The p-index: Theory and Applications

Upul Senanayake

A thesis submitted in fulfilment
of the requirements for the degree of
Master of Philosophy

School of Civil Engineering
The University of Sydney

August 2014

Declaration

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the University or other institute of higher learning, except where due acknowledgement has been made in the text.

Upul Senanayake

31 August, 2014

Abstract

Upul Senanayake
The University of Sydney

Master of Philosophy
August 2014

The p-index: Theory and Applications

This thesis defines and presents a new index to measure the scientific output of researchers, called the p-index. Typically, such indices are one-dimensional as they only consider one parameter at a time using measures like the number of citations, the number of papers and the average number of citations one receives. In that aspect, the introduction of h-index was revolutionary because it takes two parameters in to consideration. The h-index is defined as the h number of papers that have at least h citations each. As it was an improvement on the measures used at the time and it allowed one to compress the scientific achievements of a researcher in to a single metric, the h-index garnered respect and has been in constant use ever since. However many have noted the h-index has drawbacks which can lead to the possibility of exploiting the h-index for one's personal gain. The h-index does not attach weights to the citations, which may lead to potential misuse if documents are created purely to cite others. In this thesis, a new ranking metric, the p-index (pagerank-index), is introduced. This ranking is computed from the underlying citation network of papers and uses the pagerank algorithm in its computation. The index is a percentile score which can potentially be implemented in public databases such as Google scholar, and can be applied at many levels of abstraction.

In this research, we demonstrate that the p-index provides a fairer ranking of scientists than the h-index and its variants. We do this by, (i) simulating a realistic model of the evolution of citation and collaboration networks in a particular field, (ii) using real world citation datasets, and comparing the h-index and p-index of scientists under a number of scenarios. The results from the simulated system show that the p-index is immune to the author behaviors that can result in artificially bloated h-index values. Analysis is applied to two real-world datasets: a dataset of scientists from the field of quantum game theory and a dataset of scientists from the field of high energy physics - theory (HEP-TH). The first dataset is sourced for this study from Google scholar, while the second dataset is constructed from an available citation network. We show that, while the popularly used h-index rewards authors who collaborate extensively and publish in higher volumes, the p-index rewards hardworking authors who contribute more to each paper they write, as well as authors who publish in high impact and well-cited journals. As such, it could be argued that the p-index is a 'fairer' metric of the productivity and impact of scientists. The p-index relies on the actual underlying citation network to measure the real impact of each paper rather than using the impact factors of publications. Finally, this thesis investigates the effect of changing assortativity on the p-index and provides evidence to indicate that p-index changes as the assortativity is changed.

Acknowledgements

When one can see the light at the end of the tunnel, one tends to move towards it with renewed vigor, for finally the challenging times are over when one stumbles upon the light. I'm humbled that I was able to reach the light at the end of my MPhil candidature in a relatively shorter time span, giving me more time to focus on what's really important and prioritize on what should I do and what I shouldn't. It has been a wonderful journey, getting acclimated to a new country and a cutting-edge research environment. It was a pleasure to work with a very supportive supervisor who made the transition smoother. I have also learned to thrive under pressure and started to truly explore the wonderful questions in research.

However, the more I thrived, the more I found I could learn from my colleagues and that was a key factor for me to succeed in this arduous journey towards my MPhil. I owe my heartfelt gratitude to those who have been there for me through and thin, been patient with me, listening to my raves about research and keeping me afloat. My gratitude for them cannot be expressed, but let me make an attempt.

First of all, I would like to thank my supervisors, Dr. Mahendra Piraveen and Prof. Albert Zomaya, for the guidance they provided me throughout this journey. Piraveen, you've always encouraged me to pursue my goals and reminded me of the timelines to which we had to adhere. Your insights in to research as well as ethics and life were truly inspiring and one of a kind. Albert, you've provided me with numerous advises and resources needed to complete my research. You two have my utmost gratitude. I would also like to thank Dr. Kennuth Chung and Dr. Ehssan Sankhe for the valuable advice and comments you've given me.

The work environment I had in the university was productive and entertaining. I would like to thank Yamuna Subramanium, Dharshana Kasthurirathne, Gnana Kumar, Arif Khan, Rabi-ul, Andrew Leula and Chandrashekhar Bapat for their support and unwavering friendship which enabled me to endure all the hard times. I would fail my duty if I don't thank the administrative staff of the faculty including but not limited to Catherine, Christine, Madelon, Linda and Maria.

None of this wouldn't have been possible if the lecturers back in University of Peradeniya inspired me to pursue research. My sincere gratitude goes to Dr. Roshan Ragel, who has always been supportive, accessible and a really good friend rather than a supervisor. I'm grateful to Dr. Dhammika Elkaduwe and the staff of the Department of Computer Engineering, University of Peradeniya for all they've done for me. Tracing back to my school days, I'm forever grateful to the wonderful panel of teachers I had in Ranabima Royal College, Kandy.

With research, you often tend to live in your own world filled with questions, but time to time, you need to break that bubble to proceed further. For me, my friends have always been that and more. I'm thankful to Hiran Milinda and Nilusha Nanayakkara for always providing a home for me whenever I needed one. I enjoyed those lengthy discussions we have about life which is a huge distraction from research that I enjoy and look forward to. I'm thankful to Kasun Rathnayake and Kithsiri Ehelepola for understanding me, sticking with me and taking care of me when I needed it the most. Yasangie Edirisinghe and Nalika Ulapane, your fights still amaze me and I'm thankful for that. My sincere gratitude goes to Chandana De Silva, Buddhi Dayananda and their spouses for their friendship and kind advices. During my candidature, many friends helped me and I'm grateful to all of you. You all deserve to be mentioned by name, however I'm unable to do that and I hope you'll forgive me for that.

I also acknowledge and thank my thesis examiners for their valuable feedback and comments. I'm humbled by your compliments and they meant a lot to me.

Finally, I cannot thank my parents and family enough for standing by me all the time and making it possible for me to pursue this MPhil. Had it not been for you, I dread to imagine where I would be and the fact that you are always there for me no matter what, makes me strong.

To my mother and father.

Publications

The following publications and manuscripts-under-review have resulted from the candidature for this degree:

1. U. Senanayake, M. Piraveenan, A.Y Zomaya, "The p-index: Theory and Applications", journal paper in-preparation, 2014.
2. U. Senanayake, M. Piraveenan, A.Y Zomaya, "The p-index: Ranking scientists from the field of quantum game theory using p-index", pending review in *IEEE Symposium Series on Computational Intelligence*, 2014.
3. U. Senanayake, P. Szot, M. Piraveenan, D. Kasthurirathna, "The performance of page-rank algorithm under degree preserving perturbations", pending review in *IEEE Symposium Series on Computational Intelligence*, 2014.*
4. D. Kasthurirathna, H. Nguyen, M. Piraveenan, U. Senanayake, "Optimisation of strategy placements for public good in complex networks", in-press *The ASE International Conference on Social Computing*, 2014*.
5. U. Senanayake, M. Piraveenan, A.Y Zomaya, "The p-index: Ranking Scientist using Network Dynamics", *International Conference on Computational Science*, 2014.
6. G. Thedchanamoorthy, M. Piraveenan, D. Kasthururathna, U. Senanayake, "Node Assortativity in Complex Networks: An Alternative Approach", *International Conference on Computational Science*, 2014.*
7. G. Thedchanamoorthy, M. Piraveenan, S. Uddin, U. Senanayake, "Influence of Vaccination Strategies and Topology on the Herd Immunity of Complex Networks", in-press *Journal of Social Network Analysis and Mining*, 2014.*
8. Y. Sun, L. Rossi, H. Luan, C-C. Shen, J. Miller, R. Wang, J. Lizier, M. Prokopenko, U. Senanayake, "Information Transfer in Swarms with Leaders", *Collective Intelligence*, 2014.*

   The papers marked with an asterick (*) have *not* directly contributed to this thesis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

It is the human nature to compare oneself with others although it is known that comparison can be a double edged sword that can give one happiness or break one in to pieces. Whether an individual does or does not like comparison, we are inherently subject to comparison all the time. Some of these comparisons are entirely subjective, for an example, if a passerby judges one by their looks. They might decide that they are more or less handsome than oneself depending on their perception. On the other hand, some comparisons are entirely objective, for instance the marks obtained for a mathematics examination by two students. That comparison depends on two quantities and therein constitutes an objective comparison. Other comparisons involve a subjective measure and an objective measure simultaneously. The concern with using subjective comparisons is the credibility of the outcome because, if repeated, the results might not be the same. Therefore an objective evaluation system is preferred over a subjective evaluation system. This thesis aims to establish an objective evaluation system for scientists, called the p-index, who can then use this system for different purposes. This metric will aid authorities in evaluating the status of different scientists. For example, currently the success of a research grant may depend on one's h-index but if the h-index is not a fair rating, then where one stands as judged by the h-index may not necessarily be one's deserving place. Hence this research also attempts to make this new ranking system robust and resilient against manipulations.

The next sections of this chapter are organized as follows. The second section describes the research question and the third section discusses the motivation for pursuing this topic. The fourth section presents the research approach and the fifth section lists down the principle contributions of the thesis. The last section outlines the structure of this thesis.

1.2 Objectives

The ultimate outcome of this research is to develop a ranking system that can evaluate the quality of the research output of a researcher that is robust, objective and fairer than the existing methods. The new evaluation metric, which will be defined in the next chapters is known as p-index (page-rank index). The underlying complex network dynamics are used to develop the evaluation system so that it would not need any intervention and can determine the quality of one's work by the network dynamics alone. This thesis also investigates the fairness of the new evaluation system. Both synthesized networks and real networks are considered to demonstrate the suitability and feasibility of the new evaluation metric. Finally this research sheds light on how inadequate mainstream evaluation metrics can be and the fairness of the new p-index against these manipulations. Hence the research question is formulated as follows.

> What are the mechanisms by which the underlying structure of academic networks (citation and collaboration networks) can be used to determine the quality of the research output by a researcher objectively and robustly?

1.3 Motivation

The h-index is one of the three metrics used in Google scholar to quantify the impact of a scientist's research (the others are the number of citations, and the i10-index). The h-index is arguably the most sophisticated among the three, as to some extent it accounts for both the quality and the quantity of a scientists' research publications. The h-index was defined by J.E Hirsch [29] as the h number of papers that have at least h number of citations. This implies that if the scientist has N_T number of papers, $(N_T - h)$ number of papers have no more than h citations. Previous to the h-index, the common measures used to measure a researcher's impact were the number of citations, the number of papers and the number of citations per paper. These simplistic measures all had clear drawbacks. For example, the number of papers does not take into account the quality or impact of the papers, while the number of citations may be inflated by a few papers with a great number of citations co-authored by many authors, and the citations per paper measure rewards low productivity. Hirsch argued that his h-index addresses these drawbacks by drawing information from both number of papers and citation count.

The h-index has been widely accepted by the scientific community [9], and there is evidence to suggest that successful scientists have comparatively higher h-index values, at least when 'success can be measured by employability and grant success [10]. However, there has been considerable debate about the merits of the h-index. Many of its alleged shortcomings have been pointed out [22, 9] and many variants and alternatives have been suggested [29, 33, 37]. For instance, the h-index depends on the time span of the career of a scientist [29, 37]. As such, those scientists who may be brilliant but are in the early stages of their career are reflected unfavorably. Hirsch suggested using a time-compensated measure to overcome this restriction. Furthermore, L. Egghe [21] argued that, according to the definition of h-index, once a paper belongs to the top h class (among the papers ranked h or higher), it becomes unimportant whether these papers are further cited or not. Scientists who have high-impact papers are penalized, and this is evident by the difference between the lower bound of citations, h^2, proposed by the h-index, and the actual number of citations. This had been observed by Hirsch himself [29]. Egghe therefore suggested the g-index, where the g-index is defined as the highest number g of papers that together received g or more citations. Similarly, the R-index and AR-index introduced by Jin et al. [33] measure the citation intensity of the h-core, also taking in to consideration the age of the papers. They suggested that these measures could be used together with the h-index to complement each other. There are many other variants and alternative indices introduced, as summarized by [49].

A review of the existing variants of h-index by Bornmann et al. [10] suggested that all these variants fall into two types: those that describe the size of the h-core, and those that describe the impact of the h-core. The first type redefines the size of the h-core and typically makes it larger (such as the g-index) or adds additional information about it (such as the r-index), while the second type analyses the citation count of the h-core, negating the disadvantage attributed to scientists who have a few high-impact papers. However, a fundamental issue not noted or addressed by these measures is that all citations are treated equally. Yet, it is clear that a citation by a paper from a highly regarded journal, such as *Nature*, should be treated differently from a citation by a workshop paper or a technical report. If this does not happen, locally famous authors whose research does not have global impact but gets cited by their colleagues in their country or research circle can get rewarded. Moreover, if all citations are treated equally, then 'massaging' the h-index becomes possible by publishing papers in whatever available forum, purely with the intention of citing other papers by the same group of authors or collaborators. This process has the danger of encouraging the creation of a huge body of academic papers which nobody reads, let alone utilizes for further research or application. Thus, part of the motivation behind this research is to reduce/discourage this shortcoming and possible manipulative behavior by coming up with a fairer metric that is robust against manipulations.

1.4 Approach

In this thesis, therefore, it is proposed that the underlying network of citations be utilized to evolve an index which is dynamic and rewards citation by impactful documents. Our goal is to formulate and test a robust metric which rewards true excellence by giving higher weight to citations from documents with higher academic standing. Therefore, we introduce the p-index, which is designed to address the drawbacks of existing indices by utilizing the underlying citation network dynamics. The index is calculated by running the well-known pagerank algorithm [48] on the evolving citation network to give scores to papers. This has been explored before by Walker et al. and Chen et al. [57, 13] in order to find 'scientific-gems of publications which are less apparent if the number of citations is used as the ranking metric. These researchers have used a modified algorithm called 'citerank in order to rank scientific publications and their approach is summarized with merits and pitfalls by Maslov et al. [40]. This research differs in that the aim is to rank scientists as opposed to ranking publications and ranking publications is only an intermediary step. The original form of the pagerank algorithm is used to come up with pagerank values for each publication and the score of an author is calculated as a weighted sum of the scores of the papers he/she has contributed. By employing a realistic simulation system that synthesizes the evolution process of citation and collaboration networks in an emerging field, it is demonstrated that the p-index is fairer than the h-index in many scenarios in which the authors may be able to massage or manipulate their h-index. In particular, it is shown that while it is possible to massage h-index by publishing papers in low-impact forums which are used simply to cite other papers of the same group of authors, this is largely impossible with the p-index. It is also shown that while the h-index rewards collaboration and authors who focus on quantity, the p-index is much more balanced and equally rewards individual brilliance and quality of papers.

It could be asked, whether citations could have been weighed by using the impact factor of journals in which the citing documents are published. However, careful analysis reveals this method has many pitfalls. There is considerable debate about the dependability of the impact factors of journals [56], and it cannot be considered that impact factors are sacrosanct. In any case, conference papers and relatively new journals do not have impact factors. There is also no guarantee, that a paper which is published in a journal with high impact factor will itself have high impact. The intention of this research is to reward the quality of the citing document, not the quality of journal in which it is published. Even if the number of citations received by the citing document are counted and the citations are weighed accordingly in a dynamic manner, it is possible that these weights could be manipulated again by less impactful documents: thus the influence of such documents only becomes indirect and does not go away. The intention therefore, is to introduce an index which has ‘infinite levels of feedback with the impact of citing documents is factored in as much as possible. This is further elaborated in the second and third chapters, when the related work and implementation details are discussed. Feedback of this level could only be achieved by fully utilizing the overall dynamic citation network by an elegant algorithm such as pagerank, which has had proven success in measuring the impact of such feedback loops in the Google search engine.

1.5 Principal Contributions

The main contributions of this thesis are:

- Demonstrating the manipulation possibilities of the h-index. Although it has been suggested that there are many drawbacks to the h-index, this research is the first to provide systematic methodology on how to counteract these manipulations and to present analytical evidence that the said systematic methodology can indeed mitigate the undue increase in one's h-index.

- Utilizing underlying network dynamics to develop an evaluation system for a researcher's output. Not only are the impactful documents found using link analysis of the citation network, but a ranking system taking the interplay between citation networks and collaboration networks into consideration is also developed.

- Introducing a new index, known as the p-index which is fairer than the existing ranking metrics and is also robust against manipulations.

- Developing realistic evolving citation and collaboration networks that mimic the evolution of the real world academic networks.

- Presenting evidence to suggest that the p-index is robust against the manipulations that are possible with the h-index. We systematically provide evidence from three simulated scenarios and two real world examples to support this claim.

- Applying the p-index to two real world citation networks and demonstrating the characteristics of the p-index that makes it fairer and more robust than the other existing indices.

- Demonstrating the characteristics of the p-index under degree preserving perturbations to the citation network. This provides evidence to support the claim that the p-index considers the perceived importance of the citing paper without just counting the number of citations one receives.

This thesis provides a valuable insight on the evaluation systems available for researchers and how these systems can be manipulated. It also presents a new evaluation system that is both fairer and more robust and provides evidence on why it is beneficial to the scientific community. Finally, it demonstrates the suitability and feasibility of the new system on two real world networks to further establish the new index.

1.6 Thesis Structure

This thesis is organized as follows. The next chapter discusses about underlying network structure of the academic networks and the characteristics they portray. This information is critical in deriving a fairer evaluation system for researchers. The contributions several authors have already carried out in using network dynamics for different intentions which are indirectly related to the work carried out in this thesis is also analyzed. The third chapter focuses on the definition and methodology of p-index while the fourth chapter describes how a simulated system was used in order to demonstrate the drawbacks of the h-index. The same simulated system is then used to present evidence to suggest that the p-index is resilient against these drawbacks and emerges as a fairer metric than the h-index. The fifth and sixth chapters elaborate on how the p-index was applied to real academic networks and present the results we obtained. Based on the analysis of two real world systems, the p-index emerges as the fairer metric in both systems. These systems as well as the application process and analysis are also discussed in detail in those chapters. The seventh chapter explains how the p-index reacts to degree preserving perturbations, reiterating its dynamicity. The final chapter concludes this thesis and summarizes the achievements of this research and how the research question of the thesis was solved.

Chapter 2

Background

In this chapter, some basic concepts on which the rest of the thesis is based on are introduced. This chapter introduces existing concepts while the new concepts and solution formulation are addressed in later chapters.

2.1 Introduction

The indices currently used by scholarly databases, such as Google scholar, to rank scientists use citations but do not attach weights to the citations. Neither is the underlying network structure of citations considered in computing these metrics. This results in all citations being considered equal, meaning that scientists cited by well-recognized journals are not being rewarded and may lead to potential misuse if documents are created purely to cite others [9]. In order to circumvent this, this thesis introduces a new ranking metric, the p-index (pagerank-index), which is computed from the underlying citation network of papers, and uses the pagerank algorithm [48] in its computation. The p-index is a percentile score that can potentially be implemented in public databases such as Google scholar and can be applied at many levels of abstraction. The p-index metric aids in fairer ranking of scientists compared to h-index [29] and its variants [21, 33, 22, 31, 49, 50, 60]. This is demonstrated by simulating a realistic model of the evolution of citation and collaboration networks in a particular field and comparing the h-index and p-index of scientists under a number of scenarios. The results show that the p-index is immune to author behaviors that can result in artificially bloated h-index values [36]. In addition, the p-index algorithm is applied to two real world networks. An evolving citation network was constructed using ‘Quantum Game Theory Google Scholar page which is similar to the evolving citation network simulated in this research. It was necessary to construct this evolving network because the existing citation networks are only a snapshot at the end-point and do not portray the evolution of the network which is integral to verify the simulated results. Another citation network was also constructed from ‘High Energy Physics - Theory graph data that allowed a sufficiently large snapshot of a matured academic network to be provided. The analysis of application of p-index for these two academic networks could provide further proof to the validity of the p-index and would establish p-index as a mainstream ranking algorithm that can complement h-index and its variants.

The rest of this chapter is organized as follows. The second section introduces complex networks and identifies the relevance of complex networks to academic networks while the third section discusses the available methods to rank scientists. The fourth section covers the pagerank algorithm that is at the core of the new ranking metric p-index while the fifth and final section summarizes the applications of pagerank algorithm in academic networks.

2.2 Complex Networks

Complex networks are the networks that exhibit non-trivial topological characteristics [1]. It is possible to model most of real-world networks as complex networks. Some well-known examples are social networks, biological networks, citation networks, collaboration networks and the Internet [4, 1, 45]. As such, complex network analysis has become a growing research area in the recent past. The specialty in complex network analysis is that it encompasses the knowledge bodies of several prominent natural sciences like physics, mathematics, computer science, social science and psychology and integrates them to one single body of knowledge that tries to explain the phenomenon of complex networks. The work in this research directly draws from the body of knowledge of complex network analysis. In order to emphasize the importance of complex network analysis, a few well studied complex networks occurring in the nature as well as man-made systems are presented here.

- Cell: a cell is modeled as a complex network of chemicals connected to each other by chemical reactions.

- Internet: the Internet is routers and computers connected through various physical or wireless links which can be modeled as a complex network.
- World Wide Web (WWW): the WWW is a virtual collection of web pages that are linked to each other by hyperlinks and can be modeled as a complex network.
- Citation network: a citation network is research publications connected to each other by referring a previously published paper.
- Ecological networks: ecological networks (or food webs) quantify the interaction between different species using nodes as species connected to each other by predator-prey relationship.

A network is a set of nodes (or vertices) interconnected with edges that characterizes the connections between the nodes. Mathematical graph theory is at the heart of network analysis and complex network analysis is also developed from that. Another area where networks have been studied extensively is social science, where nodes represent people and edges represent the relationships between them. However, in recent years, analyzing networks has shifted from analyzing small networks and node properties to the study of large networks and large-scale statistical properties of graphs. This was due to the availability of computational power which enables the analysis of networks with millions of nodes in a matter of minutes. The analysis of complex networks became possible as a direct result of this and an arsenal of statistical measures were developed for this purpose.

The primary focus of this research is to find statistical properties that would represent the structure and the behavior of complex networks, while the secondary focus is to model networks in such a way that the meanings of these properties could be interpreted. By using the models derived through the underlying statistical properties, it is possible to come up with a system that allows the citation network to analyze its links non-intrusively and evolve a ranking system that is robust and fairer than the existing ranking systems. Before proceeding to different types of networks that are used in this research, definitions of the terms used to characterize a complex network are presented.

- Vertex: vertex is the smallest unit in any network and is commonly known as node in computer science.

- Edge: two vertices are connected with an edge. This is commonly known as link in computer science.
- Directed/Undirected: if an edge between two nodes is directed, then the connection can only run in one direction. If the connection can go both ways, the edge is known as undirected edge.
- Degree: degree is the number of edges connected to a node. A directed graph would have an in-degree and out-degree that characterize the in-coming and out-going edges between nodes.

2.3 Properties of Complex Networks

This research intends to describe some of the properties of complex networks that keep occurring in real world networks most of the time. As the study is only interested in the properties that are related to the implementation of p-index, what is covered here as properties of complex networks is not complete or exhaustive. These properties would help us to understand and divide complex networks in to different behavioral groups.

2.3.1 Small-World Effect

In 1960s, Stanley Milgram [42] carried out a famous experiment where,in order to deliver the letters to a designated individual, letters were passed from person to person. Milgram approximated that a letter could reach the designated person with around six exchanges. Milgram was able to prove that a pair of vertices are connected with each other by a short path length. This is one of the first demonstrations of the small-world effect. The same effect has been studied and verified in a large number of real world networks. In an undirected network, the average shortest distance l between vertex pairs in a network can be defined using Equation 2.1, where d_{ij} is the shortest distance from vertex i to vertex j.

$$l = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij} \tag{2.1}$$

The definition of the small-world effect may be problematic when the network has vertex pairs that are not connected to each other. When this is the case, the average shortest distance l would reach infinity, which wouldn't be an accurate representation of the network. Hence l is defined for such networks as the average shortest distance between all pairs that have a connection between them. The small-world effect is very important in studying the dynamics of complex networks. For instance, if the spread of information is considered in a social network, the small-world effect suggests that the spread would be extremely fast. If Milgram's six steps of separation is applied to the spread of rumor on a social network, at most six steps would be needed for a rumor to reach from an arbitrary node to another arbitrary node. Another example would be the Internet where $l = 3.3$, implying that only 3.3 hops are needed for a packet to travel between two arbitrary nodes.

In recent years, the small-world effect has been mathematically characterized as the value of l scaling logarithmically or slower with network size for fixed mean degree. This gives a precise meaning to small-world effect found in real world networks. Accordingly, a large number of networks such as social networks, film actor networks, collaboration networks, the Internet, train routes as well as ecological networks are found to have small-world effect.

This research is specifically interested in small-world property because the network studied exhibits the small-world property. The collaboration network built on top of the citation network is a small-world network as indicated by Barabasi [1] as well as Newman [45]. The collaboration networks built are analyzed and the results presented to concur with their indications in the next chapters.

2.3.2 Clustering

In many networks, it is identified that if two arbitrary vertices A and B are connected and if B is connected to another vertex C, then there is a heightened probability that A is also connected to C. This is known as transitivity or clustering. This relationship can be easily understood if a social network is taken in to consideration. Transitivity implies that a friend of one's friend is likely to be a friend of one as well. This is quantified by clustering coefficient C as in Equation 2.2. In terms of network topology, increased clustering means there are an increased number of triangles (three vertices connected to each other) in the network.

$$C = \frac{\text{3 x number of triangles in the network}}{\text{number of connected triples of vertices}} \tag{2.2}$$

This definition is widely used in sociology. Another definition is given by Watts and Strogatz using a local value C_i is shown in Equation 2.3,

$$C_i = \frac{\text{number of triangles connected to vertex i}}{\text{number of connected triples centered on vertex i}} \tag{2.3}$$

and for vertices having degree 0 or 1, C_i is taken as 0 in order to come up with the equation for the clustering coefficient for the whole network as in Equation 2.4.

$$C = \frac{1}{n} \sum_i C_i \tag{2.4}$$

The values of clustering coefficient calculated by these two methods can be different. Hence the method highlighted by Equation 2.4 is used in this thesis. This is because although the value obtained by Equation 2.2 is easier to calculate analytically, the value obtained by Equation 2.4 is easier to obtain programmatically. Clustering has also been known as network density in the sociological literature. Complex networks like train routes, film actors, company directors and co-author networks are found to exhibit a high clustering coefficient while networks like power grid, electronic circuits and metabolic networks are found to exhibit relatively low clustering coefficients.

As the clustering coefficient evaluates the density of the connected triads in the network, it might also be interesting to examine longer loops of length four or above. This has been examined on numerous occasions without the establishment of a concrete theory. The work related to this end can be found from the work by the following authors. [25, 6, 11, 26, 44].

Clustering in a complex network is of importance to this research because the literature suggests [1, 45] that the collaboration networks have a high clustering coefficient. We present our results to concur this statement and further utilize the clustering to show interesting behavioral patterns in collaboration networks in the fifth and sixth chapters.

2.3.3 Degree Distribution

A degree distribution is a histogram of the degrees of the nodes in a graph. Mathematically, if p_k is defined as the fraction of nodes in the network that has a degree of k, the plot for p_k is known as the degree distribution. What this implies is that there is a p_k probability for a randomly chosen node to have a degree of k.

The simplest network type is the random network and the mathematical foundation for that was laid out by Erdos and Renyi [23], hence the commonly used term E-R networks. In a random network each edge is present or absent with an equal probability, therefore making the degree distribution a binomial distribution. For sufficiently large networks, this can also be a Poisson distribution. However, most real world networks do not have the same degree distribution as the random networks, implying that the real world networks are not random at all. In fact, real world networks are usually highly-right skewed and measuring the skewness can be tricky. Hence an alternative way to represent the degree distribution is used that plots the cumulative distribution function, as shown in Equation 2.5, which portrays the probability of a degree is greater than or equal to k.

$$P_k = \sum_{k'=k}^{\infty} p_{k'} \tag{2.5}$$

Scale free degree distribution is one of the most common degree distributions in real life networks such as citation networks and WWW. In a scale-free degree distribution, $p_k \sim k^{-\alpha}$ where α is a constant exponent. Similarly, exponential degree distributions behave with exponential tails as $p_k \sim e^{-\lambda k}$. Accordingly, if the logarithmic plot of degree distribution is linear, the degree distribution can be identified as a scale-free distribution and if the semi-logarithmic plot of degree distribution is linear, degree distribution can be identified as exponential. It should also be noted that a directed complex network will have two degree distributions: in-degree distribution and out-degree distribution, while there is also an integrated degree distribution where p_{jk} becomes a function of two variables that represent the fraction of nodes that simultaneously have in-degree of j and out-degree of k.

The degree distribution is at the heart of the complex network analysis and is directly related to the networks studied in this thesis. Citation networks are scale-free networks [1, 45] and the results of this thesis concur with this conclusion. Observations on the collaboration networks, which are suggested to be networks with small-world property are also presented. The knowledge that the citation networks are indeed scale-free is used in order to compare the citation networks with the Internet to justify the application of the pagerank algorithm.

2.3.4 Network Resilience

The resilience of a network is tightly related to its degree distribution. Network resilience is identified as the resilience of a network to random removal of its vertices. For instance, if vertices are continuously removed, the network would reach a disconnected state where no node can reach another because there are no possible paths among them. Different vertice removal techniques are also studied under this category; for instance the highest degree node can be removed, or start from one end of the network, removing the vertices towards the other end. In epidemiology, removing the highest degree node is of particular importance since if one can vaccinate the highest degree node, then it is likely that the spread of the virus can be mitigated significantly. As such, network resilience is closely studied and experimented with.

