Citation Disruption (CD) Index Vs. Citation Counts: Some Improvements in the CD Index

A Final Thesis submitted for the completion of

requirements for the degree of

Master of Technology (Course Work)

by

Tejas Ashok Tonde

Post-graduate Programme

Indian Institute of Science

Under the supervision of

Prof. Murugesan Venkatapathi

Department of Computational 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.

Date: JUNE 2025 Tejas Ashok Tonde

Place: IISc Bangalore 23138

Candidate's Declaration

I hereby declare that the work carried out in this dissertation report entitled "**Citation Disruption (CD) Index Vs. Citation Counts : Some Improvements in the CD Index**" is being submitted in 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 2025
Place: IISc Bengaluru

Tejas Ashok Tonde
23138

Certificate

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

murugesan V

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

Abstract

Citation Disruption (CD) is a higher-order citation index that measures how much a scientific paper disrupts the citation network by weakening the direct linkage between its predecessors and successors, thereby capturing a unique dimension of scholarly impact: a paper's ability to shift knowledge trajectories rather than simply accumulate citations. Originally, the CD_t index ranging from -1 to +1 where t refers to the time delay between the publication and the assessment, was observed to be useful to distinguish works that consolidate existing knowledge (-1) or significantly disrupt existing knowledge by producing new ideas (+1). But it was later noted that for larger values of time t the CD_t index is quite generally an effective asymptotic indicator of the scientific significance, and not necessarily only an indicator that distinguishes disruption and consolidation for small t. First, this study investigates the limitations of relying solely on citation counts to assess scientific impact and proposes an improved CD index type metric for evaluating the significance of academic publications. Traditional citation-based indicators often fail to reflect the structural and temporal context in which citations occur.

Initially, the original CD index formulation was applied to the MAG240 dataset to analyze temporal trends and correlations between CD index values and citation counts within each subject of research. The results revealed that highly cited papers are not always the most disruptive, highlighting the limitations of citation count as a standalone metric. Building on this, we propose enhancements to the CD index by incorporating two contextual factors: the temporal gap between a focal paper and its predecessors, and the number of common predecessors between the focal paper and its successors. The improved CD index was then applied to a stratified sample of 1000 papers across various fields of study to demonstrate its suitability.

Contents

Chapter 1

Citation Count and CD Index

1.1 Introduction

In recent years, there has been growing concern that the pace of transformative scientific and technological breakthroughs is slowing down. Unlike the twentieth century, which witnessed revolutionary advances such as the discovery of quantum mechanics, the development of antibiotics, the Internet, and space exploration, the present era is marked majorly by gradual improvements than fundamental innovations. This decline is occurring despite a significant increase in research investments and the number of active scientists, suggesting that the effectiveness of producing impactful discoveries is decreasing.

This trend is visible in multiple indicators. Important scientific work now takes longer to gain recognition and fewer contributions seem to open new directions or challenge current knowledge. Many papers receive large numbers of citations, yet their influence tends to extend existing ideas rather than introduce new ones. As a result, traditional measures such as citation counts are being questioned for their ability to reflect true scientific contribution.

Research environments today are often shaped by performance metrics that reward volume and short-term influence. Metrics such as citation numbers and author indexes are frequently used to evaluate academic success, even though they may not fully represent the depth or originality of research. These indicators often ignore the structural role of a paper within the broader knowledge network and do not reflect whether the paper has changed how future work is shaped.

To address these limitations, the Citation Disruption index was introduced as a metric that measures how much a paper changes the flow of knowledge. This index calculates whether a paper weakens the direct connections between its past references and future citations if it is removed from the citation network, offering a better view of its potential to redirect research paths. It is usually measured at a fixed time after publication, with values near 1 showing disruption and values near -1 indicating knowledge reinforcement.

However, it has been seen that some papers which appear to be non-disruptive shortly after publication can become more disruptive as time passes. This finding shows that the influence of research work is not fixed and can grow significantly over time. It underlines the importance of re-examining the value of papers not only in the short-term but also over a longer period.

This thesis improves the Citation Disruption index by adding two new aspects: the time difference between a paper and its references, and the number of shared references between the paper and those that cite it. These changes help to capture how far a paper reaches across time and how it connects to future work. The improved index is tested on a sample of papers from different scientific areas, using part of the MAG240 dataset. The findings show that this enhanced method provides a better way to understand the significance of scientific contributions.

1.2 Literature survey

1.2.1 Citation Metrics and Their Limitations

Garfield's pioneering work laid the foundation for citation analysis. In 1972, [1] proposed citation counts as a systematic journal evaluation tool, which led to the development of the Journal Impact Factor. Later, [2] reflected on its misuse, emphasizing that JIF was designed for journal, not individual, assessment. [3] criticized the use of JIF for researcher evaluation, highlighting how citation distributions are skewed and unrepresentative. Similarly, [4] and [5] warned against misusing bibliometrics, with [5] proposing the Leiden Manifesto to promote responsible metric use. [6] explained the Matthew Effect in science, where recognition amplifies visibility and citations. [7] demonstrated how citation inflation over time distorts impact measures. Finally, [8] recommended combining metrics with peer review for balanced research assessment in policy-making.

1.2.2 H-index and Its Variants

H-index, combines productivity with impact by counting the number of publications with at least h citations. Recognizing its limitations, [9] proposed the g-index to give more weight to highly cited articles. [10] performed a comparative analysis of nine h-index variants using biomedical data, assessing their reliability and discriminatory power. [11] examined author-level impact over time using the Author Impact Factor, revealing dynamic patterns in citation accumulation.

1.2.3 Normalization and Field Differences

To address citation disparities across fields, [12] emphasized the need for field-normalized indicators. [13] introduced the Source Normalized Impact per Paper (SNIP), a measure designed to counter field-specific citation biases. [14] applied co-authorship network analysis to understand collaboration patterns and their influence on scientific output. International collaboration was studied [15], who explored how geographical distance and co-authorship shape research outcomes. [16] offered a country-level analysis of scientific impact, showing how national policies influence citation performance. [17] highlighted systemic advantages of US-based research in citation accrual.

1.2.4 Alternative Metrics

[18] and [19] explored the reliability of Mendeley and ResearchGate readership statistics, showing their utility as complementary metrics, though context-dependent. [20] focuses on theoretical debates about scientific contribution patterns and uses citation data to test the Ortega hypothesis.

1.2.5 Disruption and Citation Networks

[21] introduced the CD (Citation Disruption) index, which assesses how much a paper diverges from established citation patterns, distinguishing disruptive from consolidative work. [22] showed a general decline in disruptiveness in recent decades, based on CD5 metrics. [23] suggested that productivity in science is facing diminishing returns, requiring more effort per innovation. [24] explained that increasing knowledge burdens delay innovation and make it harder to achieve breakthroughs. [25] echoed this concern, ar-

guing that despite exponential paper output, true breakthroughs remain stagnant. [26] provided historical perspective on the global diffusion of science and the adoption of Western scientific practices. [27] proposes a field-normalized citation metric to enable fair comparison of citation counts across disciplines. It demonstrates that normalized citation distributions collapse into a universal form, suggesting a robust method for impact assessment.[28] outlines the principles, methodologies, and limitations of citation-based research evaluation. It highlights issues such as citation aging, field differences, and the need for responsible metric use.[6] noted how reputation and visibility heavily influence citation accumulation. [29] called for caution in interpreting CD metrics, showing how noise in citation data can distort disruptiveness scores.

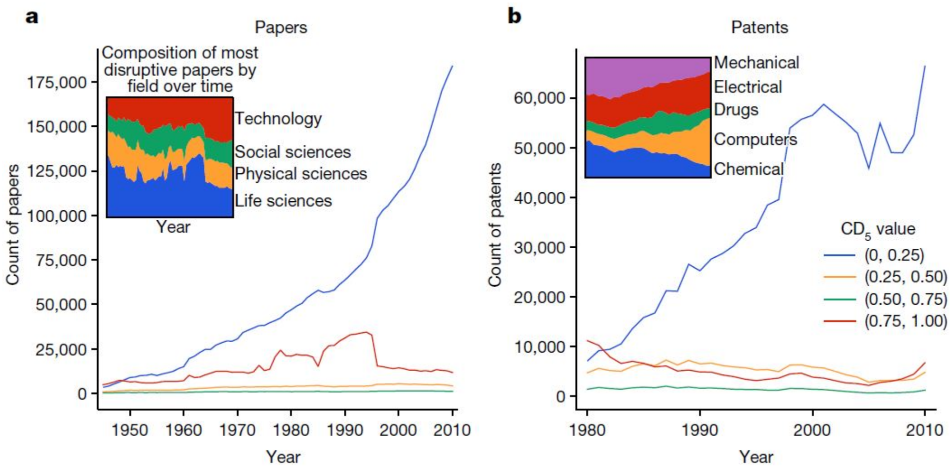


Figure 1.1: [22] The red and sky-blue lines (corresponding to higher CD5 ranges) remain relatively flat, indicating that the number of highly disruptive papers/patents remains constant, even as the total number of papers/patents (including less disruptive ones) increases exponentially.

1.3 Scope of study

This study aims to verify whether the conclusions drawn about the entire data set regarding the reliability of citation count and the correlation analysis between the CD index and citation count are applicable for individual fields of study too. Along with that, limitations of the original CD index have been identified and certain improvements have been proposed. Following are the keys points to be studied:

1. Analyze temporal dynamics of the CD index, restricting papers from the same field of study

2. Study the correlation between the CD index and citation count, within a specific field of study
3. Analyze the temporal trends of CD index of papers with high citation counts
4. Propose improvements in the CD index considering breadth and depth of the citation network
5. Evaluate the modified CD index across several fields of study

1.4 CD index

The CD (Citation Disruption) index is a quantitative metric designed to evaluate the disruptive potential of scholarly papers.[30]

CD index analyzes the citation relations of the focal paper being considered, using the papers cited by the focal paper and the papers citing the focal paper [21]. The papers citing the focal paper are called the successors and the papers cited by the focal paper are called predecessors. CD index is based on the idea that, if a paper has enough disruptive potential, successors should not have much need to cite the predecessors.[21, 22] If a paper is aggragating or summarizing the developments made by the predecessor, the successor will have to cite the focal paper as well as the predecessor. If such a citation network is a repeated for most of the successors of the focal paper, then that reflects in the CD index value of that focal paper[21, 22].

However, if the focal paper is introducing a new development in knowledge, there will be lesser chances of the successor citing the predecessor along with the focal paper. If such a citation network is observed for most of the successors of the focal paper, the CD index will have a positive value. More the successors which do not have any common predecessor, higher the magnitude of the CD index along with it being positive.

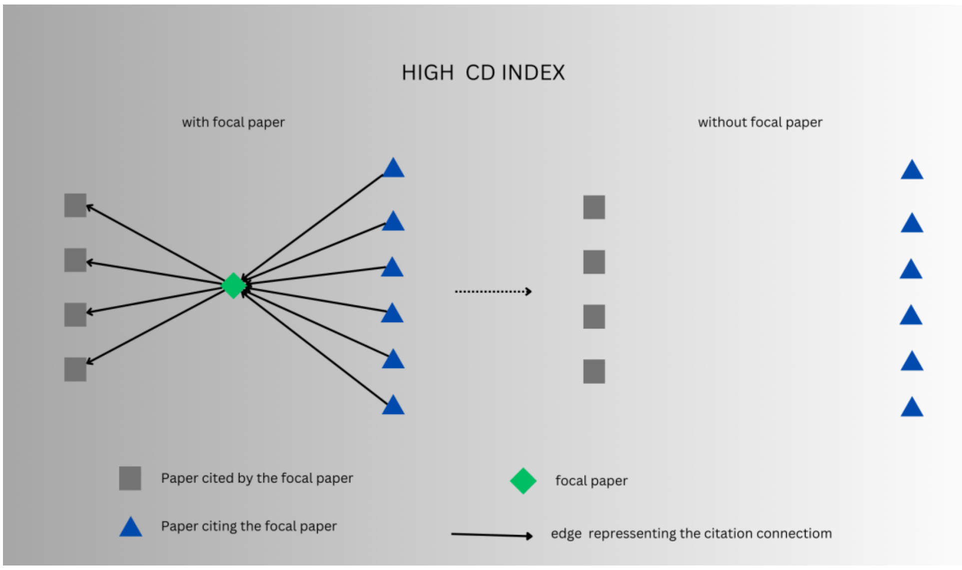


Figure 1.2: [30] Citation network of paper with high CD index

If the focal paper is removed, the citation network does not allow the citations to connect the successors to the predecessor via focal paper. Here, the focal paper is essential for the knowledge flow from the predecessors to successors.

