research objectives and available data. It allows for comparisons between individual universities, groups of universities, or across specific performance

measures. The availability and quality of data play a crucial role in determining the extent of comparison and the reliability of the stochastic dominance models. Researchers need to ensure that the data collected is comprehensive, accurate, and representative of the research performance of the universities under consideration. By considering different levels of stochastic dominance and formulating specific dominance rules, researchers can obtain a nuanced understanding of the research performance of Pakistani universities. These models allow for a systematic evaluation of dominance relationships, enabling informed decision-making and resource allocation strategies. However, it is essential to interpret the results of stochastic dominance models in conjunction with other qualitative and contextual factors to obtain a comprehensive picture of research performance.

## 5.16 Stochastic Dominance Model Validation

Stochastic dominance models for research performance assessment of HEC Pakistan can be validated through several approaches to ensure the accuracy and reliability of the results. Given the specific context of HEC Pakistan, here are some key validation methods that can be applied:

### 5.16.1 Comparison with Established Rankings:

One way to validate the stochastic dominance model is by comparing its rankings or dominance relationships with established rankings of Pakistani universities. HEC Pakistan itself publishes rankings based on various criteria, such as the Quality Research Rankings (QRR) and the General University Rankings. Comparing the model's rankings with these established rankings can help assess the model's consistency and alignment with existing assessments.

### 5.16.2 Expert Evaluation and Peer Review:

Seeking expert evaluation and peer review from researchers, academicians, and subject-matter experts in the field of higher education and research can provide valuable feedback on the model's methodology and outputs. Experts can evaluate the model's

appropriateness for capturing the nuances of research performance in HEC Pakistan and assess its alignment with the specific goals and objectives of the organization.

### 5.16.3 Empirical Validation through Case Studies:

Empirical validation through case studies can be conducted by selecting a sample of universities from HEC Pakistan and comparing the model's results with their actual research performance. This can involve examining specific indicators, such as publications, citations, grants, or awards, and comparing the model's rankings or dominance relationships with the observed performance of the selected universities. This validation approach helps assess the model's ability to capture the real-world research performance of HEC Pakistan's universities.

### 5.16.4 Longitudinal Analysis:

Validation can also be performed by conducting a longitudinal analysis using historical data. By applying the stochastic dominance model to multiple time periods, researchers can observe the consistency of the model's rankings and dominance relationships over time. If the model consistently identifies universities that maintain their performance levels or exhibit improvements or declines over time, it provides evidence of the model's validity.

## 5.17 Sensitivity Analysis and Robustness Checks:

Sensitivity analysis involves testing the model's sensitivity to changes in input parameters, such as weights assigned to performance measures or threshold values. Researchers can assess how variations in these parameters impact the model's rankings or dominance relationships. Robustness checks can also be performed by employing alternative weighting schemes or modifying the set of performance measures. If the model's results remain stable and consistent under different scenarios, it adds to its validation.

## 5.18 Stakeholder Feedback and User Satisfaction:

Validation can also be achieved through stakeholder feedback and user satisfaction surveys. Collecting feedback from HEC Pakistan officials, university administrators, and

other stakeholders involved in research performance assessment can help evaluate the model's utility, relevance, and effectiveness. Incorporating their feedback into model refinements enhances its validation and ensures it meets the needs and expectations of the intended users. It is important to note that validation is an iterative process, and ongoing feedback and continuous improvement of the stochastic dominance model are necessary to enhance its validity for research performance assessment in the context of HEC Pakistan. By combining multiple validation approaches, researchers can strengthen the reliability and credibility of the model's results.

## 5.19 Analysis and Interpretation of Research Performance

The results of the multivariate stochastic model will be analyzed and interpreted to provide insights into the research performance of degree awarding institutes in Pakistan, considering factors such as disciplinary differences, institutional strengths and weaknesses, and trends over time. The analysis will also involve benchmarking the research performance of Pakistani institutes against international standards, providing a global context for the assessment. Table 5.1 documents the research performance of two departments of Islamabad's five universities through descriptive statistics of core indicators of research performance.

The departments are named: Mathematics and Computer Sciences. The Table consists of three columns. Column 1 lists those core indicators that mainly contribute to professors' research performance. Columns 2 and 3 give the values of the core indicators for the mathematics and computer sciences departments, respectively. The table presents a concise overview of the productivity and citation impact results about the professors affiliated with two distinct academic departments.

The purpose of the productivity indicators, as shown in the upper section of Table 5.1, is to provide a comprehensive overview of the publication output, specifically focusing on the types of documents and the order of authors. The impact indicators in the lower

section of the table consist of two types: The primary data considered in this study include total citations, the proportion of self-citations, and citations per publication.

***Table 5.1: Scientific Performance of Two Departments of Five Universities of Islamabad, Pakistan***

| Indicator | Mathematics | Computer Sciences |
|---|---|---|
| Article | 5351 | 5121 |
| Proceeding Papers | 1143 | 2008 |
| Book | 1 | 4 |
| Book Chapter | 31 | 93 |
| Reviews | 90 | 232 |
| Editorial | 10 | 21 |
| Total Publications | 6626 | 7524 |
| Total documents consisting of articles, proceedings papers and reviews | 6584 | 7361 |
| Duration of Sample | 24 | 24 |
| The mean of Publications per year | 276.1 | 313.5 |
| Impact | | |
| Total citations | 142277 | 165359 |
| Mean of citations per publication | 21.4 | 21.98 |
| The ratio of self-citations to total citations. | 7.8% | 9.93% |
| Mean of h-index | 10.37 | 5.21 |

Note: Table 1 shows the descriptive statistics of the scientific research performance of two departments of the five universities of Islamabad, Pakistan. The research indicators include the core research indicators, total publications during the sample and their impact.

Additionally, the widely utilized h index is also considered. These various types of indicators offer distinct forms of information pertaining to scientific performance. The aggregate count of citations and the average number of citations per publication exhibit minimal deviation from the raw data for the given analysis.

The inclusion of the h index is necessary due to its widespread adoption within the scientific community. Due to the disparate methodologies employed by different impact indicators, there is a potential for conflicting outcomes in their assessment of research performance. We propose using advanced indicators, such as the h-index, as a recommendation. These indicators alone enable a comprehensive and equitable assessment of performance. It is crucial to consider the unique characteristics of each Department and the specific requirements of the evaluation process when interpreting the results, even when utilizing advanced indicators.

Figure 5.2 displays each Department's overall number of publications divided by document type. For further details, please refer to Table 1. It should be noted that the classification of publications by document types, as determined by Thomson Reuters, often deviates from the classification used by journals (Zhang, Wang, & Zhao, 2017). According to Costas, Van Leeuwen, and Van Raan (2010), database producers typically categorize original research findings as "Articles" and extensive literature surveys as "Reviews" upon publication. According to the data presented in Figure 1, publications classified as "Article" are the predominant document type for both departments. Proceedings papers are important in various academic departments, including computer sciences and mathematics. The mathematics department has published significantly more articles (n=7524) than its counterpart department (n=6626).

In conjunction with considerations of authorship and document classification, the timing of the issuance of publications also constitutes a noteworthy aspect of the researchers' assessment. Is there an equal or unequal distribution of publications? Is there a discernible trend in productivity over the years, indicating whether it increases or decreases? As depicted in Figure 2, the variability in article publication among the researchers of the departments examined in this study is evident (also refer to Table 1).

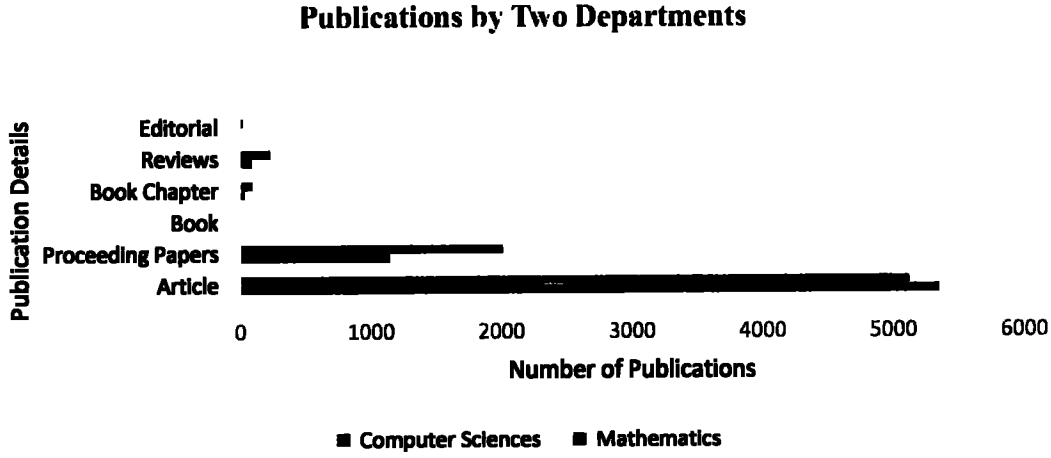



***Figure 5.2: Publications with different Document Types by Two Departments***

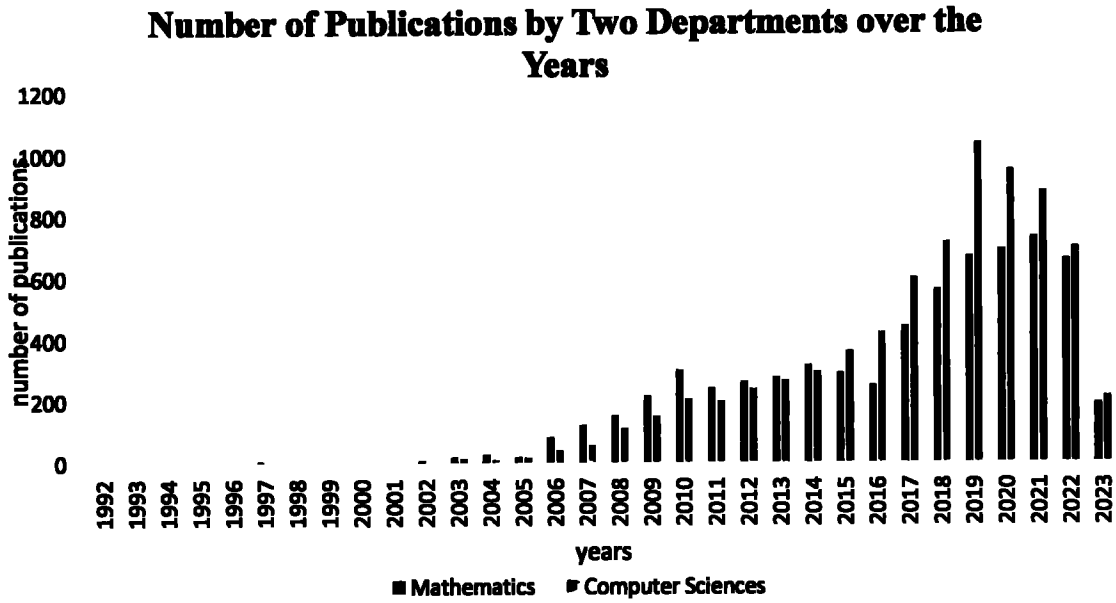



*Figure 2.3: Yearly Publications by Two Departments*

The Department of Computer Sciences attained peak levels of productivity approximately two decades into its academic trajectory, marked by the initiation of its publication endeavors. Subsequently, the Department has consistently produced a substantial volume of scholarly publications. The publications of both departments exhibited an upward trajectory since the commencement of the year, which reached a plateau between 1992 and 2017. Nevertheless, there is a noticeable disparity in publications between the mathematics department and its counterpart.

***Table 5.2: Performance of two Department Based on Publication***

| | **Mathematic *versus* Computer Sciences** | **Computer Sciences *versus* Mathematics** |
|---|---|---|
| **SD Orders** | **KS P-value** | |
| **SD1** | 0.272 | 0.009 |
| **SD2** | 0.382 | 0.003 |
| **SD3** | 0.294 | 0.001 |

**Note: SD of two pairs of publications series of mathematics and computer sciences departments of five universities of Islamabad, Pakistan. SD1, SD2, and SD3 are three p-values of stochastic orders first, second, and third respectively.**

Following the documentation of results through summaries and descriptive analysis, we employed the non-parametric method known as stochastic dominance (SD) to evaluate the performance of both departments with respect to the publications parameter. In this study, the KS test by Barrett and Donald (2003) assesses the statistical significance. Specifically, the first-, second-, and third-order standard deviations (SD1, SD2, and SD3, respectively) are utilized. Table 2 displays the p-values obtained from the Kolmogorov-Smirnov (KS) test. The purpose of this test is to assess the null hypothesis that the target publication series exhibits stochastic dominance over the other series at the $s^{th}$ order of the sample under consideration. The table displays three sets of results indicating the order of p-values for SD1, SD2, and SD3. The statistical significance of the p-values indicates that the null hypothesis, which posits that computer sciences stochastically dominate mathematics, is primarily supported. The findings of this study indicate that the Department of computer sciences exhibits stochastic dominance over other departments in its field. In a broader context, upon examining the p-values of the panel of publications series, it becomes evident that we can reject the null hypothesis that mathematics stochastically dominates computer sciences in all instances of SDs. This rejection is supported by the observation that the p-values exceed significance levels. In an alternative perspective, we posit the hypothesis that the computer sciences department exhibits stochastic dominance over mathematics, as evidenced by all observed values below any significance level.

Table 5.2 presents the descriptive statistics for the distributions of the indicators relative to the two departments under examination. By examining these descriptive statistics, one can acquire valuable insights about the measures of the dataset's central tendency, variability, symmetry, and shape. Such insights are instrumental in comprehending the characteristics and behavior exhibited by the data.

In reference to the Computer Sciences Department, the average FSS (Index) is 1.75, while the middle value, or median, is 1.25. The upper limit of the observed values is 42, while the lower limit is 0. The data set exhibits moderate variability, evidenced by a standard deviation (SD) of 3.2 and a coefficient of variation of 1.23. The obtained

skewness value of 3.29 indicates a positive skewness, implying that the distribution is skewed to the right. The kurtosis value 28.3 also suggests that the distribution has heavy tails and exhibits extreme values.

The average value for $First_A$ is 1.25, and the median, is also 1. The upper limit of the observed values is 21, while the lower limit is 0. The data set's standard deviation (SD) is 1.15, and the coefficient of variation (CV) is 1.41. The skewness value of 3.25 suggests that the distribution is skewed to the right, indicating a longer tail on the right side of the distribution. On the other hand, the kurtosis value 20.78 indicates heavy tails and the presence of extreme values in the distribution.

The average value for $Last_A$ is 2.54, and the median value is 1. The upper limit of the observed values is 84, while the lower limit is 0. The standard deviation (SD) is calculated to be 1.61, while the coefficient of variation (CV) is determined to be 5.9, suggesting a substantial level of variability. The obtained skewness value of 5.12 indicates a highly skewed distribution to the right. Additionally, the calculated kurtosis of 75.15 suggests that the distribution has heavy tails and exhibits extreme values. The average value for $HCA_{1\%}$ is 0.15, similarly, the average value for $HCA_{5\%}$ is 0.67,

Regarding the Mathematics Department, the average FSS (Index) is 2.2; the middle value is represented by a median 1. The upper limit of the observed values is 89, while the lower limit is 0. The standard deviation (SD) is calculated to be 4.27, while the coefficient of variation is determined to be 1.71, suggesting a moderate level of variability. The obtained skewness value of 6.4 indicates a highly skewed distribution to the right. At the same time, the calculated kurtosis of 80.12 suggests the presence of heavy tails and extreme values in the distribution.

The average value for $First_A$ is 1.15, and the middle value is 0. The upper limit of the observed values is 22, while the lower limit is 0. The standard deviation (SD) is 1.2, while the CV is 1.31. The observed skewness value of 3.25 suggests that the distribution

is skewed to the right, while the calculated kurtosis of 25.34 indicates the presence of heavy tails and extreme values.

The average value for $Last_A$ is 2.56, and the middle value is represented by the median, which is 1. The upper limit of the observed values is 42, while the lower limit is 0. The standard deviation (SD) is calculated to be 2.31, while the CV is determined to be 2.71. These values suggest the presence of moderate variability within the data. The obtained skewness value of 4.12 indicates that the distribution under consideration exhibits a right-skewed pattern.