Albert et al [2] have shown that networks with power-law degree distribution are extremely resilient to random attacks, but vulnerable to targeted attacks. This is intuitive because the skewed nature of the power-law distribution entails that there are a comparably larger proportion of nodes having a low-degree while only a small percentage have a larger degree. So when a random attack occurs, it is probabilistically more likely that the removed vertice is a low-degree node rather than a high degree node. However in a targeted attack scenario, if the high-degree nodes are removed, the network is bound to fail faster. Albert et al have demonstrated this for WWW and Internet experimentally. Following their footsteps, many authors have studied the phenomena and have indicated that most networks are resilient against random attacks while being vulnerable to targeted attacks. In this aspect, metabolic networks, food webs, email networks and author-generated model networks have been looked at [2, 19, 20, 30, 14, 15, 12, 32, 46].

The ability for a degree preserving perturbation in the citation network to change the p-index of an author, which is directly related to the network resilience is considered.

2.3.5 Mixing Patterns - Assortativity

Mixing patterns or assortativity considers the question of which nodes are more likely to pair up with which other nodes. For this to be defined, the network should have two or more distinguishing criterion. Almost all networks will have such distinctions. The simplest group can be high-degree nodes (hubs) and low-degree nodes (leaf nodes) and the probability of which node connects with which often depends on these groups. For instance, there are three classes of nodes for the Internet: high level connectivity providers, consumers and Internet Service Provides (ISPs). This is elaborated on the work carried out by Maslov et al [41]. Similarly, in a food web that represents the relationship of which species consume which, node groups can be plants, herbivores and carnivores. Many edges are likely to link plants and herbivores, and herbivores to carnivores but only a few if any links would exist between herbivores to herbivores or carnivores to plants. A well-known example of assortative mixing in social networks is the mixing by ethnic group [43]. It has been shown numerous times that we tend to associate ourselves with people who are in some way similar to us.

Assortative mixing is quantified by the assortativity coefficient that is defined in different ways. If E_{ij} is the number of edges in a network that connects vertices of type i and j, with $i, j = 1...N$ and E is the matrix with elements E_{ij}, a normalized mixing matrix can be defined as shown in Equation 2.6, where $\|x\|$ means the sum of all the elements of the matrix x [45].

$$e = \frac{E}{\|E\|} \tag{2.6}$$

An alternate version of the assortativity coefficient is suggested by Gupta et al. [27] that has an interesting property where perfectly assortative networks have the value of 1 and randomly mixed networks have a value of 0 is shown in the Equation 2.7. However, it also suffers from drawbacks like having two values for asymmetric matrices. Another assortativity coefficient that surpasses this issue is defined in Equation 2.8.

$$Q = \frac{\sum_i P(i|i) - 1}{N - 1} \tag{2.7}$$

$$r = \frac{Tr(e) - \|e^2\|}{1 - \|e^2\|} \tag{2.8}$$

Another definition put forth by Newman is defined in the Equation 2.9. Newman defines the assortativity coefficient as the Pearson correlation coefficient of degree between pairs of connected nodes [43]. The q_k term in Equation 2.9 refers to the remaining degree distribution which can be derived from the degree distribution p_k as $q_k = \frac{p_{k+1}}{\sum_{j \geq 1} p_j}$. The term e_{jk} refers to the joint probability distribution of the remaining degrees of the two vertices which is symmetric on an undirected graphs as it follows the sum rules of $\sum_{jk} e_j k = 1$ and $\sum_j e_j k = q_k$. According to this definition, r=+1 indicates perfectly assortative mixing in the network while r=-1 indicates disassortative mixing in the network. The network is considered neutral or non-assortative when r=0.

$$r = \frac{\sum_{jk} jk(e_{jk} - q_j q_k)}{\sigma_q^2} \tag{2.9}$$

The definition depicted in Equation 2.9 will be referred to in the next chapters when the assortativity coefficient is discussed. Degree preserving perturbations are inflicted to the citation network and measure assortativity at each instance because when a degree preserving perturbation is carried out, it essentially means the assortativity is changed. This allows for the comprehensive analysis of the behavior of the citation network and the ranking stability of the p-index to the degree preserving perturbations.

2.3.6 Centrality Measures

Centrality measures are used to identify the important vertices in a network and a range of measures were developed in social network analysis for the same purpose. Centrality is defined as a property of individual nodes rather than as a property of the network. Three key centrality measures are discussed and their respective uses in a network parlance are identified.

Degree centrality is the simplest centrality measure and is one of the oldest measures known [8]. It is defined as the number of links incident upon a node. Directed graphs can have an in-degree centrality and an out-degree centrality. Degree centrality can be used to infer the number of people one is closely associated with.

Closeness centrality is a measure that is defined as the inverse of farness which is the sum of its distances to all other nodes. By definition, the closeness centrality of an unconnected graph will be 0 which makes it difficult to calculate when a network has disjoint communities. Hence Dangalchev [17] has modified the classic closeness centrality to that depicted in Equation 2.10 where $G := (V, E)$ represents the network with $|V|$ vertices and $|E|$ edges while v, t are two nodes in the network.

$$C_C(v) = \sum_{t \in V \setminus v} 2^{-d_G(v,t)} \tag{2.10}$$

The betweenness centrality which is more complex is then examined. It quantifies the frequency that a node acts as a bridge along the shortest path between two other nodes. This measure was introduced by Linton Freeman [24] and is defined as in Equation 2.11 where σ_{st} is total number of shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of those paths that pass through v. A high betweenness centrality of a node would indicate that the node has a high number of shortest paths relaying through it, essentially implying that it is an important node in the network. If a node with a high betweenness centrality is taken out, the average shortest path may be extended and the network may divide in to two communities. In this research, the betweenness centrality measure is used in order to find out the importance of authors in collaboration networks.

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{2.11}$$

2.4 PageRank Algorithm

With the fast growth of many information networks complimented by their large size, it is largely impossible to extract coherent information from these networks without a ranking scheme. A good example for this would be the content available on the WWW. If it was not ranked depending on the search terms, it would be almost impossible for anyone to find what they are looking for. One of the established solutions for the ranking problem is Google pagerank algorithm which was introduced in 1998-1999 with the Google search engine [48]. It is a link analysis algorithm that can assign weights to each link in the worldwide web without external intervention. This ensures that the ranking system is objective and does not have a human bias. The purpose of the pagerank algorithm is to measure the relative importance of a node within a set of nodes. There are other lesser known link analysis algorithms like HITS algorithm [35] and TrustRank [28] algorithm which wouldn't be discussed in this thesis. Interested readers are referred to the original publications that are referenced here.

It is known that WWW is a scale-free network and intuitively in the case of the pagerank algorithm, each directed link from node A to node B (web page 2 to web page 1) can be considered as an endorsement coming from node A to node B. Hence the more prestige the recommending node has, the more impact the recommendation will create. However if the recommending node has given out a large number of recommendations, a single recommendation would carry less weight. Because the citations networks are similar in terms of network and behavioral properties to the WWW, the Google pagerank algorithm can be applied to citation networks in order to rank the nodes of the network (ie: research publications). The importance of using pagerank algorithm is that it can take the citing papers into consideration implicitly and self-consistently rather than pure citation counting. This enables weighted citations to be used without explicitly attaching a weight to each of the citations and letting the underlying network structure to evolve the weight dynamically. Therefore, a brief discussion about the pagerank algorithm will follow before moving forward. There have been a few successful applications of pagerank algorithm to come up with impact factors and identify impactful documents in a citation network in the past [40, 13], which will also be discussed.

$$P_t(i) = \frac{1-\alpha}{N} + \alpha \sum_j \frac{A_{ij} P_{t-1}(j)}{k_{out}(j)} \tag{2.12}$$

The Equation 2.12 represents the version of pagerank algorithm used in this research, where A_{ij} is the adjacency matrix of the network, $k_{out}(j)$ is the number of outgoing links from node j and α is a reset parameter. N is the number of nodes in the network. This equation essentially characterizes a diffusion process where $P_t(i)$ can be considered as the frequency of visitation of a node i by an agent at time step t who will move along the edges of the network with a probability of α and can also visit a randomly chosen node with a probability of $1 - \alpha$. Pagerank $P_t(i)$ of node i can be regarded as the stationary state of this diffusion process that resolves its ranking with respect to other nodes in the network.

In pagerank algorithm, an edge from a node i to node j is recognized as an endorsement and the status of the recommending node is also important. For an example, a letter of recommendation from a Nobel Laureate (essentially having a higher $P_{t-1}(i)$) carries far more weight than a letter from an amateur scientist. However it stands to reason that if the same Nobel Laureate has given another 1000 recommendations (essentially having a high $k_{out}(j)$) for different candidates, then the status of the recommendation drops significantly. Because of these reasons, the pagerank of a node (research publication in our case) can be viewed as a measure of its influence and this is then distributed to its references. Utilizing this characteristic, the intention is to introduce an index which has ‘infinite levels of feedback and the impact of citing documents is factored in as much as possible. This is portrayed in Figure 2.1. From the figure, it can be seen that the number of citations a document receives is not always proportional to its impact factor, and actual numbers of citations must be rewarded at all levels, and not the impact factors. This could only be done by fully utilizing the overall dynamic citation network, and by using an elegant algorithm such as pagerank, which has had proven success in measuring the impact of such feedback loops in the Google search engine.

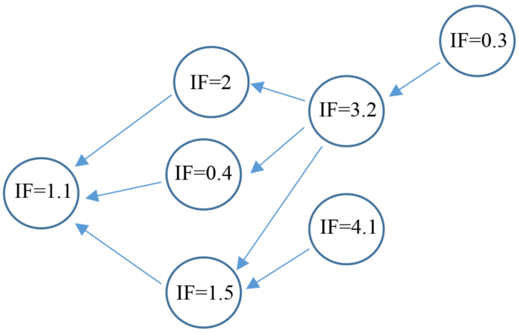


Figure 2.1: Multiple levels of citations: IF stands for the impact factor of each node (paper) and an arrow represents a citation from the node at the tail end to the node at the head of the arrow.

2.5 Ranking Scientists

Various measures are used to quantify the scientific output of a single researcher [5, 16, 31, 33, 34, 39, 49, 50, 55, 59]. Since the p-index introduced in this research also tries to do this, the existing options to quantify the scientific output of researchers and arguments against those measures will briefly be discussed.

2.5.1 Citation Count

This is the total number of citations an author has received over their entire portfolio of publications. It is a straightforward measure and is perceived as the impact of the author.

2.5.2 Number of Papers

This is the total number of papers authored/coauthored by a specific scientist. It is also perceived as the productivity of the author.

2.5.3 Citations per Paper

This is the average number of citations an author received per paper he or she has contributed to. This is an augmented measure derived from citation count and number of papers.

2.5.4 h-index

So far, the measures discussed have been straightforward without any derivation. All of these measures can be manipulated by various means simply because they are straightforward. However, when Hirsch [29] proposed the h-index, he showed that by integrating both citation count and number of papers in to one measure, he could derive a new index that would be more objective and can be used as a single measure. Formally, the h-index is defined as the h number of papers that have at least h number of citations each. By extension, if a scientist has N_T number of papers, $N_T - h$ number of papers will have less than h citations each. Hence, by previous definitions, h-index measures both the productivity and the impact of a scientist. This measure was widely accepted because the idea of being able to rank scientists by a single number was attractive to many in the academic field and Hirsch's argument about the advantages of h-index was instrumental in the adaptation. Kelly et al. [34] mention that h-index favors those authors who produce a series of influential papers rather than those authors who either produce many papers that are soon forgotten or produce a few that are uncharacteristically influential. In addition, as Hirsch pointed out himself, the h-index can be used to determine the past productivity of a scientist and it is also better equipped to predict the productivity of the scientist compared to other bibliometric indicators, which enables it to be a useful indicator of scientific quality that can be used (together with other criteria) to assist in academic appointment processes and to allocate research resources.

2.5.5 g-index

Egghe [21] argued that, according to the definition of h-index, once a paper belongs to the top h class (among the papers ranked h or higher), it becomes unimportant whether these papers are further cited or not. Scientists who have high-impact papers are penalized, as is evidenced by the difference between the lower bound of citations, h^2, proposed by the h-index, and the actual number of citations. This had been observed by Hirsch himself. Egghe therefore suggested the g-index, and defined it as the highest number g of papers that together received g or more citations. This essentially gives more weight to highly cited articles of a scientist.

2.5.6 AR-index

The AR index [22, 33] takes the age of the publications into account along with the citation intensity of the Hirsch core (the h number of papers that are included in h-index calculation). As such, this index can increase as well as decrease over time, which is a characteristic of a good research indicator sensitive to performance changes. This index is defined as the square root of the sum of the average number of citations per year of articles included in the h-core.

These are the measures that have been used as successful ranking techniques for scientists. However, the h-index prevails as the most commonly used measure to date and is adopted by major services like Google Scholar and Web of Science. Although the h-index is widely used, many critics have noted its disadvantages. One of the earliest criticisms pointed out by Hirsch himself was the time sensitivity in the measure; however this can be avoided by using a time restricted h-index as proposed by Hirsch [29]. Egghe [21] pointed out that the h-index is only weakly sensitive to the number of citations received by single publications, whereas it should be sensitive to the level of the highly cited papers. Similarly, the R-index and AR index introduced by Jin et al [33] measure the citation intensity of the h-core, also considering the age of the papers. They suggested that these measures could be used together with the h-index to complement each other. A review of the existing variants of h-index by Bornmann et al [10] suggested that all these variants fall into two types: those that describe the size of the h-core, and those that describe the impact of the h-core. The first type redefines the size of the h-core and typically makes it larger (such as the g-index), or adds additional information about it (such as the r-index) while the second type analyses the citation count of the h-core, negating the disadvantage attributed to scientists who have a few high-impact papers. The fundamental issue in these kinds of analysis is that they treat all citations equally, whereas citations should be treated differently depending on the perceived importance of the citing node. When this does not happen, one is able to massage their own h-index by various means and this creates the risk of of creating a huge body of academic publications that are hardly referred to, let alone made use of [9]. Hence, the research question is based on this premise of deriving a fairer, more robust and objective ranking system that is capable of weighting the citations coming from different platforms without our intrusion by using the underlying network dynamics.

2.6 Applications of PageRank in Academic Networks

After the pagerank algorithm was introduced to bring order to the Internet [48], it has been applied to many other ranking tasks from lung cancer research [47] to recommendation systems [3]. This review is interested in the application of the pagerank algorithm in academic networks (ie: citation networks or collaboration networks or both). The initial application of pagerank in an academic perspective can be traced back to Bollen et al. [7], who demonstrated that a weighted pagerank algorithm can be used to measure the impact of journals instead of the impact factor. They compared the differences between the impact factors and weighted pagerank impact and discovered that both measures can have overlaps and differences. Usage of the same weighted pagerank impact by Dellavalle et al. [18] to the dermatology journals proved that the weighted pagerank provided a more refined measure of the status of the respective journals.

However, the most relevant application with respect to this study was conducted by Chen et al. [13] in 2007, where they applied the pagerank algorithm to assess the relative importance of the publications in the Physical Review family of journals. They proved that pagerank algorithm can be directly applied to citation networks because the initial application base of pagerank algorithm (ie: WWW) is not very different from the citation networks. A hyperlink in a popular website would drive more traffic to one's website in the same way that a reference from a prestigious publication would refer a scientist to one's paper. This makes it possible for us to ascertain that the random surfer model used in pagerank algorithm is also relevant to citation networks. They also proved that pagerank was able to identify 'scientific gems' that were otherwise neglected by common ranking methods such as the number of citations. This in fact strengthens the initial argument of this thesis that pagerank can indeed rank the publications more fairly than the commonly used metrics, such as the number of citations for each paper. This coupled with the infinite levels of feedback is why the pagerank algorithm was chosen for ranking purposes.

Walker et al. [57] have also come up with a modified version of pagerank algorithm called 'site-rank' to rank scientific publications, which essentially works in the same way as pagerank algorithm. The difference with 'site-rank' algorithm is that it initially distributes the random surfers exponentially with age, favoring more recent publications. This is justified by noting that researchers are more likely to refer to recent publications than aged publications. They also proved that their version of pagerank algorithm outperforms classical metrics like the number of citations.

In conclusion, all of these experiments show that pagerank algorithm can be used to effectively rank scientific publications in a more fair way than conventional ranking metrics. Maslov and Redner [40] have reviewed the above literature and summarized the promises and pitfalls of using pagerank algorithms in citation networks. As it has been already iterated, they have pointed out that pagerank holds great promise for quantifying the impact of scientific publications while providing a meaningful extension to conventionally used measures to gauge impact. However, they note the potential for wrongful usage of a metric derived from pagerank algorithm exists just as with the classical metrics. While the pagerank algorithm can indeed be misused and manipulated at the same time, in the context of academic networks, the results indicate that this is increasingly difficult with the design precautions taken. In addition, this research shows by taking various pitfalls of classical measures in to consideration and presenting evidence, this design would be robust against such manipulations.

Chapter 3

Methodology

3.1 Introduction and Definition of the p-index

In this chapter, a new index, the p-index (pagerank index), is presented. This index can quantify the research output of a scientist in an objective manner. The methodology involved in deriving p-index is comprehensively described and arguments as to why it is fairer and more robust than existing ranking systems are presented. The content of this chapter builds on the theoretical foundations discussed in the second chapter.

The p-index can be defined as the individual ranking of a scientist based on their cumulative contribution to each publication they have co-authored when publications are ranked using the pagerank algorithm and the individual contribution from a publication is measured as the proportion from the pagerank value of that publication. This can be further illustrated as a process in Figure 3.1.

In this chapter, each sub-process portrayed in Figure 3.1 is discussed. The next section describes the ranking mechanism applied to the citation network while the third section details the techniques used to distribute the pagerank value of each publication to its authors. The fourth section demonstrates the p-index calculation while the dynamicity of p-index is discussed in the fifth section. The final section concludes the chapter.

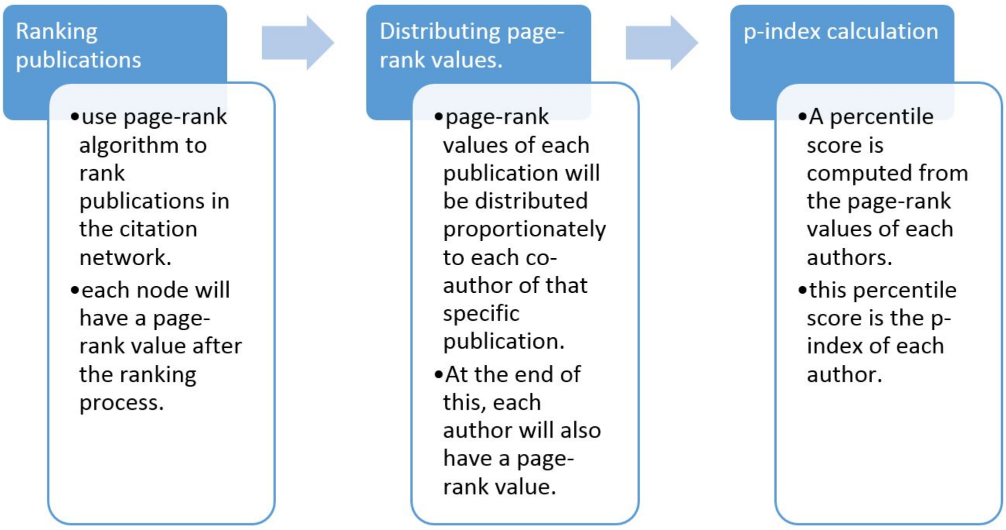

Ranking publications
•use page-rank algorithm to rank publications in the citation network.
•each node will have a page-rank value after the ranking process.
Distributing page-rank values.
•page-rank values of each publication will be distributed proportionately to each co-author of that specific publication.
•At the end of this, each author will also have a page-rank value.
p-index calculation
•A percentile score is computed from the page-rank values of each authors.
•this percentile score is the p-index of each author.

Figure 3.1: The process of computing the p-index.

3.2 Ranking Publications

The referencer-referencee relationship between publications is accurately portrayed by a citation network whose nodes are publications and edges represent referencer-referencee relationship. Each node in this network will have meta attributes such as the author or co-authors. These meta attributes can be used to build a collaboration network on top of the citation network, which represents co-author relationship. This is demonstrated in Figure 3.2. The interplay between these two networks is used to derive the p-index.

Publications are ranked using the original pagerank algorithm proposed by Larry Page and Sergei Brin [48], which can be interpreted to mimic the behavior of a random surfer in the WWW. The same interpretation holds for the referencing behavior of scientists as well, which makes pagerank algorithm the ideal link analysis platform for this research. As has been demonstrated by the works of numerous authors referenced in the second chapter [13, 40, 49], pagerank is considered better for the purpose of ranking publications than the conventional methods such as the number of citations received. Hence the pagerank algorithm as described by the Equation 3.1 is used, where A_{ij} is the adjacency matrix of the network, $k_{out}(j)$ is the number of outgoing links from node j and α is a reset parameter. N is the number of nodes in the network.

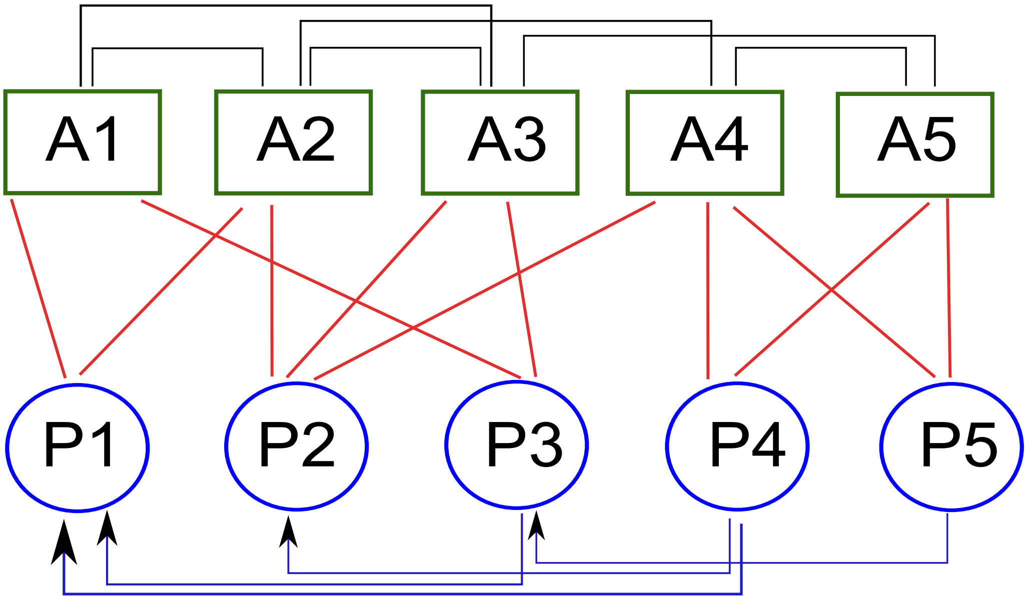


Figure 3.2: The interplay between citation network and collaboration network. Blue circles denote papers and the blue arrows denote the referencer-referencee relationship together, which creates the citation network. Each red line indicates the co-authors of each node in the citation network. The green squares represent nodes in the collaboration network while the black line represents the co-author relationship between authors.

$$P_t(i) = \frac{1-\alpha}{N} + \alpha \sum_j \frac{A_{ij} P_{t-1}(j)}{k_{out}(j)} \tag{3.1}$$

As has been discussed, the pagerank algorithm essentially measures the weight carried by each reference in the citation network discounted by the number of references given out by the referencer node and gives a decimal number to each node in the citation network, which is identified as the pagerank value of each node. The Ranking of nodes (publications) are based on the pagerank value.

Implementing this part of the process was straightforward but the pagerank calculation was decoupled from the citation network so that it can be generalized and applied to any citation network. Hence two components were implemented to execute this sub-process: the generalized pagerank algorithm and the feeding component to feed any citation network to the pagerank algorithm.

Pagerank algorithm is executed multiple times subject to the condition described in Equation 3.2. The margin-of-error (ϵ) was decided to be 0.0001, which gives us the required consistency over the multiple executions of the pagerank algorithm. After this condition is met, the pagerank value stabilizes and it is this stabilized value that is used to divide between the corresponding authors.

$$P_t(i+1) - P_t(i) \leq \epsilon \tag{3.2}$$

3.3 Distributing Pagerank Values

After a stable pagerank value has been dereived for each node in the citation network, this value was distributed among the respective authors of each publication. The initial distribution method was to distribute the pagerank value of the publication equally between the authors. Hence the pagerank value received by an author can be calculated as shown in Equation 3.3.

$$\begin{split}\text{pagerank value received by an author from a publication}\\ = \frac{\text{stable pagerank value of the publication}}{\text{the number of authors of the publication}}\end{split} \tag{3.3}$$

However this distribution was found to be flawed in that equal consideration is given to the first author and the last author whereas, the usual practice is to list the authors in the order of the contributions they made to a certain publication. Hence in order to maintain fairness and objectivity, the pagerank values were distributed proportionally as shown in Equations 3.4 and 3.5. It should be noted that $Position_{author}$ starts from 0, which means the first author's position would be 0.

$$\text{proportion} = \frac{N_{authors} - Position_{author}}{\frac{N_{authors} * (N_{authors}+1)}{2}} \tag{3.4}$$

$$\begin{split}\text{pagerank value received by an author from a publication}\\ = proportion * \text{stable pagerank value of the publication}\end{split} \tag{3.5}$$

While, as suggested by [53] it is general practice to order authors by their level of contribution, Waltman [58] notes that some disciplines still list authors in alphabetical order after the first author. Since Waltman's empirical analysis suggests that more than 80% of papers published today confirms with this standard and academia is converging towards it, this research assumes that author positions are mentioned depending on the contribution they made towards the publication until it is converged to be the universal practice.

The final step in the distribution sub-process is to aggregate the pagerank values received by each author node from each of their publications to come up with a single pagerank value for each author node in the collaboration network. Like the pagerank value in the citation network, this is also a decimal value although it was not derived by executing pagerank algorithm on top of collaboration network.

3.4 p-index Calculation

After each author has been assigned a pagerank value, their percentile rank can be calculated using the pagerank value. The percentile rank rather than using the absolute pagerank value was chosen because the percentile rank is easily interpretable compared to a decimal number. Hence not only will the p-index be a measure of a researchers' scientific output but it also will tell the researcher where exactly he/she stands among his colleagues in the scientific field [52].

If the citation network used is for a specific field, then the derived p-index would be valid for that specific field. What this means is that a researcher can have multiple p-indices depending on the field. For instance, a researcher who works in machine learning can have a p-index for machine learning, a p-index for artificial intelligence (which is the parent field of machine learning), and a p-index for physical sciences and, by extension, among the whole scientific community. This makes the interpretation easier and it would also make the index fairer. For example, if a researcher is in a field where they cannot have a high throughput, they might have a low h-index and even a low p-index as well if the whole scientific community is considered. However when an evaluation takes place, one can examine his p-index inside his scientific field, which reveals more contextual information rather than providing a universal number. This context awareness is an important characteristic of the p-index.

3.5 Dynamicity of the p-index

So far the sub-processes involved in deriving the p-index of scientists when there is a citation network have been discussed. However, the robustness of the p-index against various manipulations comes from its dynamicity. Each time a new publication is added to the citation network, the pagerank algorithm will be re-executed, the pagerank values will be re-distributed and the p-index of each scientist will be recalculated. The system was implemented so that the whole process would iterate each time a new publication enters; however this is not possible in real citation networks. Hence it is proposed that the process would be iterated each day or each week or even each month as preferred by the assessing body.

The dynamic nature of the p-index has made it possible for it to resist the manipulations that are trivially doable in h-index, which are demonstrated in the next chapter. The same dynamicity also enables the p-index to be applied to any existing citation network. However individual p-index scores cannot be calculated by the respective individuals since it needs the underlying citation network to be available, which is not the case for individuals. The p-index can be calculated by indexing services like Google Scholar, Elsevier Scopus and Thomson ISI Web of Science who have access to the underlying citation networks.

3.6 Conclusion

In this chapter, the process taken in order to calculate the p-index was described. The main process was divided into three sub-processes, which were discussed in detail. The dynamic nature of the p-index and its implications was also examined. Finally, the services that are capable of implementing p-index as a ranking metric in their domain are identified.

Chapter 4

Modeling and Analysis of the Simulated System

4.1 Introduction

The indices currently used by scholarly databases (such as Google scholar) to rank scientists do not attach weights to the citations and the underlying network structure of citations is not considered in computing these metrics. This results in scientists cited by well-recognized journals not being rewarded and may lead to potential misuse if documents are created purely to cite others. In this thesis, a new ranking metric, the p-index (pagerank-index) was introduced, which is computed from the underlying citation network of papers, and uses the pagerank algorithm in its computation. The index is a percentile score and can potentially be implemented in public databases such as Google scholar, and can be applied at many levels of abstraction. It was demonstrated that the metric aids in fairer ranking of scientists in comparison to h-index and its variants using a realistic simulation system that will be described in this chapter. It can realistically imitate the behavior of an evolving citation network of a specific field. This system was designed to simulate the exploits found against the h-index and demonstrate the drawbacks in the h-index. By employing the same realistic simulation system which synthesizes the evolution process of citation and collaboration networks in an emerging field, it is shown that the p-index is fairer than the h-index in many scenarios in which the authors may be able to massage or manipulate their h-index. In particular, while it is possible to massage the h-index by publishing papers in low-impact forums where the same group of authors continually cite each other, this is largely impossible with p-index. It is also shown that while the h-index rewards collaboration and authors who focus on quantity, the p-index is much more balanced and equally rewards individual brilliance and quality of papers. These results show that the p-index is immune to author behaviors that can result in artificially bloated h-index values.

The rest of this chapter is organized as follows. The next section revisits the methodology of deriving the p-index and the third section describes the design and implementation of the simulated academic network. The fourth section presents the results while the final section concludes the chapter.

