Reliability of citation count in predicting Scientific Impact : an estimation using higher order citation indices

A Final Thesis submitted for the completion of requirements for the degree of

Master of Technology (Course Work)

by

Naga Narasimharao Gadidamalla
Post-graduate Programme
Indian Institute of Science

Under the supervision of

Prof. Murugesan Venkatapathi
Department of Computation and Data Science

Acknowledgement

I express my sincere gratitude to my guide, Professor **Murugesan Venkatapathi**, Department of Computational and Data Science, Indian Institute of Science, for his dedicated guidance, generous help, and the precious time he gave in supervising this dissertation report. I also would like to extend my sincere thanks to other professors who assist and support all the students. I would also like to thank my fellow batch mates who helped me with their valuable suggestions throughout the thesis work.

Date: JUNE 2024
Place: IISc Bangalore

Naga Narasimharao Gadidamalla
21005

Candidate's Declaration

I hereby declare that the work carried out in this dissertation report entitled "**Reliability of citation count in predicting Scientific Impact :an estimation using higher order citation indices**" is being submitted in partial fulfilment of the requirements for the award of the degree of **"Master of Technology"** in **"Computational and Data Science"** submitted to the Department of computational and Data Science, Indian Institute of Science, Bengaluru, under the supervision of Professor **Murugesan Venkatapathi**, Computational and Data Science Department, IISc, Bengaluru.

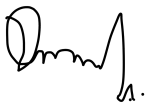
The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.

Date: JUNE 2024
Place: IISc Bengaluru

Naga Narasimharao Gadidamlla
21005

Certificate

This is to certify that the above statement made by the candidate is correct to the best of my knowledge and belief.



(Murugesan Venkatapathi)
Professor
Department of Computational & Data Sciences (CDS),
Indian Institute of Science,
Bangalore - 560012.

Abstract

This project investigates the efficacy of citation counts as a metric for evaluating the impact of scholarly works, by comparing it with the more robust Citation disruption(CD) index, which we use as a ground truth for measuring disruption in scholarly works. We begin by validating the temporal trend of the average CD5 index over several years, referencing the recent study by Park et al. (2023) . This sets the stage for a deeper analysis of how well traditional citation counts reflect the innovative impact of research, as encapsulated by the CD index.

We further analyze the correlation between citation counts and various CD indices (CD5, CD10, CD15), to understand the extent to which citations capture the disruptive nature of scholarly contributions. We observe a very low correlation between the CD index and the citation counts in general. Our study also examines the evolution of these CD indices over time, providing insights into their dynamic trends within the academic landscape. CD index is broadly based on the idea of approximating the progress of scientific literature using a process of diminishing the emerging redundant edges in the citation network as it propogates in time. During our analysis, we identify potential improvements to the CD index formula, suggesting areas for refining this metric to better capture the true impact of research. These proposed modifications aim to address any deficiencies in the current model and enhance its performance.

By conducting this comprehensive analysis, our work contributes to the broader conversation about the adequacy of single-dimensional metrics, like citation counts, in evaluating academic impact. It also reinforces the assertion that the impact of scientific work may take a few decades to fully emerge, especially the notable ones, and this is seen from the varying convergence rates of the CD indices of papers published(i.e values of CD5, CD10, CD15 etc) to an asymptotic value.

Contents

Chapter 1

Introduction & Related Work

1.1 Introduction

The last century have seen significant advances in scientific and technological knowledge. Despite these gains, there is growing concern about a slowdown in innovative activity across various fields. Research productivity appears to be waning, with fewer breakthroughs in papers, patents, and grant applications that bridge diverse knowledge domains, crucial for fostering innovation. The gap between the published year and the nobel prize awarding year of the papers also increased significantly. Several indicators of this behavior are visible and are referred to in this work. The degree to which this is due to the progressive difficulty of the scientific problems, or the professional structure of science today, is debatable.

In the modern academic landscape, the evaluation of scholarly impact plays a crucial role in determining the value and significance of research contributions. Traditional metrics such as citation counts have long been used as proxies for assessing the influence of academic works. However, these metrics often fall short in capturing the multifaceted nature of scholarly impact. The reliance on raw citation counts, for instance, does not account for the context or the manner in which citations are made, leading to potential biases and misinterpretations of a work's true influence.

Importance of Citation Metrics

Citation metrics serve multiple purposes in the academic and research community. They are used for:

1. **Evaluating Individual Researchers:** Citation counts and related indices are often used in academic evaluations, influencing hiring, promotion, and tenure decisions.

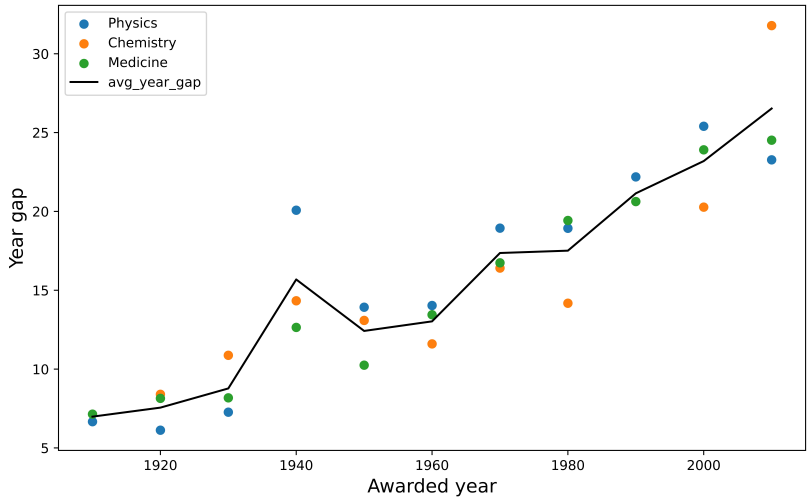


Figure 1.1: Nobel prize year gap over years

2. **Assessing Journal Impact:** Metrics like the journal impact factor are used to gauge the prestige and influence of academic journals.

3. **Allocating Funding:** Funding agencies and institutions use citation metrics to make decisions about research grants and resource allocation.

4. **Identifying Key Research Trends:** Citation analysis helps in understanding the evolution of research fields and identifying influential works that drive scientific progress.

Despite their widespread use, traditional citation metrics have several limitations. They tend to favor well-established researchers and popular fields, often overlooking groundbreaking work in emerging areas. Moreover, they do not differentiate between positive citations (which indicate recognition and validation) and negative citations (which may indicate criticism or controversy).

Need for Advanced Metrics

Recognizing these limitations, researchers have developed various advanced metrics to provide a more comprehensive assessment of scholarly impact. Some of these include:

- **h-index:** A metric that aims to measure both the productivity and citation impact of a researcher's publications. While it offers a balance between quantity and quality, it does not account for the age of citations or the context in which they are made.