***Table 5.3: Descriptive Statistics of Performance Indicators of two Departments***

| Departments | Indexes | Mean | Median | Max | Min | SD | Variant Coeff | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|---|---|
| **Computer Sciences** | $FSS$ | 1.75 | 1.25 | 42 | 0 | 3.2 | 1.23 | 3.29 | 28.3 |
| | $First_A$ | 1.25 | 1 | 21 | 0 | 1.15 | 1.41 | 3.25 | 20.78 |
| | $Last_A$ | 2.54 | 1 | 84 | 0 | 4.26 | 1.61 | 5.9 | 75.15 |
| | $HCA_{1\%}$ | 0.15 | 0 | 6 | 0 | 0.34 | 2.61 | 5.12 | 60.12 |
| | $HCA_{5\%}$ | 0.67 | 0 | 10 | 0 | 1.23 | 1.59 | 2.15 | 10.12 |
| **Mathematics** | $FSS$ | 2.2 | 1 | 89 | 0 | 4.27 | 1.71 | 6.4 | 80.12 |
| | $First_A$ | 1.15 | 0 | 22 | 0 | 1.2 | 1.31 | 3.25 | 25.34 |
| | $Last_A$ | 2.56 | 1 | 42 | 0 | 3.4 | 2.31 | 2.71 | 15.16 |
| | $HCA_{1\%}$ | 0.2 | 0 | 8 | 0 | 0.58 | 1.41 | 4.12 | 24.54 |
| | $HCA_{5\%}$ | 1.18 | 0 | 30 | 0 | 1 | 2.13 | 4.61 | 55.83 |

Note: **The table shows the descriptive statistics of five core performance indicators of two selected departments**

When examining Figure 5.3, a visual assessment can be conducted to compare the average performance indicators of the two departments. In figure 1 denotes FSS, denotes for $First_A$ 3 denotes for $Last_A$ 4 denotes $HCA_{1\%}$and 5 denotes for $HCA_{5\%}$ for In Computer Sciences, it is evident that the third performance indicator exhibits the highest mean value, followed by the first indicator. In contrast, the fourth indicator demonstrates the lowest mean value. The second and fifth indicators are situated in an intermediate position.

In the field of Mathematics, it can be observed that the initial performance indicator exhibits the highest average value, with the subsequent indicator ranking second in terms of the mean value. The mean values of the second and fifth indicators are comparatively lower, while the fourth indicator exhibits the lowest mean value.

It is important to acknowledge that this interpretation is derived exclusively from the average values presented in the bar chart. The analysis fails to consider additional variables or the wide distribution of the dataset. Examining the bar chart, a visual assessment can be made regarding the average performance indicators of the two departments. In Computer Sciences, it is evident that the third performance indicator exhibits the highest mean value, followed by the first indicator, whereas the fourth indicator demonstrates the lowest mean value. The second and fifth indicators are situated at intermediate positions.

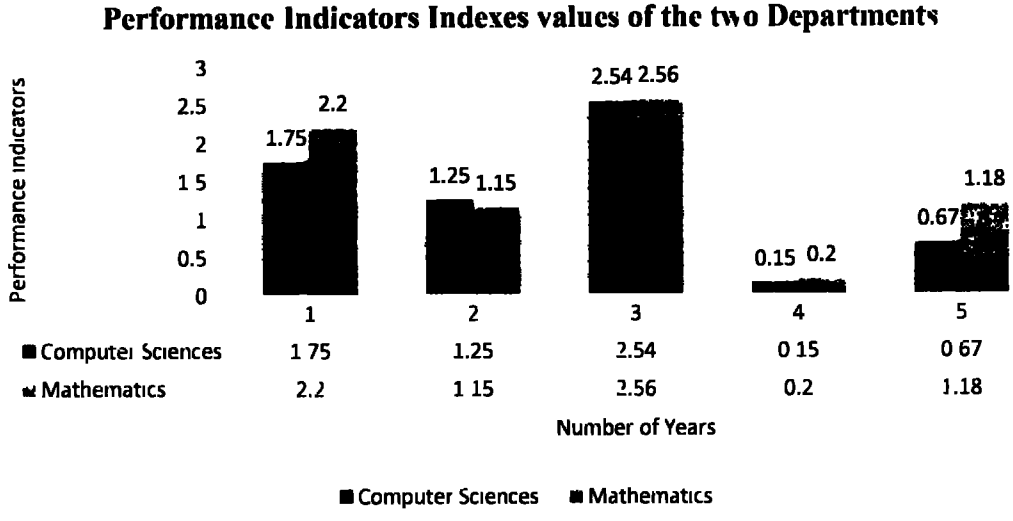



***Figure 5.4: Performance Indicators Indexed values of the two Departments.***

In the field of Mathematics, it can be observed that the initial performance indicator exhibits the highest mean value, with the subsequent indicator ranking second in terms of magnitude. The mean values of the second and fifth indicators are comparatively lower, while the fourth indicator exhibits the lowest mean value.

Table 5.3 presents descriptive statistics for core performance indexes indicator, specifically citations and h-Index, based on the$HCA_{1\%}$ index. In the Computer Sciences Department context, it is imperative to adhere to proper citation practices. The values in the dataset exhibit a range spanning from 573 to 1729. The data exhibits a rising trajectory in the number of citations, suggesting a future expansion in the research impact of the Department. The h-index values in this range span from 11 to 21. The h-index is a metric that quantifies the maximum number of papers within a dataset with equal or greater citations. The data indicates a moderate distribution of h-index values, suggesting detectable research productivity and influence level.

In the context of the Mathematics Department, the cited values exhibit a range spanning from 1277 to 21260. The data demonstrates a notable disparity in citation counts, suggesting a diverse spectrum of research influence among various publications. The h-index values in this study span a range of 19 to 84. The data indicates that the h-index values in Mathematics are relatively higher than those in the Computer Sciences Department, suggesting a potentially greater research impact and productivity in Mathematics.

**Table 5.4 : Descriptive based on (HCA)_ (1%) Performance Indicators**

| Computer Sciences | |
|---|---|
| **Citations** | **h-Index** |
| **573** | 14 |
| **666** | 15 |
| **676** | 13 |
| **703** | 11 |
| **782** | 16 |
| **1715** | 11 |
| **1729** | 21 |
| **Mathematics** | |
| **Citations** | **h-Index** |
| **1277** | 19 |
| **1865** | 24 |
| **4696** | 35 |
| **7656** | 47 |
| **8287** | 48 |
| **8862** | 54 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | 21260 |
| | | | | | | | 84 |

Figures 5.4 drawn based on Table 5.4. The performance based on the citation score of sampled departments of the top 1% of researchers. Similarly, Figure 5.4 also considered the data of the top 1% of researchers but here we show the performance based on h-index. It is seen that the citation score of the mathematics department is high compared to its counterpart. Similarly, it would be obvious that the h-index is also high for the mathematics department.

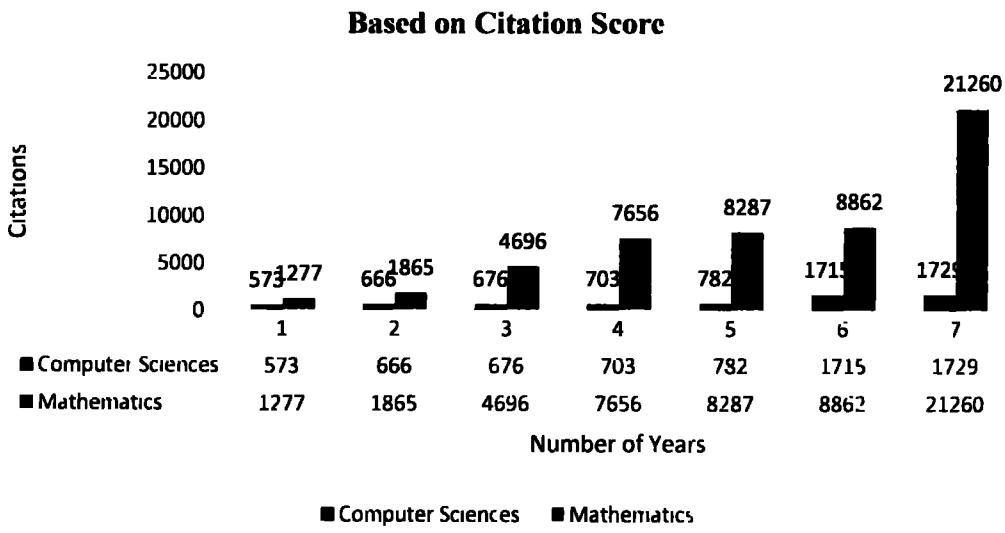



***Figure 5.5: Performance of (HCA)_(1%) based on Citation Score***

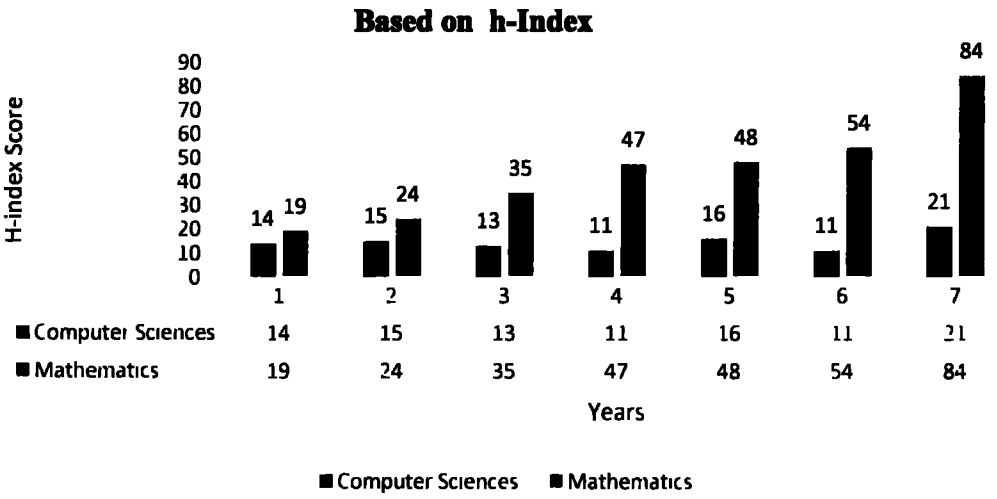



***Figure 5.6: Performance of HCA_ (1%) based on h-Index.***

# Chapter 6

# Learning to Predict Citation-Based Impact Measures: Evidence from Pakistan

## 6.1. Introduction

In today's globally connected academic world, the impact of scientific research is often gauged through several key performance indicators. Among these metrics, citation-based impact measures, such as the h-index, i10-index, and citation counts, have been recognized as reliable indicators of an institution's academic quality and research influence ***(Harzing, 2013).*** These measures provide an insight into the academic recognition and relevance of a scholar or an institution's research output, reflecting the influence they wield in their respective fields ***(Bornmann & Marx, 2015).***

However, the applicability of these citation-based measures and the ability to predict them accurately pose considerable challenges, particularly in the context of developing countries like Pakistan. The evolving nature of the academic and research landscape, the disparate research culture, the limited resources and infrastructure, and the lack of comprehensive data-driven systems are some of the factors contributing to this complexity. The Study aims to delve into this conundrum and develop a predictive model for citation-based impact measures for degree-awarding institutes in Pakistan.

The use of predictive modeling in academic research has emerged as a powerful tool to anticipate future trends based on current and historical data ***(Huang, et.al, 2018).*** Several studies have attempted to predict the citation counts of individual papers or scholars using different models ***(Wang, et al., 2013; Yan, et al., 2011).*** However, the scope of these studies is typically limited to developed nations with well-established research cultures and data-driven academic systems. In contrast, the landscape of academic research in Pakistan presents a distinct set of challenges and opportunities.

In the context of Pakistan, higher education and research have been gaining momentum, with the establishment of Higher Education Commission (HEC) in 2002 aimed at improving the quality of higher education and promoting research culture ***(Rauf, et al., 2017)***. Despite these efforts, the implementation of data-driven approaches in the academic and research sectors remains underexplored. Hence, there is a critical need to develop a predictive model for citation-based impact measures that would allow institutions to proactively enhance their research impact and quality.

The goal of this study is to apply machine learning techniques to create a predictive model that could forecast citation-based impact measures for degree-awarding institutes in Pakistan. The study will consider variables such as annual research output, total faculty size, the number of PhD faculty, funding allocation, and the research collaboration network. By doing so, this study aims to provide actionable insights that institutions can employ in their strategic planning to maximize their research impact.

The present study examines the issue of forecasting the scientific influence of particular authors and papers for up to 10 years in advance. Due to the absence of a universally agreed-upon standard for assessing scientific impact, we adhere to previous research [1] and employ the h-index to evaluate the effect of authors, while citation counts are utilized to gauge the impact of papers. The effectiveness of our approaches is evaluated using a dataset of about 35666 computer science publications authored by approximately 28825 individuals and published between 1908 and 2018. The data set, made accessible to the public, possesses distinctive characteristics in size, surpassing existing data sets by more than tenfold, and its scope.

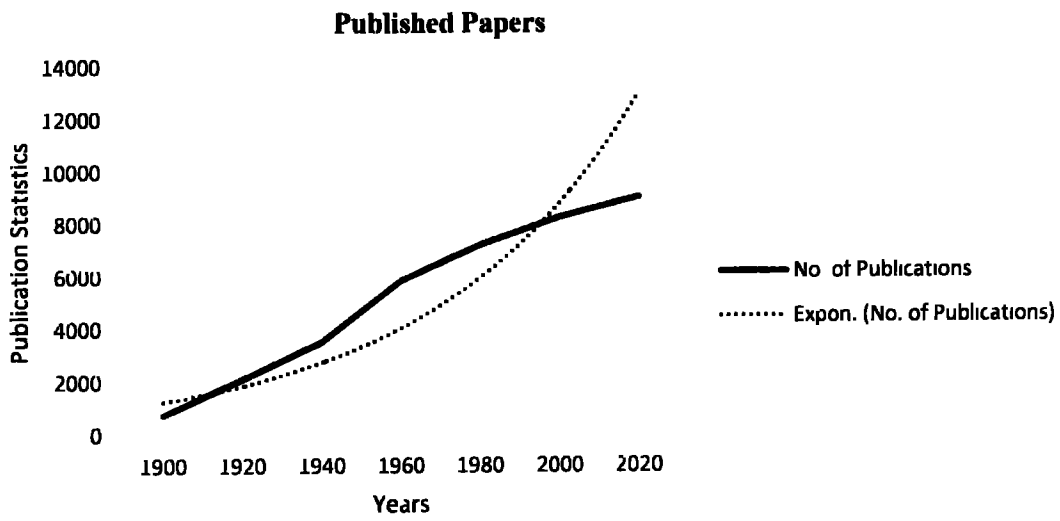



**Figure 6.2: Our dataset's published papers over time. The exponential fit predicts a huge increase in papers.**

In our predictive analysis, we utilize data from 1908 to forecast the influence during the coming decade from 2009 to 2018. Due to their inherent simplicity and widespread usage, citation counts have emerged as a widely adopted metric for assessing the effect of scientific research. Although citation counts are commonly used as a straightforward and practical measure of effect, they include certain limitations, particularly when employed to evaluate writers. An often voiced critique of citation counts pertains to their inability to account for the distribution of citations across a researcher's body of work. For example, it may be anticipated that an author with 60 citations would have a greater influence if those citations were evenly distributed throughout six papers rather than dispersed among 30 papers.

Similarly troublesome is the scenario where an author's career is brief, consisting of only one highly cited paper, such as a survey or multidisciplinary publication. In this case, the author may appear to have a greater influence than a researcher with a lengthy track record of somewhat significant publications. Recognizing these deficiencies, as initially advocated by [2], has prompted a recent surge in developing alternative indicators to assess impact. Several indicators that can be used to evaluate scholarly impact include the h-index, g-index, c-index, eigen factor, and hip-index [3-5]. The h-index is a quantitative measure employed to evaluate the impact of an author. It is determined by selecting the highest value of N, where N is the number of articles

authored by an individual, and each article has garnered N citations or more. It is important to acknowledge that the h-index is not influenced by outliers, as the inclusion of a single article can only increase the h-index by one unit. Moreover, the h-index applies a penalty to numerous papers that receive few citations, as a publication only contributes to the h-index if it obtains a substantial number of citations.

Quantifying scientific influence in the present is a topic of considerable interest. However, there are numerous issues where the significance of the future outweighs that of the present, such as the matter of awarding tenure. Furthermore, considering the rapid proliferation of scholarly articles, as depicted in Figure 6.2 , it is of utmost importance to develop automated mechanisms to identify influential research at the earliest stages. To more effectively tackle these inquiries, there has been a notable increase in scholarly investigations about the anticipation of scientific influence for authors and individual works.

There are two principal methodologies for the anticipation of impact. The initial concept builds upon the scholarly legacy of [6], wherein the statistical modelling of citation counts is informed by the principles of the preferential attachment model of network expansion and empirical investigations [7] The second approach utilizes a machine learning methodology to forecast impact. This involves utilizing considerable feature engineering and implementing supervised learning using a regression [2, 8]. In this study, we conduct a comparative analysis of two distinct ways using our dataset. Additionally, we offer a novel method to address the disparity between these approaches in predicting article citations. It is noteworthy to notice that there have been recent endeavors that do not align precisely with the aforementioned categories. In their study, [9] employ K-Spectral Clustering as a method to detect cluster centroids in author citation histories, while ensuring that the results are not influenced by scaling and shifting. The data generated from these centroids is further integrated with fundamental author-level variables in order to predict author cluster affiliation and future citation counts.