4.2 Methodology

A version of the pagerank algorithm [48] was implemented on citation networks to compute the new p-index. The premise behind pagerank is that it uses the number of links pointing to a web page, as well as the relative standing of pages from where the links originate, to determine the rank of a web page to be displayed in a Google search. Therefore, in citation network parlance, the ‘citations are ‘weighted. The pagerank of a ‘node in a network of N nodes can be calculated from the Equation 4.1:

$$P_t(i) = \frac{1-\alpha}{N} + \alpha \sum_j \frac{A_{ij} P_{t-1}(j)}{k_{out}(j)} \tag{4.1}$$

where $A_{i,j}$ is the adjacency matrix of the network, $k_{out(j)}$ is the number of outgoing links from node j and α is a reset parameter. A version of the pagerank algorithm is used to dynamically update the ‘pagerank values of each paper. Asynchronous updating is implemented such that each time a paper is uploaded into a citation database, the pagerank values of all papers are updated. The ‘p-index (pagerank index) of the author is calculated by aggregating all of the pagerank values of papers to which the author have contributed. Where there is more than one author per paper, the pagerank value of the paper is shared proportionally. The pagerank values of new papers are governed by the reset parameter α. Therefore, when an author gets cited by a paper which is just published, it will only change the author’s p-index marginally. It is only when that paper gets itself cited, that the author's p-index will begin to increase. If that paper is cited by papers which are themselves becoming well-cited, the p-index will increase considerably. Therefore, the overall dynamics of the research community determines a particular author's p-index in a non-trivial way.

Since the values generated by the pagerank algorithm are decimal numbers which does not make intuitive sense, the p-index is defined as a percentile. Therefore, for example a very famous scientist may have a p-index of 99.99%, whereas a new research student will have a value close to zero noting that since it is a percentile, the p-index can be applied to a particular field of research or the entire scientific community. As such, a computer scientist, for example, may have a p-index of 99.9887% in computer science, and 97.5675% overall. By definition, the overall score will be lower than the field specific score. It must be noted that since the p-index utilizes a citation network, it can only be implemented in a scholarly database. Individual scientists cannot calculate their own p-index based on their own publication record. Therefore, the proposed index needs to be implemented in a scholarly database to be fully tested. For this reason, a simulated system was utilized to verify its utility, as described in the next section.

4.3 Simulated System

In this section, the simulation system which was used to test the proposed index is described. The system simulates the process whereby a group of scientists produce papers in collaboration and this spawns a new field of research. Eventually, this gives rise to a citation network, where papers are nodes and citations are links, as well as a collaboration network where the scientists are the nodes and collaborations form the links. The former are directed networks and the latter are undirected. This approach was selected to demonstrate the utility of the p-index, since the researcher does not have access to a real world citation database to implement and test this index. As described below, the evolution process realistically imitates the growth of a research field and corresponding citation and collaboration networks. Three independent 'case studies' using this simulation system are considered. Each case is meant to illustrate a perceived weakness of the h-index, and how the p-index is immune to it.

In the simulation system, there are paper objects and author objects. Each of these objects have a number of variables (attributes). Table 4.1 and Table 4.2 list the attributes of these objects respectively where dynamic attributes (attributes that are changed in value in every iteration) are shown in black and static attributes are shown in blue. A short description of each attribute is provided in Tables 4.1 and 4.2 as well. The variable $is\lambda$ takes on different names for each simulation scenario, as described below.

At the beginning of the simulation, a paper is 'spawned', and some authors are spawned and assigned to it. The number of authors is randomly assigned between 2 and 5. Each spawned paper is also randomly assigned an impact factor between 0 and 20. In this initial study, the distribution of impact factors was kept linear, though clearly it is biased towards lower values. It was also assumed that conference papers and similar documents have a low impact factor. When the next papers are spawned, authors are stochastically assigned from the existing author pool or new authors are spawned. The probability of a new author being spawned decreases as the simulation progresses, so that authors are increasingly assigned from existing author pool. At steady state, the probability of an author being a new author rather than somebody from the existing author pool was fixed at 4%. Each spawned paper is randomly assigned a number between 10 and 50 for the number of references it has.

However, not all of these references necessarily apply to existing papers in the simulated system. In the real world, a paper which is a pioneer in a new field will by necessity refer to papers 'outside' its field, since there are no papers already in the emerging field to cite. Even in a saturated field, a certain proportion of the references will be outside the field anyway. To reflect this, a varying parameter ρ was set as the proportion of references which are within the simulated system. This proportion begins at zero and increases linearly with time elapsed until it settles at 70%. Once the number of internal references is determined, existing paper objects are chosen to be the references in the newly spawned paper. These are chosen by weighted preferential attachment, as described below. Thus, the more a paper is already cited, the higher its chances are to be cited by a newly spawned paper, and the impact factor of the paper also plays a part.

The references for newly spawned papers from existing papers were chosen by the well-known 'preferential attachment' method, introduced by Barabasi et al [1]. In the context of the citation network, preferential attachment works so that a new paper has a higher probability of referencing an existing paper that is highly cited rather than referencing a

Paper Object	Description
Paper ID	This attribute is a unique identifier for each paper which would act as an index.
isλ	isλ parameter will take different forms for different simulations scenarios. It is essentially like a gene defining the type of the author and the type of the paper. For instance, in the manipulative authors vs. non-manipulative authors scenario, isλ would take the form of isManipulativePaper and isManipulativeAuthor.
number of authors	This attributes dictates how many authors are assigned to each paper object.
author list	Following the number of authors attribute, author list will contain the list of authors of a paper.
Impact Factor	This represents the impact factor of the paper. The impact factor was used in order to come up with a weighted preferential attachment system when a paper decides to cite another paper. This is in congruent with the realistic behavior because one is more likely to cite a paper published in a high impact forum rather than in a low impact forum. However, it is stochastically modeled so citing a low impact paper is also possible despite with a low probability.
number of citations towards the network	This attribute quantifies how many papers are referenced by this paper inside the same citation network. Since we are considering a specific scientific field, a paper may reference other papers outside that scientific field as well. In this case, the referenced paper outside this scientific field is neglected and only internal citations are considered.
number of references	This is a total number of references a paper will have. Number of citations towards the network will be a proportion of this value.
page rank value	This is the page rank value of a paper node after the page rank algorithm has been executed and a stable value has been reached.

Table 4.1: Attributes of a paper node.

paper that is less cited or not cited at all. As such, a directed citation network and an undirected collaboration network begin to evolve. This process is continued for a fixed number of timesteps T. In this research, typically T=15000 was used. In the experiments, the number of papers: number of authors ratio was maintained at roughly 10:1. After each timestep, the pagerank algorithm was run on the citation network and the pagerank scores, and by extension the p-index, of authors were updated. The h-index of authors was also updated, using the available local information for each author (list of papers and citations for each paper). As such, it was possible to compare the evolutionary trends of h-index and p-index for each author. Below, the three simulation scenarios which highlight

Author Object	Description
Author ID	This attribute is a unique identifier for each author and would act as an index.
isλ	isλ parameter will take different forms for different simulations scenarios. It is essentially like a gene defining the type of the author and the type of the paper. For instance, in the manipulative authors vs. non-manipulative authors scenario, isλ would take the form of isManipulativePaper and isManipulativeAuthor.
paper list	This is a list of papers authored or co-authored by a specific author.
number of citations	This attribute represents the number of citations each author have for all of his papers.
h-index	This is the h-index of each author.
page rank value	This attribute is to store the page rank values of the authors.
p-index	This attribute stores the p-index of authors.

Table 4.2: Attributes of an author node.

perceived weaknesses of h-index and how the p-index overcomes them are described.

4.3.1 Manipulating the h-index using Low-impact Publications

The first simulation scenario demonstrates a known weakness in the h-index where certain authors can publish low-impact papers in order to cite their previous work and thus massage their h-index. Henceforth the authors of these low impact papers are referred to as manipulative authors because their intention is to manipulate their h-index, and the low impact publications will be known as manipulative papers, written and published with the sole intention of manipulating and massaging the h-index. A paper is randomly determined to be a manipulative document with a probability of 0.1 at the time of creation. The $is\lambda$ parameter for paper node will be named as isManipulativePaper while the $is\lambda$ parameter for authors will be isManipulativeAuthor. The probability of being a manipulative author is set to 0.2 and each new author, when created by the simulation system, will be randomly assigned as either manipulative or not manipulative. Manipulative publications are only authored by manipulative authors while both non-manipulative and manipulative authors can contribute to non-manipulative papers. The sole purpose of a manipulative document is to reference the first author's previous work therein massaging their h-index.

Simulation experiments were run to mimic a citation network of 15,000 documents and

Property	Value
Number of nodes	15000
Number of edges	67035
Average degree	8.938
Characteristic path length	1.823
Clustering coefficient	0.09
Network diameter	7
Power law fit	$P(k) \sim k^{-1.923}$

Table 4.3: Characteristics of the citation network for manipulative authors against non-manipulative authors scenario.

Property	Value
Number of nodes	1314
Number of edges	50530
Network heterogeneity	0.773
Characteristic path length	2.192
Clustering coefficient	0.210
Network diameter	5
Network Density	0.059

Table 4.4: Characteristics of the collaboration network for manipulative authors against non-manipulative authors scenario.

these documents had 1317 authors altogether among which 152 authors were manipulative authors and 1500 documents were manipulative documents. The properties of the citation network and the collaboration network are analysed in Tables 4.3 and 4.4. The degree distributions for the respective networks are also presented in Figures 4.2 and 4.3. A snapshot of the collaboration is presented in order to get a visual understanding of the network in Figure 4.1. The Results of this simulation are presented in section 4.4.1.

4.3.2 Robustness of p-index towards Preferences of the Authors

In the second simulation scenario, the intention is to demonstrate that p-index is fairer to authors who are less inclined to collaborate. Here, one sector of authors is assumed to be interested in collaboration and typically publish documents with between 1 to 9 authors per paper. The second sector of authors is relatively uninterested in collaboration and only publishes papers that have three authors at most. The implication of this scenario is that the authors who collaborate will have a better chance of getting a higher h-index because their paper count is higher regardless of how much each author contributed. Thus

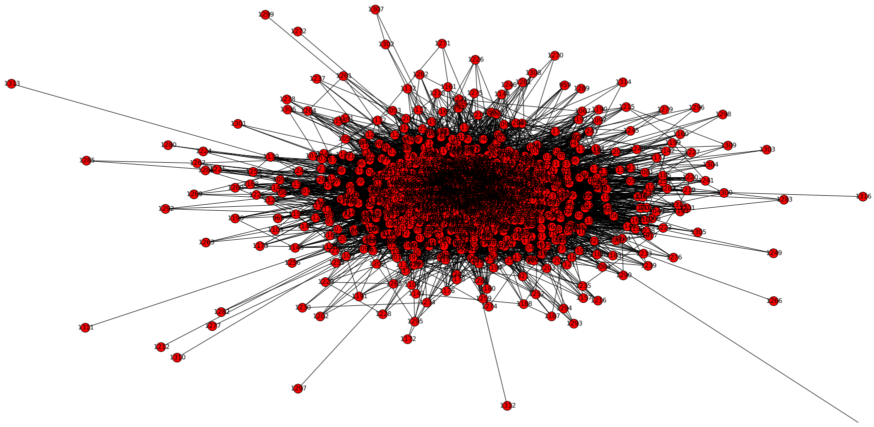


Figure 4.1: Collaboration network for manipulative and non-manipulative authors scenario.

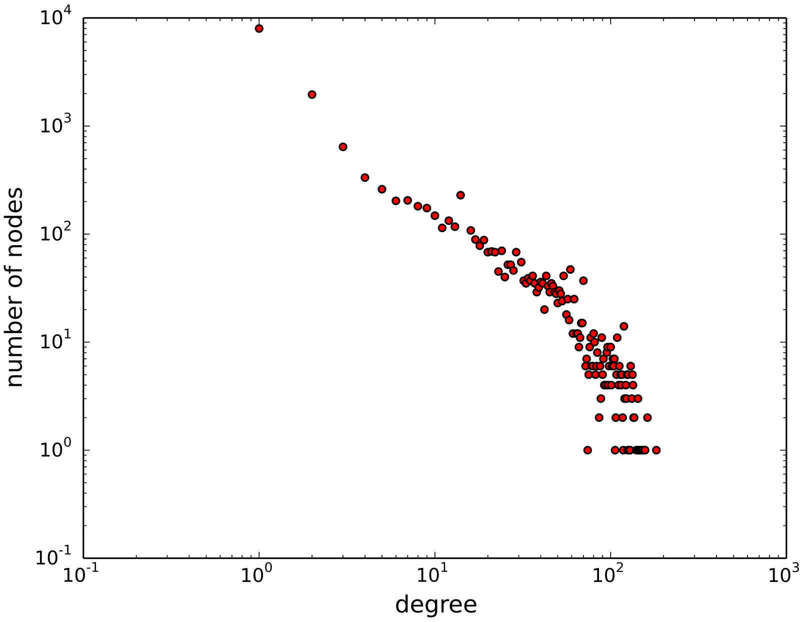


Figure 4.2: Degree distribution of the citation network for manipulative and non-manipulative authors scenario plotted in log-log scale.

it could be expected that the h-index is biased towards collaborators and this research aims to compare this feature with the p-index. As such, two types of papers were identified accordingly as collaborative papers and non-collaborative papers. The $is\lambda$ parameter here is named as isCollaborativeAuthor and isCollaborativePaper respectively for author and paper nodes. Collaborative authors in this scenario will consist of 90% of the author pool

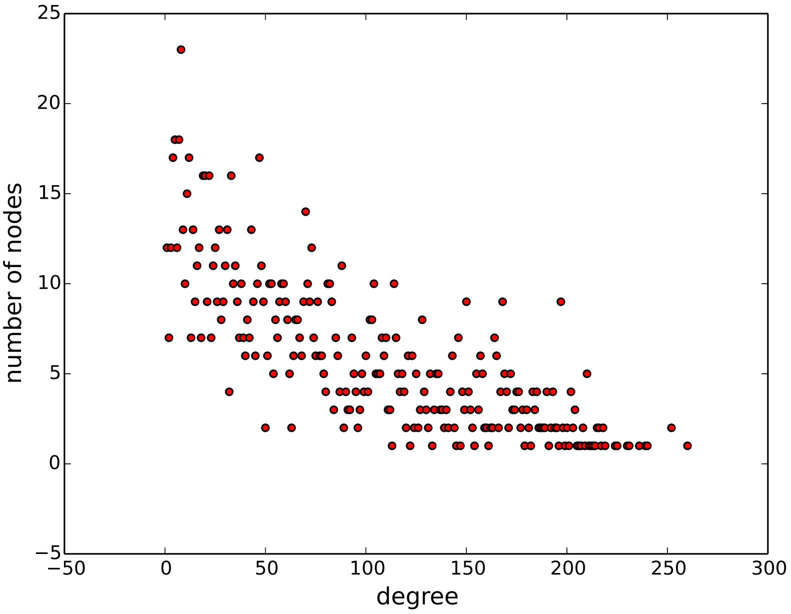


Figure 4.3: Degree distribution of the collaboration network for manipulative and non-manipulative authors scenario.

Property	Value
Number of nodes	15000
Number of edges	12105
Average degree	1.614
Characteristic path length	1.0
Clustering coefficient	0.0
Network diameter	1
Power law fit	$P(k) \sim k^{-1.444}$

Table 4.5: Characteristics of the citation network for collaborative and non-collaborative authors scenario.

Property	Value
Number of nodes	1314
Number of edges	113576
Network heterogeneity	0.723
Characteristic path length	2.114
Clustering coefficient	0.348
Network diameter	6
Network Density	0.129

Table 4.6: Characteristics of the collaboration network for collaborative and non-collaborative authors scenario.

while 10% of authors will be non-collaborative authors. The publications will be assigned as 80% collaborative publications and 20% non-collaborative publications.

The system was simulated for 15,000 papers consisting of 3006 non-collaborative papers. The author pool included 1328 authors with 126 non-collaborative authors in total. The characteristics of the citation network and the collaboration network are tabulated in Tables 4.5 and 4.6. The results of this simulation are presented in section 4.4.2.

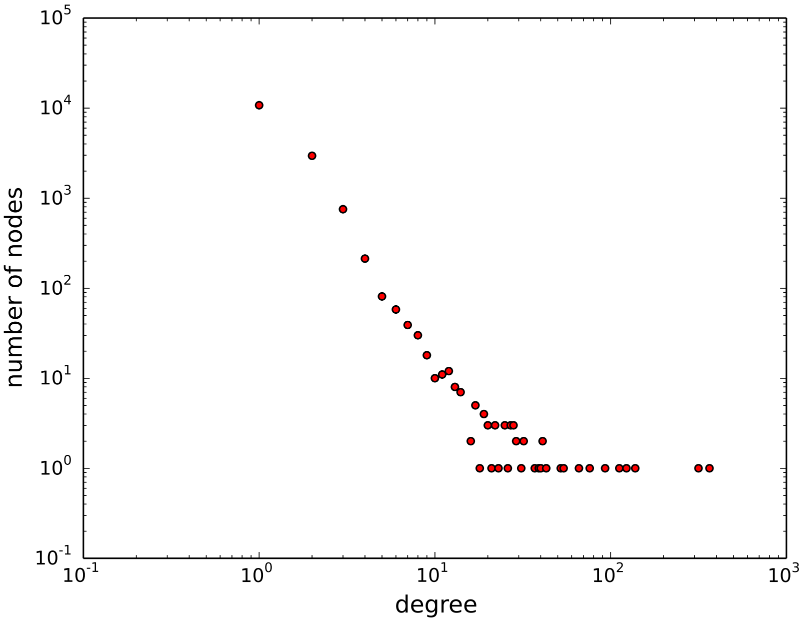


Figure 4.4: Degree distribution of the citation network for collaborative and non-collaborative authors scenario plotted in log-log scale.

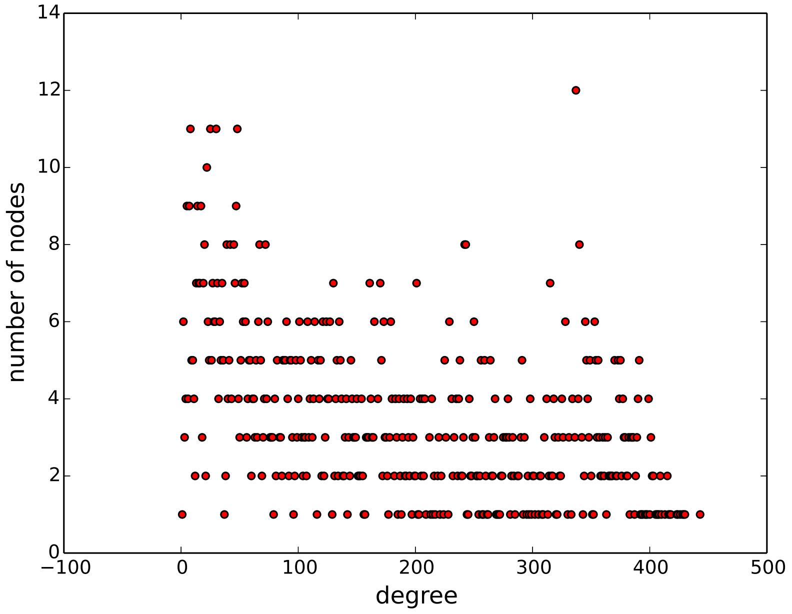


Figure 4.5: Degree distribution of the collaboration network for collaborative and non-collaborative authors scenario.

Property	Value
Number of nodes	15000
Number of edges	55878
Average degree	7.4504
Characteristic path length	1.606
Clustering coefficient	0.0
Network diameter	6
Power law fit	$P(k) \sim k^{-1.865}$

Table 4.7: Characteristics of the citation network for quantity and quality oriented scenario.

Property	Value
Number of nodes	1353
Number of edges	50747
Network heterogeneity	0.723
Characteristic path length	2.114
Clustering coefficient	0.348
Network diameter	6
Network Density	0.129

Table 4.8: Characteristics of the collaboration network for quantity and quality oriented scenario.

4.3.3 Robustness of p-index towards Scientists who are Concerned about Quality over Quantity

The third simulation scenario deals with a typical problem scientists face when being ranked by the h-index. If a scientist is only concerned about the quality of their work and does not mind how many publications they contribute to, their h-index may deteriorate due to the lack of number of papers published and cited. It has been shown that even though the papers published by these selective scientists have relatively high impact factors and a large number of citations per paper, their lack of publications challenges their standing in the scientific community [49]. As such, the following scenario was simulated in order to observe how p-index would perform comparatively.

The simulation scenario has two types of authors: those who are quality oriented and those who are quantity oriented. In order to quantitatively justify the impact, the simulations were set up such that the quantity oriented authors publish papers with an impact factor of 0-2 while quality oriented authors publish papers with an impact factor of 0-20. The system will generate roughly 10% of quantity oriented authors with 10% of quantity oriented papers. The simulation system had 15,000 papers in total of which 1511 were quantity oriented papers and had an author pool of 1357 authors with 138 quantity oriented authors. The degree distributions for the citation network and collaboration network as well as related properties respectively are presented in Figures 4.6 and 4.7 and Tables 4.7 and 4.8. The results of this simulation are presented in section 4.4.3.

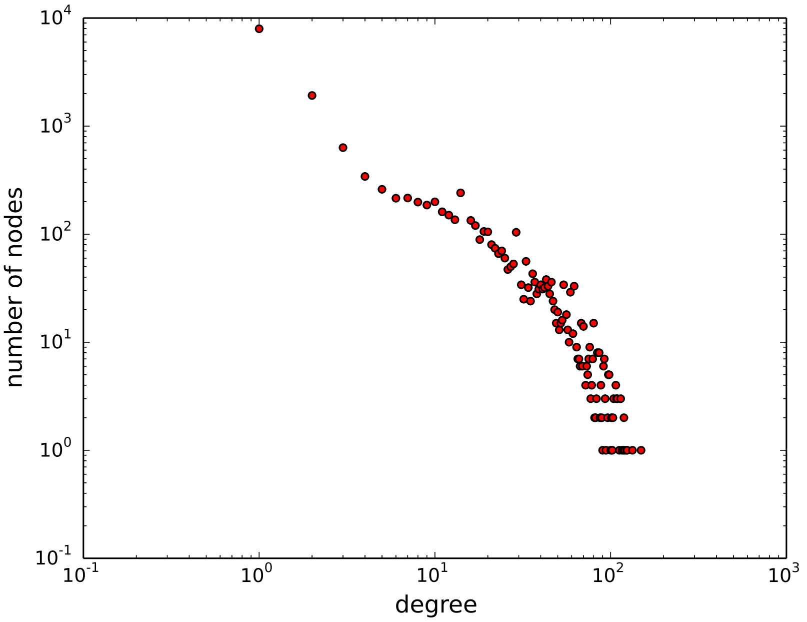


Figure 4.6: Degree distribution of the citation network for quantity and quality oriented scenario plotted in log-log scale.

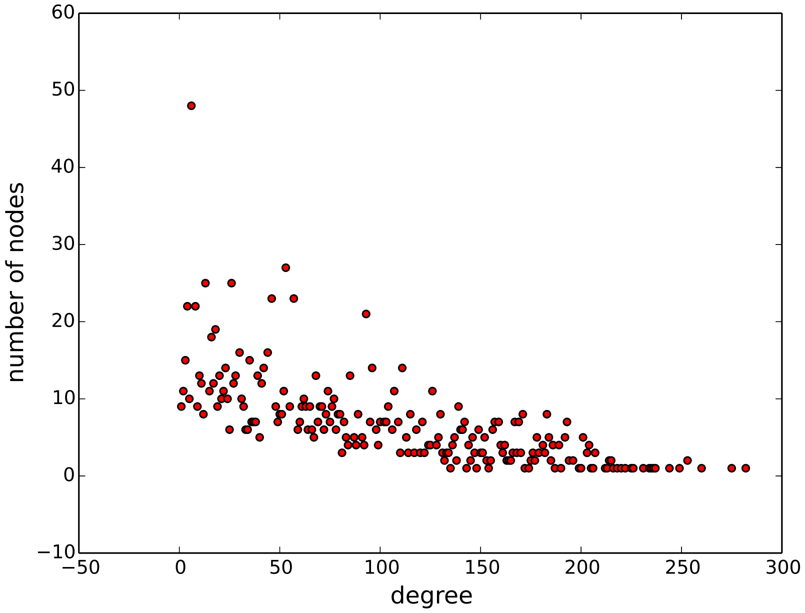


Figure 4.7: Degree distribution of the collaboration network for quantity and quality oriented scenario.

4.3.4 Measures and Visualizations

The results of the respective simulations described above and the suitability of the newly introduced p-index was evaluated using a set of heterogeneous measures which are elaborated below.

- Plots of the distribution of the h-index and p-index: These plots are used to visualize the disparity between the two classes of authors in respective scenarios. The h-index and p-index are plotted against author ID.

- Plots of the time variation of average h-index and p-index: The average values of the h-index and p-index for both classes of authors are plotted against time step.

- Plots of the time variation of Maximum h-index and p-index: The highest (best) values of the h-index and p-index for both classes of authors are plotted against time step.

- Ratio of differences: This measure is referred to as the multiplier within this paper and it signifies roughly how many times one measure is 'fairer than other. Equation 4.2 defines this measure mathematically.

$$multiplier = \Delta h - index/\Delta p - index \tag{4.2}$$

  where $\Delta h - index$ is the difference between average h-index of the two classes of authors while $\Delta p - index$ is the difference between average p-index of the two classes of authors at each timestep, for a given simulation scenario. Please note that the researcher is not claiming to introduce a measure of mathematical significance, but just using a name convention within this paper for ease of description.

4.4 Results

In this section the results for the above mentioned simulation scenarios are presented.

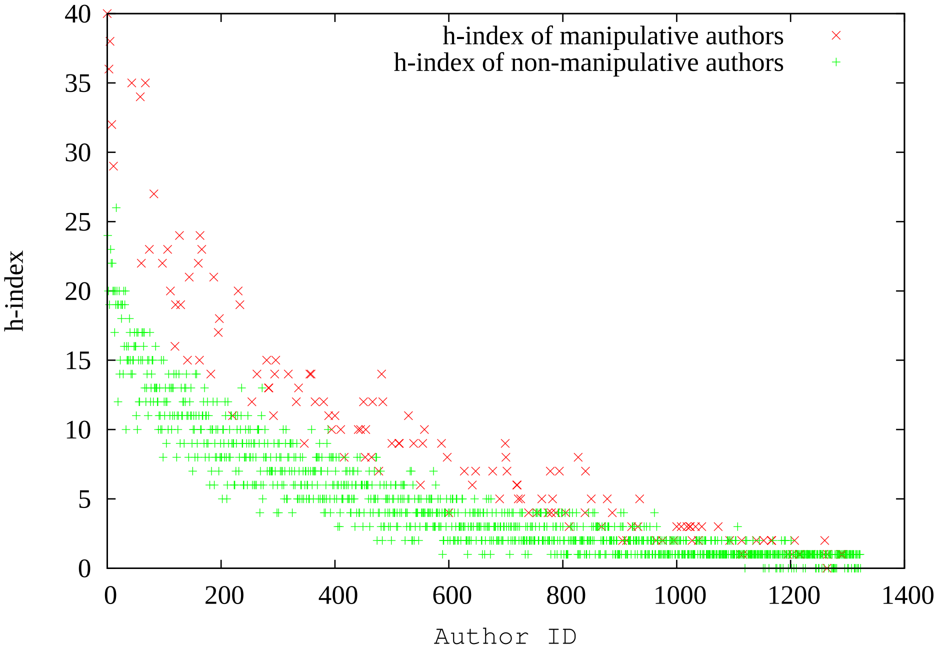


Figure 4.8: Spread of the the h-index for each manipulative and non-manipulative author (as absolute values).

4.4.1 Manipulative Authors vs Non-manipulative Authors

As portrayed in Figure 4.8, the spread of the h-index for non-manipulative authors and manipulative authors provide evidence for the fact that manipulative authors can indeed massage their h-index by publishing low impact papers with the sole purpose of referencing their previous work. On the contrary, a mixed distribution was detected for the p-index of manipulative and non-manipulative authors as seen in Figure 4.9. In Figure 4.8, there is clear separation between the two groups while in Figure 4.9 the two categories are overlapping. It should be noted that the IDs of the authors signify their duration as a researcher, thus author ID 0 is bound to get more citations than author ID 1300. This explains the lower values obtained by the authors at the right end of the Figure 4.8 and Figure 4.9. This effect has been discussed in the context of h-index and can be addressed by using a time dependent version of the same index.