- **g-index:** An extension of the h-index that gives more weight to highly cited articles, thereby addressing some of the h-index's limitations.
- **Altmetrics:** These metrics consider the attention an academic paper receives across various platforms, including social media, news outlets, and policy documents. Altmetrics aim to capture the perceived societal impact of research beyond traditional academic citations.

While these advanced metrics provide valuable insights, they still do not fully address the complexity of scholarly impact. There remains a need for metrics that can capture the disruptive potential of research – the extent to which a work changes the direction of a field by introducing novel ideas or methodologies.

The Citation Disruption (CD) index was introduced by Funk and Owen-Smith (2017) as a metric to evaluate the disruptive potential of scholarly papers. Unlike traditional metrics that focus solely on the volume of citations, the CD index measures how a paper influences the citation network of prior works. Specifically, it considers whether subsequent papers cite the new work in place of its predecessors, indicating a shift in the research landscape.

A positive CD index suggests that a paper is highly disruptive, redirecting the flow of citations away from earlier works. Conversely, a negative CD index indicates that a paper is consolidative, reinforcing and building upon existing knowledge without significantly altering the citation patterns. This distinction between disruptive and consolidative papers provides a more nuanced understanding of scholarly impact, capturing both the transformative and integrative aspects of research contributions.

Implications for research assessment

The ongoing development and refinement of citation metrics have significant implications for research assessment and academic evaluations. By adopting more sophisticated and context-sensitive metrics, institutions and funding agencies can make more informed decisions about resource allocation, hiring, and promotion. Additionally, researchers can gain a deeper understanding of the impact of their work and identify opportunities for collaboration and innovation.

Overall, this chapter sets the stage for a comprehensive examination of the reliability of citation counts in predicting the disruptive impact of scholarly papers. By situating our study within the broader context of citation metric research and identifying key gaps in existing approaches, we aim to contribute to the advancement of more accurate and meaningful methods for evaluating academic impact.

1.2 Related Work

Prior research has extensively explored various aspects of citation metrics and their efficacy in quantifying the impact and significance of scholarly works. One prominent line of inquiry focuses on the limitations of citation count as a sole measure of academic influence. Seglen (1997) highlighted several problems associated with using citation count and journal impact factors as a measure of impact. These issues include the skewed distribution of citations, where a small number of highly cited papers disproportionately influence the metric, and the variability in citation practices across different fields of study.

Building upon this, recent studies have investigated alternative metrics and indices to capture the different aspects of scientific contributions. For instance, the Hirsch index (h-index) proposes a combined measure of productivity and citation impact of a researcher (Hirsch, 2005). The h-index has gained widespread acceptance due to its simplicity and its ability to balance quantity and quality of publications. However, it too has limitations, such as not accounting for the context or significance of individual citations.

To address these limitations, newer metrics like the g-index (Egghe, 2006) and the i10-index have been proposed. The g-index gives more weight to highly cited articles, thus providing a more balanced measure of a researcher's impact. The i10-index, introduced by Google Scholar, simply counts the number of publications with at least ten citations, offering a straightforward measure of productivity and impact.

Another significant development in this field is the Altmetric score, which tracks the attention an academic paper receives across various platforms, including social media, news outlets, and policy documents. This metric aims to capture the broader impact of research beyond academic citations (Priem et al., 2010).

In the context of disruption assessment, Park et al. (2023) observed a trend of decreasing CD index over years, which can be inferred as a decrease in the percentage of destabilizing or disruptive works. Their study, published in Nature, highlighted the diminishing disruptive potential of contemporary papers and patents, suggesting a possible shift towards incremental rather than breakthrough research.

Additionally, several studies have reported a decline in research productivity over recent decades. For example, Bloom et al. (2020) found that research productivity has significantly decreased across various sectors, including pharmaceuticals, agriculture, and information technology. They argue that each successive dollar spent on research yields fewer innovations than in the past, indicating diminishing returns on investment in scientific research. Another study by Collison et al. (2018) examined trends in academic publishing and found that while the volume of publications has increased, the number of

high-impact, groundbreaking discoveries has not kept pace, suggesting a decline in the overall quality and productivity of research.

The concept of the Citation Disruption (CD) index was introduced by Funk and Owen-Smith (2017). They proposed the CD index as a measure to evaluate the disruptive potential of scholarly papers. The CD index aims to provide a more nuanced alternative to traditional citation counts by measuring the extent to which a paper disrupts existing patterns of knowledge citation within its field. This metric considers how a new publication influences the citation network of prior works, with a positive CD index indicating high disruption and a negative CD index suggesting consolidation of existing knowledge.

Our work extends this line of inquiry by directly comparing citation count with the CD index as a ground truth measure of disruption. By doing so, we aim to shed light on the reliability of citation count in capturing the disruptive nature of academic contributions and uncover potential discrepancies between traditional citation-based metrics and dynamic disruption indices.

Several other studies have contributed to this discourse. For example, Bornmann et al. (2008) conducted a comprehensive review of citation indicators, discussing their strengths and weaknesses. They emphasized the need for multi-dimensional metrics to capture the complex nature of scientific impact. Similarly, Waltman and van Eck (2012) proposed a model for assessing the citation impact of publications, arguing for normalization across fields to account for differences in citation practices.

Several researchers have explored historical contexts and country-wise differences in citation practices. Van Raan (2004) discussed variations in citation practices across different countries and the implications for international comparisons of scientific impact . Leydesdorff and Wagner (2008) examined patterns of international collaboration and how they affect citation impact, noting significant differences between countries . Glänzel and Schubert (2004) highlighted differences in citation impact and collaboration networks across countries . Hoekman et al. (2010) discussed how geographical and national factors influence collaboration and citation impact within Europe .

Moreover, Bornmann et al. (2010) explored country-specific differences in citation practices and the reliance on highly cited "giant" works . Zitt and Bassecoulard (1998) measured the internationalization of scientific journals and discussed how this varies by country, affecting citation impact . King (2004) provided a comprehensive analysis of the scientific impact of different countries, highlighting the disparities in citation impact and research output . Schubert and Braun (1996) discussed the need for normalization of scientometric indicators to account for differences in citation practices across countries and fields . Basalla (1967) provided a historical perspective on the diffusion of scientific

practices, including citation habits, from Western countries to other parts of the world .

Through this exploration of related literature, we contextualize our study within the broader landscape of scholarly impact assessment and provide a foundation for evaluating the effectiveness of citation metrics in capturing disruption within academic literature.

Objectives and Scope of the Study

This study aims to investigate the reliability of citation counts in predicting the disruptive impact of scholarly papers. By comparing citation counts with the CD index, we seek to uncover potential discrepancies between traditional citation-based metrics and dynamic disruption indices. The specific objectives of this research are:

- To analyze the correlation between citation counts and the CD index across various fields of study.
- To evaluate the temporal dynamics of the CD index, examining how the disruptive impact of papers evolves over time.
- To explore the implications of using the CD index as a complementary metric to traditional citation counts in academic evaluations and research assessments

By providing a detailed analysis of citation metrics and their ability to capture the disruptive impact of scholarly papers, this study aims to contribute to the ongoing discourse on the evaluation of academic impact and the development of more comprehensive and accurate metrics.