The subsequent sections of this work are structured in the following manner. We provide an overview of the characteristics employed to delineate individual publications and authors. Next, we illustrate how these characteristics can be employed to forecast an author's h-index and the

number of citations their papers receive. In academic articles, a comparison is made between machine learning approaches and those that draw inspiration from probabilistic modeling. It is important to note that there is currently a lack of predictive probabilistic models specifically designed for determining an author's h-index. Consequently, conducting a comparative analysis across authors using such models is impossible. The analysis concludes by examining the features that exhibit the strongest correlation with our prediction objectives, followed by a comprehensive discussion of our findings and potential avenues for further research [7].

### 6.2 Feature Selection

To effectively utilize supervised learning methods in the current study, it is imperative to initially construct a comprehensive set of features that accurately encapsulate the content of individual papers as well as the characteristics of the authors. This study primarily emphasizes extracting features from the citation network, coauthor graph, and paper metadata, such as authors and venue. However, extracting content-based features from the text of papers is deferred to future research. Table 1 contains a comprehensive list of 44 author traits, which have been influenced by previous research [10-12]. The 63 features of papers exhibit similarities and are intentionally omitted from being presented to conserve space. However, a comprehensive description of all these features is provided below. Before proceeding, it is important to acknowledge the existence of many one-to-many links across papers, authors, and venues in Table 6.1.

### 6.3 Meta Data

Certain attributes can be derived from the metadata of a research article with minimal or negligible processing. For example, the author considers many factors, such as the author's citation count, the nature of the work (whether it is a survey or not), and the number of years that have elapsed since the study's publication.

### 6.4 Impact History

Several key indicators of future influence can be derived directly from the temporal patterns of citation counts and h-indices for particular works and authors. These features encompass the cumulative number of citations, the annual variation in citation rate, and the average citation rate

over an extended period. Additionally, we consider the historical significance of venues by extracting similar attributes for venues as we do for writers. We then compile this information for specific articles and authors.

### 6.5 Citation and Coauthor Graphs

The citation and coauthor graphs' topology provides valuable insights into the centrality and influence of papers and their respective authors. It is reasonable to anticipate that coauthors will tend to cite each other. Consequently, a high degree within the coauthor graph can indicate a potential future rise in the h-index. According to [12], centrality measurements within the coauthor ship network yield significant indications of prospective achievements. To enhance computational efficiency, our main indicators of centrality consist of the in-degree and out-degree.

### 6.6 Author Impact:

In this study, we initially examine the prediction of the author's h-index using a machine learning methodology. A set of 44 distinct features is generated for each author, and these features are subsequently employed in several regression models. To make predictions for a time horizon of up to 10 years, we extract characteristics for authors who published their initials on or before 20. The author's h-indices observed from 1908 to 2018 are employed as objectives for prediction. It is important to acknowledge that we conduct many training sessions for identical models, each corresponding to one of the ten target years. We adopt the methodology [8] proposed to exclude writers not actively engaged in scholarly work. Specifically, we consider only authors with an h-index of no less than 4. The regression models we employ in our training process are arranged in ascending order of complexity, commencing with basic baselines and concluding with cutting-edge machine-learning algorithms.

#### 1. Plus-k (PK)

The proposed baseline model involves the addition of a constant value to the h-indices of authors on an annual basis. This constant, denoted as 0.402, is determined by linear regression utilizing

the Huber loss function, which is known for its robustness in handling outliers compared to the conventional squared error loss [13].

**2.Simple Markov (SM)**

The proposed model for linear regression incorporates characteristics of the author's h-index.

**3. Lasso (LAS)**

The regularized linear regression model incorporates all features, and the regularization value is determined by a 10-fold cross-validation process [14]

**4. Random forest (RF)** The utilization of randomization approaches in conjunction with an ensemble of regression trees has been proposed as a means to enhance performance [15]

**5. Gradient boosted regression trees (GBRT)** The approach involves employing a series of basic regression trees that are trained in an iterative manner using a variant of functional gradient descent [16]. Gradient-boosted trees have demonstrated high performance; nevertheless, unlike random forests, they often necessitate substantial parameter optimization.

To assess the effectiveness of our forecasts, we consider three performance indicators, which are elaborated upon in the following sections. The initial metric under consideration is widely recognized. $R^2$ measure. The coefficient of determination, denoted as $R^2$, is a statistical measure that assesses the relative effectiveness of a model compared to a predictor that outputs the average value of the labels. To assess our forecasts' effectiveness, we consider three performance measures as outlined in the following section. The initial metric under consideration is widely recognized. $R^2$measure. The coefficient of determination, denoted as $R^2$, is a statistical measure used to assess the relative effectiveness of a model in comparison to a predictor that gives the average value of the labels. Within the framework of our research, the coefficient of determination of CV $R^2$holds the same meaning as:

$$1 - \frac{\sum_{i=1}^{N} (y_{i,j} - \hat{y}_{i,j})^2}{\sum_{i=1}^{N} (y_{i,j} - \bar{y}_j)^2} \qquad (6.1)$$

In this study, $N$ represents the overall count of writers. The variable $y_{i,j}$ denotes the h-index of the $ith$ author in the $jth$ year. On the other hand, $\hat{y}_{i,j}$ represents the anticipated h-index for the author during that specific year. Additionally, $\bar{y}_j = \frac{1}{N}\sum_{i=1}^{N} y_{i,j}$ is calculated as the average h-index across all authors, computed by summing the h-indices of each author and dividing the total by $N$. The variable $N$ represents the average h-index across all authors, denoted as $\hat{y}_{i,j}$. Although $R^2$ is widely used as a metric to evaluate the performance of regression models, it has been observed that it tends to overestimate the predictive capability in the context of impact prediction [17]. The phenomenon of inflation in citation counts and h-indices arises due to their inherent inability to drop and their strong positive correlation, indicating a reliance on their previous year's values. To mitigate the presence of auto-correlation, an adjustment is made to the $R^2$measure by deducting the established count of citations in 2009 from the projected targets in the years 1908-2018.

The Past Adjusted $R^2(\text{PA} - R^2)$ is a new metric that we describe as

$$1 - \frac{\sum_{i=1}^{N} (y_{i,j} - \hat{y}_{i,j})^2}{\sum_{i=1}^{N} (z_{i,j} - \bar{z}_j)^2} \qquad (6.2)$$

where $z_{i,j} = y_{i,j} - y_{i,2005}$ and $\bar{z}_j = \frac{1}{N}\sum_{i=1}^{N} z_{i,j}$. By performing the subtraction of the known amount $y_{i,2005}$ from $y_{i,j}$ in the denominator of the PA- $R^2$equation, the denominator is effectively reduced, eliminating certain deceptive inflation in the statistic. The commonly held belief is that the $R^2$ metric typically falls within the range of 0 to 1. However, it is important to note that this assertion holds true only in certain specific scenarios, such as when calculating the training error using linear regression. Both the coefficient of determination ($R^2$) and the adjusted coefficient of determination PA- $R^2$ are bounded by a maximum value of 1, inclusive, although they can also take on negative values.

The $R^2$ and PA- $R^2$ measures are computed for all models using a test set consisting of 28825 authors. The findings are presented in Figure 2. Concerning the coefficient of determination $R^2$, it is evident from the analysis presented in Figure 2a that the first CART and then gradient-

boosted regression trees exhibit significantly superior performance compared to the simple baseline models. Among the three machine learning models, the disparities in performance are rather modest: the GBRT and RF models exhibit comparable results, while the lasso model demonstrates.

**(a) PA- *R*2, CART and baseline models with R**

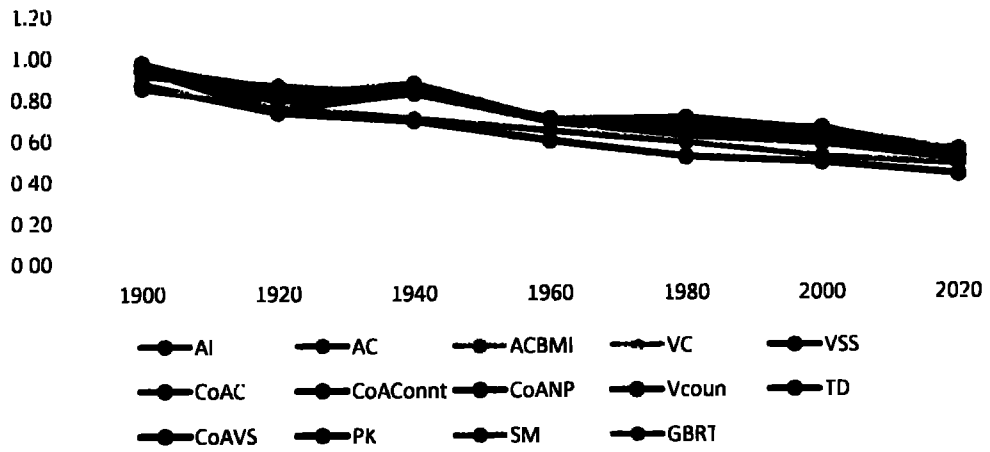



**Figure 6.3 (a) PA- *R* 2, CART and baseline models with R**

**(b) PA- *R*^2, CART and baseline models with adjusted R^2**

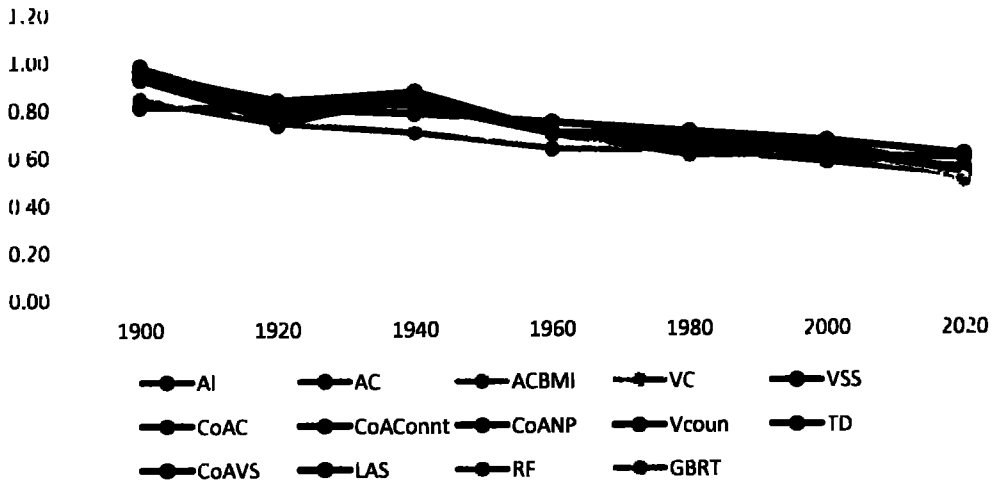



**Figure 6.3 (b) PA- *R* 2, CART and baseline models with Adjusted *R* 2**

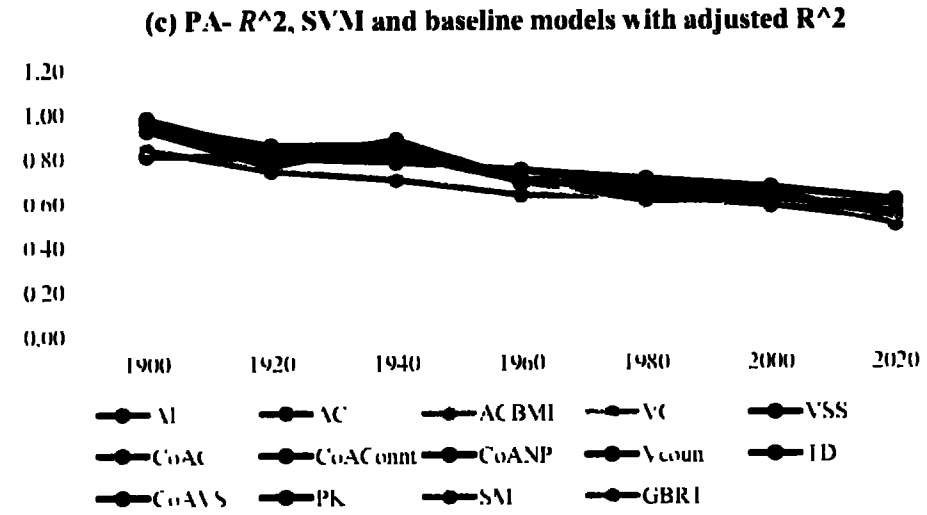



**Figure 6.3 (c) PA- *R* 2, SVM and baseline models with *R* 2**

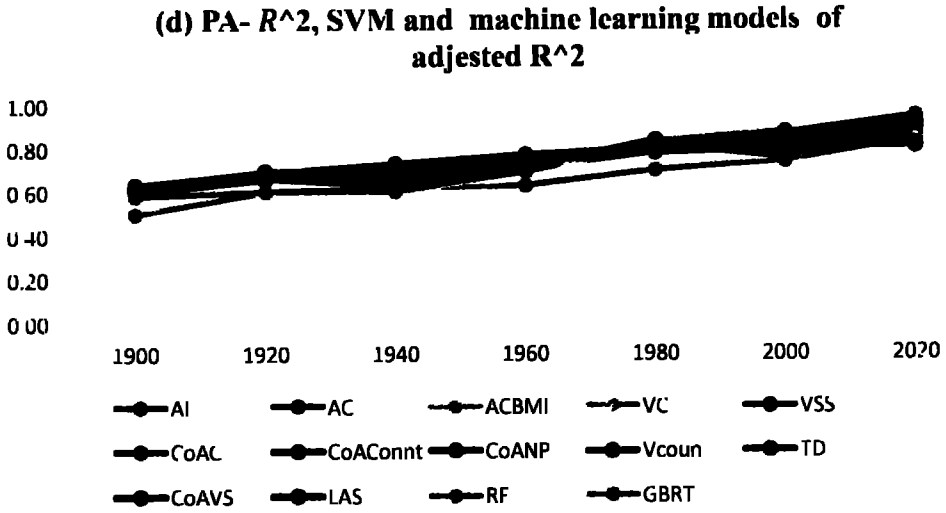



**Figure 6.3 (d) PA- *R* 2, SVM and Machine Learning Models of Adjusted *R* 2**

The utilization of the $PA - R^2$ metric sheds light on the unexpected challenges associated with accurately forecasting the h-index during relatively brief intervals of 1-2 years. This observation stands in contrast to the patterns shown in $R^2$ plots, where short-term periods exhibit a high level of predictability. The increment of an author's h-index is limited to integral values, which results in gradual growth. While the long-term cumulative impact of these increments can be anticipated, the short-term impacts are less predictable. It is worth noting that the $PA - R^2$measure indicates a significant enhancement provided by the SVM model compared to the other machine learning models (Figure 6.3 d). [8] examine the objective of predicting an index using elastic-net regularized linear regression on a dataset provided by neuroscientists. Our findings are presented in Table 6.1, which is presented alongside theirs. Our model demonstrates a significant improvement of 50% in relative. $R^2$ when making predictions for future outcomes. These data sets were combined to form a cohesive whole, and hence, our models were trained using this specific data set.

Although $R^2$-type measurements are widely used. We prefer Mean Percentage Error (MAPE), which calculates the average error percentage for each prediction. The Mean Absolute Percentage Error (MAPE) is a metric used to measure a forecasting model's accuracy.

$$\frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_{i,j}-\hat{y}_{i,j}}{y_{i,j}}\right| \tag{6.3}$$

In this context, the $i$th summand represents the absolute percentage error, denoted as $\left|(y_{i,j} - \hat{y}_{i,j})/y_{i,j}\right|$ which quantifies the deviation between the i-th prediction and the corresponding actual value. The MAPE, or Mean Absolute Percentage Error, is computed as the average of these individual mistakes. Acknowledging that a smaller Mean Absolute Percentage Error (MAPE) is considered more favorable than the coefficient of determination is important. $R^2$) Where a higher value is preferred.

***Table 6.1 The unadjusted $R^2$ values for the author's h-index predictions over 1, 5, and 10-year***

**Table 6.1:** The unadjusted $R^2$ values for the author's h-index predictions over 1, 5, and 10-year periods are reported. The $R^2$scores achieved by our CART and SVM models exhibit a notable increase compared to the findings provided by [8], particularly when forecasting a decade ahead.

| Years | 1 | 5 | 10 |
|---|---|---|---|
| Acuna et al (2012) | 0.92 | 0.67 | 0.48 |
| CART | 0.94 | 0.87 | 0.75 |
| SVM | 0.93 | 0.86 | 0.79 |
| Relative Improvement CART | 2.17% | 2.98% | 5.6% |

Upon analyzing Figure 3b, it becomes evident that the superiority of GBRT compared to the other models is further emphasized, particularly in terms of Mean Absolute Percentage Error (MAPE).In addition to its straightforward characterization, the Mean Absolute Percentage Error (MAPE) possesses two major advantages compared to the R-squared measures. Initially, the process involves standardizing the inaccuracy in each publication's forecast. In the absence of this normalization, the impact of being off by three units is deemed equally unfavorable for an author possessing an h-index of one as it is for an author with an h-index of fifty. This outcome lacks intuitive coherence. Furthermore, in contrast to the Mean Absolute Percentage Error (MAPE), $R^2$ Measurements exhibit a heightened susceptibility to outliers and are subject to significant fluctuations when encountering a limited number of inaccurate predictions. The Mean Absolute Percentage Error (MAPE) yields results that are readily comprehensible. Notably, the h-index demonstrates remarkable predictability, even when forecasting a decade ahead in Figure a.