Variation of the average h-index and p-index for non-manipulative and manipulative authors at each iteration is shown in Figure 4.10. It can be inferred from Figure 4.10 that p-index is a fairer metric to rank scientists when everyone is not manipulative. The difference between the average h-index of non-manipulative authors and manipulative authors

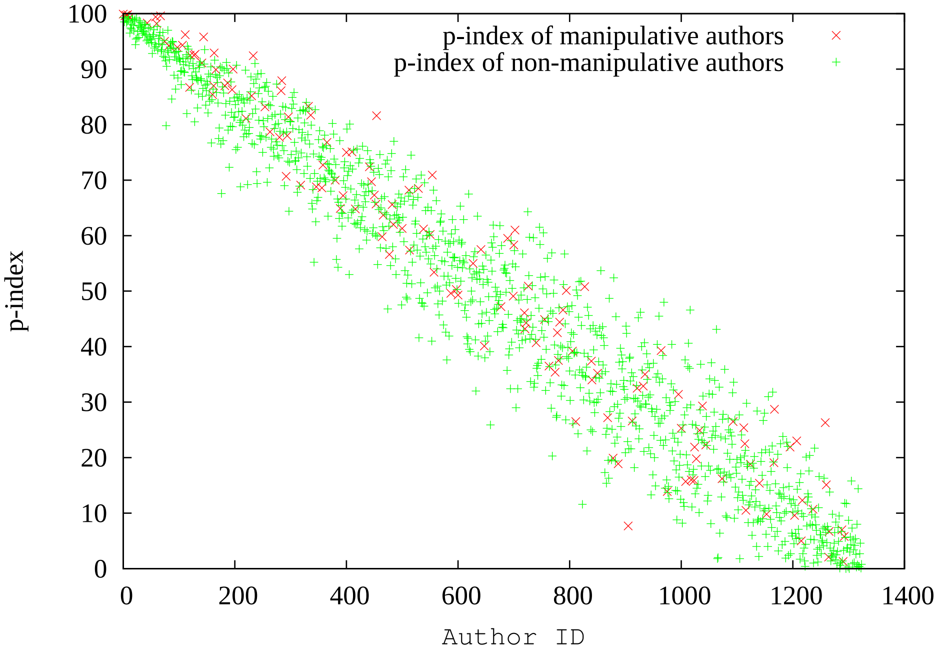


Figure 4.9: Spread of the the p-index for each manipulative and non-manipulative author (as absolute values).

is far greater than the difference between the average p-index of non-manipulative authors and manipulative authors. Quantitatively, the difference of h-index is 110.93% whereas the difference of p-index is only 10.58%, which is ten times smaller. The implication is that using h-index, a non-manipulative author will be penalized on average ten times more compared to a manipulative author. It was also tested whether the percentile ranking of p-index causes this difference. Thus, next the percentile rank of authors with respect to their h-index against the p-index was used as shown in Figure 4.11. Even here, the p-index still emerges as the fairer index but the advantage of using p-index diminishes slightly when a percentile h-index is used. The difference of the h-index percentile is 53.03% while the difference of p-index is 10.68%. Therefore, the p-index is indeed a fairer measure in comparison to the h-index in reducing manipulation by authors.

Figure 4.12 shows the h-index and p-index respectively for the highest performing authors in the simulation. Here the difference is even starker. The p-index difference between non-manipulative and manipulative authors is insignificant compared to the h-index difference: 0.60% and 66.67% respectively. Figure 4.13 characterizes the advantage of using the p-index over the h-index for an average author by using the multiplier defined before. If a

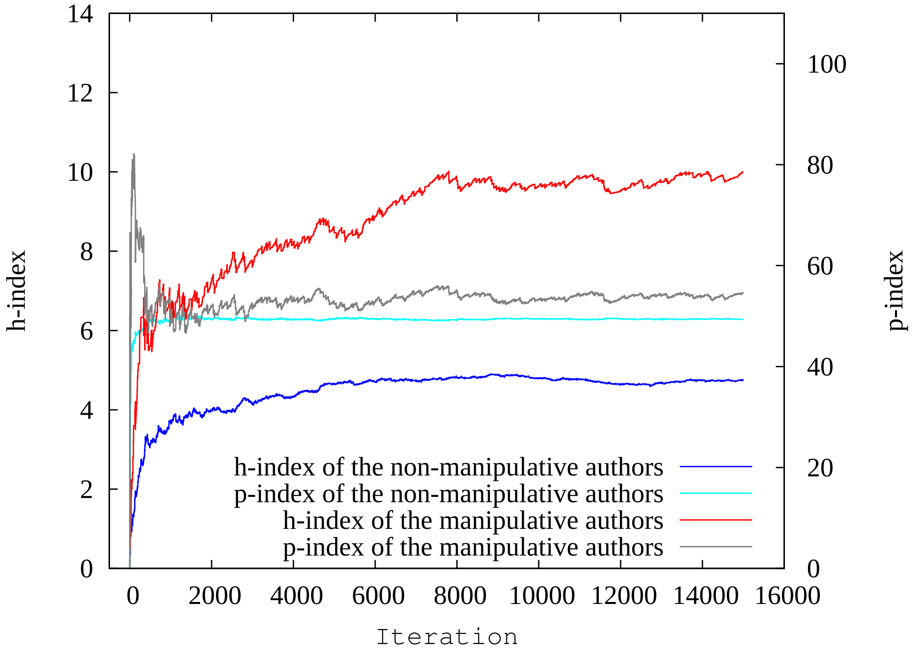


Figure 4.10: Variation of average the h-index and the p-index for non-manipulative and manipulative authors at each timestep.

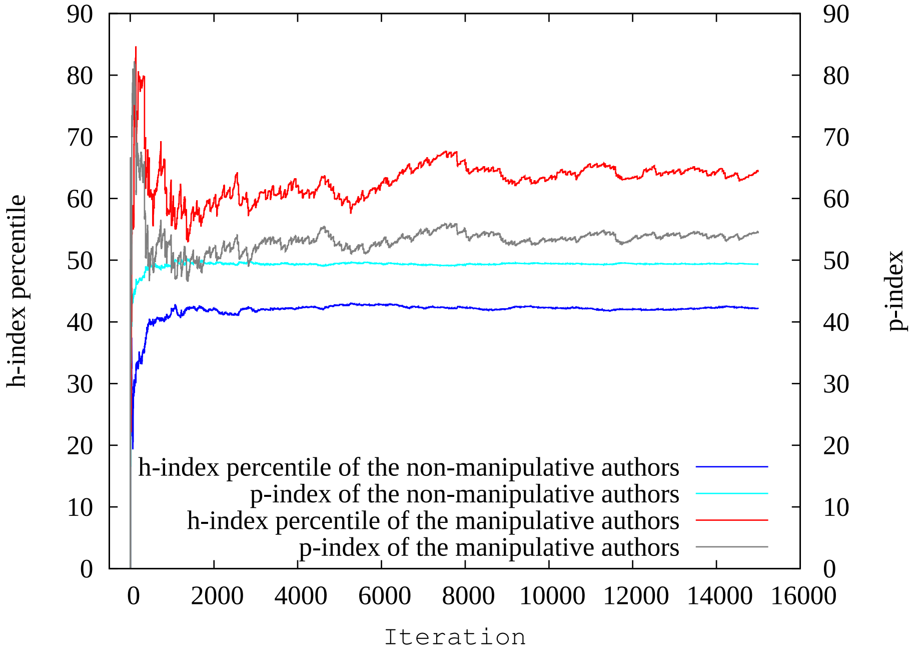


Figure 4.11: Variation of average the h-index percentile and the p-index for non-manipulative and manipulative authors at each timestep.

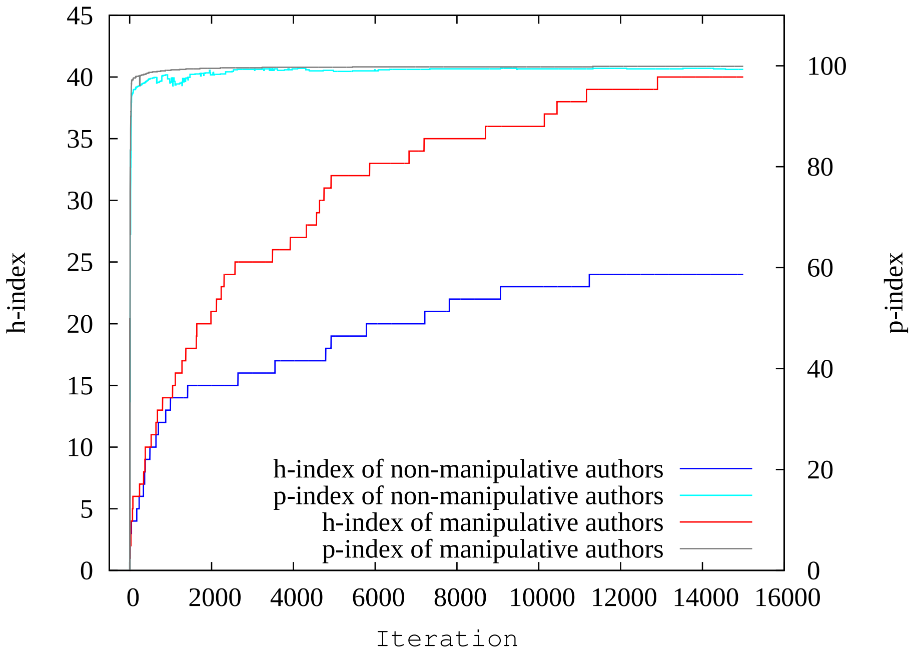


Figure 4.12: Variation of the h-index and p-index for highest ranking non-manipulative and manipulative authors at each timestep.

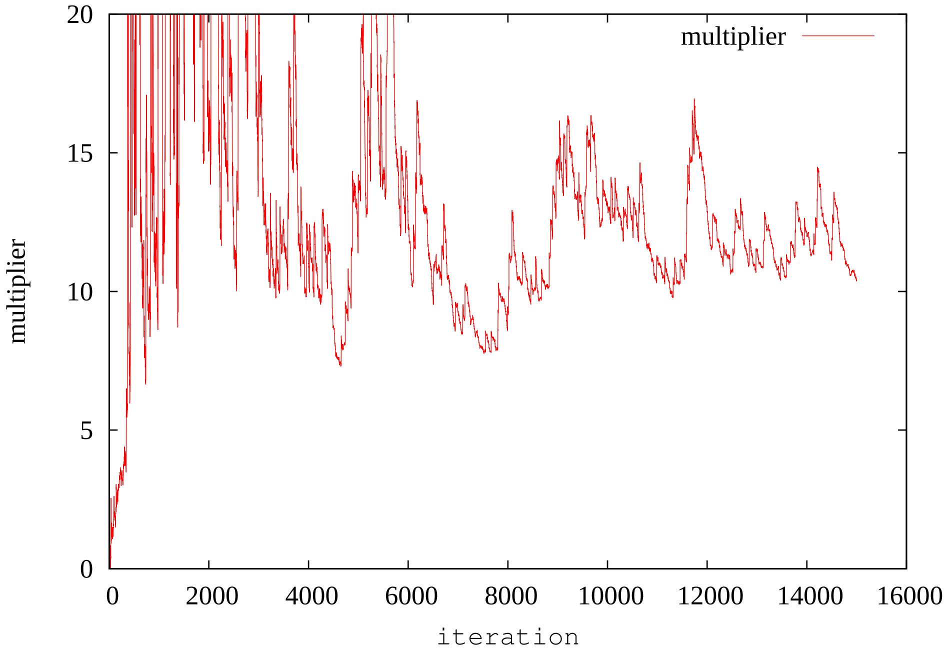


Figure 4.13: Variation of multiplier at each timestep.

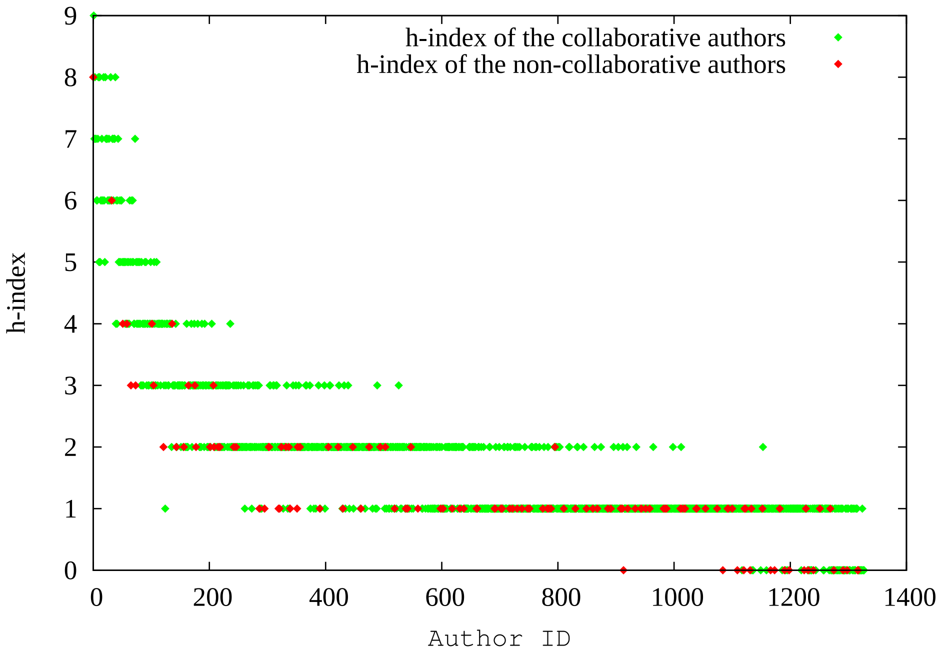


Figure 4.14: The spread of the h-index for collaborative and non-collaborative authors (as absolute values).

non-manipulative author and a manipulative author with adjacent IDs are chosen, what Figure 4.13 shows is that using h-index manipulative authors are favored ten times in average more than non-manipulative authors, compared to the p-index. Thus, for large systems of authors, the p-index is more than ten times ‘fairer.

4.4.2 Collaborative Authors vs Non-collaborative Authors

Figure 4.14 shows the spread of the h-index for collaborative and non-collaborative authors, while Figure 4.15 shows the spread of p-index for collaborative and non-collaborative authors, over author ID. It can be seen that most non-collaborative authors have relatively low h-indices, whereas in terms of p-indices the distribution is more evenly spread. As shown in Figure 4.14 and Figure 4.15, it is evident that collaborative authors usually have higher h-indices, whereas both classes of authors have similar p-indices.

Analyzing the average fluctuation of the h-index and the p-index for collaborative and non-collaborative authors reveals interesting characteristics. According to the Figure 4.16, which shows the variation of average h-index and p-index at each timestep, average h-index

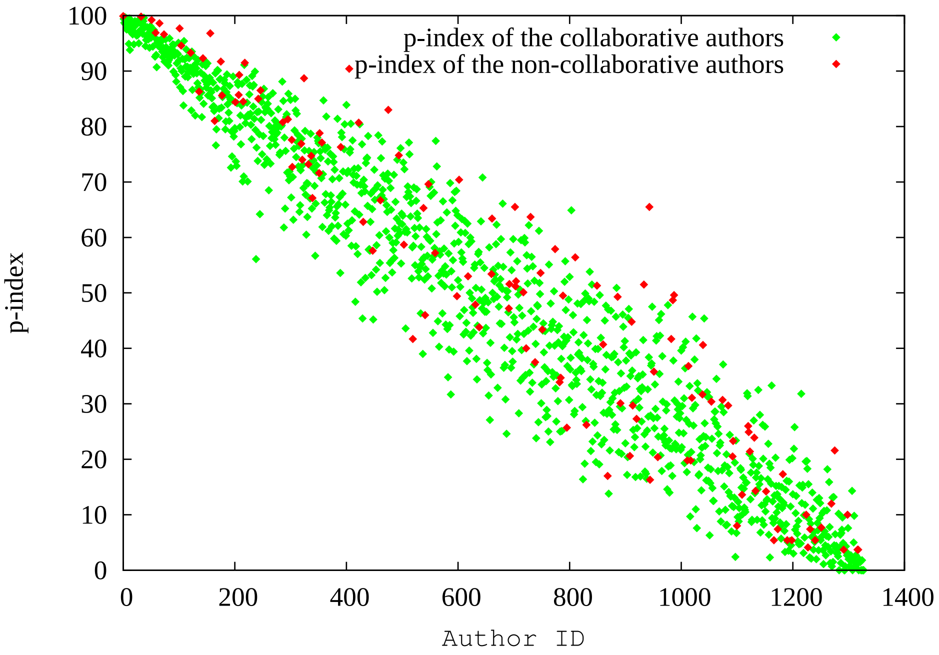


Figure 4.15: The spread of the p-index for collaborative and non-collaborative authors (as absolute values).

of the collaborative authors is 37% higher compared to non-collaborative authors. From Figures 4.14, 4.15 and 4.16, it can be inferred that p-index is robust against tendencies of manipulations in collaboration.

The h-index and p-index of the highest ranking authors of both classes are shown in Figure 4.18. This re-emphasizes that neither class of authors would gain a quantifiable benefit in rank by using the p-index, whereas it is entirely possible according to the h-index. Figure 4.18 quantifies the advantage one class of authors can have, on average, over the other class of authors by adopting the h-index instead of the p-index, by using the multiplier as defined in the previous section. This figure indicates that the p-index, at steady state, is roughly about twenty times 'fairer' than the h-index when collaborative tendencies are considered.

4.4.3 Quality-oriented Authors vs Quantity-oriented Authors

Figure 4.19 shows the spread of h-index for quality oriented and quantity oriented authors, while Figure 4.20 shows the spread of p-index for quality oriented and quantity-oriented

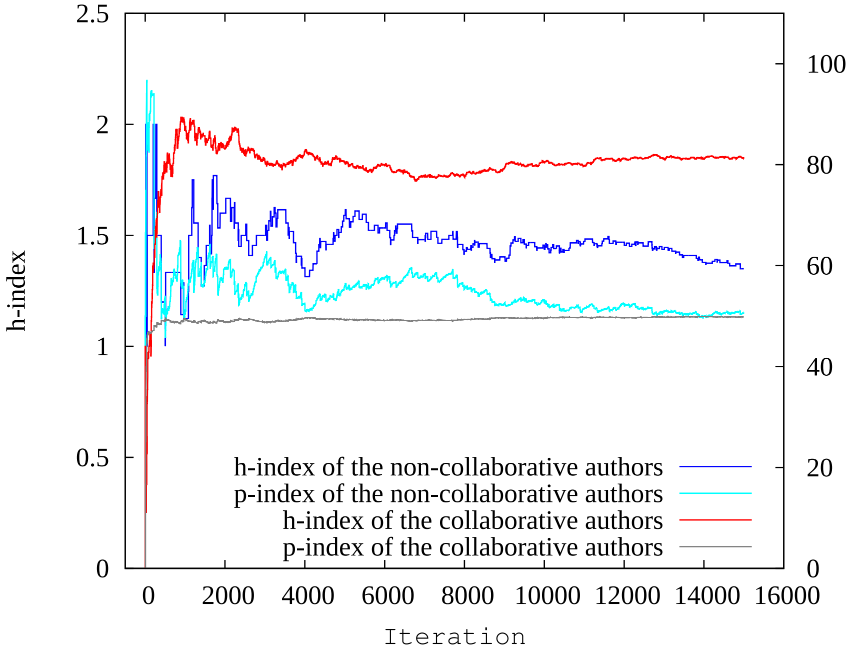


Figure 4.16: Variation of average the h-index and the p-index for collaborative and non-collaborative authors at each timestep.

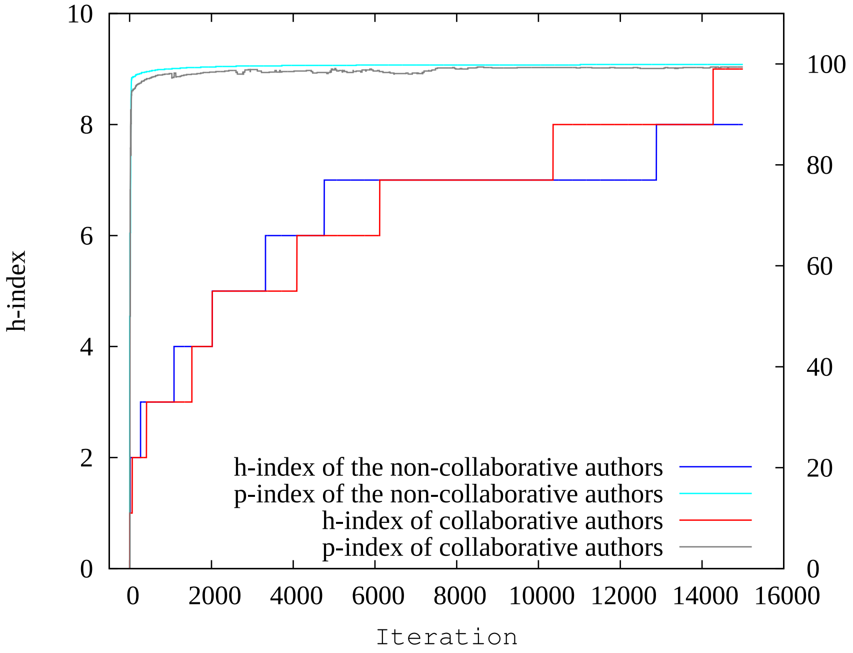


Figure 4.17: Variation of the h-index and the p-index for highest ranking collaborative and non-collaborative authors at each timestep.

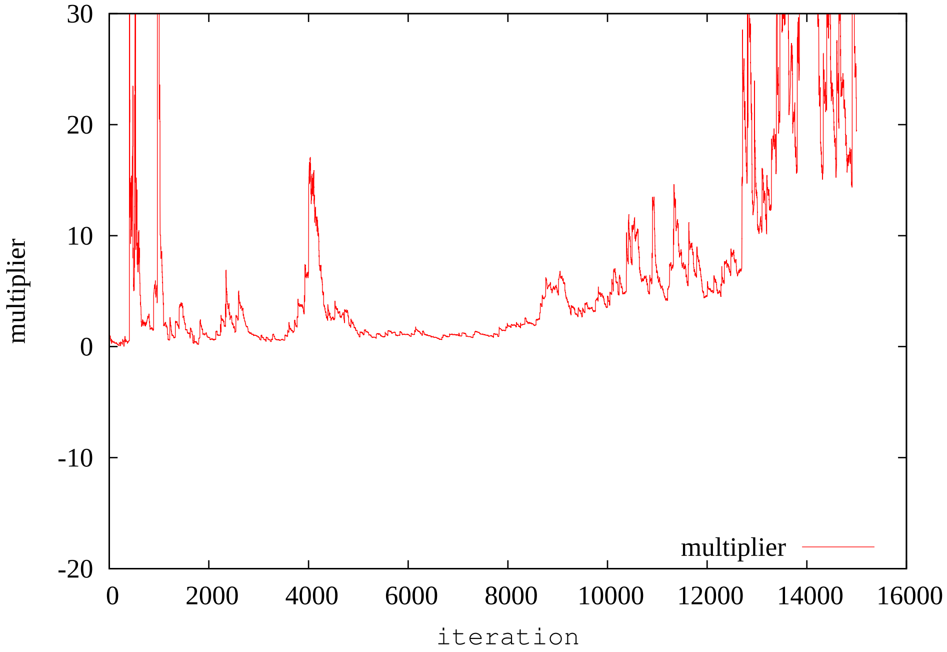


Figure 4.18: Variation of multiplier over each iteration for collaborative and non-collaborative authors over each timestep.

authors. Again it can be seen that while the h-index distribution favours quantity oriented authors, the two classes nearly overlap in p-index distribution. Figure 4.21 shows the variation of the average h-index and average p-index of the quality oriented and quantity oriented authors. From this figure, it can be inferred that average h-index of quantity oriented authors is significantly higher whereas average p-index is only marginally affected. The difference of the p-index is 7.39% whereas the difference of h-index is well over 110%, which puts the robustness of p-index into perspective in terms of quality vs quantity of publications.

Figure 4.22, on the other hand, shows the h-index and p-index variation of the highest ranked quality oriented authors and quantity oriented authors and it can be inferred that the percentage difference of h-index maybe as much as 104% while the difference of p-index is as low as 2%, showing that the differences can be amplified more than 50 times when h-index is used. This can have a clear detrimental effect on the quality oriented authors. Figure 4.23 plots the multiplier for the average quality vs quantity oriented authors in this scenario. It can be observed that the multiplier fluctuates around 15 in steady state, meaning that the advantage a quantity oriented author gains over a quality oriented author is fifteen times less when p-index is used.

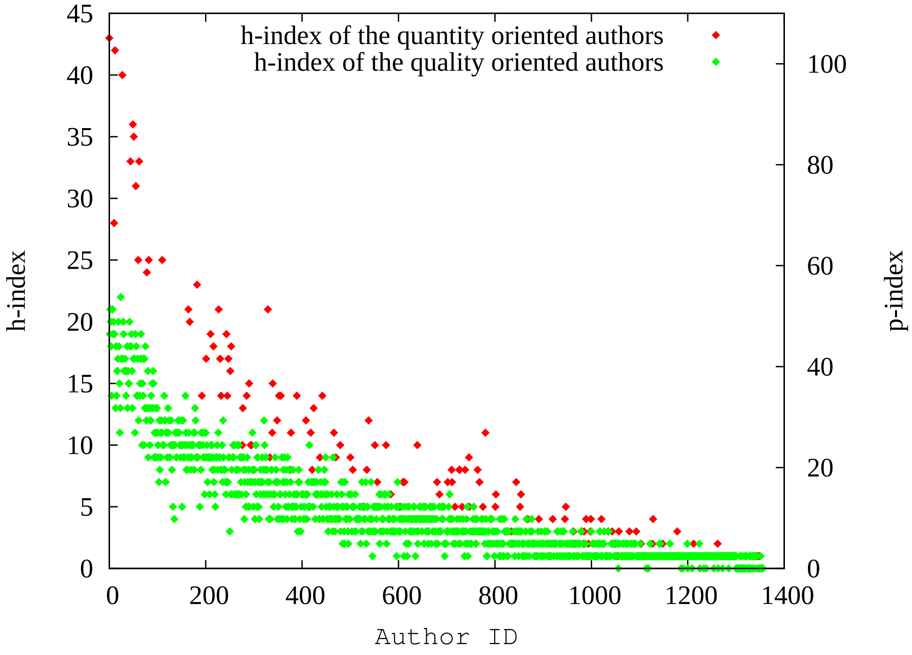


Figure 4.19: The spread of the h-index for quantity oriented authors and quality oriented authors (as absolute values).

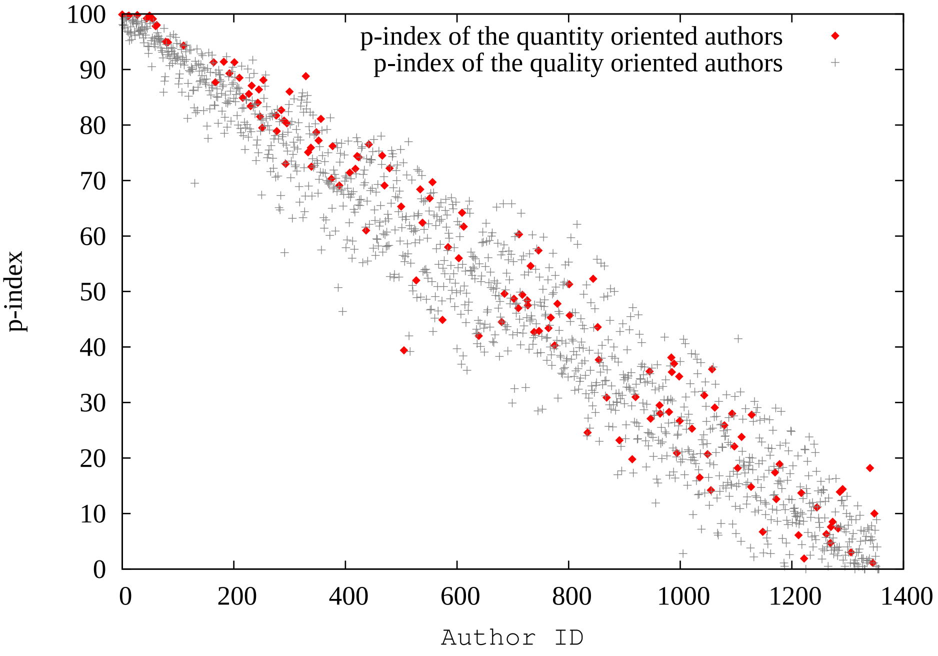


Figure 4.20: The spread of the p-index for quantity oriented authors and quality oriented authors (as absolute values).

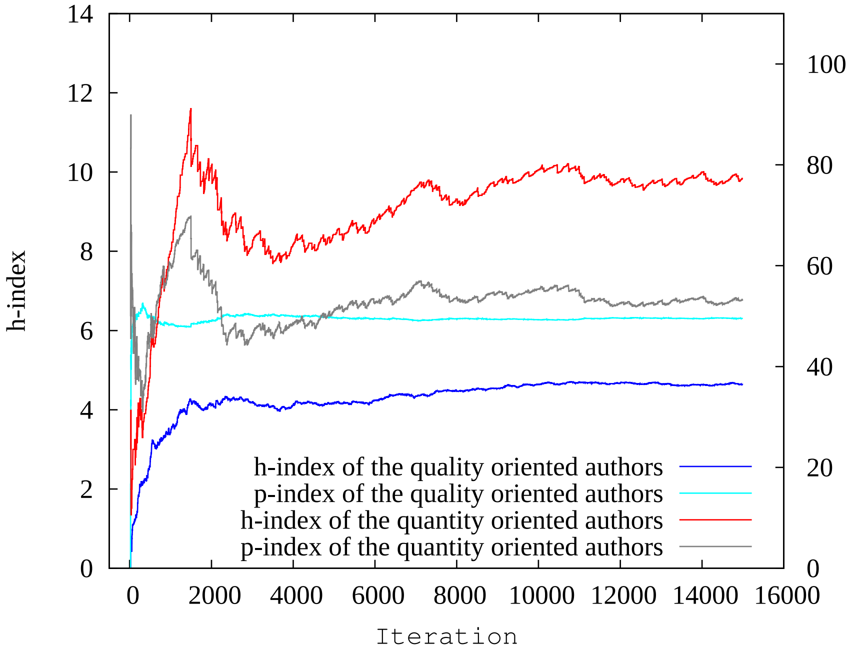


Figure 4.21: Variation of average the h-index and the p-index for quantity oriented authors and quality oriented authors at each timestep.