Chapter 2

CD Index

2.1 Introduction

The CD (Citation Disruption) index is a quantitative metric designed to evaluate the disruptive potential of scholarly papers. It provides a more nuanced alternative to traditional citation counts, specifically by measuring the extent to which a paper disrupts existing patterns of knowledge citation within its field. The index is calculated based on how a new publication influences the citation network of prior works.

A positive CD index indicates that a paper is highly disruptive, meaning it tends to be cited instead of previous literature, thereby redirecting the citation flow in its field. Such papers introduce novel ideas or methodologies that significantly shift academic perspectives or practices. Conversely, a negative CD index suggests that a paper is consolidative, building upon existing knowledge without significantly altering the citation landscape.

In calculating the CD index, each paper's citation relationships are analyzed, particularly focusing on how it cites previous works and how subsequent papers cite both the new work and its references. A key aspect of this metric involves considering the common predecessors—earlier papers cited by both the paper in question and its citing papers. The degree of disruption is gauged by the change in citation patterns, with a more disruptive work causing a decrease in citations to these common predecessors.

In our work, the CD index serves as a cornerstone for assessing the true impact of academic contributions, providing a comprehensive measure that goes beyond mere citation volume to consider the transformative influence of scholarly work.

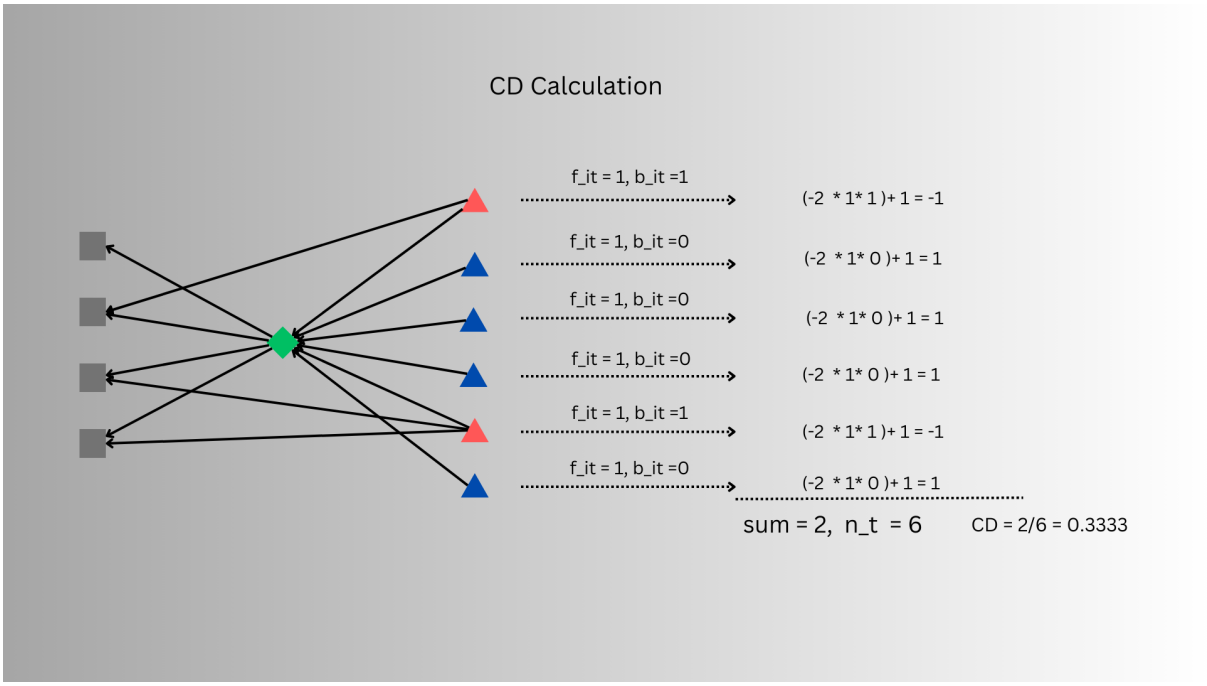


Figure 2.1: CD index

2.2 CD index formula

The calculation of the CD index is as follows:

- If a paper citing the focal paper also cites its predecessor, this citation contributes a value of -1 to the index.
- Conversely, if a paper citing the focal paper does not cite any of its predecessors, it contributes a value of +1 to the index.
- The CD index is then obtained by taking the mean of all such contributions.

The mathematical equation is as follows.

$$CD_t = \frac{1}{n_t} \sum_{i=1}^{n_t} -2f_{it} * b_{it} + f_t$$

where,

- f_{it} : 1 if i cites the focal paper; 0 if not
- b_{it} : 1 if i cites the predecessors of the focal paper; 0 if not
- n_t : number of forward cites to the focal paper as of t years after it's publication

The calculation of the CD index for a single paper as per the above formula is depicted in Figure 3.1

We can rewrite the equation as follows

$$CD_t = \frac{1}{n_t}\sum_{i=1}^{n_t} f_{it} * (1 - 2 * b_{it})$$

As we are considering only the forward citations of the focal paper, f_{it} is always 1. Therefore,

$$CD_t = \frac{1}{n_t}\sum_{i=1}^{n_t}(1 - 2 * b_{it})$$

$$CD_t = \frac{1}{n_t}(n_t - 2 * \sum_{i=1}^{n_t} b_{it})$$

$$CD_t = 1 - \frac{2}{n_t}\sum_{i=1}^{n_t} b_{it}$$

$$CD_t = 1 - 2 * \frac{n_{predecessor}}{n_t}$$

where,

- $n_{predecessor_t}$: number of papers with common predecessors as of t years after publication
- n_t: number of papers which cited the focal paper as of t years after publication

Let us calculate CD index using the deduced formula for the citation network in figure 3.1
As we can see there are total 6 papers which cited the focal paper. Among them 2 papers cited both the papers. Therefore $n_t = 6$, $n_{predecesssors_t} = 2$.