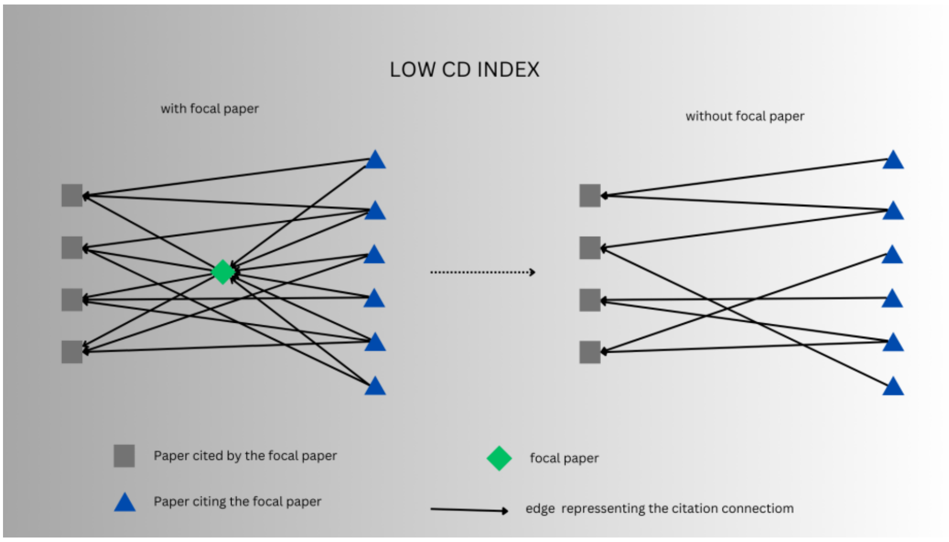


Figure 1.3: [30] Citation network of paper with low CD index

If the focal paper is removed, there isn't significant change in the connections between the successors and the predecessors. Here, the focal paper is not as essential as the previous case, for the knowledge flow from the predecessors to successors.

A paper with a positive CD index indicates that the ideas presented by the focal paper have not been much encountered in the earlier literature . This is reflected when there are less number of successors which have at least one common predecessor. Such papers introduce novel ideas and a redirect the citation flow. A paper with a negative

CD index indicates that there is consolidation or repetition of the existing ideas which does not change the citation network much[21].

1.4.1 CD index formula

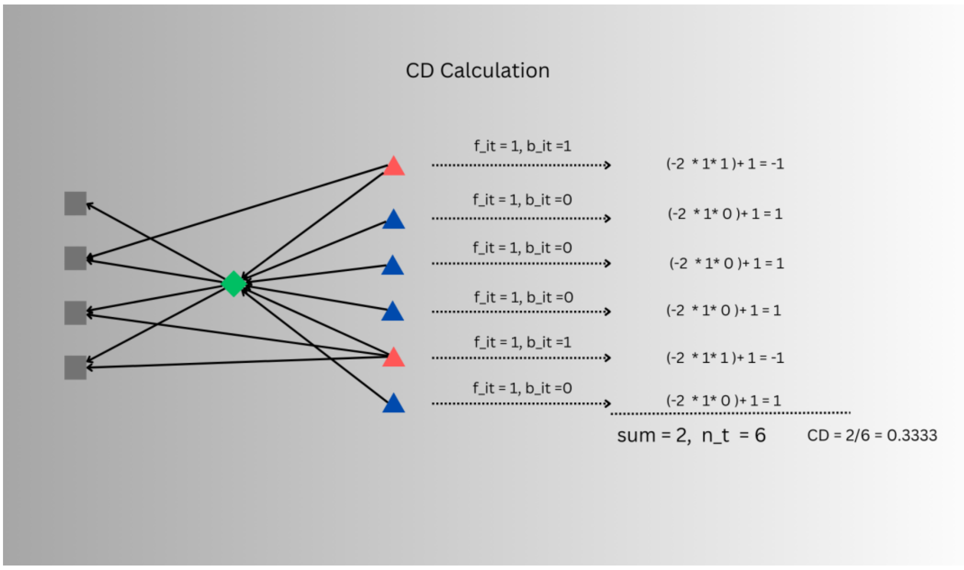


Figure 1.4: [30, 21] CD index formula

CD index is the mean of the contributions of the successors.

- If successor cites at least one predecessor and the focal paper, the successors contributes a score of -1
- If successor does not cite any predecessor but just the focal paper, the successor contributes a score of +1

Following is the formula of CD index [30, 21]:

$$CD_t = \frac{1}{n_t} \sum_{i=1}^{n_t} -2 f_{it} \cdot b_{it} + f_{it} \tag{1.1}$$

where,

- f_{it} : 1 if i cites the focal paper; 0 if not
- b_{it} : 1 if i cites at least one predecessors of the focal paper; 0 if not
- n_t : number of successors of the focal paper as of t years after its publication

Figure 1.4 depicts the calculation of the CD index.

CD index formula can be simplified if we consider only the successors which cite at least one predecessor[30]

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

f_{it} is always 1 as we are considering papers which cite at least the focal paper[21].

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

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

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

[21] $$CD_t = 1 - 2 \cdot \frac{n_{\mathrm{s}}}{n_t} \tag{1.2}$$

where,

- n_s : Number of successors which have cited at least one predecessor

1.5 Data

In this study, we utilize the Open Graph Benchmark (OGB) MAG240 dataset as our primary source of data. The MAG240 dataset, derived from the Microsoft Academic Graph (MAG), offers a rich collection of academic publications, author and institution affiliations, and citation relationships. It spans multiple disciplines, providing a diverse and comprehensive foundation for scholarly analysis.

The dataset includes approximately 1.3 billion citation links among over 121 million academic papers, offering a detailed view of scholarly interconnectedness. Each paper record also includes metadata such as publication year. This extensive coverage allows for in-depth exploration of the relationship between citation count and the CD index, along with the temporal patterns in the emergence of disruptive works among highly cited papers. Additionally, the dataset provides a 768-dimensional embedding vector for each paper, derived from its title and abstract.

Chapter 2

Improvements in the CD index

2.1 Need for improvement in the CD index

While the original Citation Disruption (CD) index, as per equation (1.1), effectively differentiates between consolidating and disruptive research, it is binary and simplistic in its formulation. The current formula assigns +1 to b_{it} when a successor cites the focal paper but not any of its predecessors, and assigns -1 to b_{it} when a successor cites both the focal paper and at least one of its predecessors.This binary approach, though intuitive, doesn't incorporate deeper structural and temporal nuances of the citation network.

2.1.1 Lack of citation breadth awareness

The original CD index only checks if a common predecessor is cited, but not how many common predecessors are cited. For example, a successor citing 1 out of 20 predecessors is treated the same as the one citing 19 out of 20. This limits the index's sensitivity to the breadth of overlap between successor and predecessor citations.

2.1.2 Neglecting temporal dynamics

The original CD index does not account for how old or recent the cited predecessors are relative to the focal paper. However, this temporal distance must be considered as an important factor for innovation and novelty. A paper that cites very old knowledge is likely to be more groundbreaking than the one slightly extending recent work. Ignoring this, the original CD index misses important indicators of disruptiveness.

2.2 Proposed improvements in the CD index

The above mentioned limitations mean that the original CD index may under represent papers that disrupt knowledge across distant or numerous citation links. To address this, we introduce an improved CD index formula:

$$CD_t = \frac{1}{n_t}\sum_{i=1}^{n_t} -2(1 - e^{-\alpha p_{it}\sum_{j=1}^{p_{it}} \frac{\lambda^{y_j} e^{-\lambda}}{y_j!}}) + 1 \tag{2.1}$$

- p_{it} : Number of common predecessors between the focal paper and a successor
- y_j : Time gap between the publication year of the focal paper and its predecessors
- λ : Mean y_j of a specific focal paper
- n_t : Number of successors of a focal paper
- α : Front factor

2.3 Rationale for Using Poisson Distribution in Improved CD Index

The traditional CD index evaluates disruptiveness by simply categorizing each successor based on whether it also cites any predecessor of the focal paper. While effective in capturing direct patterns of citation overlap, this binary treatment lacks nuance and fails to differentiate between successors that cite one predecessor versus several predecessors.To introduce a more probabilistic and refined understanding of citation behavior, we incorporate the poisson probability distribution in the improved formulation.

The Poisson distribution models the probability of a given number of events occurring in a fixed interval, given a known constant average rate. In our case, this concept is adapted to model the likelihood of citing a predecessor that is y_j years old.

$$\frac{\lambda^{y_j} e^{-\lambda}}{y_j!} \tag{2.2}$$

Older papers are less likely to be cited, not because they are irrelevant, but because newer research tends to build on more recent work. The Poisson distribution inherently captures this decay, i.e. the probability of observing a citation to a paper y_j years old declines as y_j increases. This allows the improved CD index to penalize consolidation more when the predecessor is recent, which is a strong signal of incremental work. On the other hand, citing an old predecessor (with low Poisson probability) does not strongly suggest consolidation. Rather, it may indicate foundational grounding, which should be less penalized. Thus, Poisson weighting reflects that citing very old papers is more acceptable, whereas citing many recent predecessors suggests weaker novelty.

The original CD index uses a binary logic, i.e., either the successor cites a predecessor or it doesn't, and every such citation gets a penalty of -1. In the improvement, instead of flat penalties, each predecessor's influence is modulated by its Poisson probability. This gives a interpretable measure of how expected or anomalous a citation is, based on temporal difference.

This formula defines λ as the mean of the y_j values for a focal paper. This is because each focal paper has its own temporal context. For example, in some fields, citations to older literature might be more common. By using a local λ, the Poisson probabilities are normalized to each paper's citation environment, making the metric adaptive rather than global. This prevents unfair penalization of papers in slow-moving fields or those building on older but valid literature. Also, poisson distribution is well suited as it has a single tunable parameter, λ , which is the mean and variance.

2.4 Rationale for Using a front factor α in Improved CD index

When we say a predecessor is cited at its expected age (i.e., $y_j = \lambda$), we're referring to statistical expectation of how often papers of that age are normally cited across all papers that share a similar context (here, the same focal paper's λ). So if successors of the focal paper predominantly cite predecessors whose ages exactly match this average, the behavior is statistically typical.

Novelty should deviate from pre-existing norms. Disruptive papers push the field forward. Their successors might build on much older ideas (reviving neglected knowledge), or skip over direct predecessors altogether (true innovation). If successors cite only predecessors at the average expected age, it implies they are operating well within conventional citation behavior.

The CD index tries to analyze whether the focal paper tried to divert citations away from its predecessors. If successors still cite predecessors at expected ages, then the citation trajectory is undisturbed. This makes the focal paper consolidative, not disruptive. A Poisson peak (where $y_j = \lambda$) suggests the point where we expect most citations to happen. If all cited predecessors are right at that peak, then the successors are citing as expected, showing no abnormal behavior,no shift and no disruption. If successors are citing predecessors exactly at the statistically expected age, it reflects a continuation of conventional knowledge flows, not a break from them. Therefore, there's no novelty in citation behavior, indicating no structural innovation or disruption introduced by the focal paper.