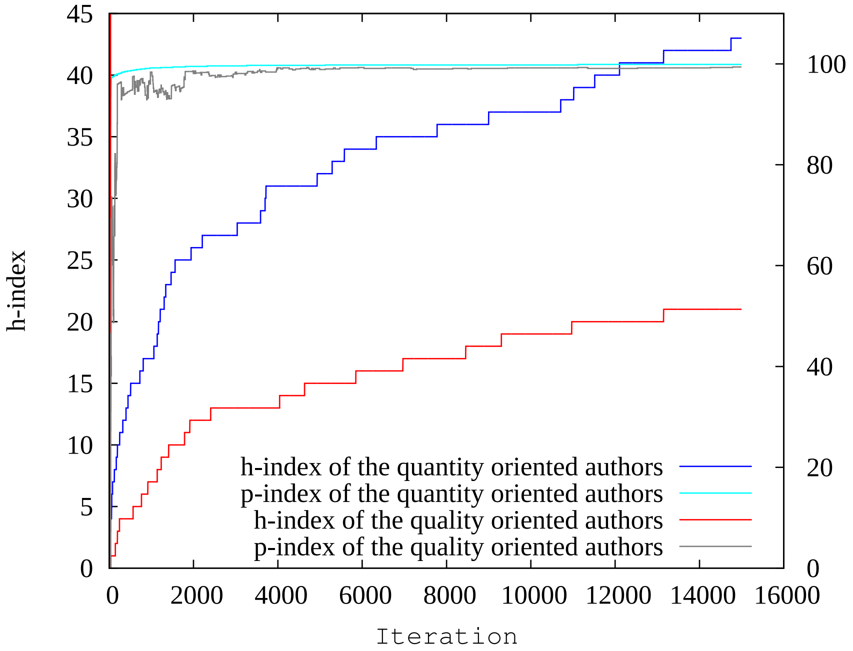


Figure 4.22: Variation of the h-index and the p-index for highest ranking quality oriented authors and quantity oriented authors at each timestep.

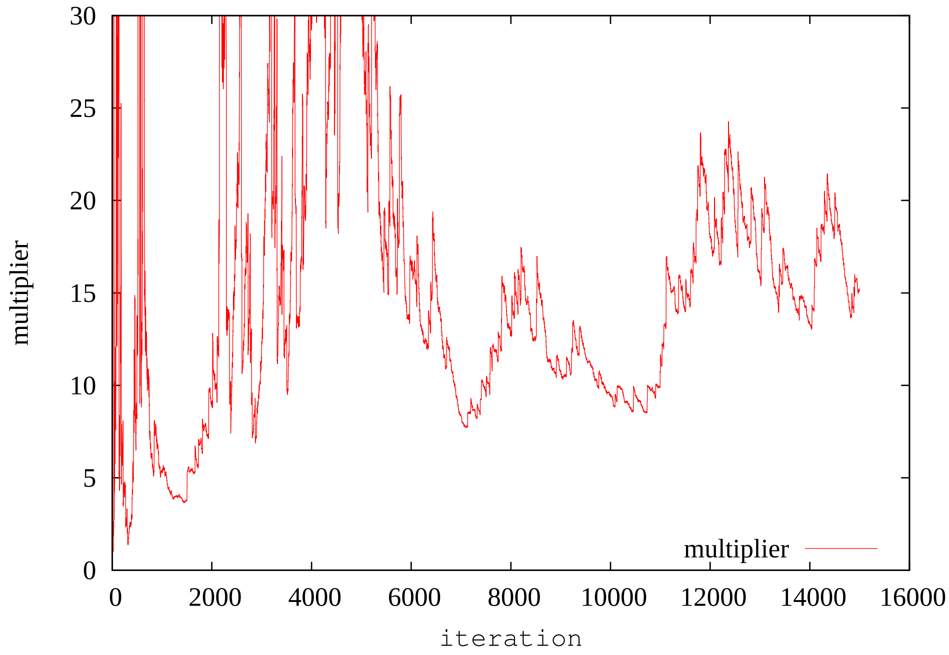


Figure 4.23: Variation of multiplier of the quantity oriented authors and quality oriented authors over each timestep.

4.5 Conclusion

This chapter has presented a new index, the p-index (pagerank-index) to quantify the scientific output of researchers. The p-index addresses the lack of weighted citations, which is a core shortcoming of h-index. Since the h-index considers all citations to be equal in nature, it is possible to manipulate the h-index for personal gain. The p-index provides a much fairer means of comparing scientists and is a lot less prone to manipulation and massaging.

The robustness and fairness of the p-index over the h-index was demonstrated using three simulated scenarios. The first scenario demonstrated that a manipulative author who is interested in publishing low impact papers in order to reference their own papers can indeed massage their h-index. In contrast, the p-index is much fairer and does not allow manipulative authors to gain an undue advantage. The second scenario revealed that the authors who are interested in contributing insignificantly to a number of papers will certainly gain a higher h-index. However, such authors cannot gain the same advantage using the p-index, which indicates that it is a fairer metric. The third scenario described a situation where authors are interested in maximizing their number of publications so that they may obtain a significant benefit by using h-index. This is not feasible if the p-index is used, as shown clearly in the results of the simulation. In conclusion, these results provide evidence to suggest that the p-index is robust against manipulations and performs fairer and more effectively in ranking scientists. The findings were quantified and it was found that on average, in each scenario the p-index reduces the unfair advantage to one group over the other by ten to twenty times, which is a significant improvement.

Even though the simulation system evolves a citation network and the underlying collaboration networks from the inception of a specific field, the p-index can be applied at any point in evolution of an existing academic database. This makes it possible to seamlessly integrate p-index into existing academic databases.

Chapter 5

Case Study I: Quantum Game Theory Dataset

5.1 Introduction

In the quest for a more objective ranking metric for scientists, the h-index was examined and several possible manipulations on it which results in the ability to massage one's h-index were exposed. This is an undesirable quality in an objective ranking metric and it creates a range of issues for the ethical and honest scientists. Hence it is proposed that the underlying citation network dynamics be utilized in order to evolve an index which is fairer and rewards scientists for their competence rather than their ability to massage the available indices. The definition of the p-index is provided and discussed in detail before moving on to a simulated system to provide evidence about possible manipulations on h-index. The presented results and analysis therein suggests that p-index is indeed fairer and more resilient to the said manipulations that are viable in the h-index. However, since a real world citation network would also give valuable insight, the p-index is applied to a citation network curated using the Quantum Game Theory Google Scholar profile, as presented here.

The rest of this chapter is organized as follows. The second section describes the dataset used in detail and how it was curated. The third section provides an overview of the results obtained by applying p-index and h-index to the citation network. The results are further analyzed under two subsections to comprehensively illustrate the characteristics of p-index. The final section summarizes and concludes this chapter.

5.2 Citation Network of Scientists in the Field of Quantum Game Theory

The p-index is designed to be implemented in a large citation database, such as Scopus, ISI Web of Science or Google Scholar, so that the entire body of scientists listed in that database can be ranked. The h-index, for example, is calculated by many of these databases (such as Google Scholar), even though the h-index displayed in Google Scholar for a scientist does not have to be the 'real' h-index of a scientist, and the h-index calculated for the same scientist using different databases could actually be different. It has become the norm to accept the 'Google Scholar h-index' of a scientist as the de-facto official h-index of that scientist. However, to demonstrate the utility of the new index, a pilot study is necessary. Such a pilot study could be undertaken using either a simulated system of authors or authors from a particular scientific field.

While using a simulated system might technically be the easier approach, it is difficult to model the publication behavior of scientists in sufficient detail for such an approach to be defensible. In any case, this approach will lack the rigor of a real world analysis, and the insights such an analysis can present. Therefore, in this chapter a second approach was chosen, by narrowing this analysis to a particular scientific field. Of course, this posts the immediate question: How is a paper defined as belonging exclusively to a particular research area? one way is to consider papers from a particular journal or group of journals. For example, many authors have considered the APS journal papers to undertake citation network analysis. However, such an approach will unnecessarily constrain the author behavior. Real world scientists typically publish in multiple journals and conferences. Furthermore, since one of the motivations is to analyze the comparative impact of low-profile papers on the h-index and the p-index, it is not ideal to limit the study to one family of journal papers. Furthermore, we are interested in reconstructing not just snapshots of citation networks but the evolutionary history of a community of authors, thus the analysis of citation networks from a particular group of journals is of limited value.

Luckily, some emerging fields of science have dedicated Google Scholar profiles, where authors can voluntarily 'attach' their papers. The field of 'Quantum Game Theory' as defined by the dedicated Google Scholar profile was chosen for this field for our analysis [51]. This field (as defined by the profile) has achieved sufficient growth to facilitate a meaningful analysis (at time of access, it contained 685 papers, 785 authors, and 3776 citations), while at the same time is sufficiently young so that the growth of the field could be traced back to the 'initial' papers. Of course, it is conceivable that some papers belonging to this field have not been added to the said profile by the authors. Despite this, the field of quantum game theory as defined by this profile is a suitable dataset to test the utility of the p-index, particularly since p-index of all participating authors could be computed 'historically' from the moment the first paper was published to the present day.

Therefore, we first generated citation networks from this dataset, after each new paper was added. The time between the additions of two papers was not considered, since only 'internal' citations (citations from papers in this particular field to papers in the same field) were studied. Any citation to external papers are ignored. Thus, the addition of a new paper corresponds to a 'timestep', and this sense of causality is sufficient for the purposes of this research - we need not worry about the actual dates of publication. The evolved citation network at four timesteps ($t = 100, t = 200, t = 400$ and $t = 685$) is shown in Figure 5.2. To generate citation networks, IDs were assigned to papers then a list of references consisting of these IDs for each paper in the field was manually prepared, since the reference list follow various conventions and sometimes do not even include the paper title and it is difficult to automate this process. Once this manual curating is done, the citation networks corresponding to each 'timestep' were generated using a computer program written in Java language, where paper IDs with reference list (also in IDs) were inputs. The IDs were in the order of publication, therefore to generate a citation network at timestep $t = t_0$, all IDs less than t_0 were considered. The naming conventions used in curating the dataset are included in Table 5.1 so that this model can be used in future experiments by different scientists as well.

As it can be seen from Table 5.1, the titles and authors have different forms and hence the first step in generating the citation network was to get them to a single conformity. This was done by converting the text files to ASCII characters and then making all the letters 'small letters', which aids us to continue with string matching in generating the citation

Section	Description
Papers in the citation network (papernet)	The title of the papers and respective authors are curated in a text file. A paper and each of its authors are separated by a comma while papers themselves are separated by a new line. An example for further preference is included here. Utility of information game theoretic approach, H Everett A GLOBAL THEORY FOR LINEAR QUADRATIC DIFFERENTIAL GAMES, DL Lukes, DL Russell A quantum shuffling game for teaching statistical mechanics, PJ Black, P Davies, JM Ogborn
What papers are referenced by each paper (reference lists)	This information is also stored using a text file. However each paper has a text file corresponding to its ID that contains a list of papers referenced by that specific paper. The list of IDs are separated using a new line as shown below. What papers are referenced by paper ID 524 45 63 466 59 197 292 225

Table 5.1: Details of the Quantum Game Theory citation network curative process.

network.

The steps required to generate the citation network and the collaboration network is demonstrated in Figure 5.1. Since it was necessary to examine the evolving citation network, the p-index and h-index were applied to the citation network after the addition of each node and the results for each author in the network up to that point in time were recorded.

5.3 Analysis and Results

First, let us make some qualitative observations which illustrate the utility of p-index and how it can help rectify biased perceptions of the scientific contributions of scientists in a field. In doing so, let us emphasize that it is not the intention of this study to judge any scientist or their contributions to science: nor is it claiming that the p-index is the

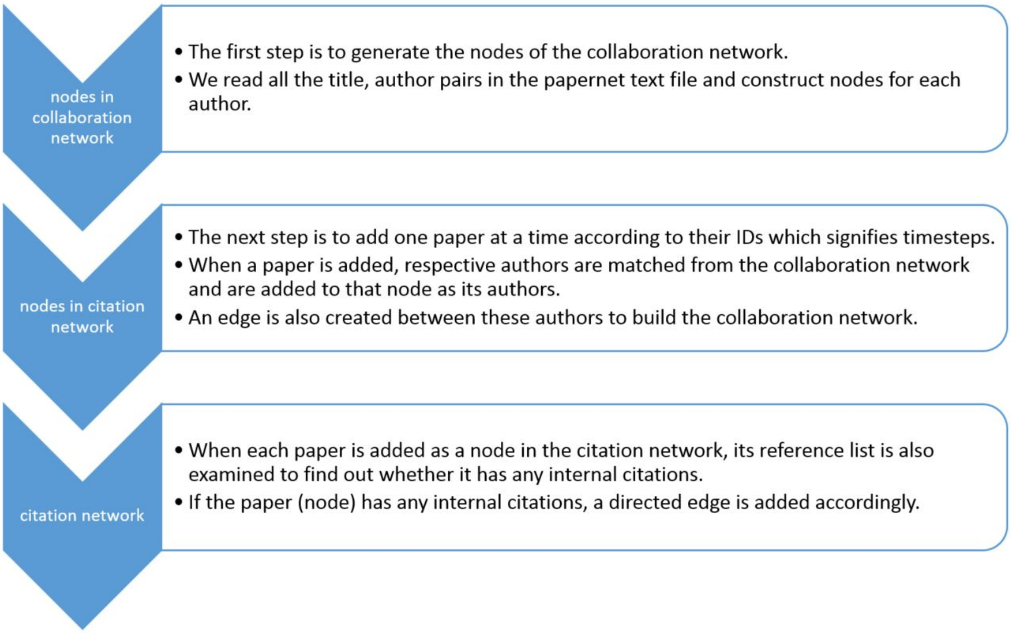
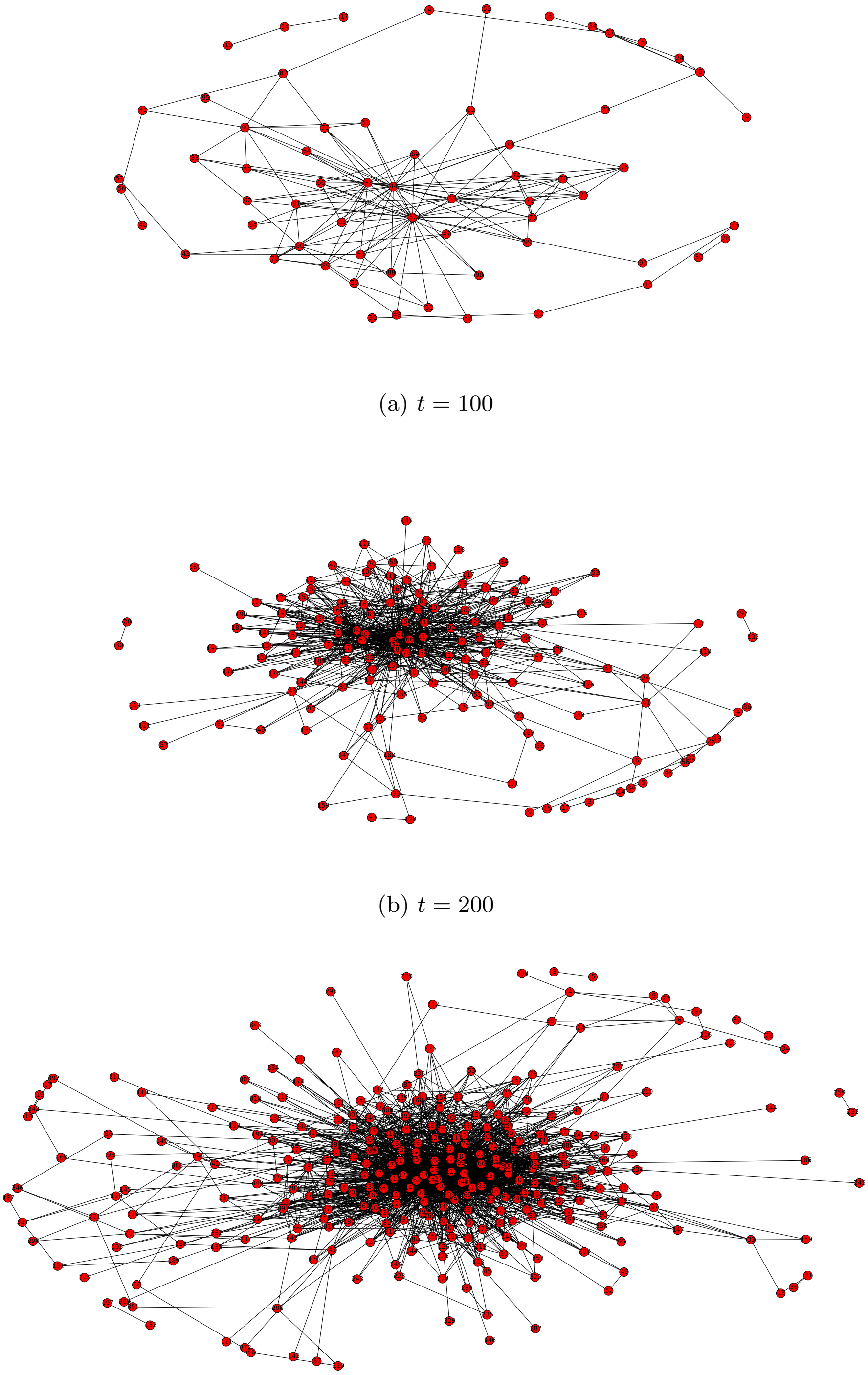

nodes in collaboration network
• The first step is to generate the nodes of the collaboration network.
• We read all the title, author pairs in the papernet text file and construct nodes for each author.
nodes in citation network
• The next step is to add one paper at a time according to their IDs which signifies timesteps.
• When a paper is added, respective authors are matched from the collaboration network and are added to that node as its authors.
• An edge is also created between these authors to build the collaboration network.
citation network
• When each paper is added as a node in the citation network, its reference list is also examined to find out whether it has any internal citations.
• If the paper (node) has any internal citations, a directed edge is added accordingly.

Figure 5.1: The construction process of the citation network and the collaboration network from the Quantum Game Theory field (according to Google Scholar).

ultimate metric for measuring scientific brilliance. We merely intend to demonstrate that, given the reality of scientists being judged, rightly or wrongly, using citation metrics, the p-index can return a fairer comparison than other metrics currently in use. Therefore, let us consider Table 5.3, which shows the top twelve scientists in the field of Quantum Game Theory, according to h-index. Note well that we are using the 'internal' h-index, which is computed using only the papers in the field and therefore typically much lower than the 'overall' h-index of scientists. We do not want to use the 'overall' h-index since that purportedly measures the scientific contributions of authors to all fields, not just quantum game theory. We also show the p-index, and ranking based on p-index, of each of these authors in Table 5.2. We can observe that some authors, such as *A. Iqbal* (h-Rank = 1, p-Rank = 1) and *A.P.Flitney* (h-Rank = 3, p-Rank = 3) maintain their positions at the top of the table with respect to both ranking systems. Other authors, such as *D.Abbott*, leap several places up (12th to 2nd) when p-index is applied. Several other authors, such as *J. Du* (h-Rank = 4, p-Rank = 14), *R. Han* (h-Rank = 7, p-Rank = 46), *X. Xu* (h-Rank = 7, p-Rank = 46), *X. Zhou* (h-Rank = 7, p-Rank = 46) drop significantly in ranking when the p-index is applied. Looking at the data available in the Google Scholar Profile, it is very easy to see how and why this happens. The top papers by D. Abbott were published in

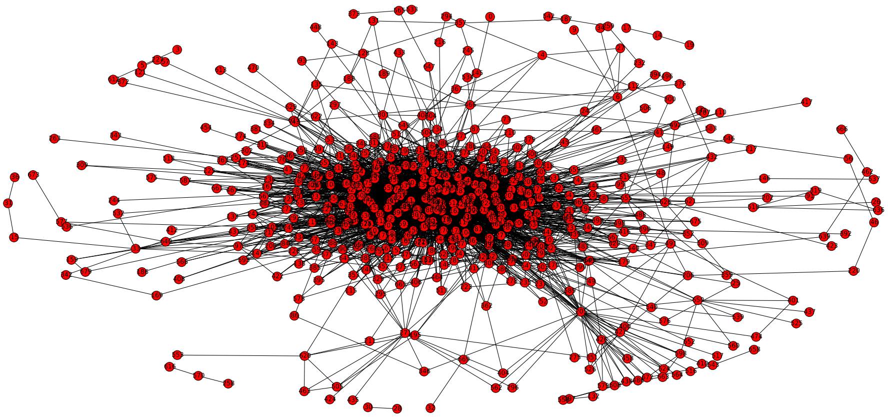


(a) $t = 100$

(b) $t = 200$

(c) $t = 400$

Figure 5.2: ...continued to next page.



(d) $t = 685$

Figure 5.1: The citation networks generated from the Quantum Game Theory field (according to Google Scholar) at $t = 100, t = 200, t = 400$ and $t = 685$. The timesteps simply correspond to the number of papers in the field at that time (e.g., at $t = 400$ there were 400 papers in the field). At the time of access there were 685 papers and 3776 intra-field citations, according to Google Scholar.

Author	h-index	h-percentile	p-index	Number of papers	Average Authors per Paper	rank-h	rank-p
A Iqbal	11	99.7	99.8	49	2.65	1	1
AP Flitney	10	99.6	99.7	26	2.34	3	2
D Abbott	7	98.4	99.6	37	2.81	12	3
EW Piotrowski	8	98.6	99.4	22	2.22	7	4
A Nawaz	4	96.6	99.3	15	1.86	21	5
DA Meyer	5	97.4	99.2	9	1.77	18	6
M Ramzan	4	96.6	99.1	14	2.42	21	7
J Sladkowski	9	99.2	98.9	23	2.26	4	8
AH Toor	11	99.7	98.8	20	2.1	1	9
T Cheon	6	97.8	98.7	16	1.93	13	10
J Du	9	99.2	98.6	15	4.66	4	11
P Frackiewicz	1	55.5	98.4	6	1.33	87	12

Table 5.2: The top 12 scientists as ordered by their p-index in the field of quantum game theory.

highly reputable journals, such as *Nature,Physical Review Letters*, and *Statistical Science*. Therefore, the chance of his citations coming from other highly reputed journals is high. More importantly, he has only an average 2.8 authors for each of his papers in this field, which means his papers had typically been authored by two or three authors, including

Author	h-index	h-percentile	p-index	Number of papers	Average Authors per Paper	rank-h	rank-p
A Iqbal	11	99.7	99.8	49	2.65	1	1
AH Toor	11	99.7	98.8	20	2.1	1	9
AP Flitney	10	99.6	99.7	26	2.34	3	2
J Du	9	99.2	98.6	15	4.66	4	11
H Li	9	99.2	97.5	12	4.66	4	19
J Sladkowski	9	99.2	98.9	23	2.26	4	8
NF Johnson	8	98.6	98.2	13	2.076	7	14
X Xu	8	98.6	95.1	8	5.5	7	38
X Zhou	8	98.6	92.1	9	5.33	7	62
R Han	8	98.6	79.59	8	5.5	7	160
EW Piotrowski	8	98.6	99.4	22	2.22	7	4
D Abbott	7	98.4	99.6	37	2.81	12	3

Table 5.3: The top 12 scientists as ordered by their h-index in the field of quantum game theory.

himself. These factors would not be taken into account by h-index, which merely counts the number of citations an author receives, but such details are rightly taken into account by p-index, which returns a relatively higher score for authors who put in a higher share of individual work into a paper and publish in and, more importantly, get cited by higher quality journals.

On the other hand, the author R. Han, while also publishing in high impact journals such as *Physical Review Letter*, often appears to be part of a large group of co-authors. The average number of authors per paper for this author is 5.5, which means the author has co-authored with typically five or six authors, which is significantly higher than authors like D. Abbott. Furthermore, this author appears as the last author in their most highly cited papers. In fact, the authors J.Du (4.7), X. Xu (5.5), X. Zhou (5.3), and R. Han (5.5) all appeared to have co-authored several papers together, as part of a relatively large pool of authors, as the average number of authors in their papers given in brackets above indicate. Some of these papers have been highly cited. While this clearly gives them all high h-index values, the p-index considers the fact that the co-author pool is relatively large and thus does not give them very high scores.

The collaboration network in Figure 5.5 lends further credence to our arguments. As

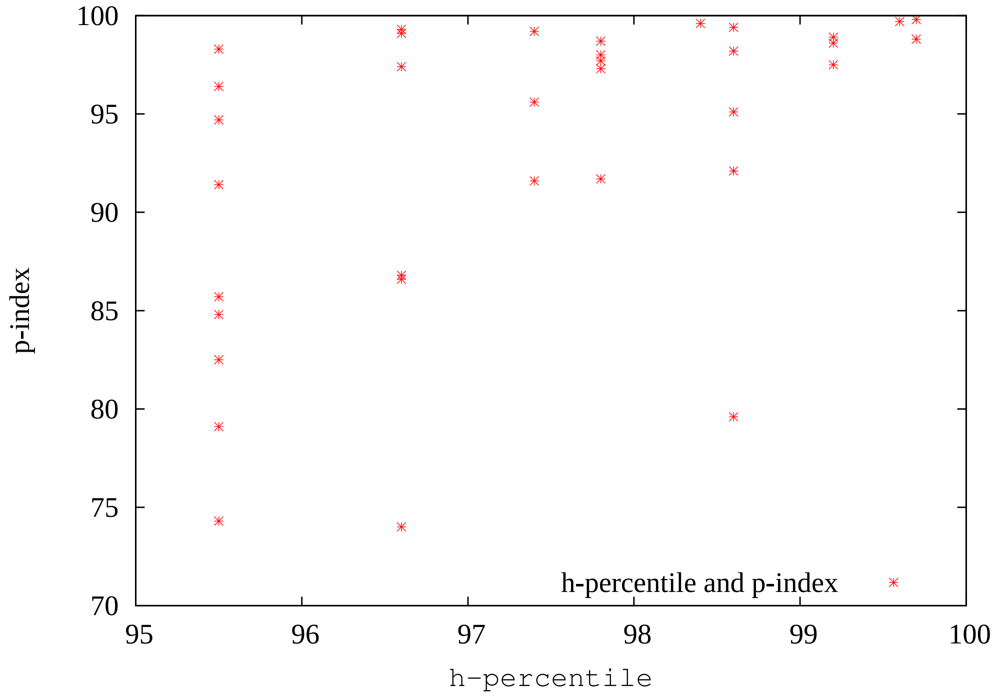


Figure 5.2: The h-index and p-index of the best 5% authors in the field of quantum game theory. Since the p-index is a percentile, percentile values were used for the h-index as well, rather than actual h-index values. Note here that the p-index value varies from 70% to 100%. That is, some authors who are among the top 5% in terms of h-index are not even among the top 25% when p-index is considered.

highlighted in the figure, it indicates that the author R. Han, for example, has only six co-authors and these authors seem to be in a almost fully connected subnetwork, meaning they wrote most of their papers in this field as a group. Particularly, authors R. Han, X. Xu, J. Wu, and M. Shi seems to have written no paper together with authors outside this subnetwork. This subnetwork has co-authored some highly cited papers, so it is not surprising that members of this subgroup, and particularly the authors mentioned above, have their ranks within the field change considerably when p-index is used (X. Xu: h-Rank = 7, p-Rank = 46, R. Han, h-Rank = 7, p-Rank = 46, M. Shi, h-Rank = 27, p-Rank = 147). Author J. Wu, who only has three papers and a h-Rank of 87, however, does not seem to be affected that much. On the author hand, author D. Abbott as highlighted is not part of a ‘dense’ subnetwork shown in Figure 5.6, but it is evident that he has high ‘centrality’ within the co-authorship network, which means his contributions help form the ‘backbone’ of the field and he has helped expand the field in various directions, which

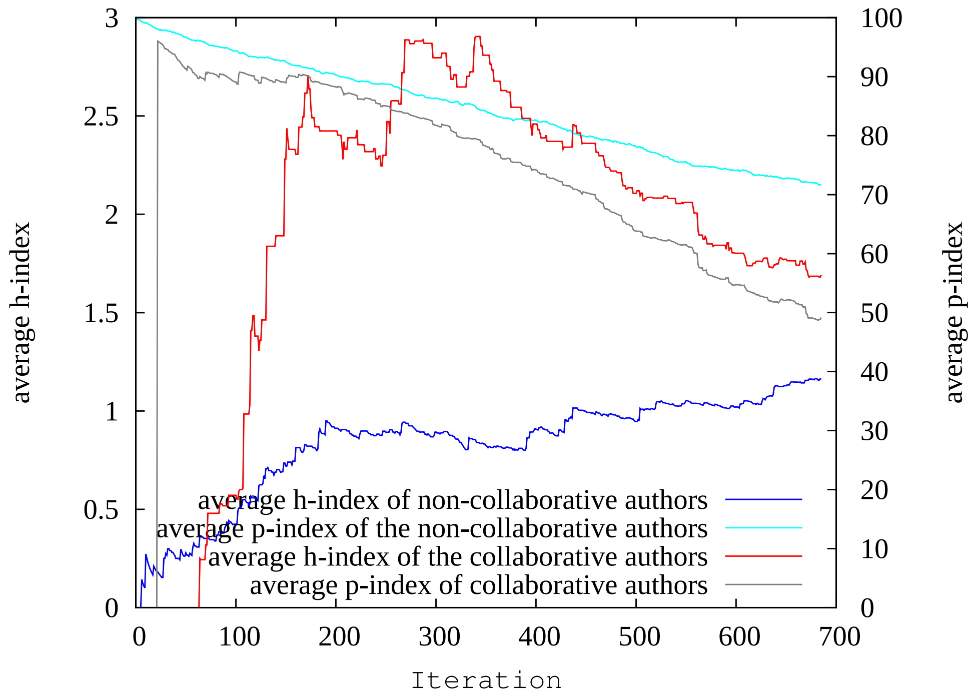


Figure 5.3: The variation of h-index and p-index for two groups of authors during the evolution of quantum game theory field. The x-axis corresponds to each new paper added and the time line of the evolution is from 1955 to 2014. One group of authors are classified as ‘collaborative’ and another group as ‘non-collaborative’. The way this classification was done is explained in the text. It is clear that while the h-index favours the ‘collaborative’ authors, the p-index, in general, tends to favour the ‘non-collaborative’ authors.

justifies by his shift from 12 to 2 when p-index is used. Another author who has obviously high centrality within the author network, performing the function of bridge between two group of authors (even though he has only two co-authors, one each from each of these groups) is P. Frackiewics, as shown in Figure 5.7, and our analysis indicated that his h-rank is 87, but his p-rank is 13. Such a jump is justified by his definitive contribution to the field. Therefore, the p-index is a lot fairer in teasing out real contributions of scientists to a field of science compared to the h-index. This is all the more interesting because the p-index does not use the co-authorship network in its calculations at all - it only uses the citation networks but, due to its in-built fairness, the authors who make definitive contributions by being the backbone of genuine collaborations are rewarded.