$$CD_t = 1 - 2 * \frac{n_{predecessor}}{n_t}$$

$$CD_t = 1 - 2 * \frac{2}{6} = \frac{2}{6} = 0.33333.$$

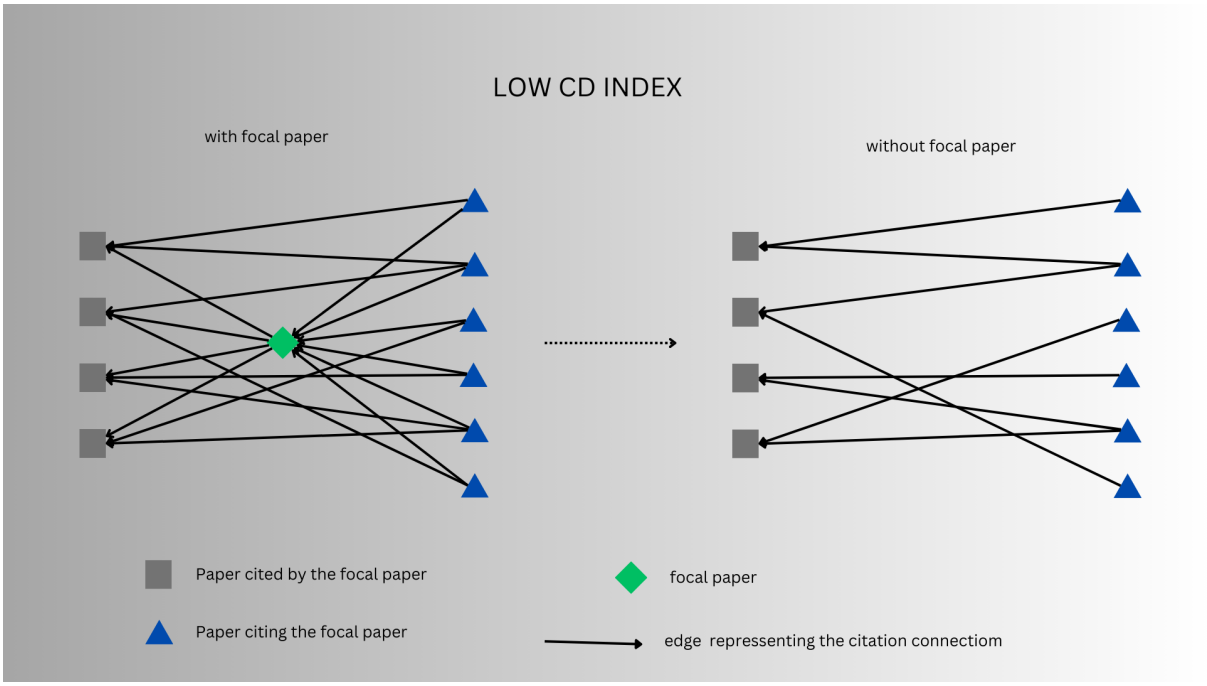


Figure 2.2: Less disruptive paper network

2.3 Interpretation of CD index

The Citation Disruption (CD) index offers a novel perspective on measuring the impact of scholarly papers. Unlike traditional citation metrics that quantify the cumulative influence of a paper based on the number of citations it receives, the CD index captures how a paper disrupts existing knowledge and redirects future research.

Disruption refers to the extent to which a paper changes the direction of research. A disruptive paper may introduce new concepts, methodologies, or paradigms that prompt subsequent research to diverge from established paths. The CD index measures this by examining how future papers cite the focal paper compared to its predecessors.

A positive CD index indicates that a paper has disrupted the existing citation network by redirecting citations away from earlier works. Conversely, a negative CD index suggests that the paper consolidates existing knowledge, reinforcing previous research without significantly altering citation patterns.

2.3.1 Less disruptive papers

Less disruptive papers are characterized by low CD index scores. These papers do not significantly affect the citation patterns of their predecessors. The citation network of the predecessors remains largely unchanged with or without the focal paper.

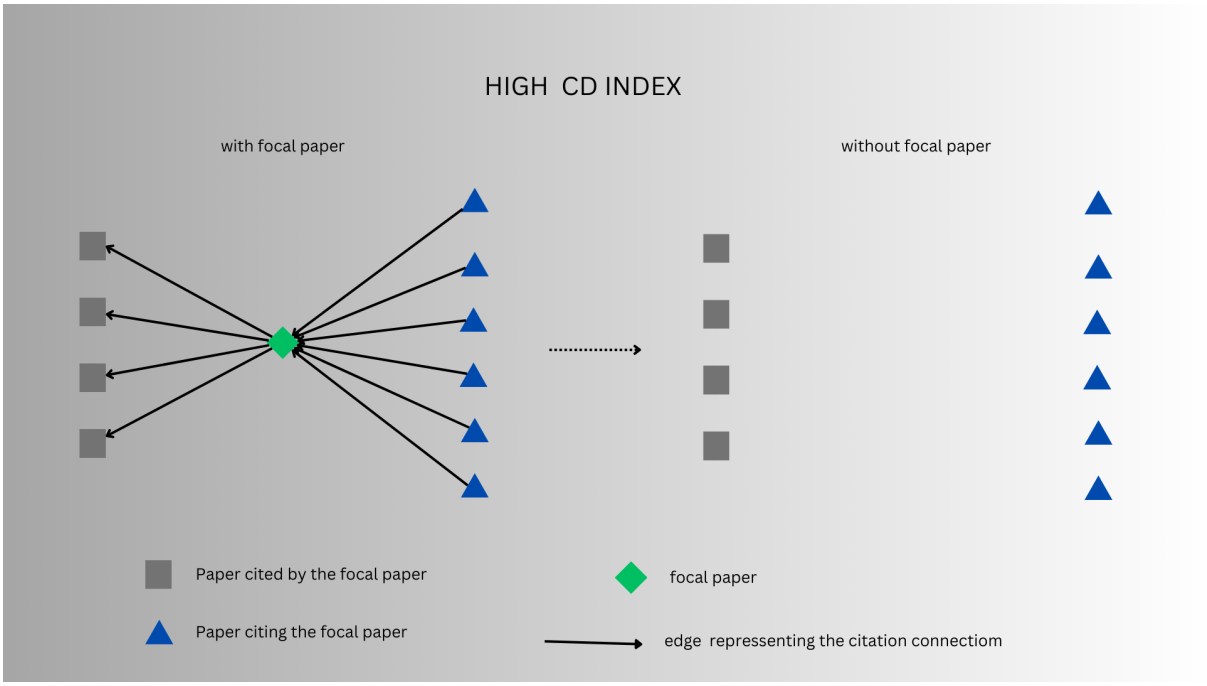


Figure 2.3: Highly disruptive paper network

- **Minimal Impact on Citation Network:** The citation patterns of preceding papers remain similar, indicating that the focal paper has not introduced significant new directions or concepts.
- **Limited Contribution to Science:** These papers do not push the field into new dimensions or create different domains of work. They primarily build on existing knowledge without causing major shifts in research focus.
- **Incremental Advancements:** The contributions are often incremental, adding to the body of knowledge without fundamentally changing the field.

Figure 2.2 illustrates a less disruptive paper network, where the citation patterns show minimal changes even after the introduction of the focal paper.

2.3.2 Highly disruptive papers

Highly disruptive papers, on the other hand, have high CD index scores. These papers significantly alter the trajectory of a field by introducing new ideas that cause subsequent research to shift focus away from older references.

Characteristics of Highly Disruptive Papers

- **Significant Shift in Citation Network:** Subsequent papers citing a disruptive

work tend to neglect older references, indicating a shift in the research focus. This is a hallmark of high disruption.