Recall the formula proposed for the improved CD index:

$$CD_t = \frac{1}{n_t} \sum_{i=1}^{n_t} -2(1 - e^{-\alpha p_{it} \sum_{j=1}^{p_{it}} \frac{\lambda^{y_j} e^{-\lambda}}{y_j!}}) + 1 \tag{2.3}$$

For the papers where above argument holds, the term inside the summation, that is, $-2(1 - e^{-\alpha p_{it} \sum_{j=1}^{p_{it}} \frac{\lambda^{y_j} e^{-\lambda}}{y_j!}}) + 1$ should tend to -1. Therefore, if $y_j = \lambda$,

$$\begin{aligned}
-2(1 - e^{-\alpha p_{it} \sum_{j=1}^{p_{it}} \frac{\lambda^{y_j} e^{-\lambda}}{y_j!}}) + 1 \to -1 & \\
\Rightarrow \left(1 - e^{-\alpha p_{it}^2 \cdot \frac{\lambda^{\lambda} e^{-\lambda}}{\lambda!}}\right) \to 1 & \\
\Rightarrow \quad -e^{-\alpha p_{it}^2 \cdot \frac{\lambda^{\lambda} e^{-\lambda}}{\lambda!}} \to 0 & \\
\Rightarrow \quad \alpha p_{it}^2 \cdot \frac{\lambda^{\lambda} e^{-\lambda}}{\lambda!} \gg 1 & \\
\Rightarrow \quad \alpha \gg \frac{\lambda! \cdot e^{\lambda}}{p_{it}^2 \cdot \lambda^{\lambda}} &
\end{aligned}$$

It has been observed that the expression $\frac{\lambda! \cdot e^{\lambda}}{p_{it}^2 \cdot \lambda^{\lambda}} \lesssim 1$ with values of λ in the range (0,5) for several fields of study. Hence, α was assumed to be 10 as a front factor in the modified CD index formula. Note that the index is only weakly dependent on the value of α used, if it satisfies the above criterion.

Chapter 3

Results and Conclusion

3.1 Results

Several observations were made to compare citation count and CD index for reliability of knowledge disruption. These observations were done across various fields of study and compared with observations considered over entire data set. In addition, the improved CD index was calculated for at most 1000 paper samples for each field of study. Few representative results have been shown and discussed below.

3.1.1 Correlation analysis

Following heatmaps show the correlation between the citation counts and the CD index for 5 different fields of study. The papers in each field of study have been divided into 9 categories, based on the citation count and the CD index.

- Low CD Index (CD < -0.3), Low Citation Count (< 10)
- Low CD Index, Mid Citation Count ($10 \leq$ Citation Count ≤ 40)
- Low CD Index, High Citation Count (> 40)
- Mid CD Index ($-0.3 <$ CD < 0.3), Low Citation Count
- Mid CD Index, Mid Citation Count
- Mid CD Index, High Citation Count
- High CD Index (CD ≥ 0.3), Low Citation Count
- High CD Index, Mid Citation Count
- High CD Index, High Citation Count

Correlation between the citations and CD index was found for each group. Majority of the groups showed non-correlated relationships. This analysis emphasizes the need to consider multiple dimensions of impact for 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.

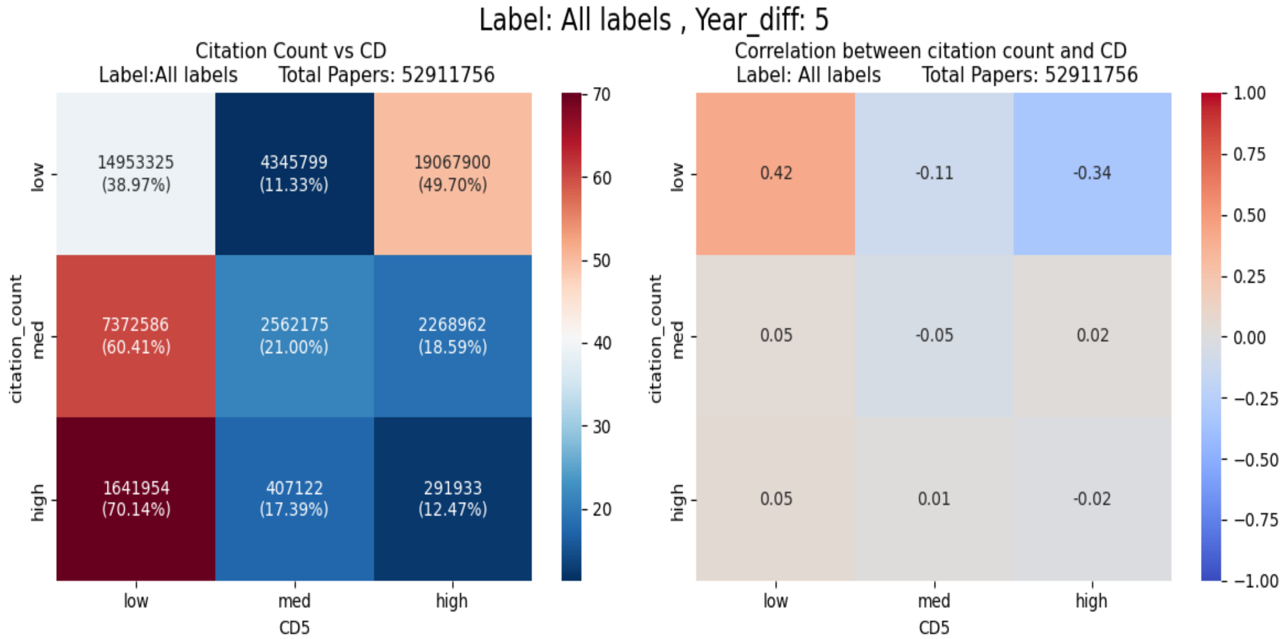


Figure 3.1: Number and percentage of papers with low, medium and high CD5 values for each category of Citation count.

High citation count category has the lowest number of papers compared to medium and low citation count category. Block with low citation count and low CD5 has the highest number of papers and block with high citation count and medium CD5 has the lowest number of papers.

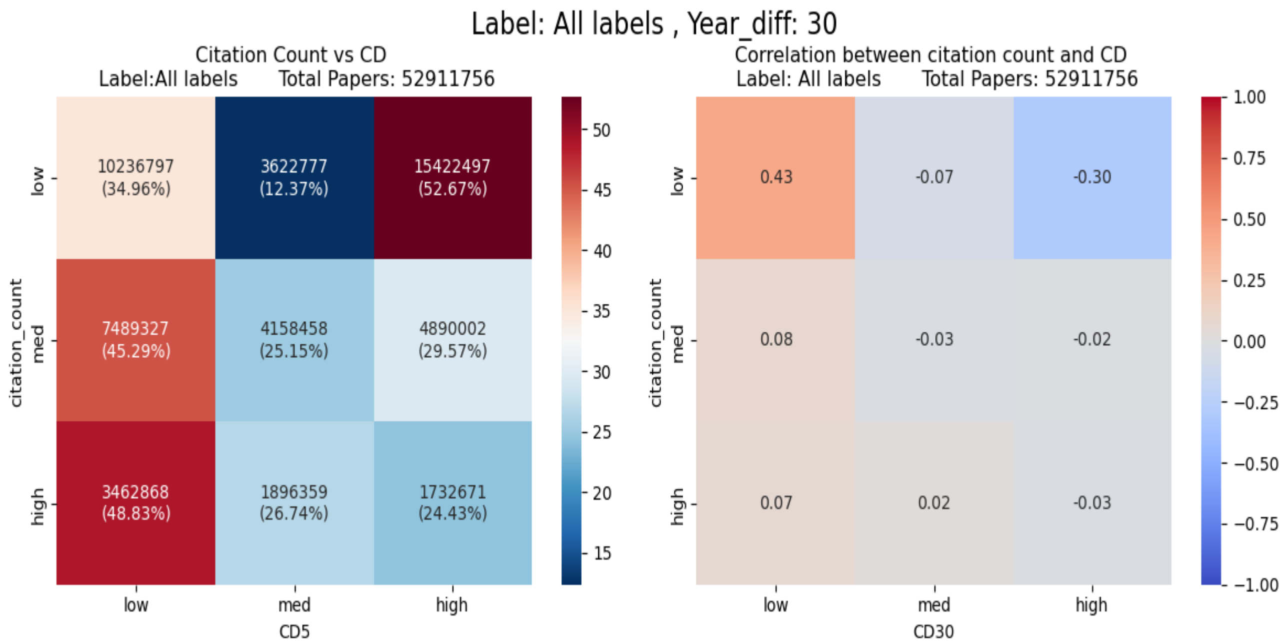


Figure 3.2: Number and percentage of papers with low, medium and high CD30 values for each category of Citation count.There is an increase in the percentage of papers with high CD30 value compared to that of CD5

Figure 3.1 and fig. 3.2 show the results about correlation between the CD index and the citation counts for the entire dataset as indicated by previous work done by Naga Narasimharao Gadidamalla[30].

We tried to analyze whether similar results are observed for papers across several fields of study. Hence,5 fields of study with labels; 14,34,50,90,97 have been chosen.

Correlation analysis between the citation counts and the CD index suggests there is low correlation coefficient in most of the 9 categories mentioned earlier. From the 9 categories, the category with high citation counts and high CD index has significantly low correlation coefficient, indicating high citation count does not imply high disruptive potential. Along with this, majority of the papers belong to the class of low citation count. This results is seen not only for the entire data set considered together, but also across 5 fields of study chosen. Refer figures fig. 3.3, fig. 3.5, fig. 3.7, fig. 3.9 and fig. 3.11.

The low correlation indicates that the citation count is not able to capture the true disruptive potential of the papers, but rather giving a benefit to the papers which summarize or consolidate the existing knowledge and don't lead way to significant improvements or innovations.

Figure 3.4 , fig. 3.6 , fig. 3.8 , fig. 3.10 and fig. 3.12 show the CD5 and CD30 distributions of papers in the category high citation count and high CD. The peak at +1 is likely an artifact of the incompleteness of the citation graph. In these figures, we can see that, apart from the peak at +1, most of the CD values have insignificant frequency.

Refer to figure fig. 3.13 which shows the distribution of CD30 of papers across all fields of study for the two categories, namely, 1)high citation count and high CD30 and 2) high citation count and low CD 30. In the fig. 3.13, histogram showing the category of high citation count and high CD30 has near-zero correlation. This shows that papers which are highly cited and have high CD30 value have near-zero correlation indicating that citations may be useless as a indicator of high impact work. In the same figure, the histogram for correlation coefficient for papers in the category of high citation count and low CD30 shows a slight positive correlation. This shows that papers which are highly cited and have low CD30 are slightly positively correlated indicating that citation counts can be counter productive as an indicator.CD5, CD10, CD15, CD20, CD25 and CD2021 present similar conclusions.

Note that few large set labels are chosen for initial detailed discussion where correlation values across all 9 blocks are presented.