Figure 5.2 shows the p-index vs percentile (internal) h-index plot for this dataset. The h-index is also converted to a percentile. This is because doing this makes it easier to

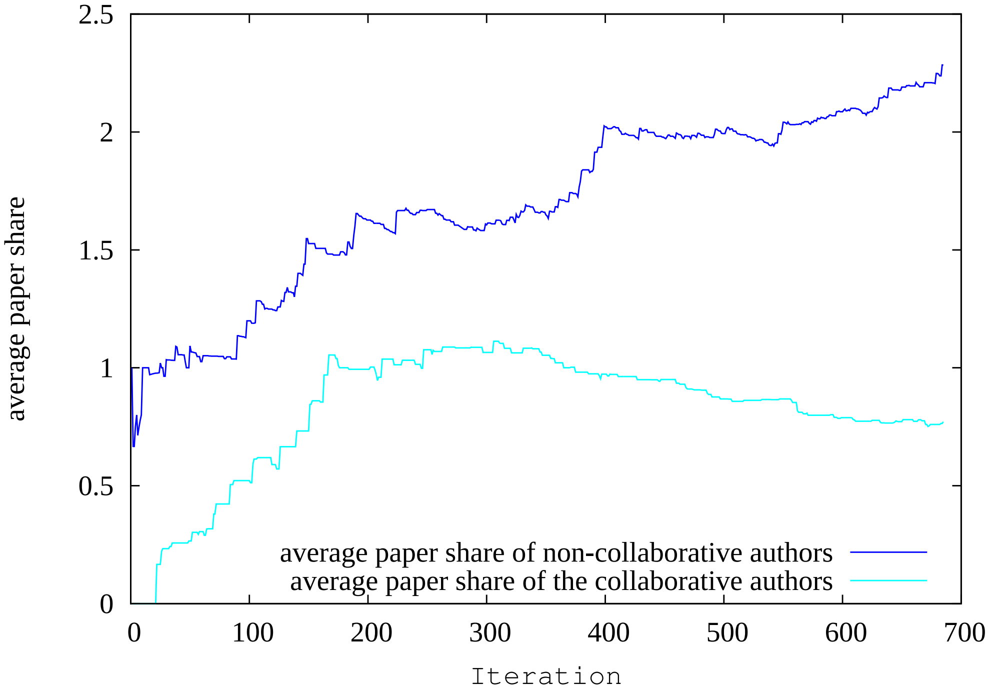


Figure 5.4: The average ‘papershare’ of collaborative and non-collaborative authors during the evolution of quantum game theory field. The ‘papershare’ is calculated as the summation of proportional contributions made to papers. For example, if an author has contributed two papers each with two other co-authors, they have a total of 4/3 papershares. It is clear that the ‘non-collaborative’ authors work harder and have more ‘papershares’ than collaborative authors. Contrasting with Figure 5.3, the p-index highlights this fact by favoring the ‘non-collaborative’ authors, while the h-index incorrectly favours collaborative authors who on average produce fewer ‘paper-shares’.

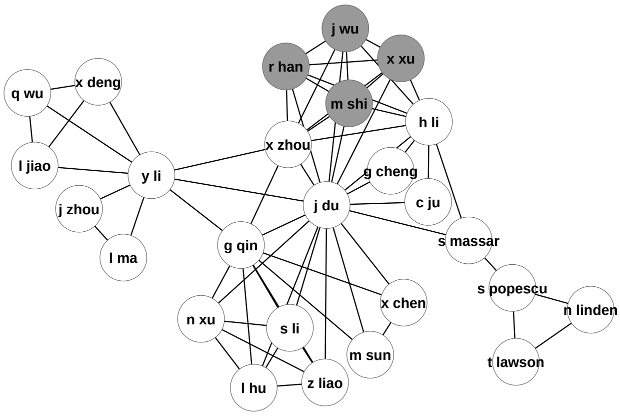


Figure 5.5: Part of the collaboration network highlighting authors R. Han, X. Xu, J. Wu, and M. Shi.

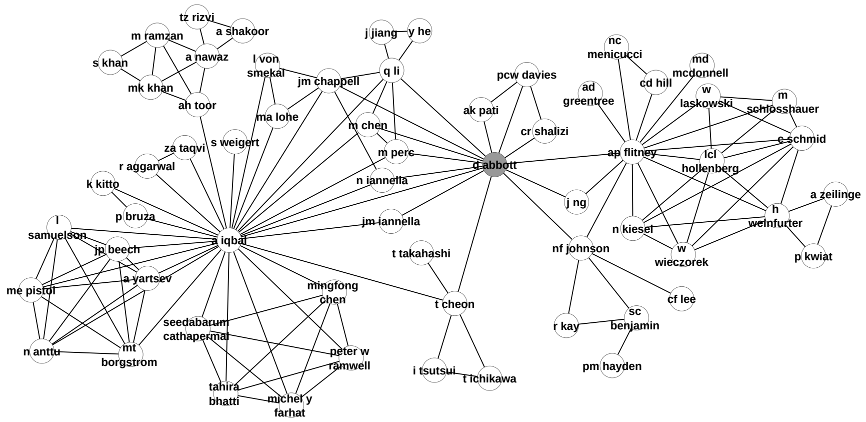


Figure 5.6: Part of the collaboration network highlighting D. Abbott.

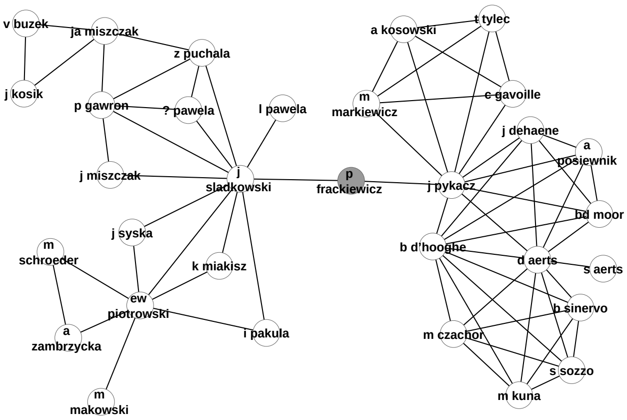


Figure 5.7: Part of the collaboration network highlighting P. Frackiewics.

compare the two metrics, and also because otherwise it might be argued that the variation is due to a percentile being compared with a 'direct' score. We again use the 'internal' h-index for the reasons described above. The h-index percentile range below 95% is not shown, since the h-index is a discrete quantity and there are only two datapoints below 95%, corresponding to h=0 (88.9%) and h=1 (55.4%), and showing these would reduce clarity. This figure illustrates that there is a lot of variation of the p-index values for each percentile h-index points. Some authors who score 95.4% h-percentile score as low as 41% in terms of p-index, while others score as high as 98%. It could be demonstrated, by individual analysis as has been done in the previous paragraph, that those authors who have low p-index value either have typically co-authored with a larger pool of authors or have received most of their citations from low-impact papers, while those authors who score highly have relatively worked with a smaller pool of authors and received their citation from more reputable sources, thus justifying their high p-index. Similarly, among authors who score 96.5% h-percentile, there is a range of p-index values from 78% to 97%. The data points at the right extreme highlighted in red represent A. Iqbal, A.H.Toor and A.P. Flitney, who have scored well both in terms of h-index and p-index as discussed above, while the datapoint highlighted in green represents D. Abbott, who is 12th in terms of h-index (98.5% h-percentile) but scores much higher in terms of p-index (99.7%).

5.3.1 Comparison of h-index and p-index

As it has been pointed out at the start of this section, evidence from the quantum game theory citation network suggests that p-index is a fairer ranking metric compared to h-index. For example, Figure 5.3 presents the average h-index and p-index of collaborative and non-collaborative authors where there is a clear propensity for the collaborative authors to score a higher h-index, which is not possible if p-index is used. Affirming the simulation evidence provided in the last chapter, it is clear that p-index emerges as a fairer metric if applied to a real life citation network as well.

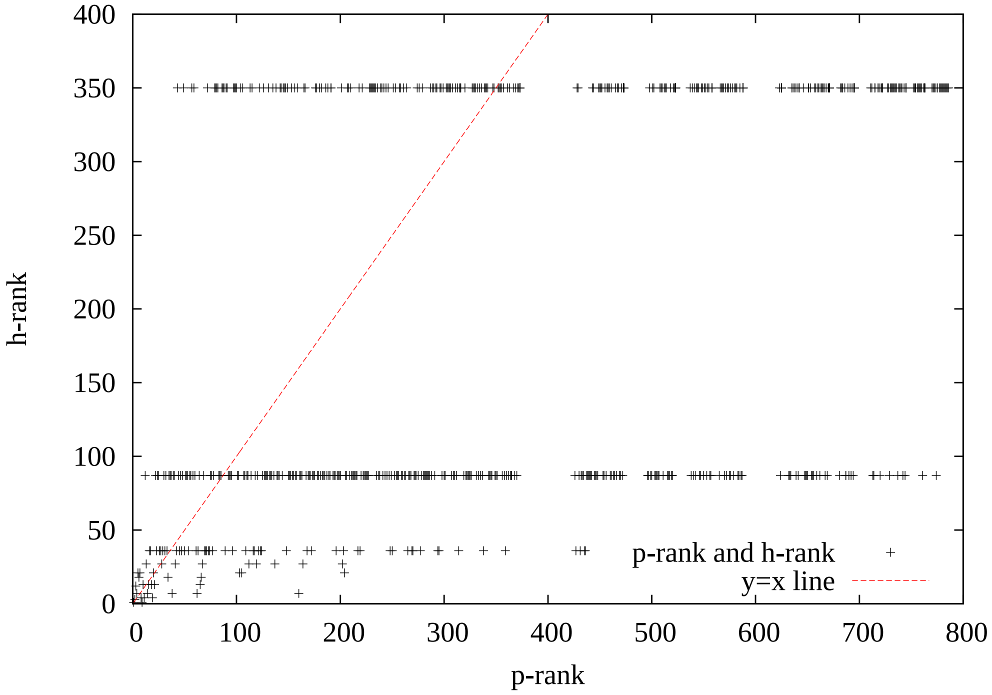


Figure 5.8: The ranks of the authors as ordered by h-index and p-index. Y=X line is also included to better visualize the authors whose ranks have not been changed.

The ranks of authors as ordered by h-index (h-rank) against the ranks of authors as ordered by p-index(p-rank) is shown in the Figure 5.8. As it can be clearly seen from this figure, the rank of some scientists have gone down when ranked using p-index and the ranks of another group of scientists have gone up when ranked using p-index. The ranks that coincides with y=x line represents the ranks of the scientists that has not been changed. Pearson's correlation between the h-ranks and the p-ranks has a value of 0.489836, signifying that the two ranks are loosely correlated.

Another interesting characteristic of p-index compared to h-index can be observed in Figure 5.9. The p-index of a scientist can increase, decrease or stay at the same level depending on their activities as opposed to one's h-index, which can only grow or stay at the same level. The variation of the p-index and h-index of author A. Iqbal, who has the highest h-index as well as the p-index at the time of obtaining data clearly demonstrates this. He enters in to the premature field with only 21 papers at which point he is able to get a p-index of 96.3% because there is only a few authors for these papers and the pagerank algorithm will treat them on equal grounds. However, when the field gets more papers and more authors and A. Iqbal does not publish until the field reaches 70 papers, his p-index goes down to 86.6%, showing that one's p-index can go down. He publishes the 71st paper and then gets citations from 73rd paper and papers from that point onwards, which in turn increases his p-index 95.1% and subsequently to 99.8%, which he manages to keep at the same level by publishing papers as well as accumulating more citations for the already published papers. His h-index, however, follows a steady growth even when he is inactive, which can create a stagnating behavior after achieving a high h-index.

Finally, the overall distribution of h-index and p-index of authors are examined in Figure 5.10. As it can be clearly seen, the p-index is more evenly spread out through the graph while h-index is not. The right end of the said figure is magnified and shown in Figure 5.11. It is evident that relatively new authors are unable to get a higher h-index, whereas when p-index was used, even relatively new authors can obtain a significant p-index. This rewards the scientists for their work and not just the time they have spent in the said scientific field. An author like M. Chen, who enters in to the field late as the 691st author, has a minimal chances of getting a high h-index. In fact, his h-index is 2 while he has a p-index of 90.5%, which is significant. He joins the citation network at the 608th paper and gains a p-index of 63.1% subsequently with the citations the paper receives.

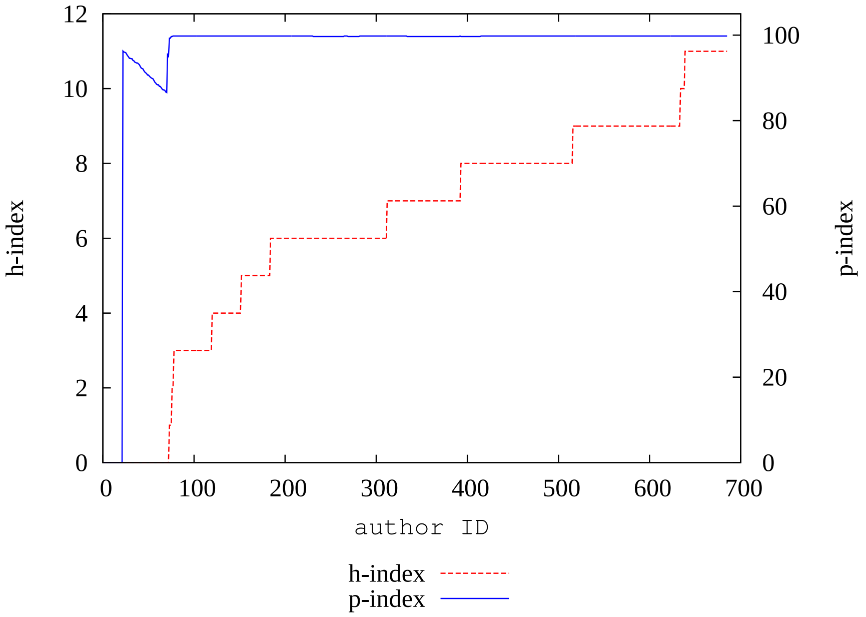


Figure 5.9: The evolution of h-index and p-index for the highest ranking author: A. Iqbal in Quantum Game Theory Google Scholar profile.

Although he co-authored 5 papers and has a total of 9 citations, he only has a h-index of 2 due to some of his papers not getting more than two citations each. In contrast, the p-index rewards him for the citations he received and shows that even a scientist who is new to a field can get a higher p-index depending on their contributions to the field and the perceived importance of their work.

5.3.2 Comparison of Author Behavior and Traits

In this subsection, the behavioral patterns of the authors of Quantum Game Theory Google Scholar profile are discussed. In general, there are communities of authors in the collaboration network as shown in Figure 5.12. There are number of collaborations of authors from no collaboration up to 27 collaborations in the network along with a number of communities. The first community considered is a small fully connected community, as seen in Figure 5.13. Table 5.4 lists the details of those authors and, as can be seen, all of them have only published one paper, which has received no citations. However, their p-index differs from one another since it uses a weighted system depending on the author position to distribute the pagerank value of that specific paper.

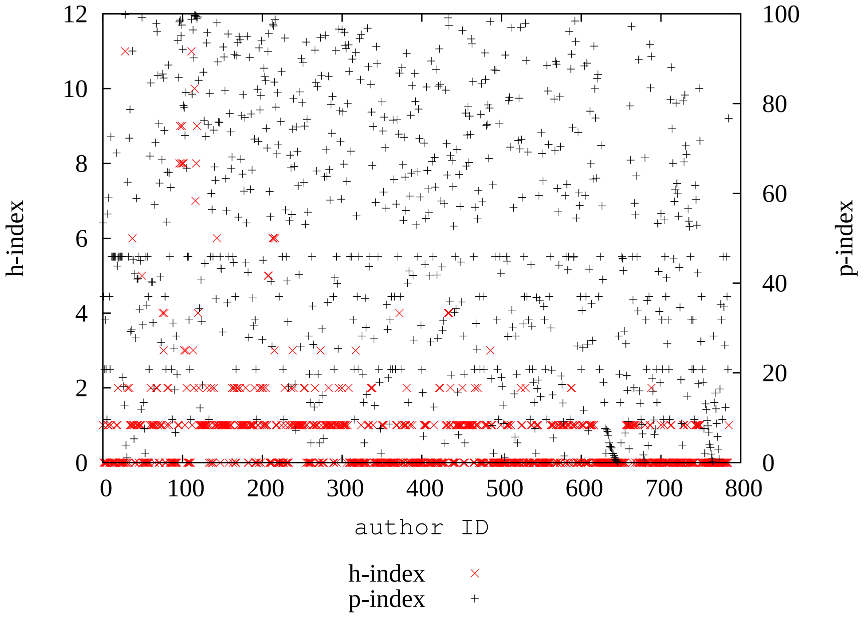


Figure 5.10: The spread of h-index and p-index of authors in Quantum Game Theory Google Scholar profile.

Author	h-index	p-index	Number of papers	Number of citations
GAD Briggs	0	1	1	0
OJE Maroney	0	9.4	1	0
R Hanson	0	0.3	1	0
ML Markham	0	4.1	1	0
LM Robledo	0	11.8	1	0
JJL Morton	0	1.9	1	0
H Bernien	0	6.8	1	0
MS Blok	0	8.2	1	0
RE George	0	13.1	1	0
DJ Twitchen	0	3.4	1	0

Table 5.4: Details of the fully connected community as demonstrated in Figure 5.13.

The next community examined is spatially more spread out than the last one. It is not a fully connected community of authors, although there are several key authors with a high betweenness centrality. This community is centric around the author J. Du, while G. Qin, Y Li, X Zhou and S. Massar act as pillars of the community by connecting isolated sub-communities to the main community. As it can be seen from their respective h-index and p-index, having a high betweenness centrality certainly helps, but that itself is not enough to score a high p-index.

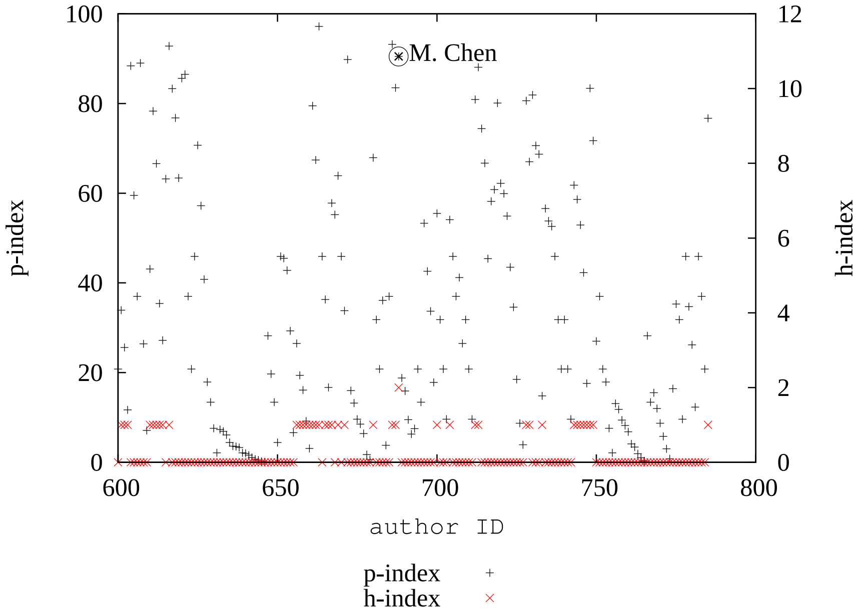


Figure 5.11: The spread of h-index and p-index of authors (ID from 600-785) in Quantum Game Theory Google Scholar profile.

Author	h-index	p-index	Number of papers	Number of citations	Betweenness centrality
J Du	9	99.2	15	343	0.000752
Y Li	1	81.6	3	3	0.000469
S Massar	2	72.3	2	41	0.000292
G Qin	2	86.2	3	9	0.000110
X Zhou	8	98.6	9	280	0.000077

Table 5.5: Ranking metrics of the community as demonstrated in Figure 5.14.

The final community looked at this subsection is the largest collection of connected communities in the collaboration network as seen in Figure 5.15. The highlighted authors in this section have a higher h-index as well as a higher p-index. They also have a high betweenness centrality, indicating that they are the authors who bridge several communities together. In fact, the author A. Iqbal has the highest betweenness centrality in the whole collaboration network. In summary, the tendency to collaborate with new authors and act as a bridge between communities can actually be beneficial in terms of p-index as well as h-index. While it is not possible to conclusively suggest that authors with a high betweenness centrality will get a high p-index, the evidence suggests that the trait of collaborating with scientists who are outside one's normal group is in fact a good choice as opposed to

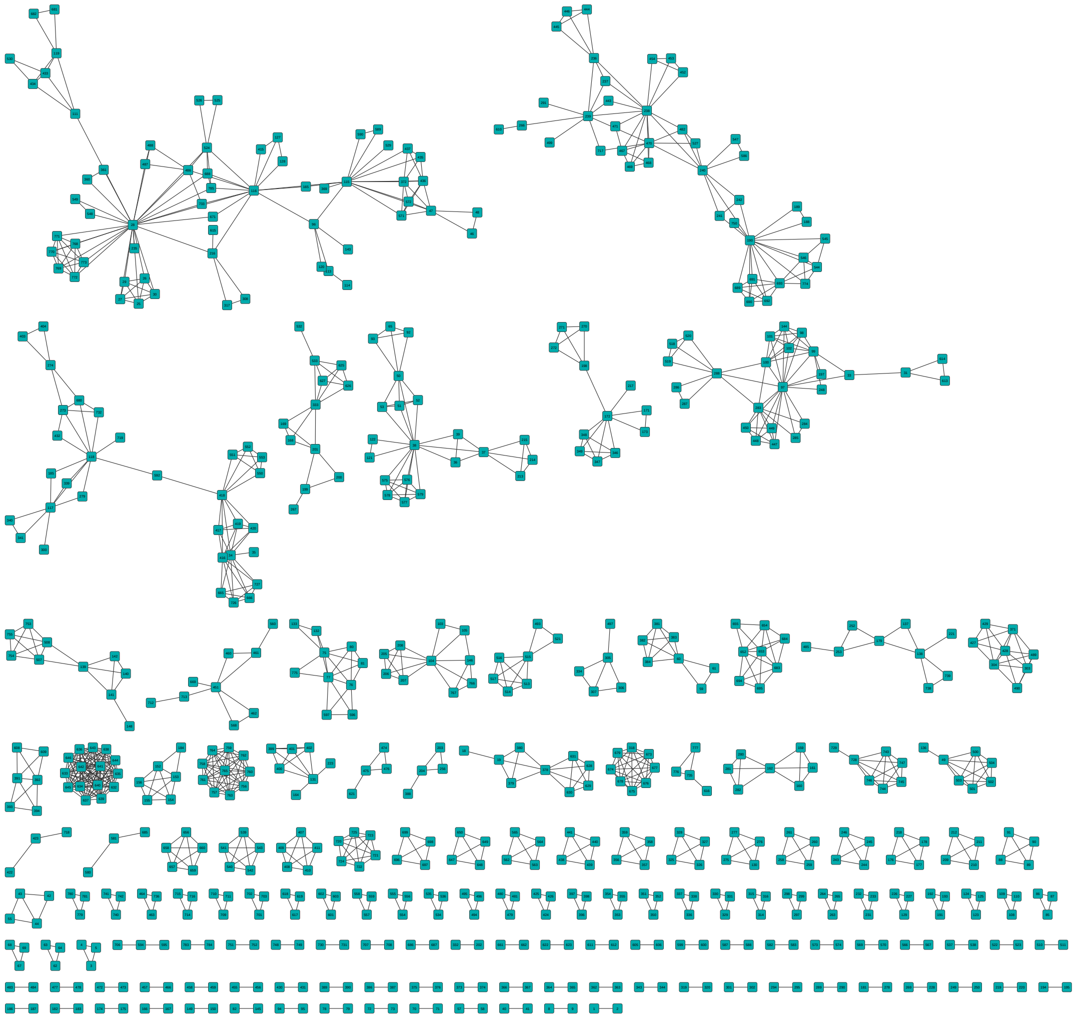


Figure 5.12: The collaboration network generated from the Quantum Game Theory Google Scholar profile.

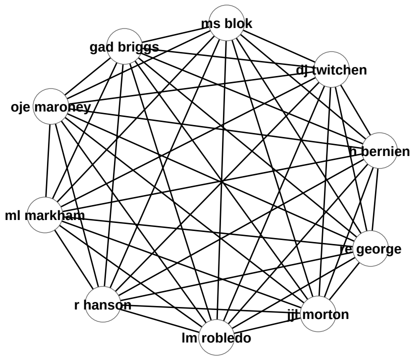


Figure 5.13: A fully connected community in the collaboration network generated from the Quantum Game Theory Google Scholar profile.

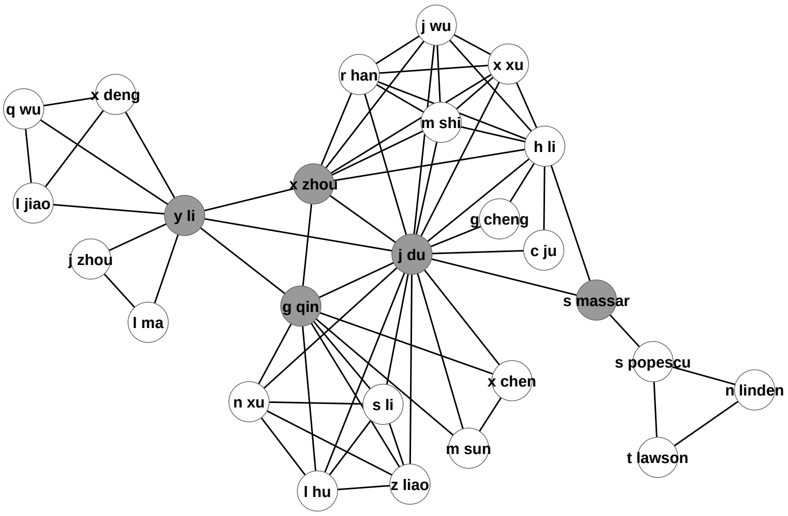


Figure 5.14: A collection of communities in the collaboration network generated from the Quantum Game Theory Google Scholar profile.

collaborating with a static group all the time as seen in the Figures 5.13, 5.14, 5.15 and Tables 5.4, 5.5 and 5.6.

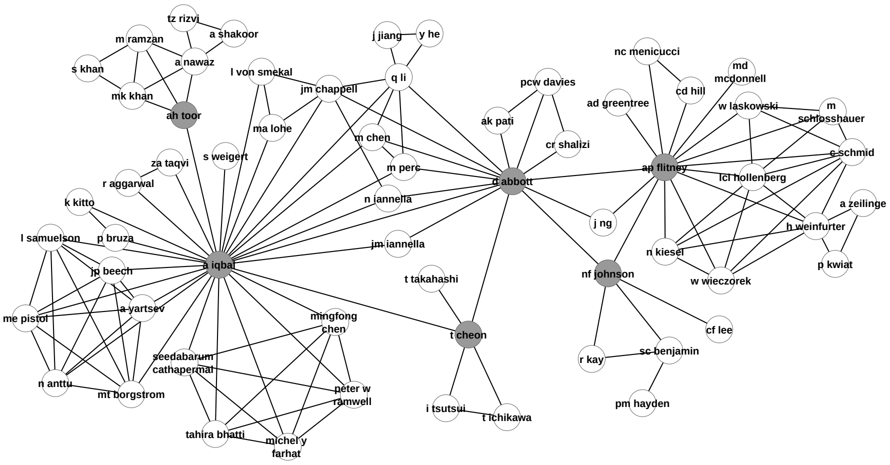


Figure 5.15: A collection of communities in the collaboration network generated from the Quantum Game Theory Google Scholar profile.

Author	h-index	p-index	Number of papers	Number of citations	Betweenness centrality
AH Toor	11	99.7	20	330	0.00146
A Iqbal	11	99.7	49	440	0.00505
D Abbott	7	98.4	37	313	0.00424
AP Flitney	10	99.6	26	332	0.00296
NF Johnson	8	98.6	13	235	0.00102
T Cheon	6	97.8	16	102	0.00078

Table 5.6: Ranking metrics of the community as demonstrated in Figure 5.15.

5.4 Conclusion

In this chapter, we use the proposed scientist-ranking system, the p-index, to quantify and compare the contributions of scientists to the field of quantum game theory (as defined by Google Scholar) as a case study. This particular field was chosen to demonstrate the utility of p-index as this is a relatively new and emerging field and has a dedicated Google Scholar page. As such, the evolution of the field could be tracked and the h-index and p-index profile of all authors could be temporally compared. It was shown that the p-index is a ‘fairer’ ranking system to measure scientific output than the h-index. It reduces the misleading effect of authors collaborating in large groups getting a higher number of citations and it ensures that the quality of the source of citation is taken into account as well as the number of citations. To achieve this, the p-index uses the pagerank algorithm, which is used in the Google search engine to rank web search results. The results provided in this chapter show that authors who are instrumental to the development of a particular field are rewarded more by the p-index.