- **Transformative Contributions:** These papers often introduce groundbreaking ideas, methodologies, or paradigms that redefine the field and open new avenues for research.
- **High Impact on Research Direction:** The influence of these papers is profound, often leading to new research questions and innovative approaches that deviate from established paths.

Figure 2.3 illustrates a highly disruptive paper network, where the citation patterns show a marked departure from older works, indicating the paper's significant influence on the field. We can infer that a subsequent papers might not have been possible without the focal paper's contribution to the field.

Interpreting the CD index involves understanding the context and citation patterns associated with a paper:

- **Addressing the circular citations:** As depicted in the figure 2.4, CD index identifies and addresses the common predecessor connections which are largely redundant and not very significant in the advancement of science.
- **Positive CD Index:** A positive value indicates disruption. The higher the positive value, the more the paper has redirected citations away from its predecessors. This suggests that the paper has introduced new ideas or methodologies that subsequent research has adopted, leading to a change in the research direction.

  Negative CD Index: A negative value indicates consolidation. The more negative the value, the more the paper has reinforced existing knowledge without significantly altering citation patterns. This suggests that the paper has contributed to the body of knowledge in a way that supports and builds upon previous research.
- **Magnitude of the CD Index:** The absolute value of the CD index reflects the extent of disruption or consolidation. Larger magnitudes (whether positive or negative) indicate a stronger influence on the citation network.
- **Temporal Dynamics:** Examining how the CD index evolves over time can provide insights into the changing impact of a paper. A paper might initially have a low disruption score but gain disruptive potential as its ideas are more widely adopted.
- **Field-Specific Considerations:** The interpretation of the CD index can vary across different fields. Some fields may naturally have higher or lower disruption

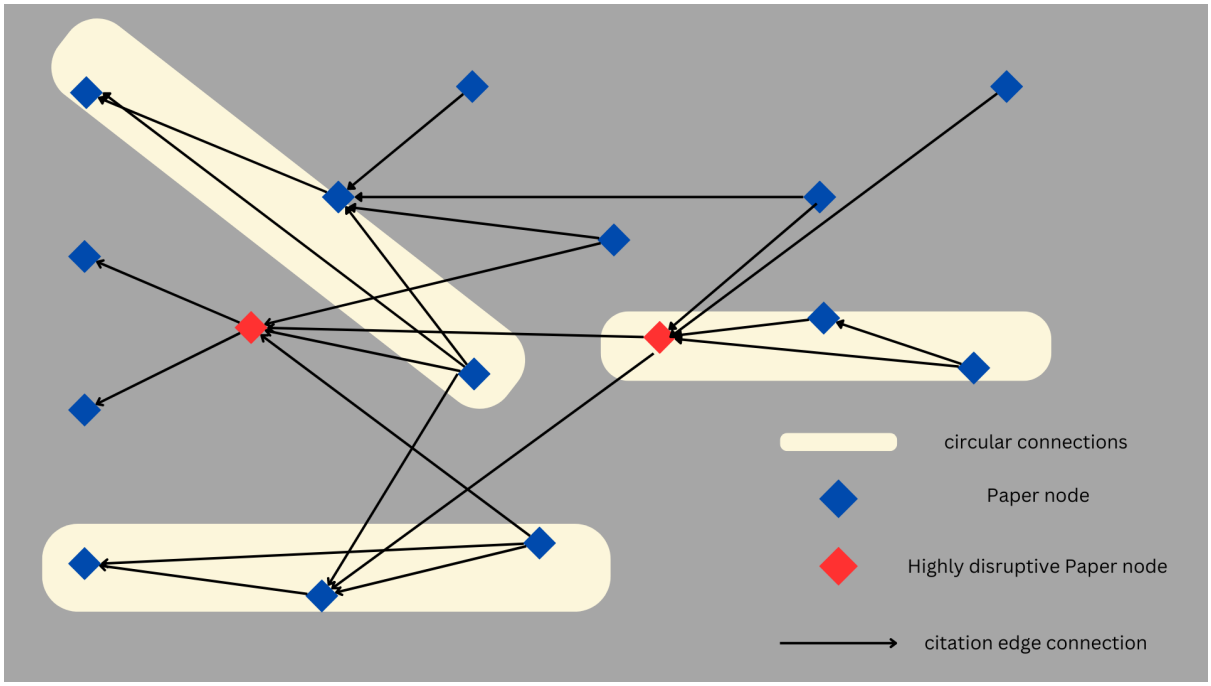


Figure 2.4: Simple citation network

levels due to their inherent research dynamics. Comparing CD indices within the same field provides a more accurate assessment.

By understanding these aspects, researchers and evaluators can use the CD index to gain a more nuanced understanding of a paper's impact, beyond what traditional citation counts can offer. This metric helps in identifying truly transformative research that drives scientific progress by shifting the focus and direction of future studies.

2.4 CD Index calculation

2.4.1 Data

In this study, we used the Open Graph Benchmark (OGB) MAG240 dataset as our primary source of empirical data. The OGB MAG240 dataset is a comprehensive repository containing information on academic publications, authors, institutions, and citation networks extracted from the Microsoft Academic Graph (MAG). Specifically, it encompasses a diverse array of scholarly works spanning various disciplines, thereby offering a rich and extensive corpus for our analyses.

The MAG240 dataset comprises a total of around 1.3 billion citations spanning over 121 million papers, providing a granular depiction of the interconnectedness within the scholarly landscape. It also consist of published year and author id for each paper. Leveraging this vast repository enables us to conduct robust investigations into the relationship

between citation count and the CD index, as well as to examine the temporal dynamics of disruptive emergence among highly cited papers. This dataset also consists of a 768 dimensional vector embedding for each paper based on its title and abstract which would be useful to calculate our modified CD index.

2.4.2 Algorithm

We have the data of the published year and the citation information of around 121 million papers. We did some data manipulation techniques on the entire data to get the CD index instead of calculating for each paper one by one. Our algorithm extensively uses Pandas library.

The algorithm is as follows

1. From the data we have, We created a data frame(df_{edges}) consisting 4 columns paper id, year of publication, cited paper id, year of publication
2. Then we merged the data frame with itself using inner join on the column 'cited paper id'. This step gives us the papers which cited the same paper. It results in all the combinations of papers with common references. Let's say this dataframe name is df_{common}
3. We merged the data frame(df_{common}) with the dataframe df_{edge} using inner join on the combination of columns ['paper id', 'cited paper id']. This step identifies the entries in the data frames which have entries of both the columns [paper id, cited paper id] same. It results in all the edge connections which have a common predecessor.
4. We need to consider only the papers which cited a focal paper within t years of publication to calculate CD_t. We have set a threshold on the difference of the published year to be t years.
5. We calculated the number of citations and the number of citations with common predecessors for each paper.
6. Then we calculated CD index by using the equation, $CD_t = 1 - 2 * \frac{n_{predecessors}}{n_t}$

2.5 Observations

During the course of our investigation, we made several significant observations that provide a deeper understanding of the relationship between traditional citation counts

and the Citation Disruption (CD) index. These observations also shed light on the dynamic trends of these metrics over time.