## 6.7 Paper Impact

The initial exploration of the citation graph of academic articles focused on the characterization of the observed power-law distribution of citation counts. The initial significant achievement in the modelling of this phenomena was accomplished by [6] who developed a model that established a proportionate relationship between the probability of a newly published paper, denoted as $p_{new}$, citing a previously published paper, denoted as $P_{\text{old}}$, and the number of citations $P_{\text{old}}$ had received at the time of p_new's publication. Price (Year) demonstrated that a network that expands using the "rich get richer" mechanism has node degrees that closely resemble the power law distribution observed in actual citation networks. The preferred attachment mechanism was later rediscovered and popularized by [18], who named this model. The preferential attachment model has lately served as a source of inspiration for the development of probabilistic models aimed at forecasting the citation counts of particular papers [7]. One particular predictive model characterizes citation trajectories through the utilization of a Reinforced Poisson Process (RPP). According to the RPP model, the act of acquiring a citation enhances the likelihood of gaining further citations, hence exhibiting a self-reinforcing mechanism akin to the concept of preference attachment as proposed by [7].

In the RPP model, obtaining a citation enhances the likelihood of receiving future citations, hence exhibiting a form of self-reinforcement that might be likened to the concept of preference attachment. Specifically, the research productivity and impact models use $C_p(t)$,, which represents the number of citations received by a paper $p$ at time $t > 0$ following its publication. This model assumes that the citation process follows a Poisson distribution with a rate function.

$$r_p(t) = \lambda_p \cdot f_p(t \mid \theta_p) \cdot (C_p(t) + m) \tag{6.4}$$

In this context, $\lambda_p$ represents a fitness parameter, $f_p$ is a non-negative temporal decay function with parameters $\theta_p$, and $m$ is a positive integer that signifies the starting visibility. The parameters of the aforementioned model can thereafter be deduced through the process of maximum likelihood estimation. Due to the susceptibility of maximum likelihood estimation to overfitting, it is advantageous to enhance this model by incorporating the Bayesian framework, wherein $\lambda_p$ is postulated to be derived from a prior distribution. In this particular scenario, we

use the assumption that the variable $\lambda_p$is created by a Gamma distribution with parameters α and β. The aforementioned approach significantly decreases the quantity of factors that need to be evaluated and aids in mitigating the impact of overfitting. Regrettably, the performance of this revised model falls short of the anticipated outcomes. To enhance the precision of this sophisticated model, we take into account the subsequent three alterations.

(i) The RPP model necessitates precise information of the specific publication dates of articles. In this study, we expand upon this model to accommodate a more realistic scenario when only the publication year is available.

(ii) Regularization is utilized as a means to alleviate the model's inclination towards overfitting.

(iii) Instead of imposing a constraint on the sharing of the Gamma prior values, α and β, across all articles, we propose an alternative approach where these parameters can be determined by a fully connected single layer neural network. This neural network would take the same attributes used in the machine learning models described below as input. This enables a more nuanced understanding of the work by incorporating more information to supplement our existing knowledge.

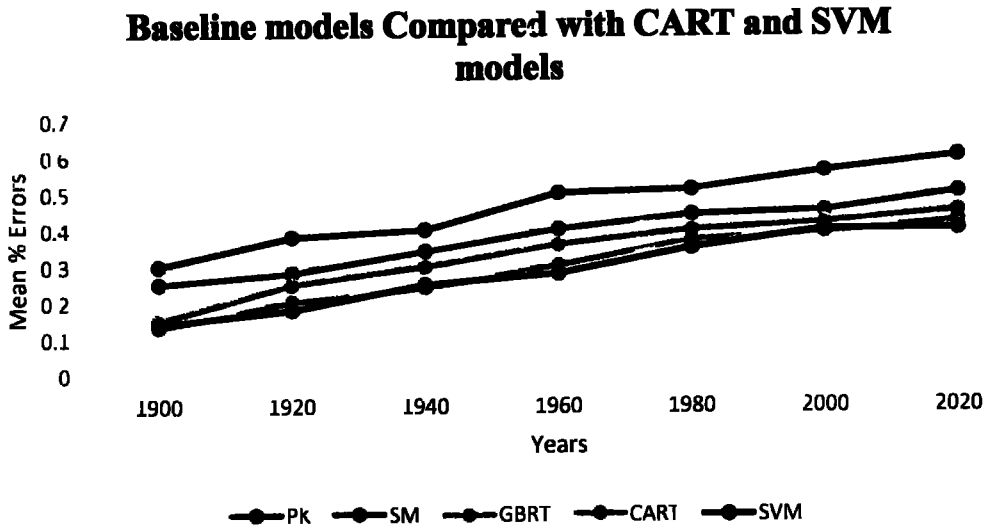
Please refer to Appendix A for a comprehensive understanding of the methods employed to achieve the aforementioned three objectives. After implementing the alterations as mentioned above, our analysis results in the existence of two potential models. The first model, referred to as an RPPNet, incorporates all three modifications by utilizing the extracted features. Conversely, the second model, known as an RPP, only incorporates the first two modifications.

In addition, a machine learning methodology is employed to extract a total of 63 distinct attributes for each publication, utilizing solely the information accessible in the year 2018. Subsequently, the aforementioned variables are employed to train the identical set of regression models, as previously outlined for the prediction of author h-index. The objective is to forecast citation counts for the period spanning from 2009 to 2018. The data is subjected to a filtration process wherein only publications that have garnered a minimum of five citations before the conclusion of 2018 are included, establishing a threshold for impact. By substituting author h-

indices with paper citations, we modify the MAPE, PA- $R^2_{,,}$ and $R^2$ metrics to align with the context of a work's influence.

Figure 3 displays the Mean Absolute Percentage Error (MAPE) for all the aforementioned models, using a test set comprising 10,000 papers. In terms of predicting author influence, it can be shown that CART and SVM models consistently beat the baseline models over several years, as depicted in Figure 3a. Among all models, the CART and SVM models consistently demonstrate superior performance. However, the observed performance primarily indicates these measures' sensitivity to outliers.

**Figure 6.5:** The Mean Absolute Percentage Error (MAPE) forecasts the author's h-index with 95% confidence intervals. It is worth noting that these intervals exhibit a high level of precision.



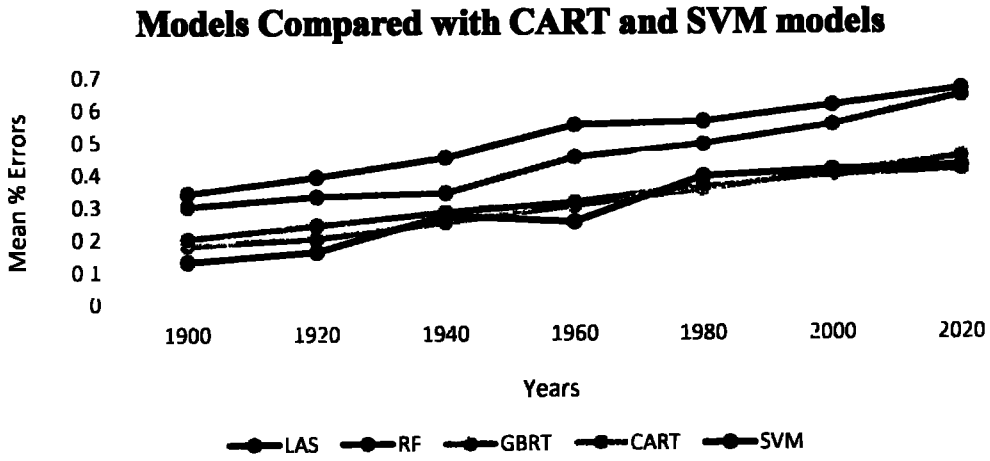



**Figure 6.6 Models Compared with CART and SVM Models**

## 6.8 Factors involved in Prediction.

One crucial element that has been neglected in our prior analysis pertains to the investigation of the influence of author career and article age on the prediction performance. It is hypothesized that the rate of citations for a scholarly article will eventually stabilize over time, hence facilitating the prediction of citation counts for older works with increased accuracy. In a similar vein, it can be posited that an author who has attained a significant level of professional accomplishment is likely to possess a more consistent and enduring h-index compared to an author who is relatively new to the field. In order to examine this inquiry, Figure 5 illustrates the percentage inaccuracy of our CART and SVM predictions for each author and publication in our test sets, stratified by age, over a period of 10 years. It is worth noting that although the inaccuracies in our forecasts appear to be primarily clustered around zero, indicating a lack of bias, we observe a significantly greater level of variability in the precision of our predictions for authors and publications of a younger age. This observation aligns with our previous conjecture that writers and publications of greater age exhibit higher predictability. Figure 6.6 also presents the Mean Absolute Percentage Error (MAPE) pertaining to authors and papers within each age group. It is noteworthy that the predictability of citations and h-indices, as quantified by the mean absolute percentage error (MAPE), exhibits a substantial initial increase for both authors and articles, followed by a subsequent leveling off.

As an illustration, those who had pursued a writing career for 25 years by 2009 exhibited a mean absolute percentage error (MAPE). This value is merely four percentage points lower than that of individuals who had been engaged in the profession for a decade. This observation implies a fundamental level of fluctuation in citation counts and h-indices that our model cannot account for, even under optimal circumstances. This observation underscores the necessity for future research endeavors to expand the existing set of features in the citation history.

In addition to comprehending the extent of variability in our predictions, we also possess a keen curiosity in discerning the primary factors that influenced those projections. To assess the significance of this particular characteristic, we employ the $t^*$ Statistic, a non-parametric correlation measure, to quantify the relationship between our features and the observed scientific impact measure [19]. The utilization of $t^*$ is advantageous as it encompasses any potential interdependence between two variables, in contrast to the Pearson correlation coefficient which solely captures linear relationships. Our analysis will only concentrate on the characteristics that contribute to the predictions of the h-index (refer to Figure 4), as the findings are comparable for predicting paper citations. Notably, the relative significance of certain factors undergoes alterations when examining forecasts throughout different forecasting periods. For instance, while making predictions on an author's h-index in 2010, it is evident that the most significant element is the author's h-index in 2009. However, in the prediction of the h-index in 2018, the significance of the h-index in 2009 is somewhat diminished compared to the quantity of articles published over the period of 200-2005 (Figure 6.5). It is reasonable to anticipate that the scholarly articles authored by an individual in the present time will require a considerable number of years to amass a sufficient number of citations that would significantly impact the author's h-index. However, once a substantial period of time has elapsed, these papers become influential in determining the author's h-index. The findings presented in this study are consistent with the patterns observed in the research conducted by [8].

### 6.9 A Modification to The Reinforced Poisson Process Model

It is crucial to acknowledge that within the framework of reinforced Poisson process models, the variable $C_p(t)$, denotes the count of citations received by a paper p at a specific time $t > 0$

following its publication. The observed variable in question adheres to a Poisson process characterized by a rate function.

$$r_p(t) = \lambda_p f(t \mid \theta_p)(C_p(t) + m) \tag{6.5}$$

The parameters $\lambda_p, \theta_p$, and $m$ are utilized in conjunction with the temporal decay function $f$. The parameters $\lambda_p$ and $\theta_p$ are estimated by maximum likelihood estimation. [7] discovered that the value of m minimally influences the performance. Therefore, we adopt the approach of [7]and use $m = 10$ as a constant value. In order to mitigate the issue of overfitting, it is possible to impose a Gamma( $\alpha, \beta$)prior distribution on the $\lambda_p$ parameters. As mentioned in Section 4, the model undergoes three adjustments, which will now be elaborated upon.

### 6.9.1 Discrete Time

As an evident expansion of the continuous time RPP model, we proceed to describe $C_p(n)$, denoting the number of citations received by an article $p$ has $n \geq 1$ years subsequent to its publication, as a discrete time Poisson process characterized by a rate function.

$$r_p(n) = \lambda_p(C_p(n-1) + m) \int_{n-1}^{n} f(t \mid \theta_p)\mathrm{d}t. \tag{6.6}$$

It should be noted that the variable denoted as $r_p(n)$ indicates the average value for the full duration between $n-1$ and $n$. In accordance with previous research, we define the function $f(t \mid \theta) = \frac{1}{\sqrt{2\pi}\sigma t} \exp\left(-\frac{1}{2\sigma^2}(\ln t - \mu)^2\right)$, where $f$ represents a log-normal probability density function for $\theta = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_{>0}$.

In order to prevent the occurrence of degenerate situations, it is imperative that we consistently make the assumption that the value of $\mu$ is more than or equal to $-1$, and the value of $\sigma$ is greater than or equal to 0.5. Consider the log-normal cumulative distribution function denoted as $\Phi_\theta$, which corresponds to the probability density function $f(t \mid \theta)$. For any $i \geq 1$, we define $\Delta\Phi_\theta(i)$ as the difference between $\Delta\Phi_\theta(i) = \Phi_\theta(i) - \Phi_\theta(i-1)$.. By employing the given notation, it is

possible to express the rate function as $r_p(n) = \lambda_p\big(C_p(n-1)+m\big)\Delta\Phi_\theta(n)$. The formulation provides a self-reinforcing relationship for any values of $n$ greater than or equal to 1.

$$C_p(n) - C_p(n-1) \mid C_p(n-1) \sim \text{Poisson}\big(\lambda_p\big(C_p(n-1)+m\big)\Delta\Phi_\theta(i)\big) \qquad (6.7)$$

Given the initial condition $C_p(0) = 0$. In order to streamline notation, we shall temporarily omit the subscripts denoted by $p$.

Let us use the per-year citation counts $C(1) - C(0) = d_1, \ldots, C(n) - C(n-1) = d_n$ and our objective is to conduct maximum likelihood estimation of $\lambda$, and $\theta$. In order to proceed, it is necessary to obtain an explicit representation of the likelihood function. According to its definition,

$$P(C(i) - C(i-1) = d_i \mid d_1, \ldots, d_{i-1}) = e^{-\lambda C_{i-1}\Delta\Phi_\theta(i)}\lambda^{d_i}\Delta\Phi_\theta(i)^{d_i}\frac{C_{i-1}^{d_i}}{d_i!}. \qquad (6.8)$$

Based on the aforementioned analysis, it becomes evident that the probability of the observations is easy.

$$\begin{aligned} L(\lambda,\theta \mid d_1,\ldots,d_n) &= \prod_{i=1}^{n} e^{-\lambda C_{i-1}\Delta\Phi_\theta(i)}\lambda^{d_i}\Delta\Phi_\theta(i)^{d_i}\frac{C_{i-1}^{d_i}}{d_i!} \\ &= \exp\Big(-\lambda\sum_{i=1}^{n} C_{i-1}\Delta\Phi_\theta(i)\Big)\lambda^{\sum_{i=1}^{n} d_i}\left(\prod_{i=1}^{n}\Delta\Phi_\theta(i)^{d_i}\frac{C_{i-1}^{d_i}}{d_i!}\right) \end{aligned} \qquad (6.9)$$

It should be noted that the $\sum_{i=1}^{n} d_i$ is equivalent to the cumulative count of citations up to time n, denoted as $N$. Considering the provided explicit form of the likelihood,

(a) The percentage inaccuracy in the forecast of an author's h-index after a period of 10 years. When we impose the criterion of include only authors with an h-index of 4 or higher, we observe that there are no authors at the age of one, and only a limited number of authors at the ages of two and three.

(b) The percentage inaccuracy in the prediction of citation count per manuscript after a period of 10 years.