Even though a particular scientific field was chosen to demonstrate the utility of p-index, it can be used in any field, just as the h-index is presently being used. However, it cannot be calculated manually and needs to be implemented within a database such as Google Scholar, to be properly used. This is one drawback of the index. However, it compensates for this drawback by being more inherently fair in comparison to the h-index.

It is natural, therefore, to compare the p-index with the author ranking systems implemented in databases such as ResearchGate. A complete comparison however is not possible because ‘ResearchGate’ and similar scientific author platforms do not completely reveal how they compute the impact of each author. However, indications are that the time an author spends within these platforms, measured by impact factors such as the number of posts, shares and downloads, influence their standing within these platforms. In this sense, these varieties could be even more misleading and subjective than the h-index, which does not require authors to spend time within a particular platform as long as they publish well. The p-index on the other hand, although it needs to be implemented within an online platform, only measures authors by their scientific output and gives weights to citations without resorting to using the impact factors. Furthermore, the output is a percentile, which means comparison is direct. It is also contextual in that the ranking of a scientists can be considered starting from their own field up to the whole academia.

Chapter 6

Case Study II: HEP-TH Dataset

6.1 Introduction

Comparison is an integral part of everyday life. Sometimes we rank entirely based on a subjective measure like comparing the taste of two brands of chocolate while other times we use an objective measure like comparing the cost of two similar items. Striking a balance between which measure or set of measures to use can be a daunting task, especially if one's reputation is at stake. In ranking scientists, there is no right answer as there is no absolute objective measure that can solely represent the brilliance of a scientist. That is why it is necessary to develop a fairer and robust metric to rank scientists and we ought to make this metric as objective as we can. As mentioned in Chapter 2, a scientist is better evaluated using a range of measures rather than a single numerical measure. The comparative measure most commonly used is the h-index. However, this research has shown that the h-index can actually be tampered with and argue against using it as a single magic number used to evaluate scientists. Instead, this research presents the p-index, which is fairer and more robust and resilient against manipulations compared to h-index. A simulated system is used to demonstrate the possible manipulations on h-index and how those manipulations can massage one's h-index. We also present evidence to indicate that p-index is remarkably resilient against these manipulations along with an explanation on the mechanisms used to circumvent these manipulations. In addition, the p-index was also applied to a real world citation network curated from the scratch that shows the evolution of the scientific field Quantum Game Theory. This dataset is of particular importance because it allows us to observe the characteristics of the h-index and p-index over an evolving citation network. The Quantum Game Theory network was relatively small (685 papers with 785 author) which made us apply p-index to the high energy physics - theory dataset (29,555 papers and 15332 authors) as described in this chapter.

The rest of this chapter is organized as follows. The second section introduces and describes the dataset used in this chapter while the third section presents a detailed analysis of the application of p-index and h-index on this dataset. After a general overview, the results obtained are further analyzed in the two subsections of the third section and the last section concludes the chapter.

6.2 Citation Network of Scientists from HEP-TH Dataset

In this chapter, the p-index is applied to a larger real world citation network and the results compared with that of h-index. It is necessary to apply the p-index to a larger network in order to examine the characteristics of the p-index when it is deployed in a citation management system like Google Scholar.

Although the Quantum Game Theory citation network enables us to observe the evolution of a citation network and how the p-index and h-index can evolve with the citation network, it is a relatively small network with 685 publications and 785 authors. Hence a bigger citation network was needed so that the p-index can be applied and the results compared with the h-index. This also demonstrated that the p-index can be applied to a citation network at any point in time and not just to an evolving network. As such, the second citation network examined is the HEP-TH dataset as provided by Leskovec et al. [38]. It is a static snapshot of the HEP-TH citation network as at April, 2003 featuring 29,555 publications over the duration of January, 1993 to April, 2003. There are 15,332 authors altogether in this dataset and it has been curated from Arxiv as a part of the KDD cup 2003 [54].

The dataset was used to construct the network using a Java program that used the provided text data to map nodes and edges of the citation network. The provided meta data was then used in order to build the collaboration network on top of the citation network because both the citation network and collaboration network are needed for this analysis. The IDs

Property	Value
Number of nodes	27770
Number of edges	352285
Average degree	25.372
Characteristic path length	6.460
Clustering coefficient	0.156
Network diameter	33
Power law fit	$P(k) \sim k^{-1.585}$

Table 6.1: Characteristics of the citation network generated from the HEP-TH dataset.

Property	Value
Number of nodes	13524
Number of edges	24883
Network heterogeneity	1.119
Characteristic path length	7.481
Clustering coefficient	0.489
Network diameter	26
Network Density	0.009

Table 6.2: Characteristics of the collaboration network generated from the HEP-TH dataset.

of the papers and authors do not correspond to the time of their entry to the dataset and only internal citations were considered. The Network level characteristics of this dataset are listed in Tables 6.1, 6.2 and the degree distribution is presented in Figure 6.1.

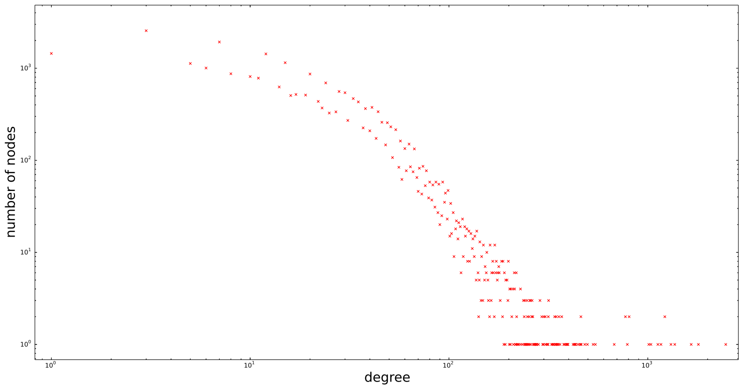


Figure 6.1: The degree distribution of the HEP-TH citation network.

As it can be seen from Table 6.3, the citation network data was readily available. It was exported as a tab separated edge list, which was directly used to establish the citation network. However, constructing the corresponding collaboration network needed significant preprocessing. The details of the authors were included as a part of the text file with the ID of the corresponding paper. Each of these text files had numerous other fields such as "Paper:", "From:", "Title". along with the abstract as well. These fields weren't uniform either because some files did not have entities such as "Date", some did not have "Comments" and some also had additional fields. On top of that, the names of the similar fields did not adhere to a particular standard. For instance, the field "Authors:" had multiple syntax, like "authors:", "Author:", "author:" etc. Sometimes the authors were separated by a comma (",") and other times, it was by a conjunction phrase such as "and". Sometimes both a comma and a conjunction were used. Hence these text files needed to be studied to understand a discernible pattern so that coherent information can be extracted from them.

First, multiple randomly picked text files were processed using Java programming language and the keywords to extract were filtered. After observing the processed text files, the ID was matched with the paper ID in the constructed citation network. Then a search was done for a field that had "author" prefix followed by a colon (":") to identify the author field. This search is presented as a regular expression in Equation 6.1. Three delimiters were used to separate each author from the author's field as described in algorithm 1.

$$REGEX = author* : \tag{6.1}$$

Algorithm 1 Separate Authors

```
procedure ObtainAuthorArray
    $author$ ← content of the author field
    if $author$ has "," then use "," as a delimiter to return array of authors
    else $author$ has " and " use " and " as a delimiter to return array of authors
    $authorList$ ← array of authors
    if elements in authorList has " and " then use " and " as a delimiter to separate
that element and add it to authorList as a new element
```

6.3 Analysis and Results

Following the analysis of Quantum Game Theory dataset, some qualitative observations are presented to illustrate the utility of p-index and how it addresses the issues associated

Section	Description
Citation network (papernet)	The citation network is available as a two column tab separated text file. The first column represents the ID of the citing paper and the second column represents the ID of the cited paper. An example of this approach is given below. 9401139 9201015 9408099 9204102 9304045 9204040
Title and authors of each paper	This information is stored in multiple text files and each text file is named according to the ID of the paper. It provides the title of the paper, authors and an abstract. An overview of a text file is given below. The content of the file ID: 9304094 Paper: hep-th/9304094 From: NARDELLI@itnvax.cineca.it Date: 21 Apr 1993 12:50:04 +0000 (12kb) Title: Canonical Analysis of Poincare' Gauge Theories for Two Dimensional Gravity Authors: G. Grignani and G. Nardelli Comments: 13 pages, plain TEX Report-no: DFUPG-76-1993/UTF-292-1993 Journal-ref: Class. Quant. Grav. 10 (1993) 2569-2580 Abstract...

Table 6.3: Details of the HEP-TH citation network curative process.

with the h-index. Table 6.4 presents the top twelve scientists of the HEP-TH dataset as ranked by h-index. Table 6.5 presents the the top twelve scientists of the HEP-TH dataset as ranked by p-index. Some authors, like A.A Tseytlin, manage to keep their position intact (p-rank= 3, h-rank= 3). Some authors who are below the top twelve in Table 6.4 come to prominence when ranked using p-index. For example, Shin'ichi Nojiri was below the top twelve if ranked using the h-index but managed to come in to the fourth position when ranked using the p-index. Likewise, E. Elizalde is ranked 229th using h-index but leaps to the 6th position when the p-index is used. These authors have published in high impact journals like *Science*, which the p-index takes in to consideration but the h-index does not. Other authors like, C.N Pope and P.K Townsend, are ranked highly when the h-index is used (6 and 10 respectively) but lose their ranking significantly when the p-index is used (25 and 47 respectively). This occurs due to the fact that both of these authors have high average authors per paper values (3.48 and 2.38 respectively) and they have been second/third authors more than they have been first authors, which the p-index

Author	h-index	h-percentile	p-index	Number of papers	Average authors per paper	rank-h	rank-p
Edward Witten	54	99.9	99.9869	94	1.5744	1	2
Ashoke Sen	47	99.9	99.9934	89	1.5505	2	1
A.A. Tseytlin	41	99.9	99.9804	109	1.7155	3	3
Cumrun Vafa	40	99.9	99.8760	77	2.0389	4	19
Michael R. Douglas	35	99.9	99.8695	52	2	5	20
H. Lu	32	99.9	99.9543	123	3.5284	6	7
C.N. Pope	32	99.9	99.8369	128	3.4843	6	25
Nathan Seiberg	32	99.9	99.5238	42	2.1666	6	73
Andrew Strominger	31	99.9	99.7586	56	2.1607	9	37
Joseph Polchinski	29	99.9	99.7912	45	1.8222	10	32
P.K. Townsend	29	99.9	99.6934	55	2.3818	10	47
Gregory Moore	29	99.9	99.6282	53	2.3396	10	57

Table 6.4: The top 12 scientists as ordered by their h-index in the HEP-TH dataset.

takes into consideration but the h-index does not.

Looking at an author like E. Bergshoeff, the effect of the p-index has is clear including how it differentiates self-citations and the high average number of authors per paper. This specific author cites his own papers 86% of the time, which is a very high percentage and also has a value of 3.34 as the average number of authors per paper, which is relatively high for this dataset. Although he has a median value for average number of citations per paper of 34.89, that can only inflate his h-index, giving him a h-rank of 29 while his p-rank is 70, since p-index takes all of the above factors in to consideration rather than counting all the citations.

Figure 6.2 presents the plot of p-index vs h-index. In order to compare the p-index directly, we convert the h-index values to a percentile and use the h-percentile in the plot as the

Author	h-index	h-percentile	p-index	Number of papers	Average authors per paper	rank-h	rank-p
Ashoke Sen	47	99.9	99.9934	89	1.5505	2	1
Edward Witten	54	99.9	99.9869	94	1.5744	1	2
A.A. Tseytlin	41	99.9	99.9804	109	1.7155	3	3
Shin'ichi Nojiri	20	99.7	99.9739	94	2.4574	29	4
Nathan Berkovits	23	99.8	99.9673	59	1.3898	20	5
E. Elizalde	11	98.1	99.9608	77	2.5064	229	6
H. Lu	32	99.9	99.9543	123	3.5284	6	7
Zurab Kakushadze	21	99.8	99.9478	63	1.7777	25	8
Donam Youm	18	99.5	99.9412	55	1.2909	53	9
Sergei V. Ketov	10	97.5	99.9347	46	1.2173	291	10
Itzhak Bars	18	99.5	99.9217	44	1.25	53	11
Valeri V. Dvoeglazov	7	94.8	99.9217	41	1.0487	606	11

Table 6.5: The top 12 scientists as ordered by their p-index in the HEP-TH dataset.

p-index is a percentile and not a direct score. The top 95% of the h-index is shown for the purpose of clarity and as can be seen, the corresponding p-index has a range from 65%-100%, implying that some scientists who are in the top 5% of the h-index may not even make it to the top 25% if ranked using the p-index. An author who scores 99% using the h-index may score as high as 99.9% and as low as 91.8% if p-index is used. These variations of the authors can be individually explained as before using various factors like their percentage of self-citation, their number of collaborations, the impact of their publications and the centrality they have. If they categorized, the authors who tend to co-author with a small pool of authors, publish in high impact platforms and have a high centrality may score higher when ranked using p-index and conversely, authors who tend to significantly self-cite, co-author papers with a large pool of authors and publish in low impact forums may score lower when ranked using p-index. The top right end of the same

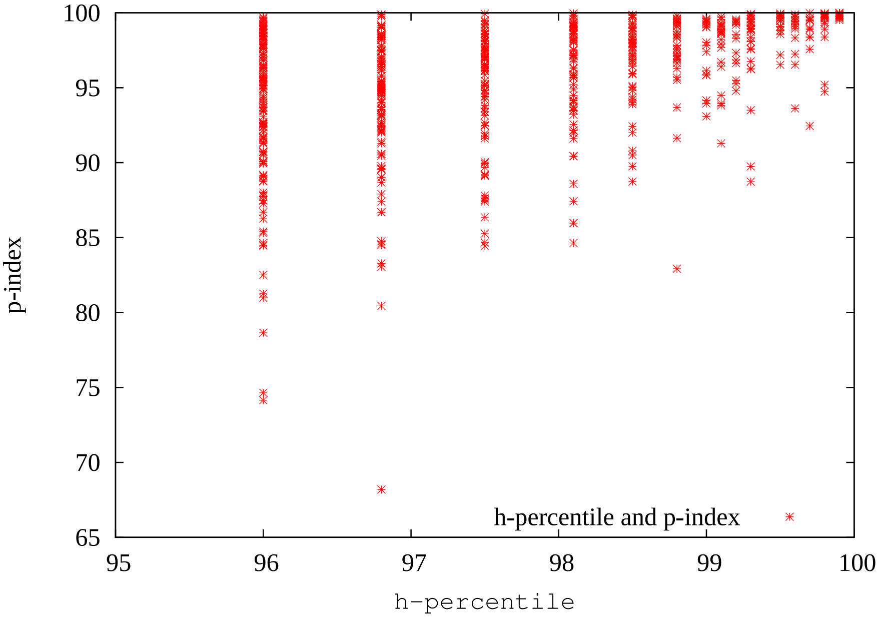


Figure 6.2: The h-index and p-index of the best 5% authors in the HEP-TH dataset. Since the p-index is a percentile, percentile values were used for the h-index as well, rather than actual h-index values. Note here that the p-index value varies from 65% to 100%. That is, some authors who are among the top 5% in terms of h-index are not even among the top 25% when p-index is considered.

figure represents the scientists who are ranked at the top in both h-index and p-index. A few examples are Ashok Sen, Edward Witten and A.A. Tseytlin, who are all prominent and well respected scientists in the field of theoretical high energy physics.

6.3.1 Comparison of h-index and p-index

Figure 6.3 shows the ranks of authors as ordered by p-index and h-index plotted along with the $Y + X$ line. Some scientists have managed to keep their ranks at the same level while the ranks of some scientists exhibit drastic variations. The $Y = X$ line coincides with the authors who managed to keep their ranks at the same level. Pearson's correlation between the h-ranks and the p-ranks has a value of 0.477849, which implies that the two ranks are loosely correlated. This is consistent with the findings from the last chapter where the said two ranks were loosely correlated and had a similar Pearson's correlation (0.489836).

In order to get a general overview of h-index and p-index, the spread of both were plotted

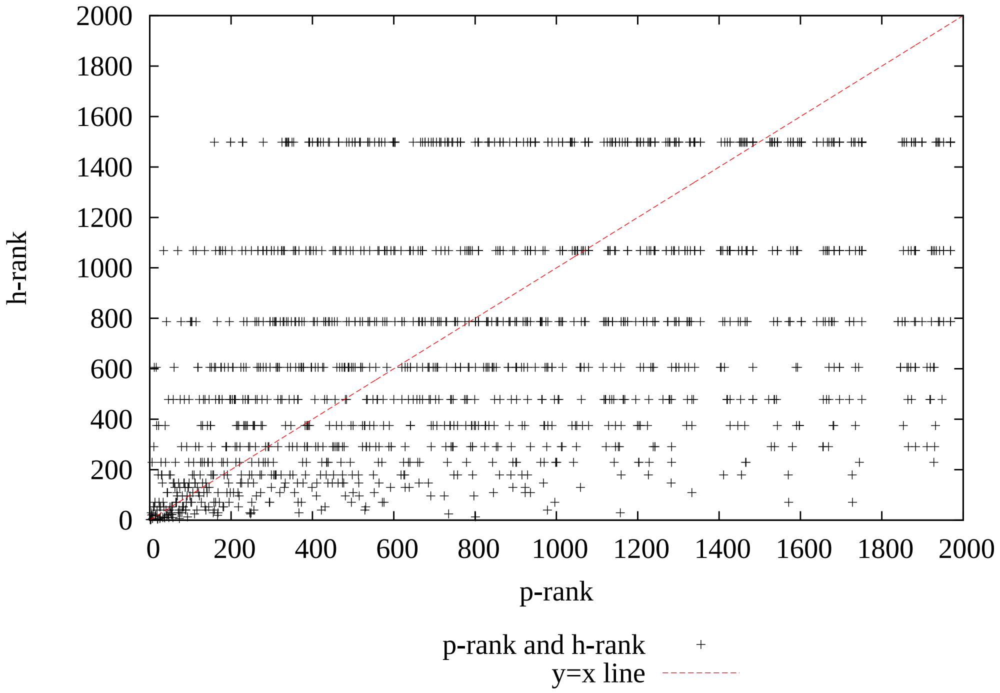


Figure 6.3: The ranks of the authors as ordered by h-index and p-index. Y=X line is also included to better visualize the authors whose ranks have not been changed. The plot only shows the ranks from 1-2000 for the purpose of clarity.

against the author ID as shown in Figure 6.4. As was the case in the Quantum Game Theory profile dataset, it is clear that the h-index has a limited spread while the p-index tend to spread throughout the plot, indicating a better coverage. The first 200 author IDs were also zoomed and highlighted R.S Ward who entered to the HEP-TH citation network in January 2000, which means he is relatively new to the author pool. He has five publications in total with 12 citations. Although he has a h-index of 2, he manages to get a p-index of 90.87% which is significant. Because R.S Ward began publishing quite recently, it is difficult for him to obtain a high h-index over the span of two years whereas his effort is rewarded by p-index regardless of his time spent in the field. This example illustrates that even authors who began publishing recently can obtain a high p-index if they are excellent in their particular field.

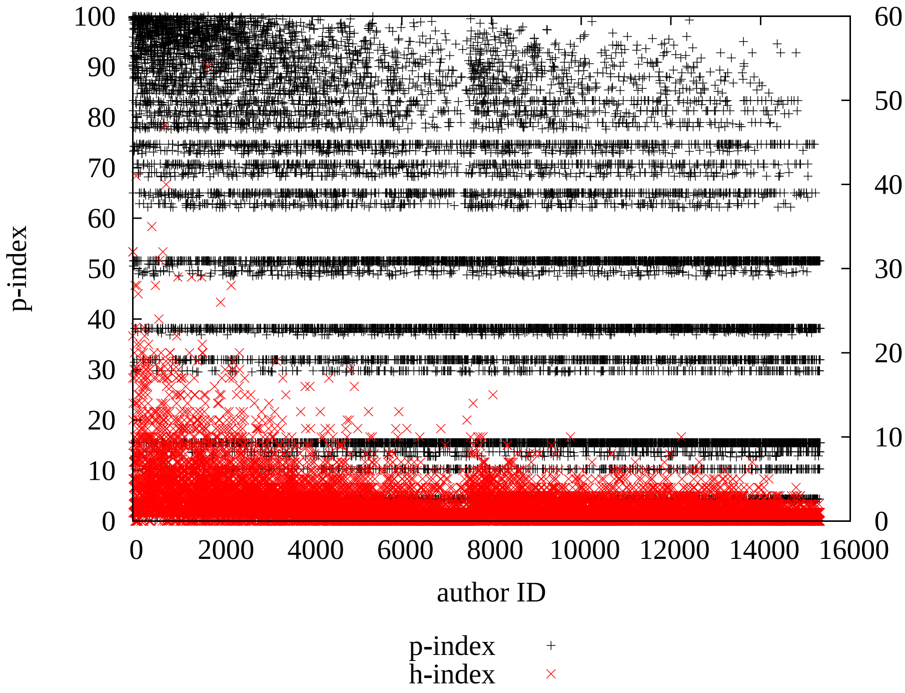


Figure 6.4: The spread of h-index and p-index of authors in HEP-TH dataset. Author IDs does not have a causal identity.

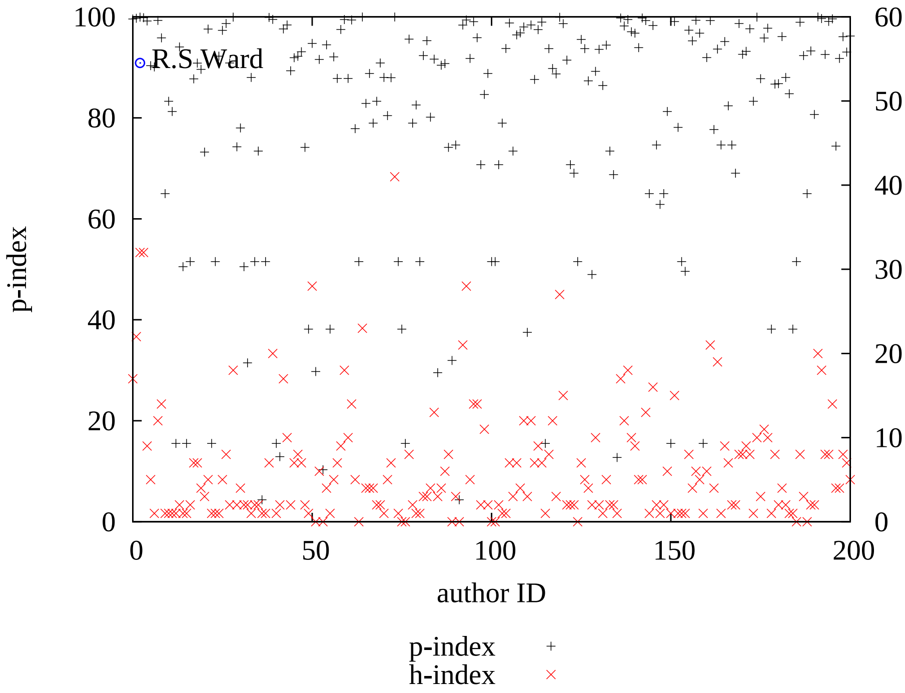


Figure 6.5: The spread of h-index and p-index of authors in HEP-TH dataset (author ID from 0 - 200). Author IDs does not have a causal identity.

Author	h-index	p-index	Number of papers	Number of citations
Y Abe	0	15.50	1	0
Y Higashide	0	10.30	1	0
M Matsunaga	0	13.04	1	0
N Haba	0	12.89	1	0
K Kobayashi	0	3.82	1	0

Table 6.6: Details of the fully connected community as demonstrated in Figure 6.6.

6.3.2 Comparison of Author Behavior and Traits

The HEP-TH dataset was examined to identify discernible patterns of behavior of authors. Since it is a large collaboration network, it is difficult to focus individual attention on a few authors. Instead, the details about a fully connected community and a more spread out community that are part of the collaboration network are presented. These sub communities provide evidence of author traits and the effects those traits will have on their p-index. The first community is presented in Figure 6.6 and their details are listed in Table 6.6. As can be seen, all of the authors have only one published paper and they all collaborated in that paper. This specific paper received no citations, which is partly responsible for their low p-index. Additionally, each author has different p-indices depending on their position in the author list, where K. Kobayashi was listed as the final author and Y. Abe was listed as the first author.

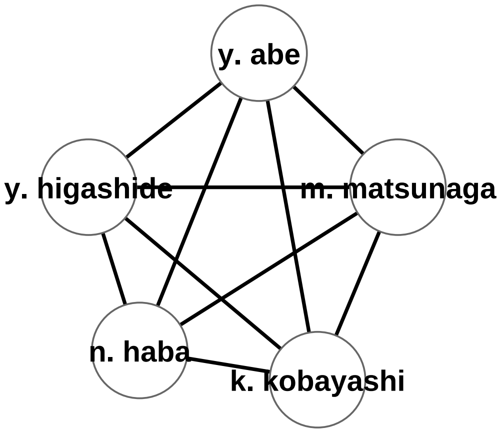


Figure 6.6: A fully connected community of authors in HEP-TH dataset.

The next community analyzed is more spatially spread out and diverse, as shown in Figure 6.7. The corresponding details of highlighted nodes (authors) are listed in Table 6.7

and demonstrates that p-index not only takes the citation and paper count in to consideration but also various other factors like where the paper was published and from which papers the citations are coming, as well as how many authors co-authored the paper, how much self-citing an author has done and the behavioral patterns of authors. While it is possible to do a case-by-case analysis of why the p-index of each scientist has its value but instead, a selected sample of authors from this community were examined and their collective traits are analyzed.

Author Klaus Frednhagen has a h-index of 2 but he manages to score a p-index of 65%, which is a better score than his h-index. He has co-authored four papers and received 13 citations. In addition, his position in the community indicates a relatively high centrality compared to a leaf node, which can be inferred from his betweenness centrality. All these factors affect his p-index and are reflected by it. An author like Detlev Buchholz has a high betweenness centrality and acts as a hub in this community among his other achievements like publishing high quality papers and getting referenced by high impact papers. This is reflected by his p-index.

Sergio Doplicher also has a similar number of citations as Klaus Frednhagen. However, he manages to score a higher p-index than Klaus because of his high betweenness centrality serving as a bridge connecting a hub and another clique in the community. A leaf node like Katsunori Kawamura receives a low p-index due to his inactivity (one publication and no citations) as well as his position in the community, with only one connection to Mitsuo Abe. Reiterating what was suggested in the previous chapter, this evidence suggests a tendency for the p-index to take the collaboration network into consideration implicitly which would mean author behaviors and traits are taken into consideration. What makes p-index so versatile is that these are not taken into consideration explicitly; rather, the interplay between the citation network and the collaboration network ensures that these traits are taken in to consideration when p-index is calculated. This suggests that it is indeed fairer, more robust and resilient against manipulative behavior.

6.4 Conclusion

The p-index was applied to a large real world citation network that was constructed using the high energy physics theory section from Arxiv containing publications from 1992 -

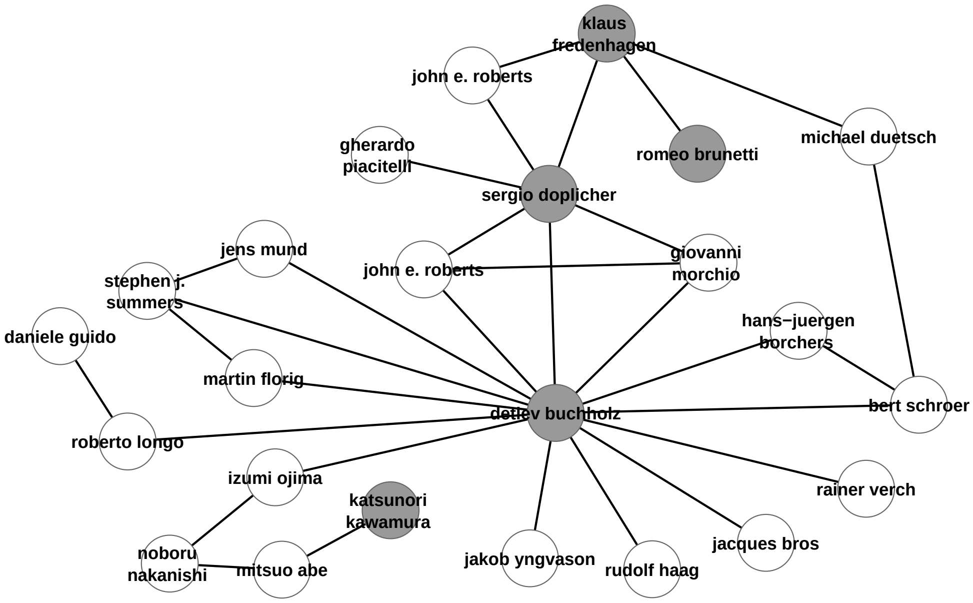


Figure 6.7: A spread out community of authors in HEP-TH dataset.

Author	h-index	p-index	Number of papers	Number of citations	Betweenness centrality
Klaus Fredenhagen	2	65.00	4	13	0.00002
Sergio Doplicher	2	80.36	4	20	0.00042
Romeo Brunetti	1	38.16	1	1	0.0
Detlev Buchholz	7	98.53	22	116	0.00239
Katsunori Kawamura	0	15.50	1	0	0.0

Table 6.7: Details of the community as demonstrated in Figure 6.7.

2003. Although we only have access to the snapshot of the network as at 2003, p-index can be applied to this dataset in the hope of observing the differences between p-index and h-index in citation network that is closer to reality. The ideal entities on which to apply the p-index as a ranking metric would be Google Scholar or Web of Science and these entities typically have 20,000 to 500,000 publications in their subsections. As such, applying the p-index to a citation network with 29,555 publications can be a sufficient justification in terms of differences it would make compared to the h-index. The results presented suggests that the p-index is indeed a fairer metric compared to than h-index. The p-index is also robust and takes a lot of factors in to consideration as opposed to just the number of citations and number of papers.