1. **Temporal Trends in Average CD Index:** - We tracked the average CD index value over several years to observe its temporal trends. This analysis revealed consistent patterns and fluctuations in the CD index, indicating how the disruptive impact of scholarly works evolves over time. The validation of these trends with reference to the study by Park et al. (2023) helped confirm the reliability and robustness of the CD index as a metric for scholarly impact.

2. **Correlation Between Citation Counts and CD Index**: - Our examination of the correlation between citation counts and the CD index for all papers provided insights into how well traditional citation counts reflect the innovative impact of research. We found that while there is some level of correlation, it varies significantly across different papers, suggesting that citation counts alone may not fully capture the disruptive nature of scholarly contributions.

3. **Group-wise Analysis of Correlation**: - To delve deeper into the nuances of the relationship between citation counts and the CD index, we divided the papers into nine groups based on thresholds for low, mid, and high values of both metrics. Specifically, we categorized the papers as follows:

- Low CD Index($CD < 0$), Low Citation Count(< 50)
- Low CD Index, Mid Citation Count($50 <=$ Citation count $<= 1000$)
- Low CD Index, High Citation Count(Citation Count > 1000)
- Mid CD Index($0 < CD < 0.35$), Low Citation Count
- Mid CD Index, Mid Citation Count
- Mid CD Index, High Citation Count
- High CD Index($CD >= 0.35$), Low Citation Count
- High CD Index, Mid Citation Count
- High CD Index, High Citation Count

- Within each group, we calculated the correlation between the CD index and citation counts. This stratified analysis highlighted the variations in correlation across different levels of scholarly impact, revealing that papers with high CD indices and high citation counts exhibit a stronger correlation, whereas other groups showed more diverse and less consistent relationships.

4. **Trends in CD Indices for High-Cited Papers:** - We observed the trends of CD5, CD10, ..., CD30 index values for a selection of high-cited papers over time. This detailed analysis provided insights into how the disruptive nature of highly cited papers evolves, showing that while some papers maintain high CD values consistently, others experience a decline in disruptive impact over time. This observation underscores the dynamic nature of research impact and the importance of longitudinal analysis in understanding scholarly influence.

These observations collectively emphasize the complexity and multi-dimensionality of evaluating scholarly impact. They highlight the limitations of relying solely on citation counts and advocate for the incorporation of more nuanced metrics like the CD index to better capture the true innovative and disruptive impact of academic research.

Chapter 3

Results & Discussions

3.1 Results

Our comprehensive analysis yielded several significant findings regarding the reliability of citation counts as a measure of scholarly impact and the advantages of using the Citation Disruption (CD) index. These results provide valuable insights into the dynamics of academic research and its influence over time.

Average CD Index Trends

The trend of the average CD index over the years reaffirmed the notion that papers and patents are becoming less disruptive over time. Specifically, our analysis showed a consistent decline in the CD index, indicating that recent publications are more likely to build upon existing knowledge rather than introducing groundbreaking innovations. This finding aligns with the observations made by Park et al. (2023), suggesting a shift towards incremental improvements in research.

This decline in disruptiveness can be attributed to several factors, including increased specialization in research fields, funding and publication pressures favoring incremental work, and the collaborative nature of modern research. As researchers focus on narrower domains and seek to secure funding and publications, the likelihood of producing highly disruptive work diminishes.

Correlation Analysis

Our examination of the correlation between citation counts and the CD index revealed a generally low correlation, indicating that high citation counts do not necessarily equate to high disruptive impact. This suggests that while citation counts serve as a useful proxy for gauging the popularity of scholarly works, they are not sufficient for capturing the disruptiveness of research.

The low correlation highlights the limitations of traditional citation metrics in assess-

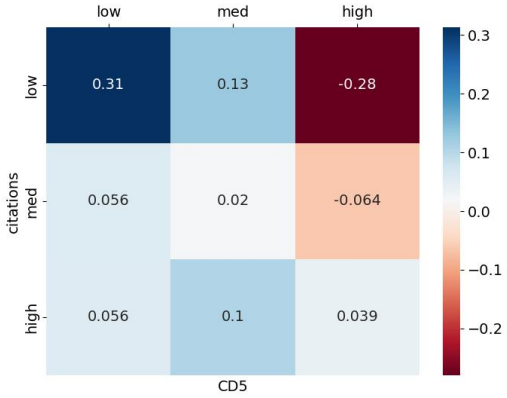


Figure 3.1: This figure depicts the correlation of CD5 and Citation count in 9 categories of papers based on high,medium and low citations and high,medium and low CD5 index. We can see that there is almost zero correlation in most cases

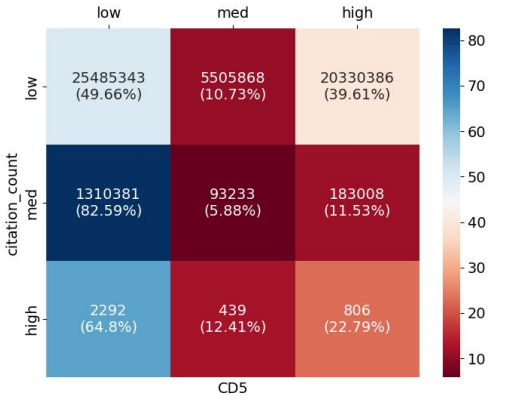


Figure 3.2: % This figure depicts the number and percentage of papers with low, medium and high CD5 values for each category of Citation count

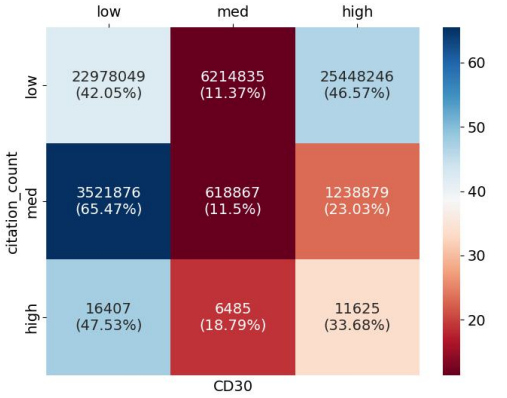


Figure 3.3: This figure depicts the number and percentage of papers with low, medium and high CD30 values for each category of Citation count. We can see a clear increase in the % of papers with high CD30 value compared to that of CD5

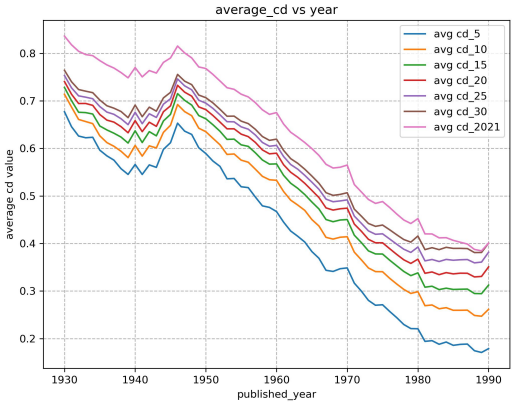


Figure 3.4: In this figure, we have shown how the trends of CD5, CD10, CD15, CD20, CD25, CD30, CD2021 varies over the years. We can see a clear decreasing trend in the temporal patterns and also a clear increase from CD5 to CD10, CD10 TO CD15 and so on.