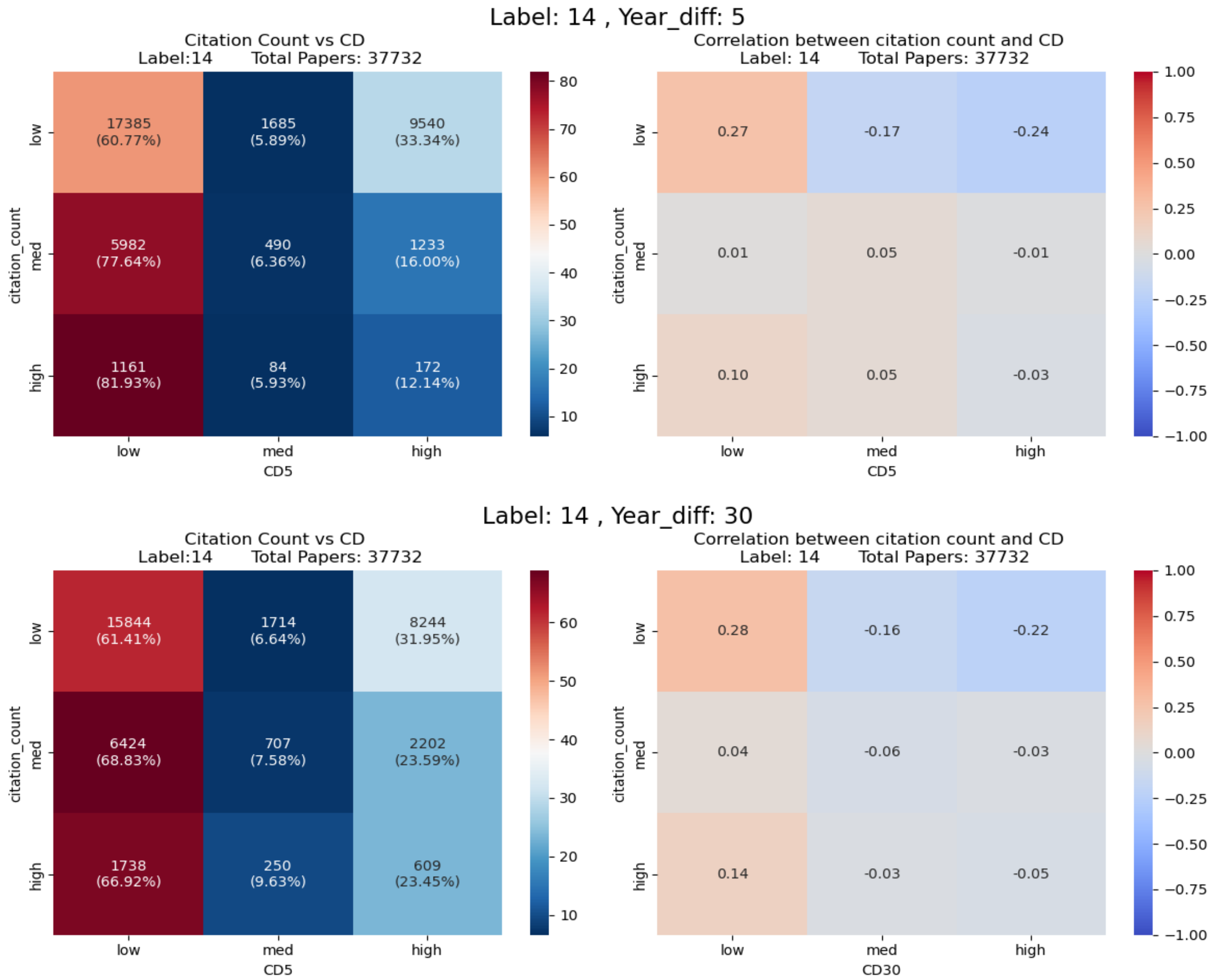
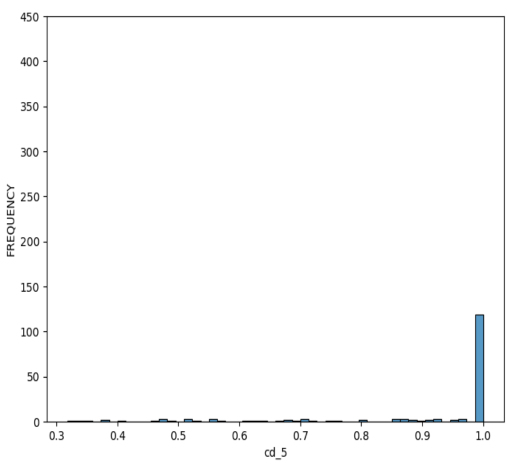
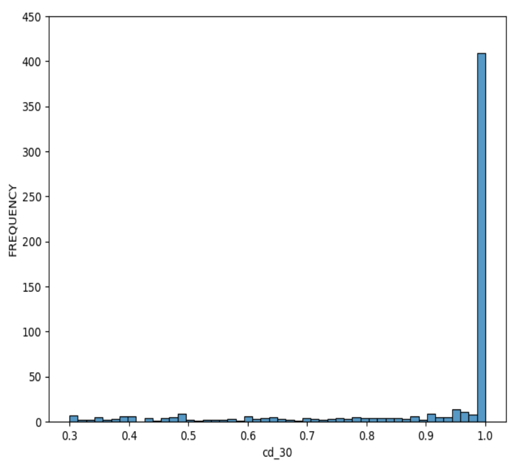


Figure 3.3: Label:14 - Number and percentage of papers as per the 9 categories and correlation coefficient of each category for CD5 (top) and CD30 (bottom)



CD5 | Number of papers: 172



CD30 | Number of papers: 609

Figure 3.4: Label:14 CD of papers in high CD and high citation category

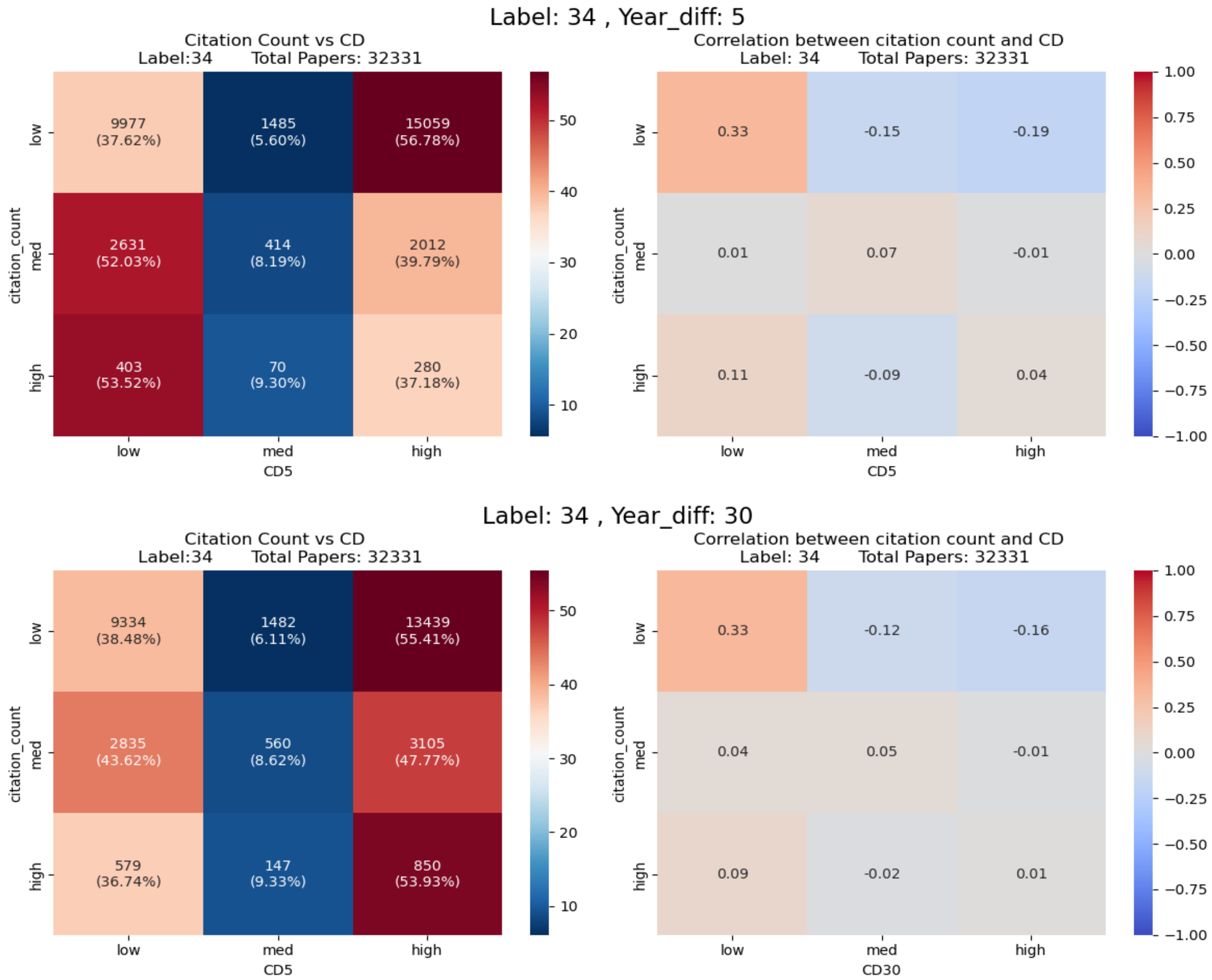
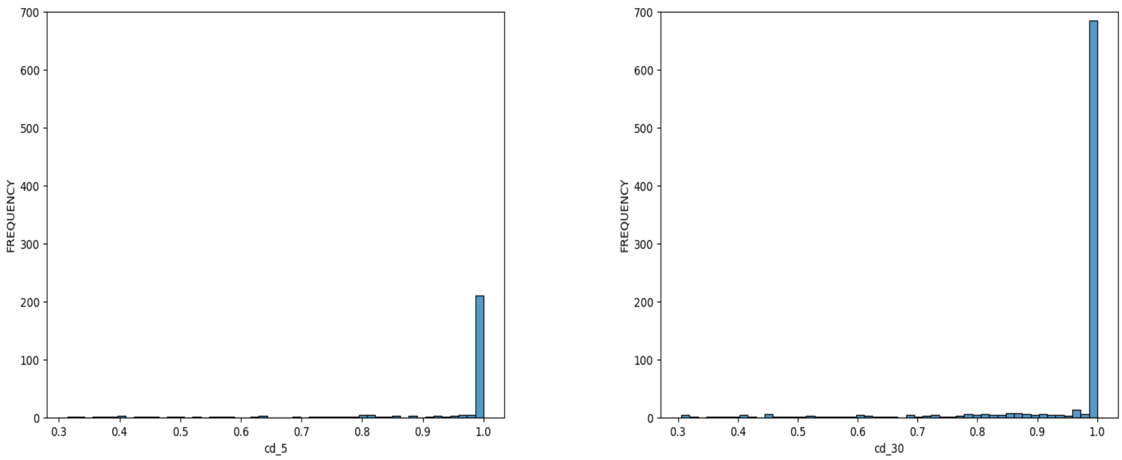


Figure 3.5: Label:34 - Number and percentage of papers as per the 9 categories and correlation coefficient of each category for CD5 (top) and CD30 (bottom)



CD5 | Number of papers: 280 CD30 | Number of papers: 850

Figure 3.6: Label:34 CD of papers in high CD and high citation category

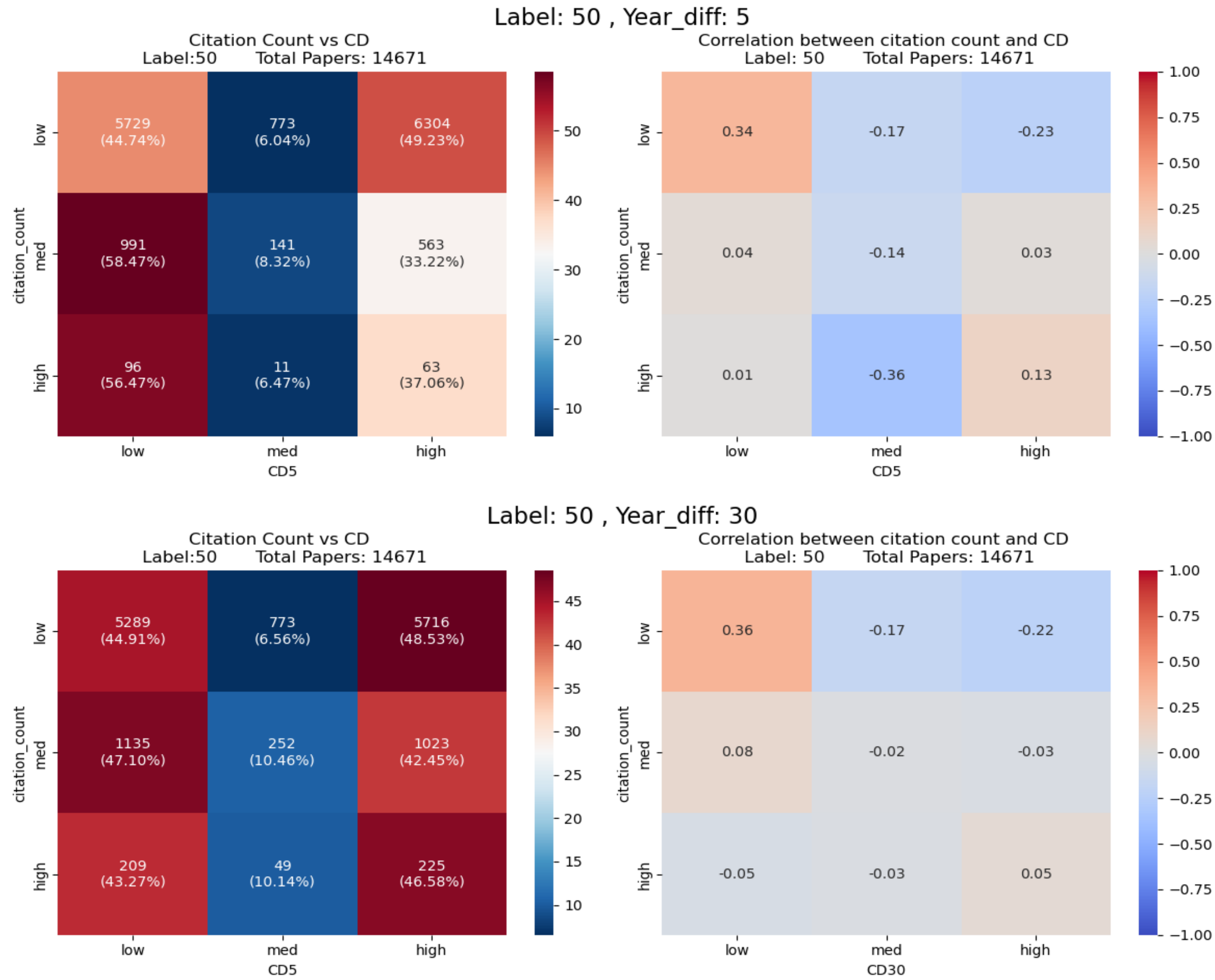
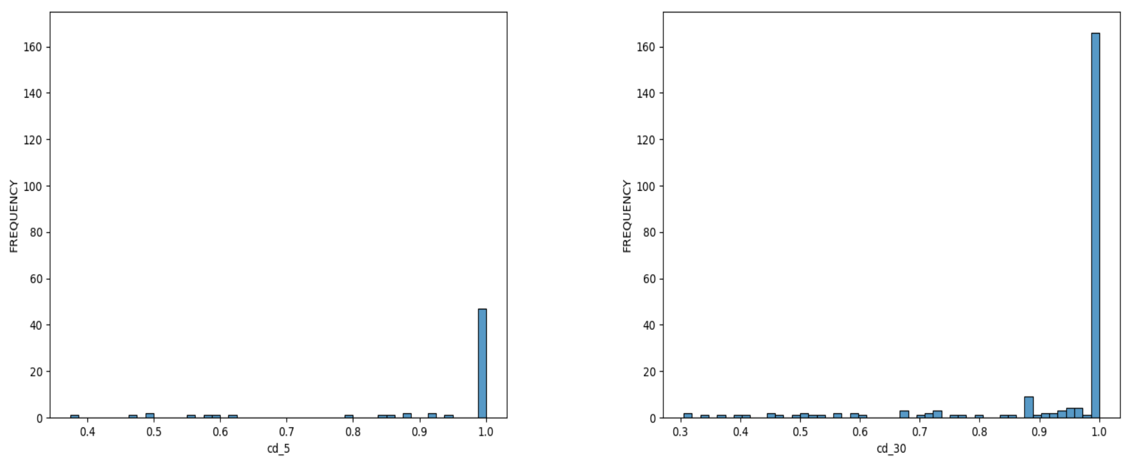