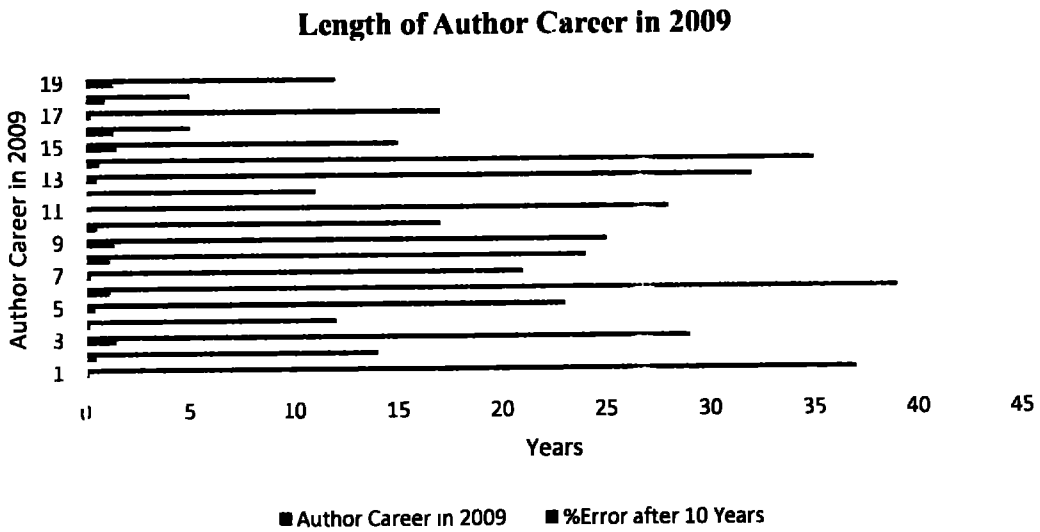



**Figure 6.6 Length of Author Career in 2009**

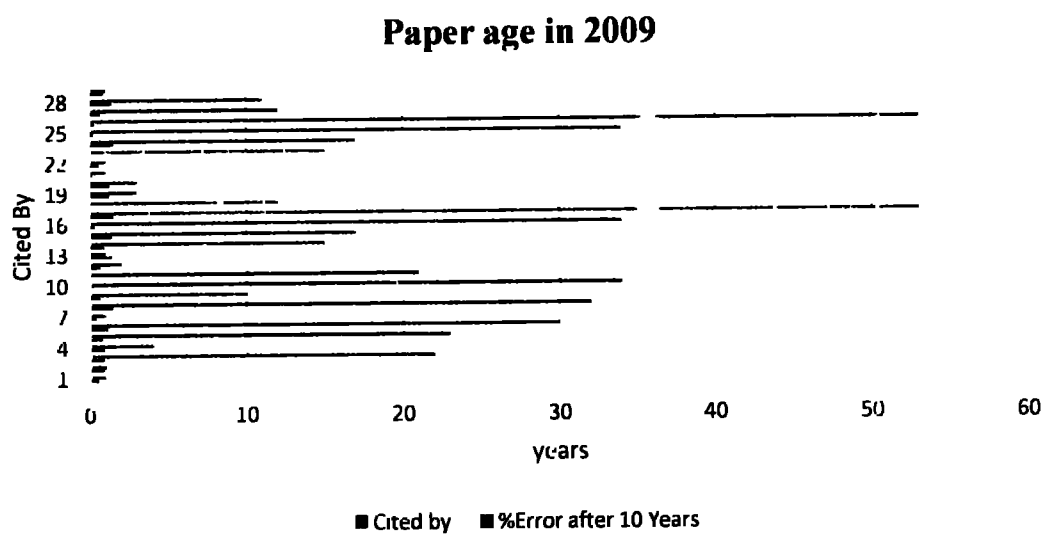



**Figure 6.7 : The percentage error of the GBRT forecasts after a decade is plotted for each author and paper in our test sets. The authors and publications have been categorized into groups depending on their ages as of the conclusion of 2009. After applying restrictions on the publications and authors within that group, we have also included the Mean Absolute Percentage Error (MAPE) for each group.**

The estimation of $\lambda$ and $\theta$ can be performed directly using maximum likelihood estimation. It is important to observe that the log-likelihood exhibits the following structure.

$$\mathcal{L}(\lambda,\theta) = \log L(\lambda,\theta \mid d_1,\dots,d_n)$$
$$= -\lambda\sum_{i=1}^{n} C_{i-1}\Delta\Phi_\theta(i) + N\log\lambda + \sum_{i=1}^{n} d_i\log(\Delta\Phi_\theta(i)) + \sum_{i=1}^{n}\log\left(\frac{C_{i-1}^{d_i}}{d_i!}\right), \quad (6.10)$$

differentiating we find that:

$$0 = \frac{\partial}{\partial\lambda}\mathcal{L}(\lambda,\theta) \Leftrightarrow \lambda = \frac{N}{\sum_{i=1}^{n} C_{i-1}\Delta\Phi_\theta(i)}. \quad (6.11)$$

Plugging this optimum value, $\lambda^* = \frac{N}{\sum_{i=1}^{n} C_{i-1}\Delta\Phi_\theta(i)}$, into $\mathcal{L}$ gives

$$\mathcal{L}(\lambda^*,\theta) = \sum_{i=1}^{n} d_i\log(\Delta\Phi_\theta(i)) - N\log(\sum_{i=1}^{n} C_{i-1}\Delta\Phi_\theta(i)) + \text{const.} \quad (6.12)$$

The differentiation of the above expression is now performed with respect to $\mu,\sigma$. Let $\phi_\theta$ represent the probability density function associated with a Normal Distribution with mean $\mu$ and variance $\sigma^2$. Subsequently, it may be inferred that:

$$\frac{\partial}{\partial\mu}\mathcal{L}(\lambda^*,\theta) = \sum_{i=1}^{n}\left(\lambda^* C_{i-1} - \frac{d_i}{\Delta\Phi_\theta(i)}\right)(\phi_\theta(\log(i)) - \phi_\theta(\log(i-1))) \quad (6.13),$$

$$\frac{\partial}{\partial\sigma}\mathcal{L}(\lambda^*,\theta) = \sum_{i=1}^{n}\left(\lambda^* C_{i-1} - \frac{d_i}{\Delta\Phi_\theta(i)}\right)\left(\frac{\log(i)-\mu}{\sigma}\phi_\theta(\log(i)) - \frac{\log(i-1)-\mu}{\sigma}\phi_\theta(\log(i-1))\right). \quad (6.14)$$

By employing derivative-based optimization methods, the parameters $\mu,\text{and}\,\sigma$ can be determined based on the information above. By utilizing the estimated parameters $\theta = (\mu,\sigma)$, we can make future predictions based on the mean of the Poisson process, taking into account the previous data. The expression $c_n(t \mid \lambda,\theta)$ represents the conditional expectation of $E[C(n+t) \mid C(1),\dots,C(n)]$,

A positive correlation exists between feature values and observed h-indices within the period of 2009-2018. Higher values of feature variables indicate a greater degree of dependence. All features utilize data that was accessible in the year 2005. The evaluated characteristics encompass the Hind (h-index), Cites (aggregate number of citations), Ave. Cites (average number of citations per year), Cites '05 (number of citations in the year 2009), Papers '04-'05 (quantity of papers published between 2004 and 2005),

The number of published papers and the author's PageRank inside the coauthor network are two important metrics to consider in academic research. The data indicates a decrease in the importance of authors' h-indices in 2009, but there is an increase in the value of other predictors, such as the number of papers produced by an author between 2004 and 2005.

Next, we can employ a recursive approach to calculate $c_n(t \mid \lambda, \theta)$. This allows us to determine the value of $c_n(t \mid \lambda, \theta)$for t greater than or equal to 1.

$$c_n(t \mid \lambda, \theta) = (C(n) + m) \prod_{i=1}^{t} (1 + \lambda \Delta \Phi_\theta(n + i)) - m. \quad (6.15)$$

**Table 6.2: Learning to Predict Citation-Based Impact Measures**

| Years | h Index | Cites | Avg Cites | Cites '09 | Paper '08-09" | Papers | Page Rank |
|---|---|---|---|---|---|---|---|
| 2009 | 0.210 | 0.200 | 0.115 | 0.150 | 0.062 | 0.103 | 0.052 |
| 2010 | 0.177 | 0.201 | 0.195 | 0.129 | 0.063 | 0.109 | 0.055 |
| 2011 | 0.239 | 0.173 | 0.107 | 0.151 | 0.065 | 0.109 | 0.058 |
| 2012 | 0.146 | 0.191 | 0.186 | 0.210 | 0.068 | 0.116 | 0.058 |
| 2013 | 0.121 | 0.160 | 0.123 | 0.161 | 0.077 | 0.120 | 0.083 |
| 2014 | 0.261 | 0.146 | 0.181 | 0.126 | 0.092 | 0.122 | 0.090 |
| 2015 | 0.199 | 0.132 | 0.106 | 0.171 | 0.116 | 0.123 | 0.104 |
| 2016 | 0.263 | 0.211 | 0.117 | 0.180 | 0.121 | 0.134 | 0.115 |
| 2017 | 0.274 | 0.191 | 0.158 | 0.232 | 0.135 | 0.138 | 0.135 |
| 2018 | 0.196 | 0.134 | 0.107 | 0.203 | 0.138 | 0.149 | 0.146 |

The table 6.2 provides the learning predictions citations based on impact factors measure such as h index, cities, average cites, cites on year 2009 . papers on year 2008-09 and page rank.

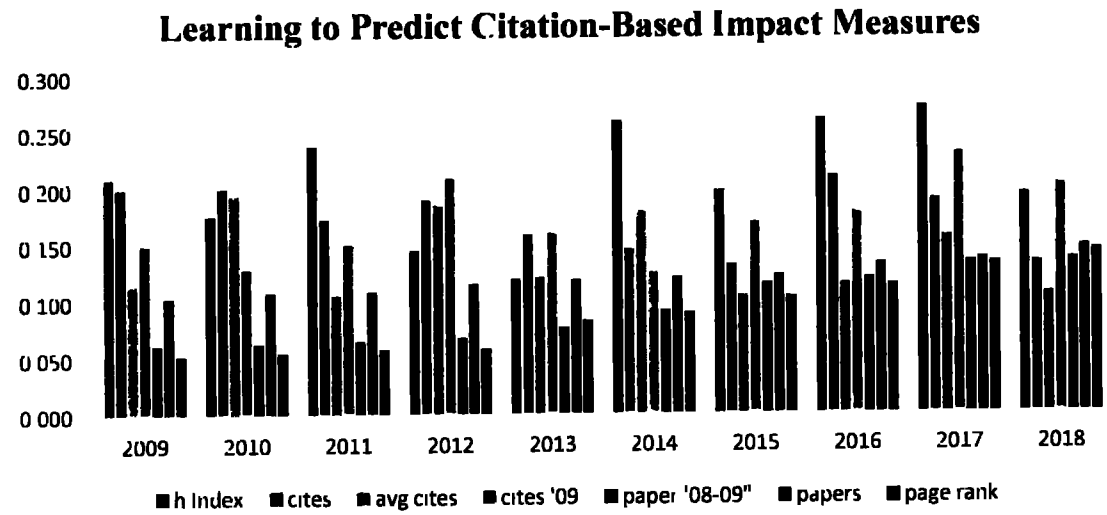



**Figure 6.8 Learning to Predict Citation-Based Impact Measure**

### 6.9.2 Prior Extensions and Regularization

A Gamma distribution with parameters α and β is employed as a prior distribution for the parameter λ. Subsequently, the marginalization likelihood is computed. In order to streamline the representation, we shall denote A as the product of the iterated differential operator $\Delta\Phi_\theta$ raised to the power of $d_i$, divided by the factorial of $d_i$, and evaluated at $C_{i-1}$ raised to the power of d_i, that is $A = (\prod_{i=1}^{n} \Delta\Phi_\theta(i)^{d_i} \frac{C_{i-1}^{d_i}}{d_i!})$ where i ranges from 1 to n. Similar to the preceding section, it is assumed that we have made observations of $C(1) - C(0) = d_1, \ldots, C(n) - C(n-1) = d_n$, for which we may express the marginalization likelihood.

$$\begin{aligned} L(\theta, \alpha, \beta \mid d_1, \ldots, d_n) &= \int_0^\infty L(\lambda, \theta \mid d_1, \ldots, d_n) \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} d\lambda \\ &= A \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+N)}{(\beta+\sum_{i=1}^{n} C_{i-1}\Delta\Phi_\theta(i))^{\alpha+N}}. \end{aligned} \tag{6.16}$$

Upon careful examination, it becomes apparent that the posterior distribution of the parameter λ, given the observed data $\mid d_1, \ldots, d_n$, follows a Gamma distribution with parameters $(\alpha + N, \beta + \sum_{i=1}^{n} C_{i-1}\Delta\Phi_\theta(i))$. This distribution enables us to readily calculate the posterior mean.

$$\lambda = E[\lambda \mid d_1, \ldots, d_n] = \frac{\alpha+N}{\beta+\sum_{i=1}^{n} C_{i-1}\Delta\Phi_\theta(i)}. \tag{6.17}$$

The logarithm of the minimized likelihood is obtained by applying the natural logarithm function.

$$\begin{aligned}\mathcal{L}(\theta,\alpha,\beta) &= B + \sum_{i=1}^{n} d_i \log(\Delta\Phi_\theta(i)) + \alpha\log\beta - \log\Gamma(\alpha) \\ &+\log\Gamma(\alpha+N) - (\alpha+N)\log(\beta+\sum_{i=1}^{n} C_{i-1}\Delta\Phi_\theta(i))\end{aligned} \quad (6.18)$$

where $B = \sum_{i=1}^{n}\log\left(\frac{C_{i-1}^{d_i}}{d_i!}\right)$ is constant with respect to the parameters $\theta, \alpha, \beta$. Now letting $\psi$ be the digamma function we have

$$\frac{\partial}{\partial\alpha}\mathcal{L}(\theta,\alpha,\beta) = \log\beta - \psi(\alpha) + \psi(\alpha+N) - \log(\beta+\sum_{i=1}^{n} C_{i-1}\Delta\Phi_\theta(i) \quad (6.19)$$

$$\frac{\partial}{\partial\beta}\mathcal{L}(\theta,\alpha,\beta) = \frac{\alpha}{\beta} - \frac{\alpha+N}{\beta+\sum_{i=1}^{n} C_{i-1}\Delta\Phi_\theta(i)} = \frac{\alpha}{\beta} - \bar{\lambda}, \quad (6.20)$$

that

$$\frac{\partial}{\partial\mu}\mathcal{L}(\theta,\alpha,\beta) = \sum_{i=1}^{n}\left(\bar{\lambda}C_{i-1} - \frac{d_i}{\Delta\Phi_\theta(i)}\right)(\phi_\theta(\log(i)) - \phi_\theta(\log(i-1))), \quad (6.21)$$

$$\begin{aligned}\frac{\partial}{\partial\sigma}\mathcal{L}(\theta,\alpha,\beta) = \sum_{i=1}^{n}&\left(\bar{\lambda}C_{i-1} - \frac{d_i}{\Delta\Phi_\theta(i)}\right)\left(\frac{\log(i)-\mu}{\sigma}\phi_\theta(\log(i))\right. \\ &\left.-\frac{\log(i-1)-\mu}{\sigma}\phi_\theta(\log(i-1))\right)\end{aligned} \quad (6.22)$$

It is noteworthy that, when comparing the derivatives concerning μ and σ to those computed in the preceding section, the optimal $\lambda^*$ in that context has been substituted with the posterior mean $\bar{\lambda}$.

In the previous section, when conducting maximum likelihood inference, treating each document as an independent entity was deemed satisfactory, as there were no shared parameters among any pair of papers. In the present context, it is notable that all papers possess identical α and β parameters, necessitating the execution of maximum likelihood estimation for all papers simultaneously. To this end let $\mathcal{P}$ be a collection of papers. For each $p \in \mathcal{P}$ suppose we have observations $C_p(1) = d_1^p, C_p(2) - C_p(1) = d_2^p, \ldots, C_p(n_p) - C_p(n_p-1) = d_{n_p}^p$, and let $\mathcal{L}_p(\theta_p,\alpha,\beta)$ calculate the log-likelihood for the specific paper, denoted as p. The log-likelihood of all papers concurrently can be expressed as the summation.

$$\mathcal{L}_{\mathcal{P}}\left(\alpha,\beta,\{\theta_p\}_{p\in\mathcal{P}}\right) = \sum_{p\in\mathcal{P}}\mathcal{L}_p(\theta_p,\alpha,\beta) \quad (6.23)$$

The computation of gradients for $\mathcal{L}_p$ can be easily achieved by leveraging the derivatives that have been previously computed for each individual $\mathcal{L}_p$. Hence, the procedure of performing maximum likelihood estimation for the parameters $\alpha, \beta$, and $\{\theta_p\}_{p\in\mathcal{P}}$ can be carried out by utilizing any gradient-based optimization algorithm that is accessible.

Let us consider a scenario where $\alpha, \beta, and\ \theta_p$ are held constant. The acquisition of future forecasts can be easily accomplished through the utilization of posterior techniques. Through the process of iterated conditioning, one may readily verify that.

$$c_{p,n_p}(t \mid \alpha, \beta, \theta_p) = E[c_n(n_p + t \mid \lambda_p, \theta_p) \mid C_p(1), \ldots, C_p(n_p)] \tag{6.24}$$

Given a constant value of $\lambda_p$, we can calculate $c_n(n_p + t \mid \lambda_p, \theta_p)$ by utilizing the findings presented in the preceding section. Therefore, it is possible to estimate the aforementioned expectation with arbitrary precision by employing a Monte-Carlo approach. Specifically, we extract samples from the posterior distribution of $\lambda_p$, which follows a Gamma distribution with parameters $\alpha + N$ and $(\beta + \sum_{i=1}^{n} C_{i-1}\Delta\Phi_\theta(i))$, Subsequently, we calculate $c_n(n_p + t \mid \lambda_p, \theta_p)$for each of these samples and then compute the average of the obtained results.