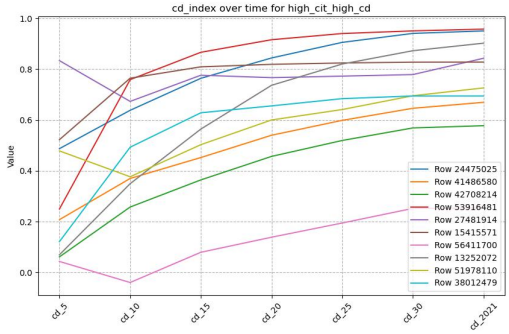


Figure 3.5: In this figure, we can see how CD values are evolving for a few randomly selected highly cited papers with positive CD5. We can see they are consistently increasing for most of the papers

ing the true impact of scholarly contributions. For instance, highly cited review articles may consolidate existing knowledge without introducing significant innovations, thereby receiving a low CD index despite their high citation counts.

Group-wise Analysis of Correlation

To delve deeper into the nuances of the relationship between citation counts and the CD index, we divided the papers into nine groups based on thresholds for low, mid, and high values of both metrics.

Within each group, we calculated the correlation between the CD index and citation counts. This stratified analysis highlighted the variations in correlation across different levels of scholarly impact, revealing that papers with high CD indices and high citation counts exhibit a stronger correlation, whereas other groups showed more diverse and less consistent relationships.

This group-wise analysis underscores the importance of considering multiple dimen-

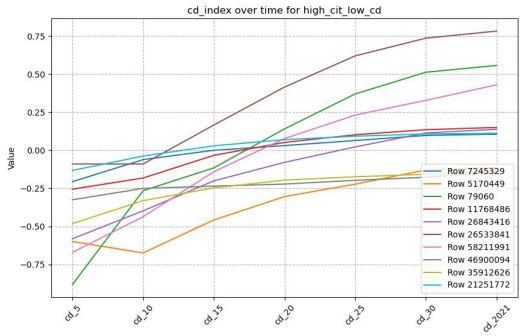


Figure 3.6: In this figure, we can see how CD values are evolving for a few randomly selected highly cited papers with negative CD5. We can see they are consistently increasing for most of the papers

sions of impact when evaluating scholarly contributions. Papers with high disruptiveness and high citation counts tend to be truly groundbreaking, whereas those with high citation counts but low disruptiveness may represent well-regarded but incremental work.

Longitudinal CD Index Trends

Our investigation into the evolution of CD indices after 5, 10, 15, 20, 25, and 30 years of publication showed an interesting trend of increasing CD values over time. This pattern was particularly evident among highly cited papers, revealing that even initially less disruptive papers have the potential to evolve and exhibit greater disruptive impact over the years.

This finding suggests that the disruptive potential of research can change over time, influenced by subsequent developments and shifts in the academic landscape. It highlights the dynamic nature of scholarly impact and the need for longitudinal analysis to fully understand the influence of academic contributions.

3.2 Conclusions

Based on our findings, we draw the following important conclusions:

1. **Decline in Disruptiveness:** The observed decline in the CD index over time suggests that contemporary research is becoming less disruptive, favoring incremental improvements over groundbreaking innovations. This trend has significant implications for the future of scientific progress and the development of new knowledge.

2. **Limitations of Citation Counts:** Traditional citation metrics, while useful for measuring popularity and recognition, fall short in capturing the disruptive impact of scholarly works. The low correlation between citation counts and the CD

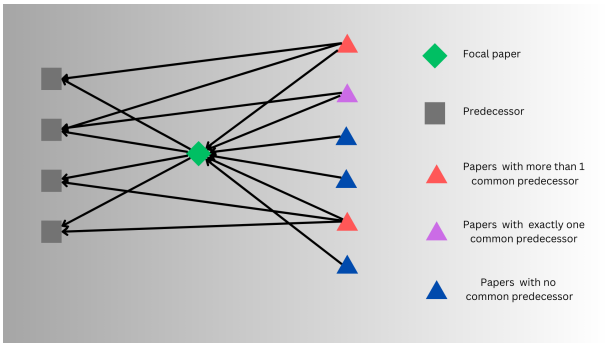


Figure 3.7: Multiple predecessors

index highlights the need for more nuanced metrics that can accurately assess the transformative influence of research.

3. **Dynamic Nature of Scholarly Impact:** The longitudinal analysis of CD indices demonstrates that the disruptive potential of research can evolve over time. This dynamic nature of scholarly impact calls for ongoing assessment and reevaluation of academic contributions, considering both immediate and long-term effects.

4. **Need for Comprehensive Metrics:** Our study underscores the importance of incorporating multiple dimensions of impact, including both citation counts and disruption indices, to provide a holistic evaluation of academic contributions. The CD index serves as a valuable complement to traditional metrics, offering deeper insights into the transformative influence of research.

3.3 Future Work

To build upon these findings, future research should focus on:

1. **Developing a Modified CD Index:** We identified a few limitations of the CD index. A few of them are :

   - CD index doesn't consider the effect of multiple common predecessors. This phenomenon is shown in figure 3.7.
   - The penalty for having a common predecessor is always -1. We must consider the following factors:

     (a) **The similiarity between the common predecessor and focal paper:** If the focal paper is very similar to the common predecessor, it should be penalized with a higher value because it is consolidating. If the focal paper and the common predecessor are not closely related, it is not fair to penalize the focal paper with a high value.

(b) **The similiarity between the focal paper and the citing paper:** If the subsequent work is very similar to the focal paper, it will likely cite the predecessors of the focal paper. In this situation, it is not fair to penalize the focal paper heavily because the common predecessors reflect more on the nature of the subsequent works than on the focal paper.

The penalty should be a function of the two similiarity scores mentioned above to fairly capture the consolidative nature of having a common predecessor.

2. **Expanding the Dataset:** Incorporate additional datasets to validate the findings across different disciplines and improve the robustness of the analysis.

3. **Similarity based Clustering:** Cluster the papers based on the similarity so that we can set field specific thresholds for analysing the correlation characteristics with citation counts better.