Figure 3.7: Label:50 - Number and percentage of papers as per the 9 categories and correlation coefficient of each category for CD5 (top) and CD30 (bottom)



CD5 | Number of papers: 63 CD30 | Number of papers: 225

Figure 3.8: Label:50 CD of papers in high CD and high citation category

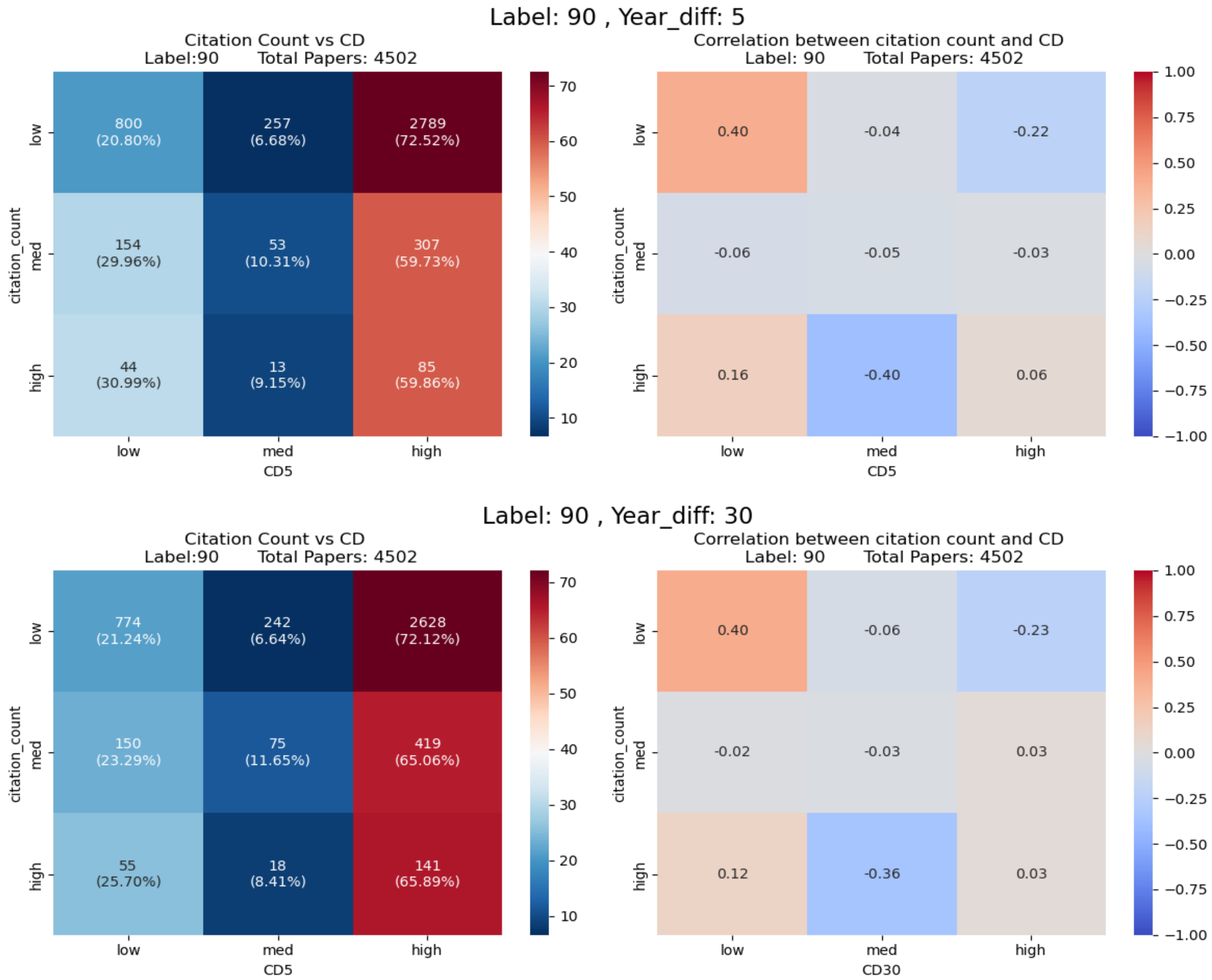
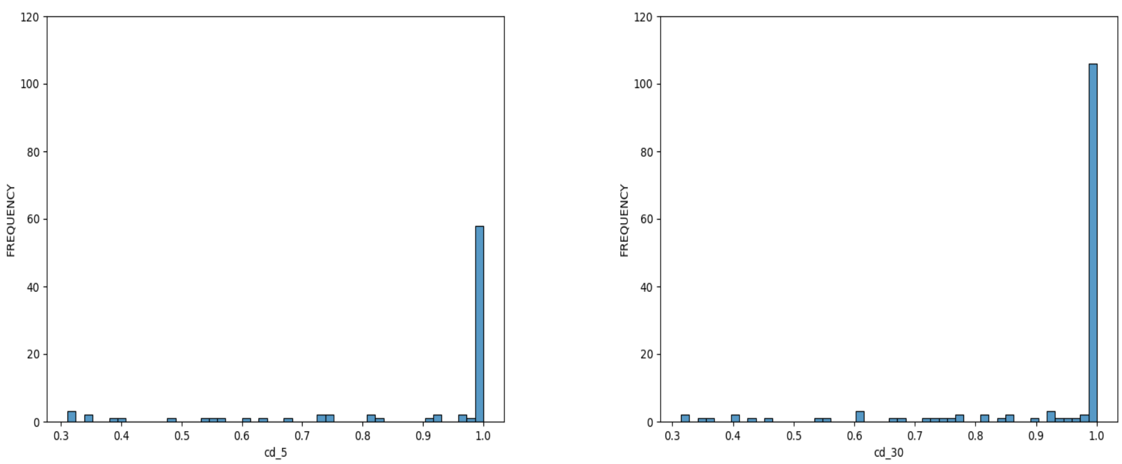


Figure 3.9: Label:90 - Number and percentage of papers as per the 9 categories and correlation coefficient of each category for CD5 (top) and CD30 (bottom)



CD5 | Number of papers: 85 CD30 | Number of papers: 141

Figure 3.10: Label:90 CD of papers in high CD and high citation category

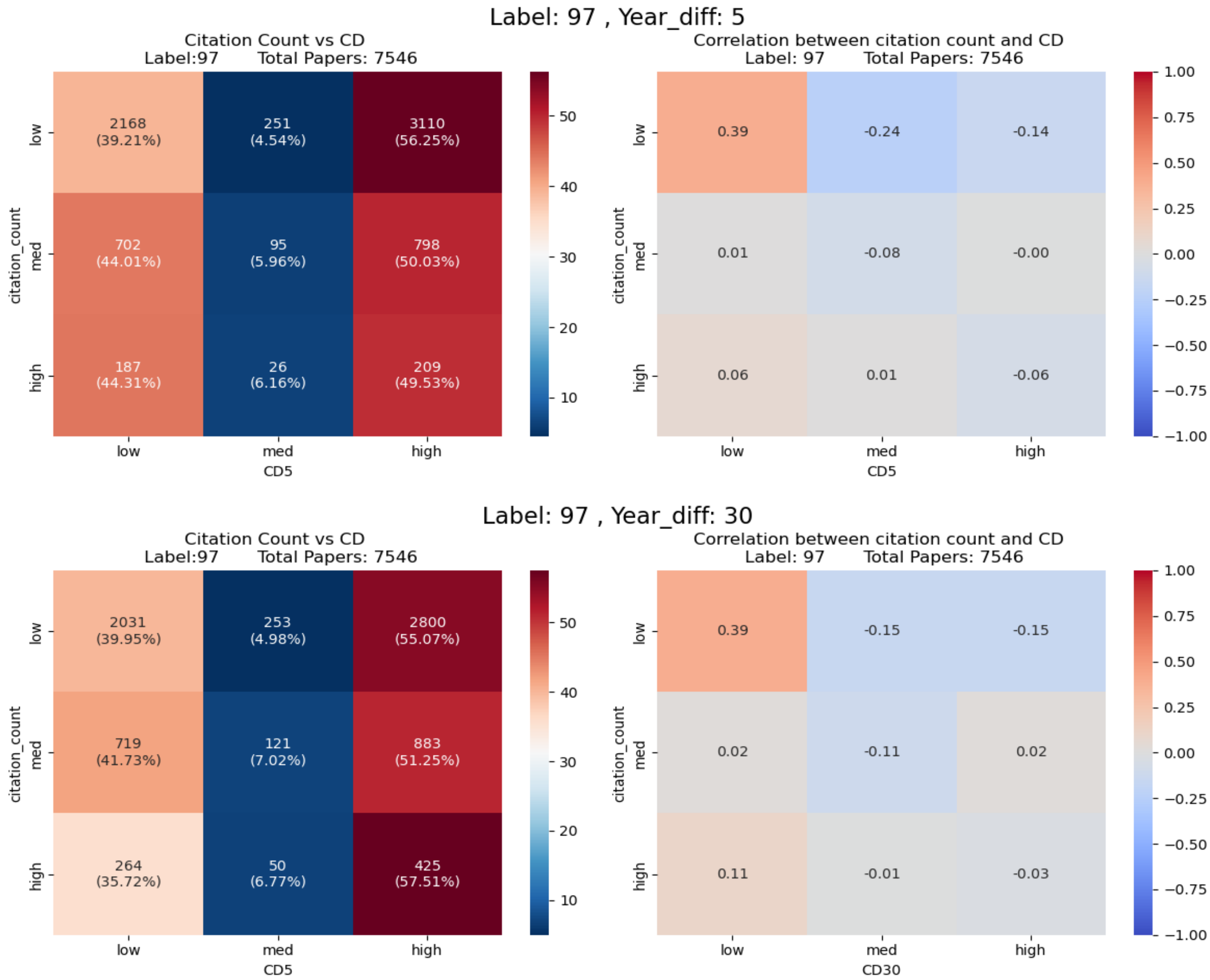
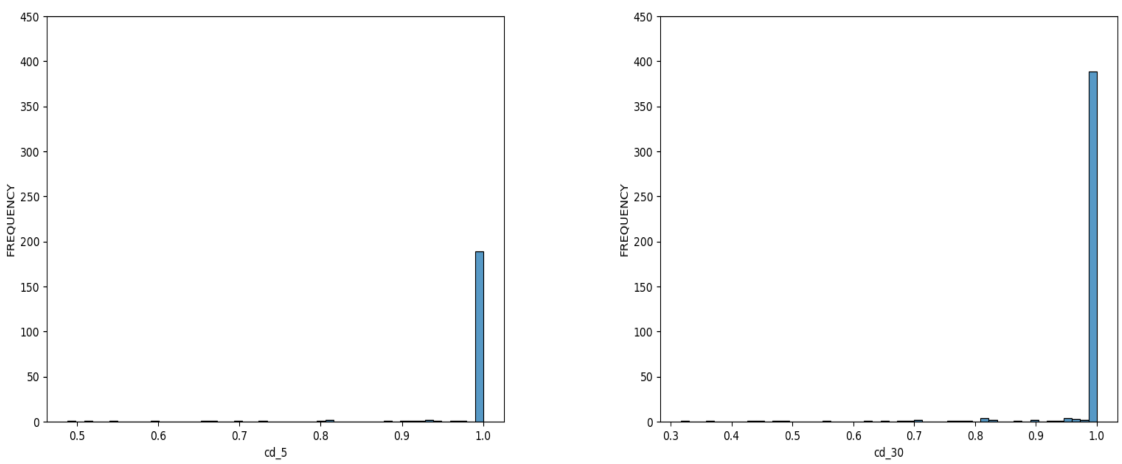
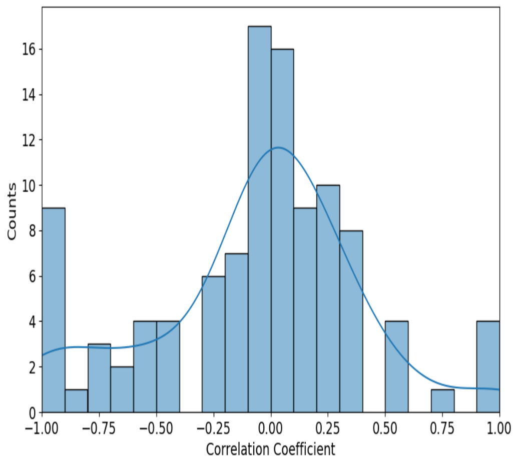