In light of the observed issue of overfitting associated with utilizing maximum likelihood inference, we propose incorporating a regularization penalty into the optimization process. Specifically, we suggest the inclusion of a penalty term of the form $-\gamma\left(\frac{\alpha}{\beta}\right)^2, \gamma \geq 0$ is a hyper parameter. Instead of aiming to maximize $\mathcal{L}_p$, our objective is to maximize.

$$\mathcal{L}_{\mathcal{P}}^{*}\left(\alpha, \beta, \{\theta_p\}_{p\in\mathcal{P}}\right) = \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}} \mathcal{L}_p(\theta_p, \alpha, \beta) - \gamma\left(\frac{\alpha}{\beta}\right)^2 \tag{6.25}$$

The value of $\alpha/\beta$ corresponds to the average of the previous distribution Gamma$(\alpha, \beta)$, so serving as a deterrent against excessively large prior distributions. In practical applications, the value of $\gamma$ is typically determined through the process of cross validation.

### 6.9.3 Neural Network Structure

To facilitate the incorporation of nuanced information into our prior distribution, we additionally contemplate the possibility of α and β being learnt functions derived from attributes retrieved from each manuscript. For every paper $p$ let $x_p \in \mathbb{R}^k$ be a set of features that correspond to the paper. Next, we denote α and β as functions of the variable $x_p$, represented as $\alpha(x_p), \beta(x_p)$, respectively. Consequently, the distribution $\lambda_p$ is approximately Gamma-distributed with parameters $\left(\alpha(x_p), \beta(x_p)\right)$. The functions α and β are acquired as the output of a neural network with a single layer that is fully connected and utilizes soft plus non-linearities. This is achieved by maximizing the penalized log-likelihood.

$$\mathcal{L}_{\mathcal{P}}^{**}\left(\alpha, \beta, \{\theta_p\}_{p\in\mathcal{P}}\right) = \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}}\left(\mathcal{L}_p\left(\theta_p, \alpha(x_p), \beta(x_p)\right) - \gamma\frac{\alpha(x_p)^2}{\beta(x_p)^2}\right) \qquad (6.26)$$

The empirical findings presented in the main study demonstrate that performance can be enhanced by allowing α and β to vary based on $x_p$.

# Chapter 7 Conclusion and Future Work

The use of bibliometric research performance evaluation to evaluate degree awarding institutes' research output is a commonly utilized approach since it offers essential information for policymakers and organizations to allocate resources and money to encourage research. In this work, we employed stochastic models to analyze and evaluate the bibliometric data of Pakistani institutes.

Data found that established and larger institutes produce more research than younger and smaller institutes. This is most likely owing to established and larger institutes having more resources, financing, and research infrastructure. We also discovered that the bulk of Pakistani institutes' research output is in the field of natural sciences, followed by social sciences and engineering. This means that study in other subjects, such as humanities and arts, should be encouraged.

Thirdly, we found that international collaborations significantly impact the research output of Pakistani institutes, with institutes having higher numbers of international collaborations producing more research output. This highlights the importance of international collaborations in enhancing research output and promoting knowledge exchange.

While our analysis gives significant insights into the research performance of Pakistani degree granting institutes, there are several constraints to consider. For starters, our approach solely considers bibliometric data and ignores other factors that may impact research output, such as institutional policies, research infrastructure, and funding sources. As a result, future study might look at these characteristics to present a more complete picture of research performance.

Second, our analysis is limited to Pakistani institutes and does not take into account other nations or areas. Future study might compare Pakistani institutes to those in other nations to uncover parallels and variations in research performance.

There are various options for additional exploration in terms of future research proposals. Investigating the variables that lead to the difference in research output among Pakistani institutes is one such option. This might entail investigating different institutes' funding sources, research facilities, and institutional rules. Another area of future research would be to look at the effect of research production on the reputation and ranking of Pakistani institutes. This might entail investigating the relationship between research production and ranks in international university rankings like the QS World University ranks. The effect of international cooperation in increasing research output is required. Analyzing the nature and effect of various sorts of international partnerships, such as collaborative research projects, joint publications, and joint funding applications, might be part of this. Overall, our analysis emphasizes the importance of better collaboration and resource sharing among Pakistani institutes to boost research output and performance. The use of bibliometric research performance evaluation to evaluate degree awarding institutes' research output is a commonly utilized approach since it offers essential information for policymakers and organizations to allocate resources and money to encourage research. In this work, we employed stochastic models to analyze and evaluate the bibliometric data of Pakistani institutes. The skewness of research output is most seen in the amount of citations obtained by Pakistani institutes, with a small handful of institutes garnering the vast bulk of citations. This suggests that there is room for improvement in Pakistani institutes' research output. This means that study in other subjects, such as humanities and arts, should be encouraged.

Third, we discovered that foreign collaborations had a considerable influence on Pakistani institutes' research production, with institutes with a larger number of international collaborations creating more research output. This emphasizes the significance of international cooperation in improving research output and fostering knowledge exchange.

While our analysis gives significant insights into the research performance of Pakistani degree granting institutes, there are several constraints to consider. For starters, our approach solely considers bibliometric data and ignores other factors that may impact research output, such as institutional policies, research infrastructure, and funding sources. As a result, future study might look at these characteristics to present a more complete picture of research performance.

Second, our analysis is limited to Pakistani institutes and does not consider other nations or areas. Future study might compare Pakistani institutes to those in other nations to uncover parallels and variations in research performance. Finally, our analysis only analyses research output in the form of publications and citations, and does not evaluate other types of research output, such as patents or policy effect. Future studies might investigate how research output affects non-academic sectors like industry and policy.

.

## References:

[1] K. W. Boyack and R. Klavans, "Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately?," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 12, pp. 2389–2404, Dec. 2010, doi: https://doi.org/10.1002/asi.21419.

[2] W. Glänzel and B. Thijs, "Using 'core documents' for detecting and labelling new emerging topics," *Scientometrics*, vol. 91, no. 2, pp. 399–416, Dec. 2011, doi: https://doi.org/10.1007/s11192-011-0591-7.

[3] L. Leydesdorff and I. Rafols, "A global map of science based on the ISI subject categories," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 2, pp. 348–362, Feb. 2009, doi: https://doi.org/10.1002/asi.20967.

[4] L. Meng, K.-H. Wen, R. Brewin, and Q. Wu, "Knowledge Atlas on the Relationship between Urban Street Space and Residents' Health—A Bibliometric Analysis Based on VOSviewer and CiteSpace," *Sustainability*, vol. 12, no. 6, p. 2384, Mar. 2020, doi: https://doi.org/10.3390/su12062384.

[5] L. Waltman and N. J. van Eck, "A smart local moving algorithm for large-scale modularity-based community detection," *The European Physical Journal B*, vol. 86, no. 11, Nov. 2013, doi: https://doi.org/10.1140/epjb/e2013-40829-0.

[6] W. Peng, J. Wang, W. Wang, Q. Liu, F.-X. Wu, and Y. Pan, "Iteration method for predicting essential proteins based on orthology and protein-protein interaction networks," *BMC Systems Biology*, vol. 6, no. 1, Jul. 2012, doi: https://doi.org/10.1186/1752-0509-6-87.

[7] D. Hicks, P. Wouters, L. Waltman, S. de Rijcke, and I. Rafols, "Bibliometrics: The Leiden Manifesto for research metrics," *Nature*, vol. 520, no. 7548, pp. 429–431, Apr. 2015, doi: https://doi.org/10.1038/520429a.

[8] H. F. Moed, "Statistical relationships between downloads and citations at the level of

individual documents within a single journal," *Journal of the American Society for Information Science and Technology*, vol. 56, no. 10, pp. 1088–1097, 2005, doi: https://doi.org/10.1002/asi.20200.

[9] W. Marx, L. Bornmann, A. Barth, and L. Leydesdorff, "Detecting the historical roots of research fields by reference publication year spectroscopy (RPYS)," *Journal of the Association for Information Science and Technology*, vol. 65, no. 4, pp. 751–764, Nov. 2013, doi: https://doi.org/10.1002/asi.23089.

[10] L. Waltman, "A review of the literature on citation impact indicators," *Journal of Informetrics*, vol. 10, no. 2, pp. 365–391, May 2016, doi: https://doi.org/10.1016/j.joi.2016.02.007.

[11] I. Rafols, L. Leydesdorff, A. O'Hare, P. Nightingale, and A. Stirling, "How journal rankings can suppress interdisciplinary research: A comparison between Innovation Studies and Business & Management," *Research Policy*, vol. 41, no. 7, pp. 1262–1282, Sep. 2012, doi: https://doi.org/10.1016/j.respol.2012.03.015.

[12] Z. Mahmood, R. Kouser, W. Ali, Z. Ahmad, and T. Salman, "Does Corporate Governance Affect Sustainability Disclosure? A Mixed Methods Study," *Sustainability*, vol. 10, no. 1, p. 207, Jan. 2018, doi: https://doi.org/10.3390/su10010207.

[13] L. I. Meho and K. Yang, "Impact of data sources on citation counts and rankings of LIS faculty: Web of science versus scopus and google scholar," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 13, pp. 2105–2125, 2007, doi: https://doi.org/10.1002/asi.20677.

[14] D. M. Boyd and N. B. Ellison, "Social Network Sites: Definition, History, and Scholarship," *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210–230, Oct. 2008, doi: https://doi.org/10.1111/j.1083-6101.2007.00393.x.

[15] S. Saeed, S. Y. Yousafzai, M. Yani-De-Soriano, and M. Muffatto, "The Role of Perceived University Support in the Formation of Students' Entrepreneurial Intention," *Journal of Small Business Management*, vol. 53, no. 4, pp. 1127–1145, Dec. 2013, doi: https://doi.org/10.1111/jsbm.12090.

[16] A. Abrizah, A. N. Zainab, K. Kiran, and R. G. Raj, "LIS journals scientific impact and subject categorization: a comparison between Web of Science and Scopus,"

*Scientometrics*, vol. 94, no. 2, pp. 721–740, Jul. 2012, doi: https://doi.org/10.1007/s11192-012-0813-7.

[17] S. K. U. Shah Bukhari, H. Said, R. Gul, and P. M. Ibna Seraj, "Barriers to sustainability at Pakistan public universities and the way forward," *International Journal of Sustainability in Higher Education*, vol. 23, no. 4, pp. 865–886, Sep. 2021, doi: https://doi.org/10.1108/ijshe-09-2020-0352.

[18] M. Kousar and K. Mahmood, "Perceptions of Faculty about Information Literacy Skills of Postgraduate Engineering Students," *International Information & Library Review*, vol. 47, no. 1–2, pp. 52–57, Apr. 2015, doi: https://doi.org/10.1080/10572317.2015.1055694.

[19] A. Charnes, W. W. Cooper, and E. Rhodes, "Measuring the efficiency of decision making units," *European Journal of Operational Research*, vol. 2, no. 6, pp. 429–444, Nov. 1978, doi: https://doi.org/10.1016/0377-2217(78)90138-8.

[20] L. Egghe and R. Rousseau, "An informetric model for the Hirsch-index," *Scientometrics*, vol. 69, no. 1, pp. 121–129, Oct. 2006, doi: https://doi.org/10.1007/s11192-006-0143-8.

[21] A. Charnes, W. W. Cooper, and E. Rhodes, "Measuring the efficiency of decision making units," *European Journal of Operational Research*, vol. 2, no. 6, pp. 429–444, Nov. 1978, doi: https://doi.org/10.1016/0377-2217(78)90138-8.

[22] M. Kousar and K. Mahmood, "Perceptions of Faculty about Information Literacy Skills of Postgraduate Engineering Students," *International Information & Library Review*, vol. 47,

[23] L. Egghe and R. Rousseau, "An informetric model for the Hirsch-index," *Scientometrics*, vol. 69, no. 1, pp. 121–129, Oct. 2006, doi: https://doi.org/10.1007/s11192-006-0143-8

[24] W. Peng, J. Wang, W. Wang, Q. Liu, F.-X. Wu, and Y. Pan, "Iteration method for predicting essential proteins based on orthology and protein-protein interaction networks," *BMC Systems Biology*, vol. 6, no. 1, Jul. 2012, doi: https://doi.org/10.1186/1752-0509-6-87.

[25] L. Leydesdorff and I. Rafols, "A global map of science based on the ISI subject categories," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 2, pp. 348–362, Feb. 2009, doi: https://doi.org/10.1002/asi.20967.

[25] K. W. Boyack and R. Klavans, "Co-citation analysis, bibliographic coupling, and direct

citation: Which citation approach represents the research front most accurately?," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 12, pp. 2389–2404, Dec. 2010, doi: https://doi.org/10.1002/asi.21419.

[26] W. Glänzel and B. Thijs, "Using 'core documents' for detecting and labelling new emerging topics," *Scientometrics*, vol. 91, no. 2, pp. 399–416, Dec. 2011, doi: https://doi.org/10.1007/s11192-011-0591-7.

[27] L. Leydesdorff and I. Rafols, "A global map of science based on the ISI subject categories," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 2, pp. 348–362, Feb. 2009, doi: https://doi.org/10.1002/asi.20967.

[28] L. Meng, K.-H. Wen, R. Brewin, and Q. Wu, "Knowledge Atlas on the Relationship between Urban Street Space and Residents' Health—A Bibliometric Analysis Based on VOSviewer and CiteSpace," *Sustainability*, vol. 12, no. 6, p. 2384, Mar. 2020, doi: https://doi.org/10.3390/su12062384.

[29] L. Waltman and N. J. van Eck, "A smart local moving algorithm for large-scale modularity-based community detection," *The European Physical Journal B*, vol. 86, no. 11, Nov. 2013, doi: https://doi.org/10.1140/epjb/e2013-40829-0.

[30] W. Peng, J. Wang, W. Wang, Q. Liu, F.-X. Wu, and Y. Pan, "Iteration method for predicting essential proteins based on orthology and protein-protein interaction networks," *BMC Systems Biology*, vol. 6, no. 1, Jul. 2012, doi: https://doi.org/10.1186/1752-0509-6-87.

[31] D. Hicks, P. Wouters, L. Waltman, S. de Rijcke, and I. Rafols, "Bibliometrics: The Leiden Manifesto for research metrics," *Nature*, vol. 520, no. 7548, pp. 429–431, Apr. 2015, doi: https://doi.org/10.1038/520429a.

[32] H. F. Moed, "Statistical relationships between downloads and citations at the level of

individual documents within a single journal," *Journal of the American Society for Information Science and Technology*, vol. 56, no. 10, pp. 1088–1097, 2005, doi: https://doi.org/10.1002/asi.20200.

[33] W. Marx, L. Bornmann, A. Barth, and L. Leydesdorff, "Detecting the historical roots of research fields by reference publication year spectroscopy (RPYS)," *Journal of the Association for Information Science and Technology*, vol. 65, no. 4, pp. 751–764, Nov. 2013, doi: https://doi.org/10.1002/asi.23089.

[34] L. Waltman, "A review of the literature on citation impact indicators," *Journal of Informetrics*, vol. 10, no. 2, pp. 365–391, May 2016, doi: https://doi.org/10.1016/j.joi.2016.02.007.

[35] I. Rafols, L. Leydesdorff, A. O'Hare, P. Nightingale, and A. Stirling, "How journal rankings can suppress interdisciplinary research: A comparison between Innovation Studies and Business & Management," *Research Policy*, vol. 41, no. 7, pp. 1262–1282, Sep. 2012, doi: https://doi.org/10.1016/j.respol.2012.03.015.

[36] Z. Mahmood, R. Kouser, W. Ali, Z. Ahmad, and T. Salman, "Does Corporate Governance Affect Sustainability Disclosure? A Mixed Methods Study," *Sustainability*, vol. 10, no. 1, p. 207, Jan. 2018, doi: https://doi.org/10.3390/su10010207.

[37] L. I. Meho and K. Yang, "Impact of data sources on citation counts and rankings of LIS faculty: Web of science versus scopus and google scholar," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 13, pp. 2105–2125, 2007, doi: https://doi.org/10.1002/asi.20677.

[38] D. M. Boyd and N. B. Ellison, "Social Network Sites: Definition, History, and Scholarship," *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210–230, Oct. 2008, doi: https://doi.org/10.1111/j.1083-6101.2007.00393.x.

[39] S. Saeed, S. Y. Yousafzai, M. Yani-De-Soriano, and M. Muffatto, "The Role of Perceived University Support in the Formation of Students' Entrepreneurial Intention," *Journal of Small Business Management*, vol. 53, no. 4, pp. 1127–1145, Dec. 2013, doi: https://doi.org/10.1111/jsbm.12090.