4. **Exploring country/region specific metrics:** Several studies have indicated significant differences in citation behaviors and research output between countries. We aim to Develop metrics that account for variations in citation practices, research funding levels, and collaboration patterns across different geographical areas.

5. **Investigating Temporal Dynamics Further:** Conduct more in-depth studies on the temporal dynamics of disruption to understand how the disruptive impact of papers evolves over longer periods.

6. **Exploring the correlation with other popular metrics:** We aim to examise the correlation of CD index with other popular metrics like h-index, g-index etc to get an idea about which popular metric captures the disruptiveness better.

By advancing these research directions, we can enhance our understanding of scholarly impact and develop more effective tools for evaluating the transformative influence of academic contributions.

References:

1. Basalla, G. (1967). The spread of Western science. *Science*, 156(3775), 611-622.

2. Bloom, N., Jones, C. I., Van Reenen, J., & Webb, M. (2020). Are ideas getting harder to find? *American Economic Review*, 110(4), 1104-1144.

3. Bornmann, L., de Moya Anegón, F., & Leydesdorff, L. (2010). Do scientific advancements lean on the shoulders of giants? A bibliometric investigation of the Ortega hypothesis. *PLoS ONE*, 5(10), e13327.

4. Bornmann, L., Mutz, R., & Daniel, H. D. (2008). Are there better indices for evaluation purposes than the h-index? A comparison of nine different variants of the h-index using data from biomedicine. *Journal of the American Society for Information Science and Technology*, 59(5), 830-837.

5. Collison, P., Nielsen, M., & Michael, N. (2018). Science is getting less bang for its buck. *The Atlantic*. Retrieved from [https://www.theatlantic.com/science/archive/2018/03/sci getting-less-bang-for-its-buck/555173/](https://www.theatlantic.com/science/archive/2018/03/sc getting-less-bang-for-its-buck/555173/)

6. Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, 69(1), 131-152.

7. Funk, R. J., & Owen-Smith, J. (2017). A dynamic network measure of technological change. *Management Science*, 63(3), 791-817.

8. Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, 178(4060), 471-479.

9. Garfield, E. (2006). The history and meaning of the journal impact factor. *JAMA*, 295(1), 90-93.

10. Glänzel, W., & Schubert, A. (2004). Analyzing scientific networks through co-authorship. *Handbook of Quantitative Science and Technology Research*, 257-276.

11. Harzing, A. W., & van der Wal, R. (2008). Google Scholar as a new source for citation analysis. *Ethics in Science and Environmental Politics*, 8, 61-73.

12. Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, 520(7548), 429-431.

13. Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46), 16569-16572.

14. Hoekman, J., Frenken, K., & Tijssen, R. J. (2010). Research collaboration at a distance: Changing spatial patterns of scientific collaboration within Europe. *Research Policy*, 39(5), 662-673.

15. Jones, B. F. (2009). The burden of knowledge and the 'death of the Renaissance man': Is innovation getting harder? *The Review of Economic Studies*, 76(1), 283-317.

16. King, D. A. (2004). The scientific impact of nations. *Nature*, 430(6997), 311-316.

17. Kostoff, R. N. (1998). The use and misuse of citation analysis in research evaluation. *Scientometrics*, 43(1), 27-43.

18. Larivière, V., & Gingras, Y. (2010). The impact factor's Matthew Effect: A natural experiment in bibliometrics. *Journal of the American Society for Information Science and Technology*, 61(2), 424-427.

19. Leydesdorff, L., & Milojević, S. (2015). Scientometrics. In *International Encyclopedia of the Social & Behavioral Sciences* (2nd ed., pp. 322-327).

20. Leydesdorff, L., & Opthof, T. (2010). Scopus's source normalized impact per paper (SNIP) versus a journal impact factor based on fractional counting of citations. *Journal of the American Society for Information Science and Technology*, 61(11), 2365-2369.

21. Leydesdorff, L., & Wagner, C. S. (2008). International collaboration in science and the formation of a core group. *Journal of Informetrics*, 2(4), 317-325.

22. Merton, R. K. (1968). The Matthew effect in science: The reward and communication systems of science are considered. *Science*, 159(3810), 56-63.

23. Moed, H. F. (2005). *Citation Analysis in Research Evaluation*. Springer.

24. Pan, R. K., & Fortunato, S. (2014). Author impact factor: Tracking the dynamics of individual scientific impact. *Scientific Reports*, 4, 4880.

25. Park, M., Leahey, E., & Funk, R. J. (2023). Papers and patents are becoming less disruptive over time. *Nature*, 613(7942), 138-144.

26. Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). Altmetrics: A manifesto. Retrieved from http://altmetrics.org/manifesto

27. Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, 105(45), 17268-17272.

28. Schubert, A., & Braun, T. (1996). Cross-field normalization of scientometric indicators. *Scientometrics*, 36(3), 311-324.

29. Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ*, 314(7079), 498-502.

30. Thelwall, M., & Kousha, K. (2015). ResearchGate: Disseminating, communicating, and measuring Scholarship? *Journal of the Association for Information Science and Technology*, 66(5), 876-889.

31. Thelwall, M., & Wilson, P. (2014). Mendeley readership altmetrics for medical articles: An analysis of 45 fields. *Journal of the Association for Information Science and Technology*, 65(8), 1627-1638.

32. Van Raan, A. F. J. (2004). Measuring science: Capita selecta of current main issues. *Handbook of Quantitative Science and Technology Research*.

33. Waltman, L., & van Eck, N. J. (2012). The inconsistency of the h-index. *Journal of the American Society for Information Science and Technology*, 63(2), 406-415.

34. Wouters, P., & Costas, R. (2012). Users, narcissism and control – tracking the impact of scholarly publications in the 21st century. *SURFfoundation*.

35. Zitt, M., & Bassecoulard, E. (1998). Internationalization of scientific journals: A measurement based on publication and citation scope. *Scientometrics*, 41(1-2), 255-271.

36. Li, J., Yin, Y., Fortunato, S. et al. A dataset of publication records for Nobel laureates. Sci Data 6, 33 (2019). https://doi.org/10.1038/s41597-019-0033-6

Appendix

Code Availability

The code used for calculating the Citation Disruption (CD) index in this project is available upon request. You may find the code at the link **https://github.com/naga0808/CD-Index-calculation** . The code is provided under an open-source license, allowing for further development and adaptation in related research projects. Detailed documentation is included to facilitate understanding and replication of the CD index calculations presented in this study.

Data Availability

The datasets generated and analyzed during the current study are available for open access. You can directly download the nodeyear.csv, edgeindex.csv data using the following links: **https://storage.cloud.google.com/mag240dataset/CD_calc_data/edge_index.csv** , **https://storage.cloud.google.com/mag240dataset/CD_calc_data/node_year.csv** respectively . This includes the data used for calculating the Citation Disruption (CD) index. The data is provided under an open-access license, ensuring transparency and reproducibility. Detailed descriptions and metadata are included to facilitate understanding and replication of the analyses presented in this study.