As noted previously, although this research has applied the p-index to a specific scientific field (high energy particle physics - theory), it can be easily applied to any number of scientific fields separately or to the academic community as a whole. This make it extremely versatile and context-aware because each scientific field may be able to derive their own p-index for their scientists, which is more important than the scientists' standing out of the whole academic community amongst a plethora of disjointed fields.

Chapter 7

Ranking Stability of the Authors

7.1 Introduction

The intention of this thesis was to develop a ranking system that is fairer and more resilient against manipulations than the h-index. This was done by utilizing the underlying dynamics of the complex network in order to evolve a ranking metric that is aware of the perceived importance of each node in a network. This perceived importance was used to rank the nodes in the citation network and therein derive a measure that can be used to rank scientists. So far, evidence has been provided that this new index, the p-inde, is fairer and more resilient against manipulations. In this chapter, the preliminarily assumption made about the p-index that p-index takes the perceived importance of the citing paper in to consideration is reexamined. If this is the case, when the citing paper is changed without changing the degree of the cited paper, there should a difference in the p-index. A methodology to do this is systematically designed and evidence provided to support this claim in this chapter.

The rest of this chapter is organized as follows. The second section presents the methodology by which the citation network is perturbed while the third section presents the results and the analysis. The last section summarizes and concludes the chapter.

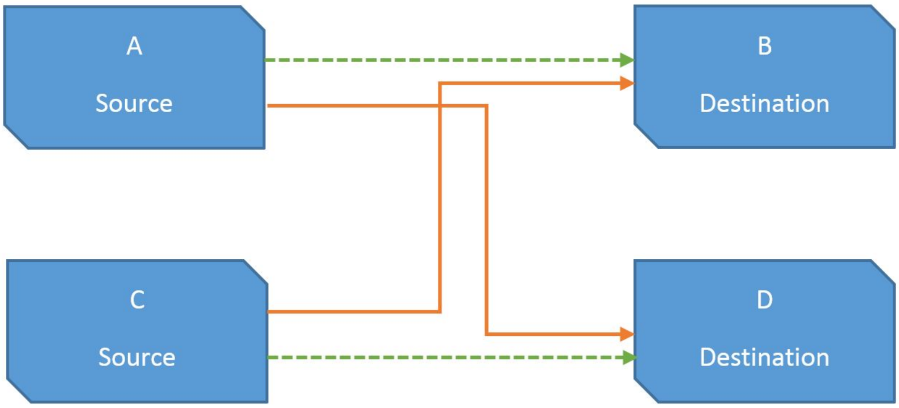


Figure 7.1: Degree Preserving Perturbation: Green dotted arrows represent the source-destination relationship before rewiring and orange continuous arrows represent the source-destination relationship after rewiring.

7.2 Methodology

The HEP-TH citation network is perturbed while preserving the degree distribution in order to understand how that would affect the ranks of the authors. The proportion of the perturbations are in the range of 0.1% to 1.0% of the size of the network (29,555). We then plot the author ID against the different p-indices that we calculate for different proportions of perturbations. The assortativity of the said perturbed networks is presented in order to ascertain that the mixing patterns of the network have been changed. It is hypothesized that the ranks of the authors should not be stable as mixing patterns play a significant role in calculating the p-index. Changing the assortativity would essentially mean that each node in the citation network would have the same degree as before, but who cites them is going to change. As p-index takes this information in to consideration when it analyses the edges of the citation network, the perceived weights attached to each node will change and in turn their page rank value will change. Because their page rank value is changed, the p-index of the authors will also have to change. Essentially, the perturbations done to the citation network will propagate. On the contrary, the h-index of the authors will not change with degree preserving perturbations because the h-index would only take the degree to consideration.

Figure 7.1 shows the technique employed to perturb the citation network while preserving the degree distribution. The same process is repeated over a randomly selected quadruple of nodes until the perturbation proportion is met. For instance, this process is repeated

Proportion	Assortativity
0.0	-0.030312
0.001	-0.030318
0.002	-0.030325
0.003	-0.030330
0.004	-0.030331
0.005	-0.030335
0.006	-0.030354
0.007	-0.030359
0.008	-0.030363
0.009	-0.030375
0.01	-0.030380

Table 7.1: The assortativity of the perturbed citation networks.

for 29 times when the perturbation proportion is 0.1% ($29555 * 0.1\% \sim 29$). After this has been carried out, we have the original network and ten perturbed networks with the proportions including (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0%).

7.3 Results and Analysis

The assortativity of the perturbed networks are listed which provides evidence to the changed mixing patterns of the network. The initial assortativity of the network is slightly negative indicating a disassortative mixing pattern as seen in Table 7.1. However, it is more inclined towards a neutral pattern as it is closer to zero. When the citation network is perturbed to an increasing proportion, the disassortative nature increases as indicated by the increase in assortativity.

When the variation of the p-index is examined against the proportion of perturbation as shown in Figure 7.2, we can clearly see that the ranking becomes unstable, implying that the p-index in fact takes the mixing patterns into consideration, as opposed to the h-index. It clearly shows that one's p-index comes down when the perceived importance of a citation changes. The variation of p-index for authors who are ranked at 8000-8500 is examined in Figure 7.3. It can be seen that an author who initially had a p-index of 38% can obtain a p-index as much as 69% when the network is perturbed. These two plots suggest that p-index takes the mixing patterns into consideration, unlike the h-index.

The variation of the p-index of the ten highest ranking authors of the network is examined

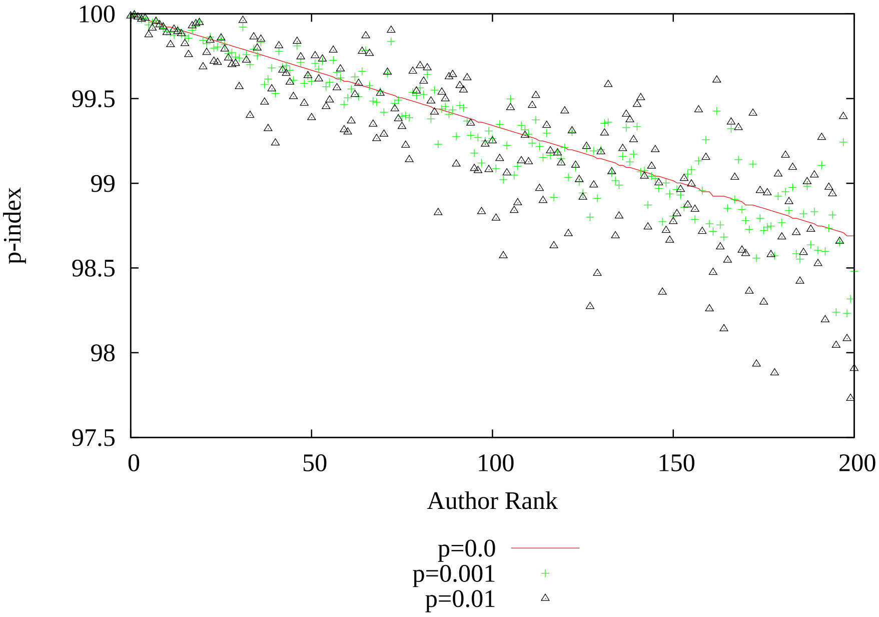


Figure 7.2: The variation of p-index of the authors in HEP-TH dataset against author rank. The authors are ranked using the original p-index (that of the unperturbed citation network) and only the first 200 authors included to avoid clutter. In addition, the original p-index, 0.1% perturbed p-index and 1.0% perturbed p-index were included for ease of visualization.

in Figure 7.4. As it can be seen, the p-index of the highest ranking authors remains remarkably stable despite the perturbations. This is because they tend to have a high number of papers and a high number of citations per paper. This makes it possible for them to have a stable p-index despite the change in mixing patterns. However, we can see that if the network is perturbed further, even the top ranking authors may lose their ranking stability.

7.4 Conclusion

In this chapter, the behavior of the p-index when the underlying citation network is perturbed was examined. A degree preserving perturbation technique was employed to perturb the citation network to varying proportions. The p-index was then applied to each of these networks and the results obtained compared with that of the original unperturbed citation network. The objective in analyzing these results is to demonstrate that the p-index takes a range of factors into consideration, whereas h-index only takes the number

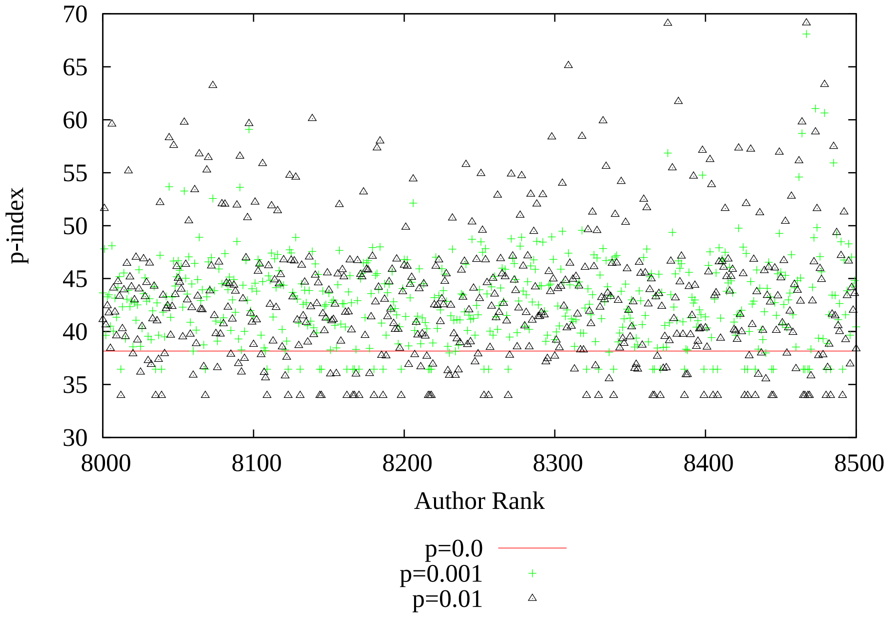


Figure 7.3: The variation of p-index of the authors in HEP-TH dataset against author rank. The authors are ranked using the original p-index (that of the unperturbed citation network) and only authors ranking 8000 to 85000 are included to avoid clutter. In addition, the original p-index, 0.1% perturbed p-index and 1.0% perturbed p-index were included for ease of visualization.

of papers and number of citations into account. As it has been pointed out, by changing which paper cites what paper, we can essentially change the p-index of an author. This is what was initially predicted in the third chapter. By analyzing the links in the citation network and weighting each citation, it is possible to derive a dynamic index that is sensitive to the source of the citation. As such, when the source of the citation changes from a high ranking publication to a low ranking publication, one's p-index goes down. Likewise, when the source of the citation changes from a low ranking publication to a high ranking publication, the respective p-index goes up. Evidence to this phenomenon has been provided in this chapter that shows that if one has a sufficiently large amount of papers and citations, they can stay immune to the perturbations to a certain point, essentially making their ranks stable. However, this is only possible when one is ranked at the top and has a relatively large number of papers and citations.

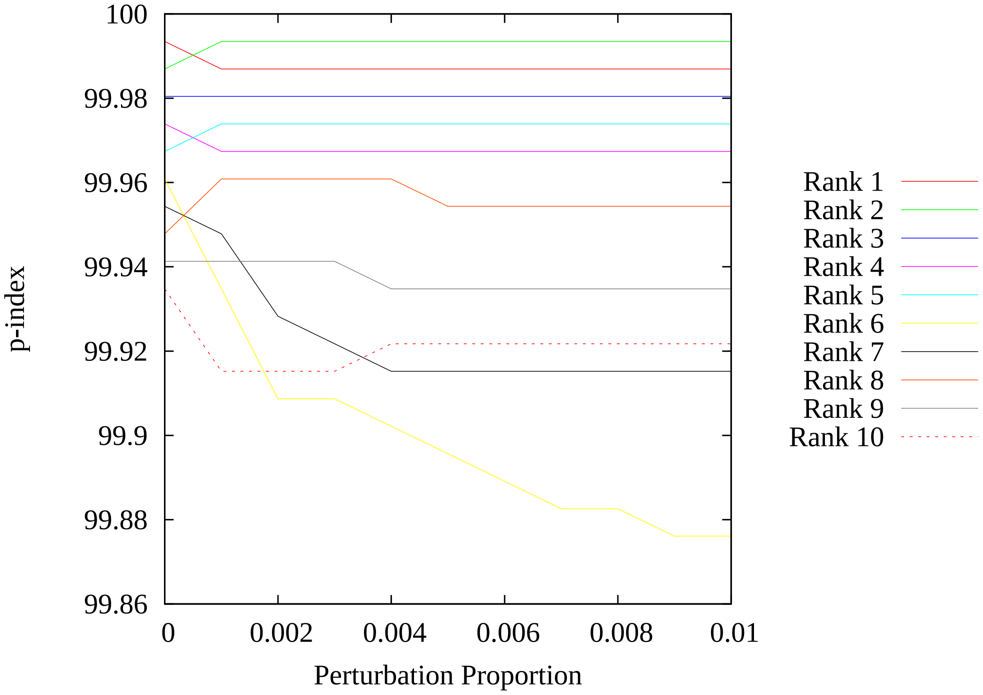


Figure 7.4: The variation of p-index of the top ten authors in HEP-TH dataset against the perturbation proportion. The authors are ranked using the original p-index (that of the unperturbed citation network).

Chapter 8

Conclusion

This thesis introduced a new ranking system for scientists that is fairer, more robust and resilient to manipulations when compared to the h-index. This new index was named the p-index and used the underlying citation network structure to evolve this measure. Comprehensive evidence has been provided on the manipulations possible with the h-index and demonstrated them using a simulated publication and ranking environment. The same simulated system was used to demonstrate that the p-index is immune to these manipulations. The application of the p-index to two real world citation networks was investigated as case studies. The p-index was then applied over the HEP-TH dataset while it was subjected to degree preserving perturbations in order to further demonstrate that the p-index considers the underlying dynamics of the network.

This chapter summarizes the content presented in this thesis. The first section discusses the summary of contributions and the second section proposes directions for future research.

8.1 Summary of Contributions

8.1.1 Possible Manipulations of the h-index

Although it has been suggested that there are many drawbacks to the h-index, this research is the first to provide systematic methodology on how to circumvent these manipulations and to present analytical evidence that the said methodology can reduce the undue advantage gained by using h-index.

8.1.2 Definition of p-index

The utilization of underlying network dynamics to come up with an evaluation system for a researcher's output is a novel contribution made by this thesis. Impactful documents were found using link analysis of the citation network and a ranking system taking the interplay between citation networks and collaboration networks into consideration was developed. Therein, a new index called p-index was introduced. This index is fairer than the existing ranking metrics and is also robust against manipulations.

8.1.3 Developing a Realistic Evolving Citation Network

This research has developed realistic citation and collaboration networks that mimic the evolution of real world academic networks. The manipulations possible in the h-index and the circumventing techniques available against them were demonstrated using this system.

8.1.4 Robustness of the p-index

Results were presented that suggest the p-index is robust against the manipulations that are possible with the h-index. Evidence was systematically provided from three simulated scenarios and two real world examples to support this claim.

8.1.5 Application of the p-index in Real Citation Networks

The p-index was applied to two real world citation networks and the characteristics of p-index that make it fairer, context-aware and more robust than the other existing indices were demonstrated. This is a novel contribution included in this thesis. Both a smaller evolving citation network (Quantum Game Theory) and a larger snapshot of a citation network (HEP-TH dataset) were used for this investigation.

8.1.6 Ranking Stability of the p-index

The characteristics of the p-index under degree preserving perturbations to the citation network were demonstrated. This provides evidence to support the claim that the p-index considers the perceived importance of the citing publication without just counting the number of citations one receives.

8.2 Directions for future work

A comprehensive range of experiments were carried out to test the new index and evidence was provided to suggest that it is a fairer index than the existing ranking metrics. The researcher believes the scientific community can utilize this knowledge and expand upon it. Specifically, the following directions may be pursued.

8.2.1 Adopting the p-index to Indexing Services

Although one can calculate their own h-index, one needs to use indexing services like Google Scholar or Web of Science to calculate their p-index since calculating the p-index needs the underlying citation network data. As such, adopting and deploying the p-index in an indexing service is a possible next step as it will ensure that a fairer index like p-index reaches a wider audience and achieves ubiquitous utility.

8.2.2 Discovering New Methods of Manipulations

After identifying and demonstrating how the h-index can be manipulated, it was shown that the p-index is resilient to those manipulations. While thorough testing was carried out, discovering new ways to manipulate the p-index may be a worthwhile direction to pursue.

8.2.3 Applying the p-index to New Citation Networks

This research applied the p-index to two real world citation networks and provided evidence to support our claims of fairness and robustness. A future direction for research would be to apply the p-index to alternate citation networks and observe the changes in the characteristics.

8.3 Epilogue

This thesis defines a new index to rank scientists, the p-index, and provides comprehensive evidence to suggest that p-index is fairer, more robust and resilient than the existing indices like h-index and its variants. Further evidence is drawn from the application of the p-index to two real world academic networks followed by inflicting degree preserving perturbations to a citation network and observing the characteristic change in p-index. It is hoped that this new index will be utilized by the scientific community as an ubiquitous metric to measure the scientific competence of researchers.

Bibliography

[1] R. Albert and A.-L. Barabási, “Statistical mechanics of complex networks,” *Rev. Mod. Phys.*, vol. 74, no. 1, pp. 47–97, Jan. 2002. [Online]. Available: http://link.aps.org/doi/10.1103/RevModPhys.74.47

[2] R. Albert, H. Jeong, and A. Barabasi, “Error and attack tolerance of complex networks,” *Nature*, vol. 406, no. 6794, pp. 378–382, 2000.

[3] B. Bahmani, A. Chowdhury, and A. Goel, “Fast Incremental and Personalized PageRank,” *ArXiv e-prints*, June 2010.

[4] A. L. Barabasi and R. Albert, “Emergence of scaling in random networks,” *Science*, vol. 286, no. 5439, pp. 509–512, 1999. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10521342

[5] P. D. Batista, M. G. Campiteli, and O. Kinouchi, “Is it possible to compare researchers with different scientific interests?” *Scientometrics*, vol. 68, no. 1, pp. 179–189, 2006. [Online]. Available: http://dx.doi.org/10.1007/s11192-006-0090-4

[6] G. Bianconi and A. Capocci, “Number of loops of size h in growing scale-free networks,” *Phys. Rev. Lett.*, vol. 90, p. 078701, Feb 2003. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevLett.90.078701

[7] J. Bollen, M. A. Rodriguez, and H. V. de Sompel, “Journal status,” *CoRR*, vol. abs/cs/0601030, 2006.

[8] P. Bonacich, “Power and centrality: A family of measures,” *American Journal of Sociology*, vol. 92, no. 5, pp. 1170–1182, 1987.

[9] L. Bornmann and H.-D. Daniel, "What do we know about theh index?" *Journal of the American Society for Information Science and Technology*, vol. 58, no. 9, pp. 1381–1385, 2007.

[10] L. Bornmann, R. Mutz, and H.-D. Daniel, "Are there better indices for evaluation purposes than theh index? A comparison of nine different variants of theh index using data from biomedicine," *Journal of the American Society for Information Science and Technology*, vol. 59, no. 5, pp. 830–837, 2008.

[11] G. Caldarelli, R. Pastor-Satorras, and A. Vespignani, "Structure of cycles and local ordering in complex networks," *The European Physical Journal B - Condensed Matter and Complex Systems*, vol. 38, no. 2, pp. 183–186, 2004. [Online]. Available: http://dx.doi.org/10.1140/epjb/e2004-00020-6

[12] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, and D. J. Watts, "Network robustness and fragility: Percolation on random graphs," *Phys. Rev. Lett.*, vol. 85, pp. 5468–5471, Dec 2000. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevLett.85.5468

[13] P. Chen, H. Xie, S. Maslov, and S. Redner, "Finding Scientific gems with Google's PageRank algorithm," *Journal of Informetrics*, 2007.

[14] R. Cohen, K. Erez, D. ben Avraham, and S. Havlin, "Resilience of the internet to random breakdowns," *Phys. Rev. Lett.*, vol. 85, pp. 4626–4628, Nov 2000. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevLett.85.4626

[15] R. Cohen, K. Erez, D. ben Avraham, and S. Havlin, "Breakdown of the internet under intentional attack," *Phys. Rev. Lett.*, vol. 86, pp. 3682–3685, Apr 2001. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevLett.86.3682

[16] R. Costas and M. Bordons, "The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level," *The Hirsch Index*, vol. 1, no. 3, pp. 193–203, 2007. [Online]. Available: http://dx.doi.org/10.1016/j.joi.2007.02.001

[17] C. Dangalchev, "Residual closeness in networks," *Physica A Statistical Mechanics and its Applications*, vol. 365, pp. 556–564, June 2006.

[18] R. P. Dellavalle, L. M. Schilling, M. A. Rodriguez, H. V. de Sompel, and J. Bollen, "Refining dermatology journal impact factors using pagerank," *Journal of the American Academy of Dermatology*, vol. 57, no. 1, pp. 116 – 119, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0190962207005026

[19] J. A. Dunne, R. J. Williams, and N. D. Martinez, "Food-web structure and network theory: The role of connectance and size," *Proceedings of the National Academy of Sciences*, vol. 99, no. 20, pp. 12 917–12 922, 2002. [Online]. Available: http://www.pnas.org/content/99/20/12917.abstract

[20] J. A. Dunne, R. J. Williams, and N. D. Martinez, "Network structure and biodiversity loss in food webs: robustness increases with connectance," *Ecology Letters*, vol. 5, no. 4, pp. 558–567, 2002. [Online]. Available: http://dx.doi.org/10.1046/j.1461-0248.2002.00354.x

[21] L. Egghe, "Theory and practise of the g-index," *Scientometrics*, vol. 69, no. 1, pp. 131–152, 2006. [Online]. Available: http://dx.doi.org/10.1007/s11192-006-0144-7

[22] L. Egghe and R. Rousseau, "An h-index weighted by citation impact," *Information Processing & Management*, vol. 44, no. 2, pp. 770–780, 2008.

[23] P. Erdös and A. Rényi, "On random graphs, I," *Publicationes Mathematicae (Debrecen)*, vol. 6, pp. 290–297, 1959. [Online]. Available: http://www.renyi.hu/~{}p_erdos/Erdos.html#1959-11

[24] L. C. Freeman, "A Set of Measures of Centrality Based on Betweenness," *Sociometry*, vol. 40, no. 1, pp. 35–41, Mar. 1977. [Online]. Available: http://links.jstor.org/sici?sici=0038-0431%28197703%2940%3A1%3C35%3AASOMOC%3E2.0.CO%3B2-H

[25] A. Fronczak, J. A. Holyst, M. Jedynak, and J. Sienkiewicz, "Higher order clustering coefficients in Barabasi-Albert networks," *Physica A*, Dec. 2002. [Online]. Available: http://arxiv.org/abs/cond-mat/0212237

[26] P. M. Gleiss, P. F. Stadler, A. Wagner, and D. A. Fell, "Relevant cycles in chemical reaction network," *Adv. Complex Syst*, vol. 4, pp. 207–226, 2001. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.20.9298

[27] S. Gupta, R. M. Anderson, and R. M. May, "Networks of sexual contacts: implications for the pattern of spread of HIV," *Aids*, vol. 3, 1989.

[28] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen, "Combating web spam with trustrank," in *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*, ser. VLDB '04, pp. 576–587. VLDB Endowment, 2004. [Online]. Available: http://dl.acm.org/citation.cfm?id=1316689.1316740

[29] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 46, pp. 16 569–16 572, 2005.

[30] P. Holme, B. J. Kim, C. N. Yoon, and S. K. Han, "Attack vulnerability of complex networks," *Phys. Rev. E*, vol. 65, p. 056109, May 2002. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevE.65.056109

[31] J. Iglesias and C. Pecharromán, "Scaling the h-index for different scientific ISI fields," *Scientometrics*, vol. 73, no. 3, pp. 303–320, 2007. [Online]. Available: http://dx.doi.org/10.1007/s11192-007-1805-x

[32] H. Jeong, S. P. Mason, A. L. Barabasi, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41–42, May 2001. [Online]. Available: http://dx.doi.org/10.1038/35075138

[33] B. Jin, L. Liang, R. Rousseau, and L. Egghe, "The R- and AR-indices: Complementing the h-index," *Chinese Science Bulletin*, vol. 52, no. 6, pp. 855–863, 2007.

[34] C. D. Kelly and M. D. Jennions, "The h index and career assessment by numbers," *Trends Ecol Evol*, vol. 21, no. 4, pp. 167–170, 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16701079

[35] J. M. Kleinberg, "Hubs, authorities, and communities," *ACM Comput. Surv.*, vol. 31, no. 4es, Dec. 1999. [Online]. Available: http://doi.acm.org/10.1145/345966.345982

[36] K. Konstantin, "Self-citation can inflate h -index," *Scientometrics*, vol. 77, no. 2, pp. 373–375, 2008.

[37] E. A. Leicht, G. Clarkson, K. Shedden, and M. E. J. Newman, “Large-scale structure of time evolving citation networks,” *The European Physical Journal B*, vol. 59, no. 1, pp. 75–83, 2007.

[38] J. Leskovec, J. Kleinberg, and C. Faloutsos, “Graphs over time: Densification laws, shrinking diameters and possible explanations,” in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, ser. KDD ’05, pp. 177–187. New York, NY, USA: ACM, 2005. [Online]. Available: http://doi.acm.org/10.1145/1081870.1081893

[39] B. R. Martin, “The use of multiple indicators in the assessment of basic research,” *Scientometrics*, vol. 36, no. 3, pp. 343–362, 1996.

[40] S. Maslov and S. Redner, “Promise and Pitfalls of Extending Google’s PageRank algorithm to citation networks,” *The Journal of Neuroscience*, 2009.

[41] S. Maslov, K. Sneppen, and A. Zaliznyak, “Detection of topological patterns in complex networks: correlation profile of the internet,” *Physica A: Statistical Mechanics and its Applications*, vol. 333, no. 0, pp. 529 – 540, 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378437103008409

[42] S. Milgram, *The Individual in a Social World: Essays and Experiments.* McGraw-Hill, New York, 1992.

[43] M. Newman, “Mixing patterns in networks,” *Phys. Rev. E*, vol. 67, no. 2, p. 026126, Feb. 2003.

[44] M. E. J. Newman, “Ego-centered networks and the ripple effect,” *Social Networks*, vol. 25, no. 1, pp. 83–95, Jan. 2003. [Online]. Available: http://dx.doi.org/10.1016/s0378-8733(02)00039-4

[45] M. E. J. Newman, “The structure and function of complex networks,” *SIAM REVIEW*, vol. 45, pp. 167–256, 2003.

[46] M. E. J. Newman, S. Forrest, and J. Balthrop, “Email networks and the spread of computer viruses,” *Phys. Rev. E*, vol. 66, p. 035101, Sep 2002. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevE.66.035101

[47] P. K. Newton, J. Mason, K. Bethel, L. A. Bazhenova, J. Nieva, and P. Kuhn, "A stochastic markov chain model to describe lung cancer growth and metastasis," *PLoS ONE*, vol. 7, no. 4, p. e34637, 04 2012. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pone.0034637

[48] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," 1998.

[49] A. J. Raan, "Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups," *Scientometrics*, vol. 67, no. 3, pp. 491–502, 2006. [Online]. Available: http://dx.doi.org/10.1556/Scient.67.2006.3.10

[50] C. Rodrigo and B. Maria, "Is g-index better than h-index? An exploratory study at the individual level," *Scientometrics*, vol. 77, no. 2, pp. 267–288, 2008.

[51] G. Scholar, "Quantum game theory google scholar page," 2008, http://scholar.google.com.au/citations?user=wkfPcaQAAAAJ&hl=en.

[52] U. Senanayake, M. Piraveenan, and A. Zomaya, "The p-index: Ranking scientists using network dynamics," *Procedia Computer Science*, vol. 29, no. 0, pp. 465 – 477, 2014, 2014 International Conference on Computational Science. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050914002191

[53] T. Tscharntke, M. E. Hochberg, T. A. Rand, V. H. Resh, and J. Krauss, "Author sequence and credit for contributions in multiauthored publications," *PLoS Biol*, vol. 5, no. 1, p. e18, 01 2007. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pbio.0050018

[54] C. University, "Kdd cup 2003," 2003, http://www.cs.cornell.edu/projects/kddcup/.

[55] T. N. van Leeuwen, M. S. Visser, H. F. Moed, T. J. Nederhof, and A. F. J. van Raan, "Holy Grail of science policy: Exploring and combining bibliometric tools in search of scientific excellence," *Scientometrics*, vol. 57, no. 2, pp. 257–280, 2003. [Online]. Available: #

[56] E. van Nierop, "Why do statistics journals have low impact factors?" *Statistica Neerlandica*, vol. 63, no. 1, pp. 52–62, 2009.

[57] D. Walker, H. Xie, K.-K. Yan, and S. Maslov, "Ranking scientific publications using a model of network traffic," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 6, p. 10, June 2007.

[58] L. Waltman, "An empirical analysis of the use of alphabetical authorship in scientific publishing," *CoRR*, vol. abs/1206.4863, 2012. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr1206.html#abs-1206-4863

[59] L. Waltman, R. Costas, and N. Jan van Eck, "Some Limitations of the H Index: A Commentary on Ruscio and Colleagues' Analysis of Bibliometric Indices," *Measurement*, vol. 10, no. 3, pp. 172–175, 2012.

[60] C.-T. Zhang, "The h'-Index, Effectively Improving the h-Index Based on the Citation Distribution," *PLoS ONE*, vol. 8, no. 4, pp. 1–8, 2013.