[40] A. Abrizah, A. N. Zainab, K. Kiran, and R. G. Raj, "LIS journals scientific impact and subject categorization: a comparison between Web of Science and Scopus," *Scientometrics*, vol. 94, no. 2, pp. 721–740, Jul. 2012, doi: https://doi.org/10.1007/s11192-012-0813-7.

[41] A. S. Spanias, "Solar energy management as an Internet of Things (IoT) application," *IEEE Xplore*, Aug. 01, 2017. https://ieeexplore.ieee.org/abstract/document/8316460 (accessed Mar. 24, 2021).

[42] S. K. U. Shah Bukhari, H. Said, R. Gul, and P. M. Ibna Seraj, "Barriers to sustainability at Pakistan public universities and the way forward," *International Journal of Sustainability in Higher Education*, vol. 23, no. 4, pp. 865–886, Sep. 2021, doi: https://doi.org/10.1108/ijshe-09-2020-0352.

[43] M. Kousar and K. Mahmood, "Perceptions of Faculty about Information Literacy Skills of Postgraduate Engineering Students," *International Information & Library Review*, vol. 47, no. 1–2, pp. 52–57, Apr. 2015, doi: https://doi.org/10.1080/10572317.2015.1055694.

[44] A. Charnes, W. W. Cooper, and E. Rhodes, "Measuring the efficiency of decision making units," *European Journal of Operational Research*, vol. 2, no. 6, pp. 429–444, Nov. 1978, doi: https://doi.org/10.1016/0377-2217(78)90138-8.

[45] L. Egghe and R. Rousseau, "An informetric model for the Hirsch-index," *Scientometrics*, vol. 69, no. 1, pp. 121–129, Oct. 2006, doi: https://doi.org/10.1007/s11192-006-0143-8.

[46] P. R. C. Rahul, "r-index: Quantifying the quality of an individual's scientific research output," *Journal of Scientometric Research*, vol. 2, no. 1, p. 80, 2013, doi: https://doi.org/10.4103/2320-0057.115867.

[47] X. Shuai, A. Pepe, and J. Bollen, "How the Scientific Community Reacts to Newly Submitted Preprints: Article Downloads, Twitter Mentions, and Citations," *PLoS ONE*, vol. 7, no. 11, p. e47523, Nov. 2012, doi: https://doi.org/10.1371/journal.pone.0047523.

[48] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, Dec. 1974, doi: https://doi.org/10.1109/tac.1974.1100705.

[49] T. R. Johnson, "Hensher, D.A., Rose, J.M., & Greene, W.H. (2005). Applied choice analysis: A primer. Cambridge: Cambridge University Press. 742+xxiv pp. US$60.00. ISBN: 0-521-60577-6.," *Psychometrika*, vol. 72, no. 3, pp. 449–450, Sep. 2007, doi: https://doi.org/10.1007/s11336-007-9029-9.

[50] B. Karrer and M. E. J. Newman, "Stochastic blockmodels and community structure in networks," *Physical Review E*, vol. 83, no. 1, Jan. 2011, doi: https://doi.org/10.1103/physreve.83.016107.

[51] V. Lyzinski, M. Tang, Avanti Athreya, Y. Park, and C. E. Priebe, "Community Detection and Classification in Hierarchical Stochastic Blockmodels," *IEEE Transactions on Network Science and Engineering*, vol. 4, no. 1, pp. 13–26, Jan. 2017, doi: https://doi.org/10.1109/tnse.2016.2634322.

[52] G. Scutari, D. Palomar, F. Facchinei, and J. Pang, "Convex Optimization, Game Theory, and Variational Inequality Theory," *IEEE Signal Processing Magazine*, vol. 27, no. 3, pp. 35–49, May 2010, doi: https://doi.org/10.1109/msp.2010.936021.

[53] J. Trowsdale and J. C. Knight, "Major Histocompatibility Complex Genomics and Human Disease," *Annual Review of Genomics and Human Genetics*, vol. 14, no. 1, pp. 301–323, Aug. 2013, doi: https://doi.org/10.1146/annurev-genom-091212-153455.

[54] M. W. Leung, "Community based participatory research: a promising approach for increasing epidemiology's relevance in the 21st century," *International Journal of Epidemiology*, vol. 33, no. 3, pp. 499–506, May 2004, doi: https://doi.org/10.1093/ije/dyh010.

[55] C. C. Holt, F. Modigliani, and H. A. Simon, "A Linear Decision Rule for Production and Employment Scheduling," *Management Science*, vol. 2, no. 1, pp. 1–30, Oct. 1955, doi: https://doi.org/10.1287/mnsc.2.1.1.

[56] A.-L. Barabási, R. Albert, and H. Jeong, "Mean-field theory for scale-free random networks," *Physica A: Statistical Mechanics and its Applications*, vol. 272, no. 1–2, pp. 173–187, Oct. 1999, doi: https://doi.org/10.1016/s0378-4371(99)00291-5.

[57] A. C. Worthington and B. L. Lee, "Efficiency, technology and productivity change in Australian universities, 1998–2003," *Economics of Education Review*, vol. 27, no. 3, pp. 285–298, Jun. 2008, doi: https://doi.org/10.1016/j.econedurev.2006.09.012.

[58] T. A. B. Snijders, G. G. van de Bunt, and C. E. G. Steglich, "Introduction to stochastic actor-based models for network dynamics," *Social Networks*, vol. 32, no. 1, pp. 44–60, Jan. 2010, doi: https://doi.org/10.1016/j.socnet.2009.02.004.

[59] Syed Hafeez Ahmad and F. A. Junaid, "Higher Education in Federally Administered Tribal Areas of Pakistan after 9/11: Problems and Prospects.," *US-China education review*, vol. 7, no. 5, pp. 55–65, May 2010.

[60] L.-H. N. Chiang, "Luo Di Sheng Gen (落地生根): Early Taiwanese-Chinese Immigrants in Canada and Guam," *Journal Of Chinese Overseas*, vol. 8, no. 2, pp. 169–204, 2012, doi: https://doi.org/10.1163/17932548-12341236.

[61] M. Ramzan, M. A. Munir, N. Siddique, and M. Asif, "Awareness about plagiarism amongst university students in Pakistan," *Higher Education*, vol. 64, no. 1, pp. 73–84, Sep. 2011, doi: https://doi.org/10.1007/s10734-011-9481-4.

[62] B. Latour, "Why Has Critique Run out of Steam? From Matters of Fact to Matters of Concern," *Critical Inquiry*, vol. 30, no. 2, pp. 225–248, Jan. 2004, doi: https://doi.org/10.1086/421123.

[63] G. Khan and R. Bhatti, "An analysis of collection development in the university libraries of Pakistan," *Collection Building*, vol. 35, no. 1, pp. 22–34, Jan. 2016, doi: https://doi.org/10.1108/cb-07-2015-0012.

[64] G. Myhre *et al.*, "Radiative forcing of the direct aerosol effect from AeroCom Phase II simulations," *Atmospheric Chemistry and Physics*, vol. 13, no. 4, pp. 1853–1877, Feb. 2013, doi: https://doi.org/10.5194/acp-13-1853-2013.

[65] A. Clauset, S. Arbesman, and D. B. Larremore, "Systematic inequality and hierarchy in faculty hiring networks," *Science Advances*, vol. 1, no. 1, p. e1400005, Feb. 2015, doi: https://doi.org/10.1126/sciadv.1400005.

[66] Z. Yang, Juan Ignacio Perotti, and C. J. Tessone, "Hierarchical benchmark graphs for testing community detection algorithms," *Physical review*, vol. 96, no. 5, Nov. 2017, doi: https://doi.org/10.1103/physreve.96.052311.

[67] F. Zeng *et al.*, "Association of inflammatory markers with the severity of COVID-19: A meta-analysis," *International Journal of Infectious Diseases*, vol. 96, pp. 467–474, Jul. 2020, doi: https://doi.org/10.1016/j.ijid.2020.05.055.

[68] G. Abramo, A. C. D'Angelo, and G. Murgia, "The relationship among research productivity, research collaboration, and their determinants," *Journal of Informetrics*, vol. 11, no.

4, pp. 1016–1030, Nov. 2017, doi: https://doi.org/10.1016/j.joi.2017.09.007.

[69] E. Yan and Y. Ding, "Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 7, pp. 1313–1326, May 2012, doi: https://doi.org/10.1002/asi.22680.

[70] X. Li, J. Feng, Y. Meng, Q. Han, F. Wu, and J. Li, "A Unified MRC Framework for Named Entity Recognition," *arXiv.org*, Nov. 22, 2022. https://arxiv.org/abs/1910.11476 (accessed Jul. 24, 2023).

[71] L. Bornmann, R. Mutz, and H.-D. Daniel, "A Reliability-Generalization Study of Journal Peer Reviews: A Multilevel Meta-Analysis of Inter-Rater Reliability and Its Determinants," *PLoS ONE*, vol. 5, no. 12, p. e14331, Dec. 2010, doi: https://doi.org/10.1371/journal.pone.0014331.

[72] R. Van Noorden, "The science that's never been cited," *Nature*, vol. 552, no. 7684, pp. 162–164, Dec. 2017, doi: https://doi.org/10.1038/d41586-017-08404-0.

[73] J. IntHout, J. P. A. Ioannidis, M. M. Rovers, and J. J. Goeman, "Plea for routinely presenting prediction intervals in meta-analysis," *BMJ Open*, vol. 6, no. 7, p. e010247, Jul. 2016, doi: https://doi.org/10.1136/bmjopen-2015-010247.

[74] J. P. A. Ioannidis, N. A. Patsopoulos, and E. Evangelou, "Uncertainty in heterogeneity estimates in meta-analyses," *BMJ*, vol. 335, no. 7626, pp. 914–916, Nov. 2007, doi: https://doi.org/10.1136/bmj.39343.408449.80.

[75] R. M. Lang *et al.*, "Recommendations for Cardiac Chamber Quantification by Echocardiography in Adults: An Update from the American Society of Echocardiography and

the European Association of Cardiovascular Imaging," *Journal of the American Society of Echocardiography*, vol. 28, no. 1, pp. 1-39.e14, Jan. 2015, doi: https://doi.org/10.1016/j.echo.2014.10.003.

[76] P. Gendreau, T. Little, and C. Goggin, "A meta-analysis of the predictors of adult offender recidivism: What works!," *Criminology*, vol. 34, no. 4, pp. 575–608, Nov. 1996, doi: https://doi.org/10.1111/j.1745-9125.1996.tb01220.x.

[77] J. Niu, W. Tang, F. Xu, X. Zhou, and Y. Song, "Global Research on Artificial Intelligence from 1990–2014: Spatially-Explicit Bibliometric Analysis," *ISPRS International Journal of Geo-Information*, vol. 5, no. 5, p. 66, May 2016, doi: https://doi.org/10.3390/ijgi5050066.

[78] S. L. Bressler and V. Menon, "Large-scale brain networks in cognition: emerging methods and principles," *Trends in Cognitive Sciences*, vol. 14, no. 6, pp. 277–290, Jun. 2010, doi: https://doi.org/10.1016/j.tics.2010.04.004.

[79] G. Abramo, C. A. D'Angelo, M. Ferretti, and A. Parmentola, "An individual-level assessment of the relationship between spin-off activities and research performance in universities," *R&D Management*, vol. 42, no. 3, pp. 225–242, May 2012, doi: https://doi.org/10.1111/j.1467-9310.2012.00680.x.

[80] C. R. Sugimoto, S. Work, V. Larivière, and S. Haustein, "Scholarly use of social media and altmetrics: A review of the literature," *Journal of the Association for Information Science and Technology*, vol. 68, no. 9, pp. 2037–2062, Jun. 2017, doi: https://doi.org/10.1002/asi.23833.

[81] A. K. Fortunato *et al.*, "Strength Training Session Induces Important Changes on Physiological, Immunological, and Inflammatory Biomarkers," *Journal of Immunology Research*, vol. 2018, pp. 1–12, Jun. 2018, doi: https://doi.org/10.1155/2018/9675216.

[82] W. D. Travis *et al.*, "International Association for the Study of Lung Cancer/American

Thoracic Society/European Respiratory Society International Multidisciplinary Classification of Lung Adenocarcinoma," *Journal of Thoracic Oncology*, vol. 6, no. 2, pp. 244–285, Feb. 2011, doi: https://doi.org/10.1097/jto.0b013e318206a221.

[83] S. Lee and B. Bozeman, "The Impact of Research Collaboration on Scientific Productivity," *Social Studies of Science*, vol. 35, no. 5, pp. 673–702, Oct. 2005, doi: https://doi.org/10.1177/0306312705052359.

[84] D. G. Hackam *et al.*, "The 2013 Canadian Hypertension Education Program Recommendations for Blood Pressure Measurement, Diagnosis, Assessment of Risk, Prevention, and Treatment of Hypertension," *Canadian Journal of Cardiology*, vol. 29, no. 5, pp. 528–542, May 2013, doi: https://doi.org/10.1016/j.cjca.2013.01.005.

[85] G. Abramo, C. A. D'Angelo, and F. Di Costa, "Identifying interdisciplinarity through the disciplinary classification of coauthors of scientific publications," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 11, pp. 2206–2222, Oct. 2012, doi: https://doi.org/10.1002/asi.22647.

[86] M. E. J. Newman, "Clustering and preferential attachment in growing networks," *Physical Review E*, vol. 64, no. 2, Jul. 2001, doi: https://doi.org/10.1103/physreve.64.025102.

[87] L. Q. Uddin, K. Supekar, and V. Menon, "Reconceptualizing functional brain connectivity in autism from a developmental perspective," *Frontiers in Human Neuroscience*, vol. 7, 2013, doi: https://doi.org/10.3389/fnhum.2013.00458.

[88] S. Mallapaty, "Predicting scientific success," *Nature*, vol. 561, no. 7723, pp. S32–S33, Sep. 2018, doi: https://doi.org/10.1038/d41586-018-06627-3.

[89] N. C. Kawa, J. A. Clavijo Michelangeli, J. L. Clark, D. Ginsberg, and C. McCarty, "The Social Network of US Academic Anthropology and Its Inequalities," *American Anthropologist*, vol. 121, no. 1, pp. 14–29, Dec. 2018, doi: https://doi.org/10.1111/aman.13158.

[90] M. D. J. Peters, C. M. Godfrey, H. Khalil, P. McInerney, D. Parker, and C. B. Soares, "Guidance for Conducting Systematic Scoping Reviews," *International Journal of Evidence-Based Healthcare*, vol. 13, no. 3, pp. 141–146, Sep. 2015, doi: https://doi.org/10.1097/xeb.0000000000000050.

[91]F. Radicchi, S. Fortunato, and C. Castellano, "Universality of citation distributions: Toward an objective measure of scientific impact," *Proceedings of the National Academy of Sciences*, vol. 105, no. 45, pp. 17268–17272, Oct. 2008, doi: https://doi.org/10.1073/pnas.0806977105.

[92] R. Sinatra, D. Wang, P. Deville, C. Song, and A.-L. Barabási, "Quantifying the evolution of individual scientific impact," *Science*, vol. 354, no. 6312, Nov. 2016, doi: https://doi.org/10.1126/science.aaf5239.

[93] P. Gong *et al.*, "Annual maps of global artificial impervious area (GAIA) between 1985 and 2018," *Remote Sensing of Environment*, vol. 236, p. 111510, Jan. 2020, doi: https://doi.org/10.1016/j.rse.2019.111510.

[94] Y. Zhu, J. Xie, F. Huang, and L. Cao, "Association between short-term exposure to air pollution and COVID-19 infection: Evidence from China," *Science of The Total Environment*, vol. 727, p. 138704, Jul. 2020, doi: https://doi.org/10.1016/j.scitotenv.2020.138704.

[95] Y. Liu *et al.*, "Clinical and biochemical indexes from 2019-nCoV infected patients linked to viral loads and lung injury," *Science China Life Sciences*, vol. 63, no. 3, pp. 364–374, Feb. 2020, doi: https://doi.org/10.1007/s11427-020-1643-8.

[96] S. H. Strogatz, "Exploring complex networks," *Nature*, vol. 410, no. 6825, pp. 268–276, Mar. 2001, doi: https://doi.org/10.1038/35065725.

[97] Yu. Ts. Oganessian *et al.*, "Publisher's Note: Measurements of cross sections and decay

properties of the isotopes of elements 112, 114, and 116 produced in the fusion reactionsU233,238,Pu242, andCm248+Ca48[Phys. Rev. C 70, 064609 (2004)]," *Physical review*, vol. 71, no. 2, Feb. 2005, doi: https://doi.org/10.1103/physrevc.71.029902.

[98] E. Ravasz, "Hierarchical Organization of Modularity in Metabolic Networks," *Science*, vol. 297, no. 5586, pp. 1551–1555, Aug. 2002, doi: https://doi.org/10.1126/science.1073374.

[99] "Quantifying scientific collaboration," *Physics Today*, Aug. 2015, doi: https://doi.org/10.1063/pt.5.7195.