Figure 3.11: Label:97 - Number and percentage of papers as per the 9 categories and correlation coefficient of each category for CD5 (top) and CD30 (bottom)

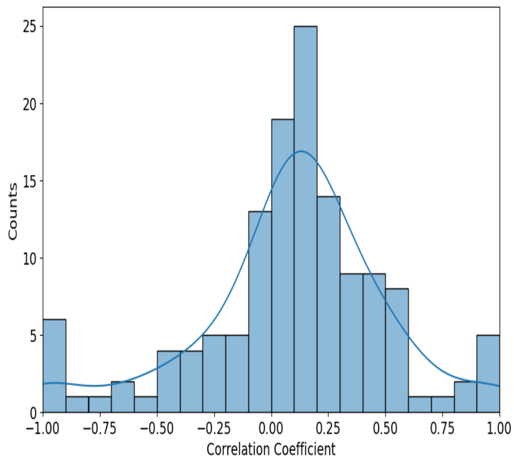


CD5 | Number of papers: 209 CD30 | Number of papers: 425

Figure 3.12: Label:97 CD of papers in high CD and high citation category



Distribution of correlation values among labels: High CD30 Vs. High Citations



Distribution of correlation values among labels: Low CD30 Vs. High Citations

Figure 3.13: Note that the smaller peaks observed at +1 and -1 is likely an artifact of the incompleteness of citation graph.

3.1.2 Average CD Index Trends

The trends of CD5, CD10, CD15, CD20, CD25, CD30 , CD2021 vary over the year, showing a decreasing trend. Figure 3.14 shows that, considering the entire data set, not only there is a decreasing trend but also a significant jump as we go from CD5 to CD10 , CD10 to CD15 and so on. Considering the 5 fields of study, labelled, 14,34,50,90,97, similar trends have been observed in the figures fig. 3.15, fig. 3.16, fig. 3.17, fig. 3.18 and fig. 3.19 . However,the jump does not seem to be significant as we go from CD5 to CD10 and so on.

The analysis of average CD index values across years reinforces the idea that scientific papers and patents are becoming progressively less disruptive. Our results indicate a steady decline in CD index scores, suggesting that more recent publications tend to reinforce existing knowledge frameworks rather than introduce transformative innovations. This trend is consistent with the findings of [22], who observed a growing preference for incremental advancements in academic research. Also, for most of the duration, CD5 is the bottom trendline , but there are instances where CD10 or CD15 cross-over CD5 and are below CD5.

In contrast to the overall dataset, the domain-specific plots (e.g., for labels 14, 34, 50, etc.) show more compact or overlapping trends across different CD index windows (such as CD5, CD10, CD15, etc.). This compactness arises primarily due to the homogeneity of citation behavior within a single field of study.

Each academic discipline typically has its own standardized pace of research, citation patterns and collaboration norms. For example, in fast-evolving fields like computer science or biomedicine, papers tend to get cited relatively quickly, while in more disciplines like mathematics or economics, citation accumulation is more gradual. However, within any one field, the variation in these patterns is limited compared to the global dataset.

Due of this internal consistency, the effect of extending the CD index window from 5 to 10 or 15 years doesn't drastically alter the structural role of the focal paper in the citation network. Whether disruptiveness is measured over 5 or 30 years, the same core group of successor papers continues to define the citation dynamics of the focal paper. Hence, the values of CD5, CD10, and CD15 tend to lie closer together, leading to overlapping or narrowly spaced trendlines.

In narrow domains, shorter-term CD metrics might be nearly as effective as longer-term ones. However, in interdisciplinary or general datasets, longer-term metrics (like CD30) are more effective at smoothing out noise and capturing the true structural impact of a paper.

Along with this, in specialized fields, successors are more likely to cite the same foundational works (predecessors) and new work is often built incrementally. This further reduces the variability in disruptiveness over different time windows, as the successors' citation patterns remain relatively stable. As a result, the variation in disruptiveness across different time windows is less noticeable within individual fields, unlike in the overall dataset where the presence of diverse research domains leads to a wider spectrum of citation behaviors.

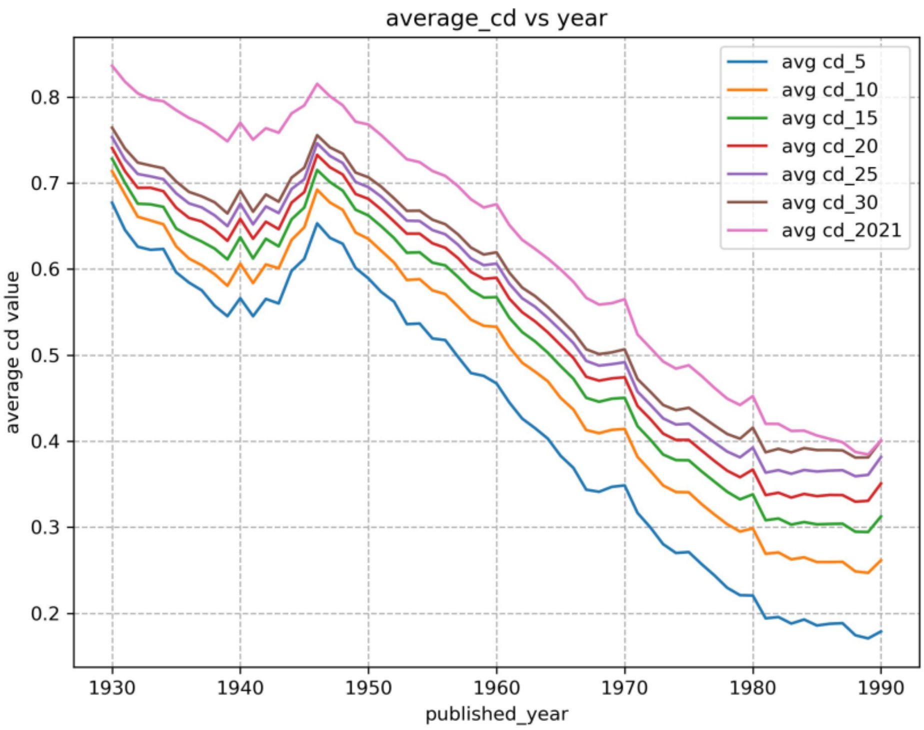


Figure 3.14: Trends of CD5, CD10, CD15, CD20, CD25, CD30, CD2021 varies over the years

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.

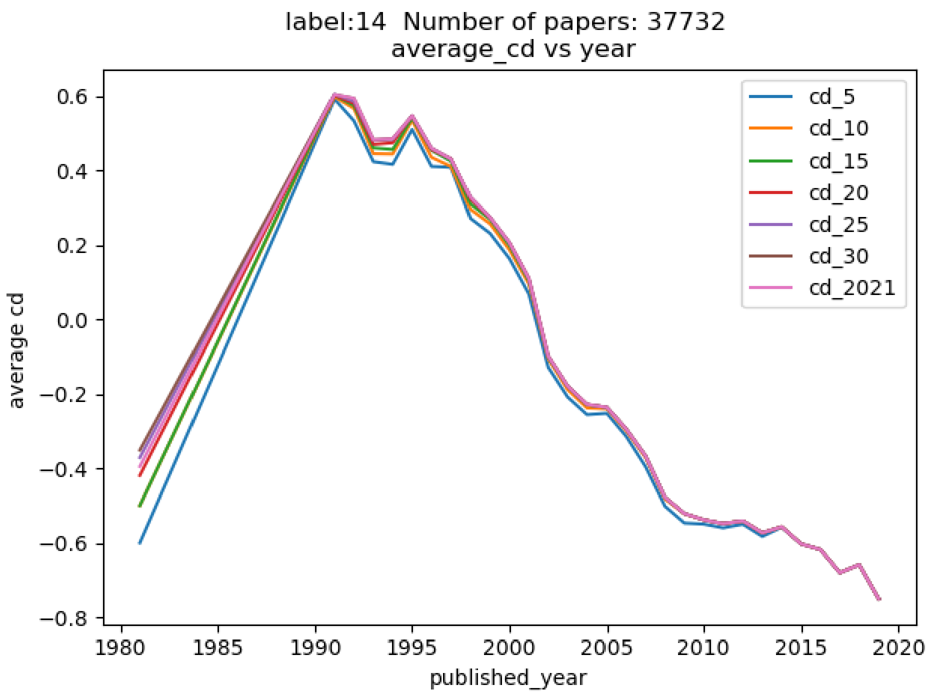


Figure 3.15: Label 14: Trends of CD

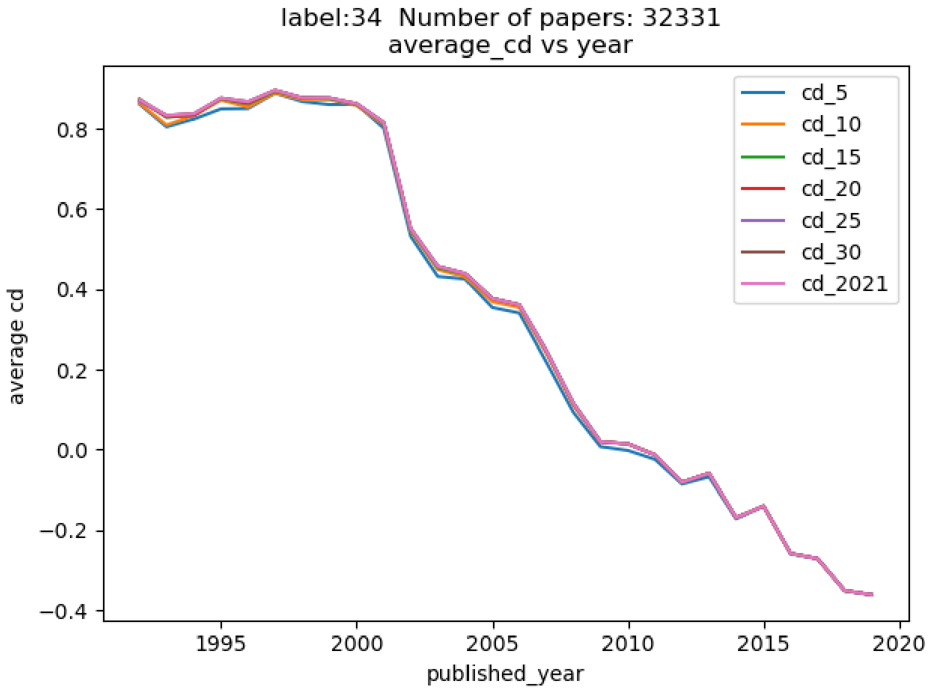


Figure 3.16: Label 34: Trends of CD

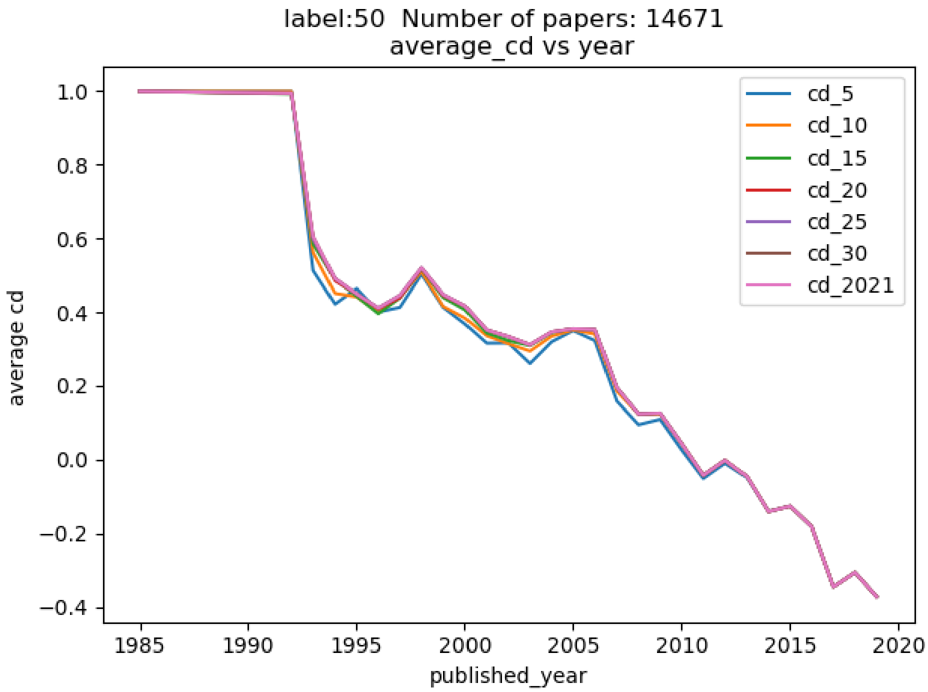


Figure 3.17: Label 50: Trends of CD

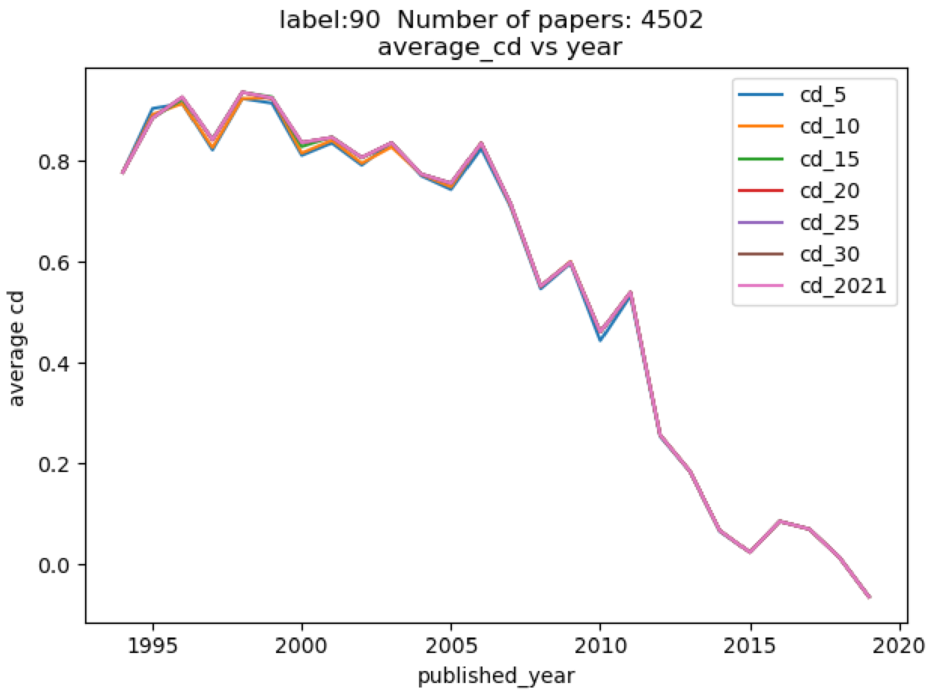


Figure 3.18: Label 90: Trends of CD

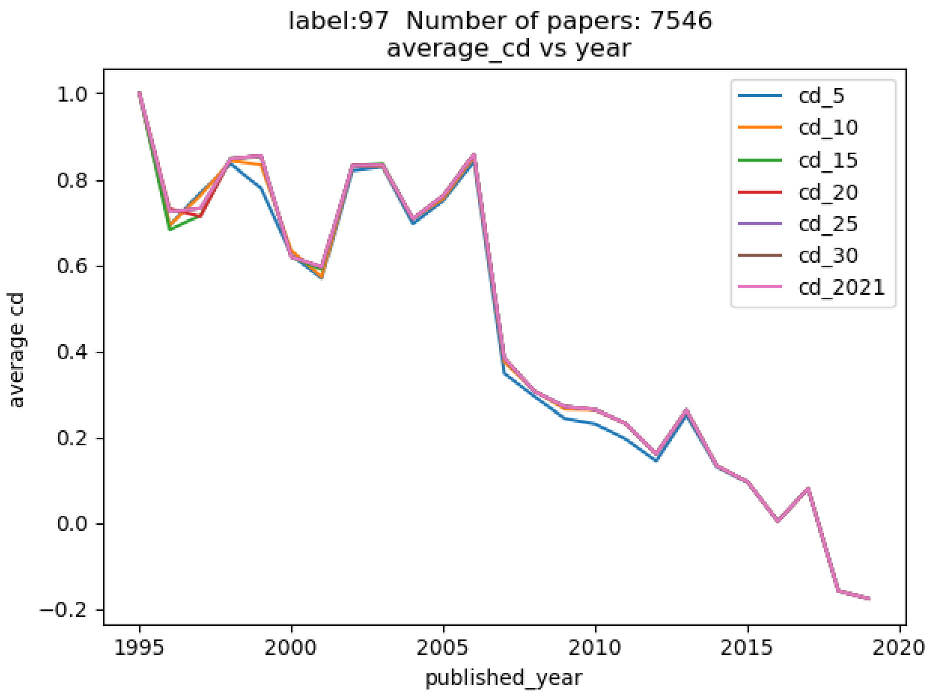


Figure 3.19: Label 97: Trends of CD

3.1.3 CD Index trends with time

Referring to fig. 3.20, our analysis of CD_t index values measured at t values 5, 10, 15, 20, 25, and 30 years post-publication revealed a notable trend: CD values tend to increase over time. This upward progression was especially prominent in the case of highly cited papers, indicating that works which initially appeared less disruptive can, over time, demonstrate significant disruptive influence. Along with this, in fig. 3.20 , we can see that the 3 out of 4 papers started with a negative CD index but asymptotically reached a significant positive value.

This observation underscores the idea that the impact of research is not static but can evolve as it interacts with future developments and shifts within the scholarly community. It emphasizes the importance of adopting a longitudinal perspective to accurately capture the true influence and disruptiveness of academic work.

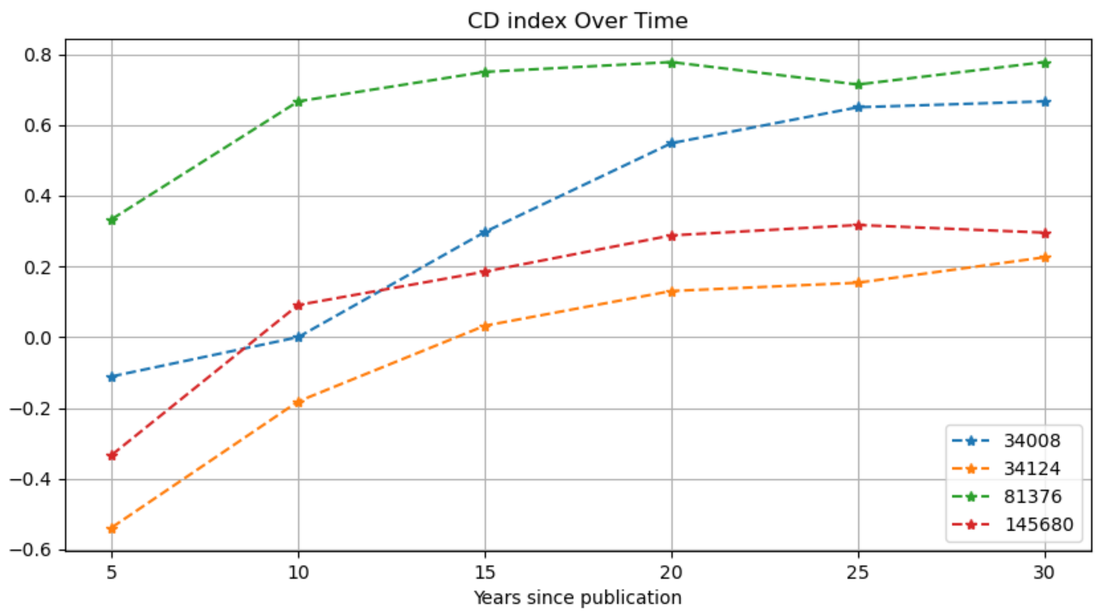


Figure 3.20: Above papers have citations greater than 100. CD index increases with time

3.1.4 Improved CD index

Improved CD index was applied on 1000 papers, sampled at random, from several fields of study. It was observed that the data set has papers with interesting citation network i.e. there exist such focal papers which have many successors which do not cite any of the focal papers. Results for papers from the field of study with labels 0, 14, 28 are shown in fig. 3.21 , fig. 3.23 and fig. 3.25. Its been observed that nearly 20% of the papers have all successors with no citation to predecessors. This effect can be seen in the histogram plots in the figures fig. 3.21 , fig. 3.23 and fig. 3.25 which shows high frequency for improved CD index value of 1. The papers which don't have any successor are assigned improved CD index value of 0. As their is a significant number of such papers, a slight peak in the frequency is seen at 0.

The correlation plots can be seen in figures fig. 3.22, fig. 3.24 and fig. 3.26. For papers with label 0 , fig. 3.22 , it can be seen that 7 out of 9 categories have negligible correlation coefficient between the citation count and the modified CD index. Papers belonging to the category with low citation count and high modified CD index have negatively correlated modified CD index and citation count, whereas papers belonging to low citation count and low modified CD index have a positive correlation coefficient. The skewness for these papers is 0.1091 .

For papers with field of study labeled 14, the skewness of the histogram plot is 0.286. As per previous field of study's observation, correlation heatmap fig. 3.24 shows low citation count and high modified CD are negatively correlated whereas papers with low modified CD index and low citation counts are positively correlated. Although the magnitude of correlation coefficient of these 2 categories is larger than other 7 categories, the magnitude is less than 0.5.