[100] B. L. Ponomariov and P. C. Boardman, "Influencing scientists' collaboration and productivity patterns through new institutions: University research centers and scientific and technical human capital," *Research Policy*, vol. 39, no. 5, pp. 613–624, Jun. 2010, doi: https://doi.org/10.1016/j.respol.2010.02.013.

[101] James D. Lewis *et al.*, "Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease," *Cell Host & Microbe*, vol. 18, no. 4, pp. 489–500, Oct. 2015, doi: https://doi.org/10.1016/j.chom.2015.09.008.

[102] W. Glänzel and A. Schubert, "Domesticity and internationality in co-authorship, references and citations," *Scientometrics*, vol. 65, no. 3, pp. 323–342, Dec. 2005, doi: https://doi.org/10.1007/s11192-005-0277-0.

[103] A. M. Abbasi, M. A. Khan, N. Khan, and M. H. Shah, "Ethnobotanical survey of medicinally important wild edible fruits species used by tribal communities of Lesser Himalayas-Pakistan," *Journal of Ethnopharmacology*, vol. 148, no. 2, pp. 528–536, Jul. 2013, doi: https://doi.org/10.1016/j.jep.2013.04.050.

[104] E. Yan and Y. Ding, "Applying centrality measures to impact analysis: A coauthor ship network analysis," *Journal of the American Society for Information Science and Technology*,

vol. 60, no. 10, pp. 2107–2118, Oct. 2009, doi: https://doi.org/10.1002/asi.21128.

[105] L. Leydesdorff and C. S. Wagner, "International collaboration in science and the formation of a core group," *Journal of Informetrics*, vol. 2, no. 4, pp. 317–325, Oct. 2008, doi: https://doi.org/10.1016/j.joi.2008.07.003.

[106] B. D'Ippolito and C.-C. Rüling, "Research collaboration in Large Scale Research Infrastructures: Collaboration types and policy implications," *Research Policy*, vol. 48, no. 5, pp. 1282–1296, Jun. 2019, doi: https://doi.org/10.1016/j.respol.2019.01.011.

[107] J. Sylvan. Katz and B. R. Martin, "What is research collaboration?," *Research Policy*, vol. 26, no. 1, pp. 1–18, Mar. 1997, doi: https://doi.org/10.1016/s0048-7333(96)00917-1.

[108] A. Gazni and F. Didegah, "Investigating different types of research collaboration and citation impact: a case study of Harvard University's publications," *Scientometrics*, vol. 87, no. 2, pp. 251–265, Jan. 2011, doi: https://doi.org/10.1007/s11192-011-0343-8.

[109] E. FRICK, C. RIEDNER, M. J. FEGG, S. HAUF, and G. D. BORASIO, "A clinical interview assessing cancer patients' spiritual needs and preferences," *European Journal of Cancer Care*, vol. 15, no. 3, pp. 238–243, Jul. 2006, doi: https://doi.org/10.1111/j.1365-2354.2005.00646.x.

[110] E. Jacob and P. Mörters, "Spatial preferential attachment networks: Power laws and clustering coefficients," *The Annals of Applied Probability*, vol. 25, no. 2, Apr. 2015, doi: https://doi.org/10.1214/14-aap1006.

[111] O. Cimenler, K. A. Reeves, and J. Skvoretz, "A regression analysis of researchers' social network metrics on their citation performance in a college of engineering," *Journal of Informetrics*, vol. 8, no. 3, pp. 667–682, Jul. 2014, doi: https://doi.org/10.1016/j.joi.2014.06.004.

[112] J. A. Salomon *et al.*, "Disability weights for the Global Burden of Disease 2013 study," *The Lancet Global Health*, vol. 3, no. 11, pp. e712–e723, Nov. 2015, doi: https://doi.org/10.1016/s2214-109x(15)00069-8.

[113] M. H. Dueck, M. Klimek, S. Appenrodt, C. Weigand, and U. Boerner, "Trends but Not Individual Values of Central Venous Oxygen Saturation Agree with Mixed Venous Oxygen Saturation during Varying Hemodynamic Conditions," *Anesthesiology*, vol. 103, no. 2, pp. 249–257, Aug. 2005, doi: https://doi.org/10.1097/00000542-200508000-00007.

[114] V. Goyal, O. Pandey, A. Sahai, and B. Waters, "Attribute-based encryption for fine-grained access control of encrypted data," *Proceedings of the 13th ACM conference on Computer and communications security - CCS '06*, 2006, doi: https://doi.org/10.1145/1180405.1180418.

[115] Raija Laukkanen, P. Oja, M. Pasanen, and Ilkka Vuori, "Validity of a two kilometre walking test for estimating maximal aerobic power in overweight adults.," *PubMed*, vol. 16, no. 4, pp. 263–8, Apr. 1992.

[116] M. G. I. Langille *et al.*, "Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences," *Nature Biotechnology*, vol. 31, no. 9, pp. 814–821, Aug. 2013, doi: https://doi.org/10.1038/nbt.2676.

[117] C. Olmeda□Gómez, A. Perianes□Rodriguez, M. Antonia Ovalle□Perandones, V. P. Guerrero□Bote, and F. de Moya Anegón, "Visualization of scientific co□authorship in Spanish universities," *Aslib Proceedings*, vol. 61, no. 1, pp. 83–100, Jan. 2009, doi: https://doi.org/10.1108/00012530910932302.

[118] A. Perianes-Rodriguez, L. Waltman, and N. J. van Eck, "Constructing bibliometric networks: A comparison between full and fractional counting," *Journal of Informetrics*, vol. 10, no. 4, pp. 1178–1195, Nov. 2016, doi: https://doi.org/10.1016/j.joi.2016.10.006.

[119] B. S. Smith, D. S. Murray, J. D. Green, W. M. Wanyahaya, and D. L. Weeks, "Interference of Three Annual Grasses with Grain Sorghum (Sorghum bicolor)," *Weed Technology*, vol. 4, no. 2, pp. 245–249, Jun. 1990, doi: https://doi.org/10.1017/s0890037x00025343.

[120] J. Bai and P. Perron, "Computation and analysis of multiple structural change models," *Journal of Applied Econometrics*, vol. 18, no. 1, pp. 1–22, Jan. 2003, doi: https://doi.org/10.1002/jae.659.

[121] W. Glänzel and B. Thijs, "Does co-authorship inflate the share of self-citations?," *Scientometrics*, vol. 61, no. 3, pp. 395–404, 2004, doi: https://doi.org/10.1023/b:scie.0000045117.13348.b1.

[122]A. Abbasi, L. Hossain, and L. Leydesdorff, "Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks," *Journal of Informetrics*, vol. 6, no. 3, pp. 403–412, Jul. 2012, doi: https://doi.org/10.1016/j.joi.2012.01.002.

[123] X. Wang, H. Bai, Z. Yao, A. Liu, and G. Shi, "Electrically conductive and mechanically strong biomimetic chitosan/reduced graphene oxide composite films," *Journal of Materials Chemistry*, vol. 20, no. 41, p. 9032, 2010, doi: https://doi.org/10.1039/c0jm01852j.

[124] M. A. Moritz *et al.*, "Climate change and disruptions to global fire activity," *Ecosphere*, vol. 3, no. 6, p. art49, Jun. 2012, doi: https://doi.org/10.1890/es11-00345.1.

[125] P. Zimmermann *et al.*, "ExpressionData - A public resource of high quality curated datasets representing gene expression across anatomy, development and experimental conditions," *BioData Mining*, vol. 7, no. 1, Aug. 2014, doi: https://doi.org/10.1186/1756-0381-7-18.

[126] C. Arrighi *et al.*, "multi-risk assessment in a historical city," *Natural Hazards*, Jan. 2022, doi: https://doi.org/10.1007/s11069-021-05125-6.

[127] Riccardo Dondi, Mohammad Mehdi Hosseinzadeh, G. Mauri, and Italo Zoppis, "Top-k overlapping densest subgraphs: approximation algorithms and computational complexity," *Journal of Combinatorial Optimization*, vol. 41, no. 1, pp. 80–104, Nov. 2020, doi: https://doi.org/10.1007/s10878-020-00664-3.

[128] A. Timmermann *et al.*, "El Niño–Southern Oscillation complexity," *Nature*, vol. 559, no. 7715, pp. 535–545, Jul. 2018, doi: https://doi.org/10.1038/s41586-018-0252-6.

[129] S. Navlakha and C. Kingsford, "The power of protein interaction networks for associating genes with diseases," *Bioinformatics*, vol. 26, no. 8, pp. 1057–1063, Feb. 2010, doi: https://doi.org/10.1093/bioinformatics/btq076.

[130] M. Hrachowitz *et al.*, "A decade of Predictions in Ungauged Basins (PUB)—a review," *Hydrological Sciences Journal*, vol. 58, no. 6, pp. 1198–1255, Jun. 2013, doi: https://doi.org/10.1080/02626667.2013.803183.

[131] S. Borgatti, "The Network Paradigm in Organizational Research: A Review and Typology," *Journal of Management*, vol. 29, no. 6, pp. 991–1013, Dec. 2003, doi: https://doi.org/10.1016/s0149-2063(03)00087-4.

[132] B. Corominas-Murtra, J. Goñi, R. V. Solé, and C. Rodríguez-Caso, "On the origins of hierarchy in complex networks," *Proceedings of the National Academy of Sciences*, vol. 110, no. 33, pp. 13316–13321, Jul. 2013, doi: https://doi.org/10.1073/pnas.1300832110.

[133] A. Bukhari and X. Liu, "A Web service search engine for large-scale Web service discovery based on the probabilistic topic modeling and clustering," *Service Oriented Computing and Applications*, vol. 12, no. 2, pp. 169–182, Mar. 2018, doi: https://doi.org/10.1007/s11761-018-0232-6.

[134] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tompkins, and E. Upfal, "The Web as a graph," *Symposium on Principles of Database Systems*, May 2000, doi: https://doi.org/10.1145/335168.335170.

[135] E. Nathan, A. Zakrzewska, J. Riedy, and D. Bader, "Local Community Detection in Dynamic Graphs Using Personalized Centrality," *Algorithms*, vol. 10, no. 3, p. 102, Aug. 2017, doi: https://doi.org/10.3390/a10030102.

[136] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, no. 2, Feb. 2004, doi: https://doi.org/10.1103/physreve.69.026113.

[137] C. Monroe and J. Newman, "The Impact of Elastic Deformation on Deposition Kinetics at Lithium/Polymer Interfaces," *Journal of The Electrochemical Society*, vol. 152, no. 2, p. A396, 2005, doi: https://doi.org/10.1149/1.1850854.

[138] P. Bonacich, "Power and Centrality: A Family of Measures," *American Journal of Sociology*, vol. 92, no. 5, pp. 1170–1182, Mar. 1987, doi: https://doi.org/10.1086/228631.

[139] Marie-Catherine de Marneffe and C. D. Manning, "The Stanford typed dependencies representation," Aug. 2008, doi: https://doi.org/10.3115/1608858.1608859.

[140] M. Craven and C. D. Page, "Big Data in Healthcare: Opportunities and Challenges," *Big Data*, vol. 3, no. 4, pp. 209–210, Dec. 2015, doi: https://doi.org/10.1089/big.2015.29001.mcr.

[141]B. Zetsche *et al.*, "Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System," *Cell*, vol. 163, no. 3, pp. 759–771, Oct. 2015, doi: https://doi.org/10.1016/j.cell.2015.09.038.

[142] M. Tatliyer and N. Gur, "Individualism and Working Hours: Macro-Level Evidence,"

*Social Indicators Research*, Aug. 2021, doi: https://doi.org/10.1007/s11205-021-02771-y.

[143]S. K. Brooks *et al.*, "The Psychological Impact of Quarantine and How to Reduce it: Rapid Review of the Evidence," *The Lancet*, vol. 395, no. 10227, pp. 912–920, Feb. 2020, doi: https://doi.org/10.1016/S0140-6736(20)30460-8.

[144] J. Adams, "Early citation counts correlate with accumulated impact," *Scientometrics*, vol. 63, no. 3, pp. 567–581, Jun. 2005, doi: https://doi.org/10.1007/s11192-005-0228-9.

[145] N. Maflahi and M. Thelwall, "How quickly do publications get read? The evolution of mendeley reader counts for new articles," *Journal of the Association for Information Science and Technology*, vol. 69, no. 1, pp. 158–167, Aug. 2017, doi: https://doi.org/10.1002/asi.23909.

[146] É. Archambault and V. Larivière, "History of the journal impact factor: Contingencies and consequences," *Scientometrics*, vol. 79, no. 3, pp. 635–649, Jan. 2009, doi: https://doi.org/10.1007/s11192-007-2036-x.

[147]J. Bollen, H. Van de Sompel, A. Hagberg, and R. Chute, "A Principal Component Analysis of 39 Scientific Impact Measures," *PLoS ONE*, vol. 4, no. 6, p. e6022, Jun. 2009, doi: https://doi.org/10.1371/journal.pone.0006022.

[148]L. Engqvist and J. G. Frommen, "The h-index and self-citations," *Trends in Ecology & Evolution*, vol. 23, no. 5, pp. 250–252, May 2008, doi: https://doi.org/10.1016/j.tree.2008.01.009.

[149]N. Carayol and M. Matt, "Individual and collective determinants of academic scientists' productivity," *Information Economics and Policy*, vol. 18, no. 1, pp. 55–72, Mar. 2006, doi: https://doi.org/10.1016/j.infoecopol.2005.09.002.

[150]R. COSTAS and M. BORDONS, "The h-index: Advantages, limitations and its relation

with other bibliometric indicators at the micro level," *Journal of Informetrics*, vol. 1, no. 3, pp. 193–203, Jul. 2007, doi: https://doi.org/10.1016/j.joi.2007.02.001.

[151]L. Egghe, "Theory and practise of the g-index," *Scientometrics*, vol. 69, no. 1, pp. 131–152, Oct. 2006, doi: https://doi.org/10.1007/s11192-006-0144-7.

[152]D. J. Lang *et al.*, "Transdisciplinary research in sustainability science: practice, principles, and challenges," *Sustainability Science*, vol. 7, no. S1, pp. 25–43, Feb. 2012, doi: https://doi.org/10.1007/s11625-011-0149-x.

[153]W. Glänzel, "On the h-index - A mathematical approach to a new measure of publication activity and citation impact," *Scientometrics*, vol. 67, no. 2, pp. 315–321, May 2006, doi: https://doi.org/10.1007/s11192-006-0102-4.

[154]L. Šubelj, N. J. van Eck, and L. Waltman, "Clustering Scientific Publications Based on Citation Relations: A Systematic Comparison of Different Methods," *PLOS ONE*, vol. 11, no. 4, p. e0154404, Apr. 2016, doi: https://doi.org/10.1371/journal.pone.0154404.

[155] Loet Leydesdorff, "On the normalization and visualization of author co-citation data: Salton's Cosineversus the Jaccard index," vol. 59, no. 1, pp. 77–85, Jan. 2007, doi: https://doi.org/10.1002/asi.20732.

[156] H. F. Moed, "New developments in the use of citation analysis in research evaluation," *Archivum Immunologiae et Therapiae Experimentalis*, vol. 57, no. 1, pp. 13–18, Feb. 2009, doi: https://doi.org/10.1007/s00005-009-0001-5.

[157]S. Alonso, F. J. Cabrerizo, E. Herrera-Viedma, and F. Herrera, "h-Index: A review focused in its variants, computation and standardization for different scientific fields," *Journal of Informetrics*, vol. 3, no. 4, pp. 273–289, Oct. 2009, doi: https://doi.org/10.1016/j.joi.2009.04.001.

[158]Y.-W. Chang, "Examining interdisciplinarity of library and information science (LIS) based on LIS articles contributed by non-LIS authors," *Scientometrics*, vol. 116, no. 3, pp. 1589–1613, Jun. 2018, doi: https://doi.org/10.1007/s11192-018-2822-7.

[159]J. J. Louviere, "What If Consumer Experiments Impact Variances as well as Means? Response Variability as a Behavioral Phenomenon," *Journal of Consumer Research*, vol. 28, no. 3, pp. 506–511, Dec. 2001, doi: https://doi.org/10.1086/323739.

[160]L. Waltman *et al.*, "The Leiden ranking 2011/2012: Data collection, indicators, and interpretation," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 12, pp. 2419–2432, Nov. 2012, doi: https://doi.org/10.1002/asi.22708.

[161]K. Mahmood and F. Shafique, "Changing research scenario in Pakistan and demand for research qualified LIS professionals," *Library Review*, vol. 59, no. 4, pp. 291–303, Apr. 2010, doi: https://doi.org/10.1108/00242531011038596.

[162]M. Petri *et al.*, "Derivation and validation of the Systemic Lupus International Collaborating Clinics classification criteria for systemic lupus erythematosus," *Arthritis & Rheumatism*, vol. 64, no. 8, pp. 2677–2686, Jul. 2012, doi: https://doi.org/10.1002/art.34473.