For papers with field of study labeled 28, the skewness of the histogram plot fig. 3.25 is 0.237. An interesting observation is that papers in the category of high citation count and high modified CD index are negatively correlated with highest magnitude of correlation coefficient among the other categories. Low citation count and high modified CD index papers are negatively correlated whereas low citation count and low modified CD index are positively correlated but both having correlation coefficient less than 0.5

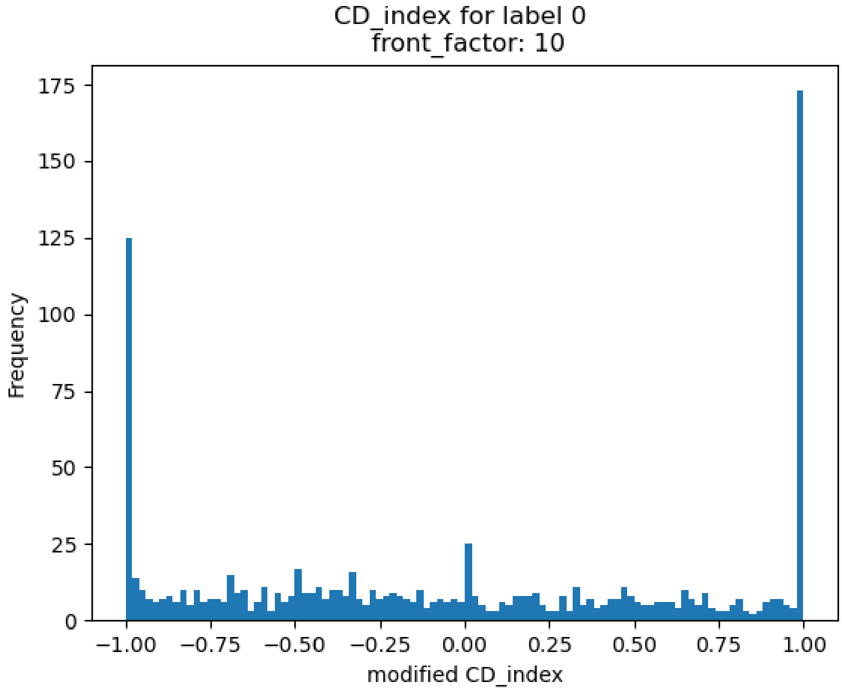


Figure 3.21: Label 0: Improved CD index values histogram.The peaks observed at +1 and -1 are likely an artifact of the incompleteness of citation graph. For the peak at +1, the key citation connections between the successors and predecessors might be absent while for -1, successors which actually don't have common predecessors might not be connected to the focal paper.

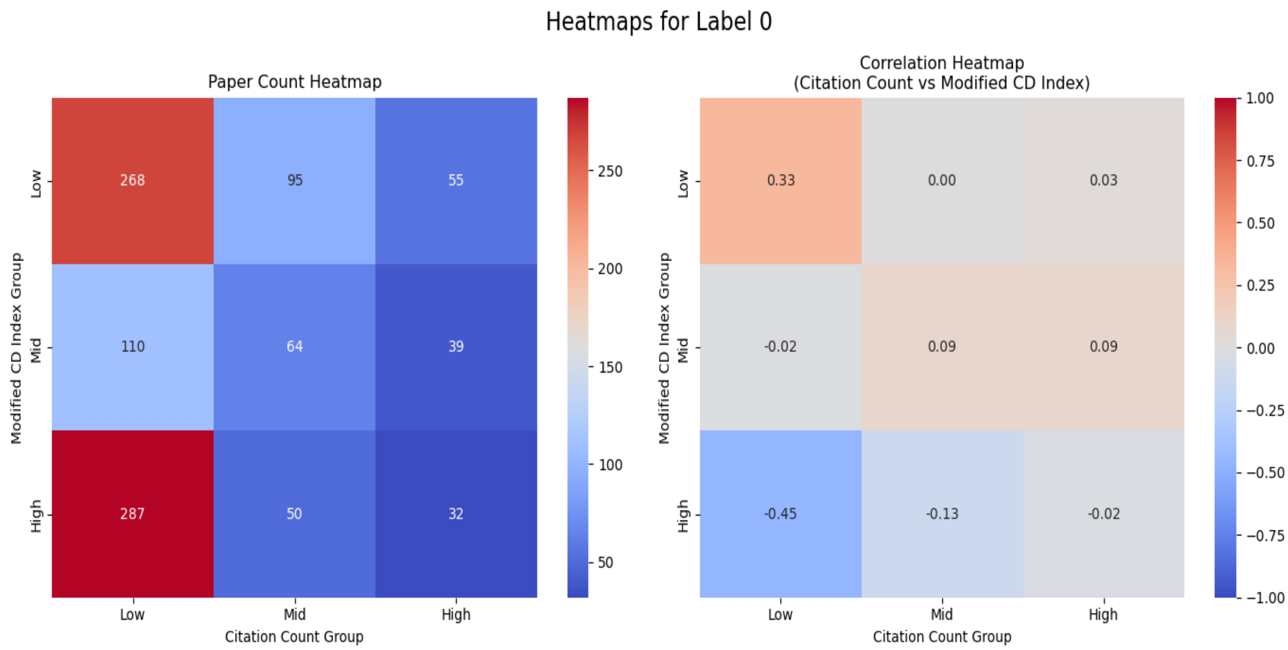


Figure 3.22: Label 0: Number of papers as per 9 categories and correlation coefficient

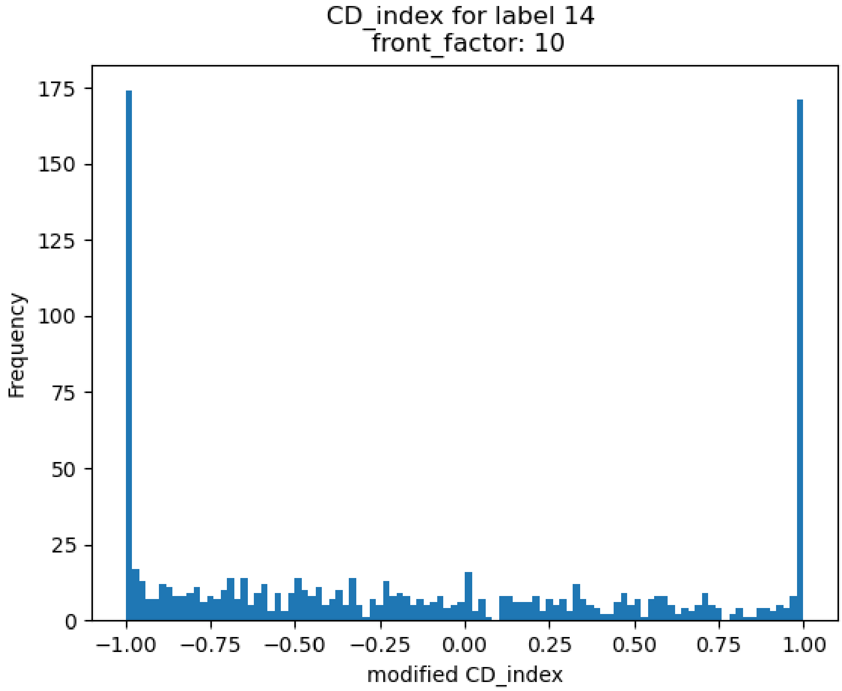


Figure 3.23: Label 14: Improved CD index values histogram.The peaks observed at +1 and -1 are likely an artifact of the incompleteness of citation graph. For the peak at +1, the key citation connections between the successors and predecessors might be absent while for -1, successors which actually don't have common predecessors might not be connected to the focal paper.

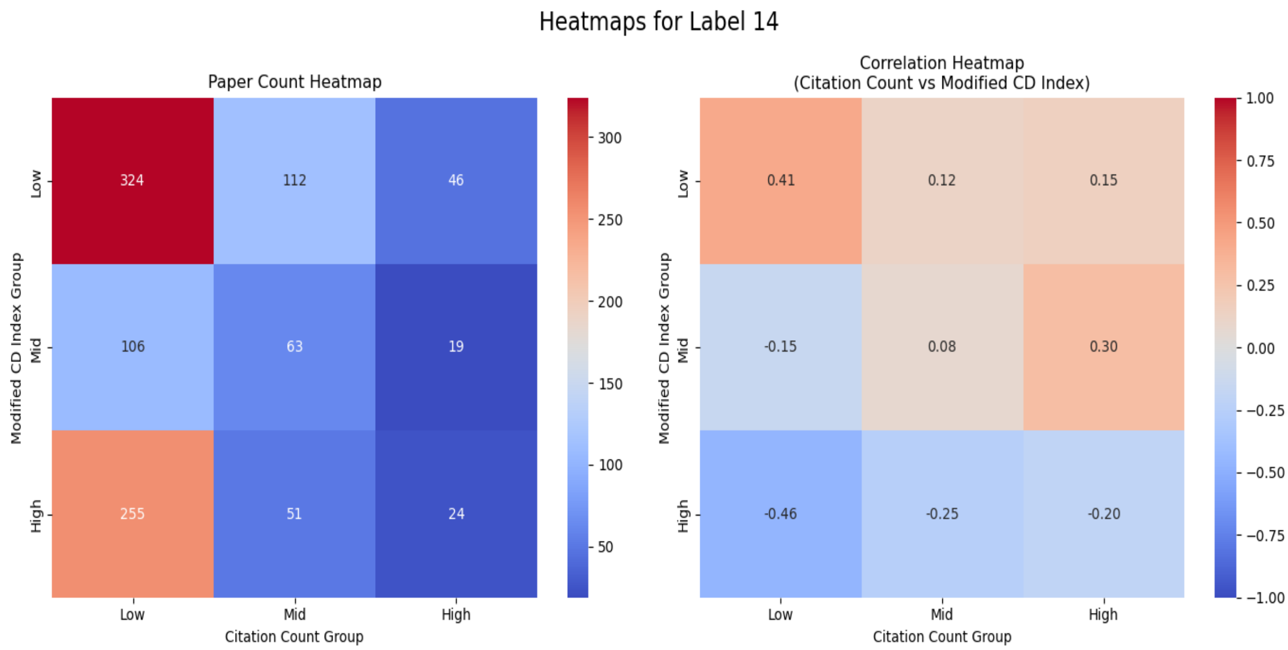


Figure 3.24: Label 14: Number of papers as per 9 categories and correlation coefficient

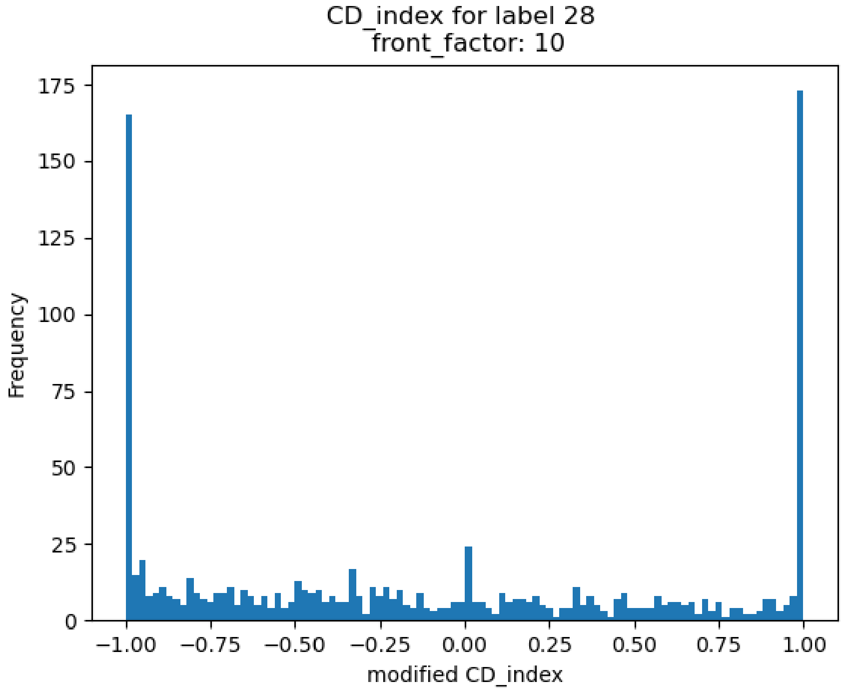


Figure 3.25: Label 28: Improved CD index values histogram. The peaks observed at +1 and -1 are likely an artifact of the incompleteness of citation graph. For the peak at +1, the key citation connections between the successors and predecessors might be absent while for -1, successors which actually don't have common predecessors might not be connected to the focal paper.

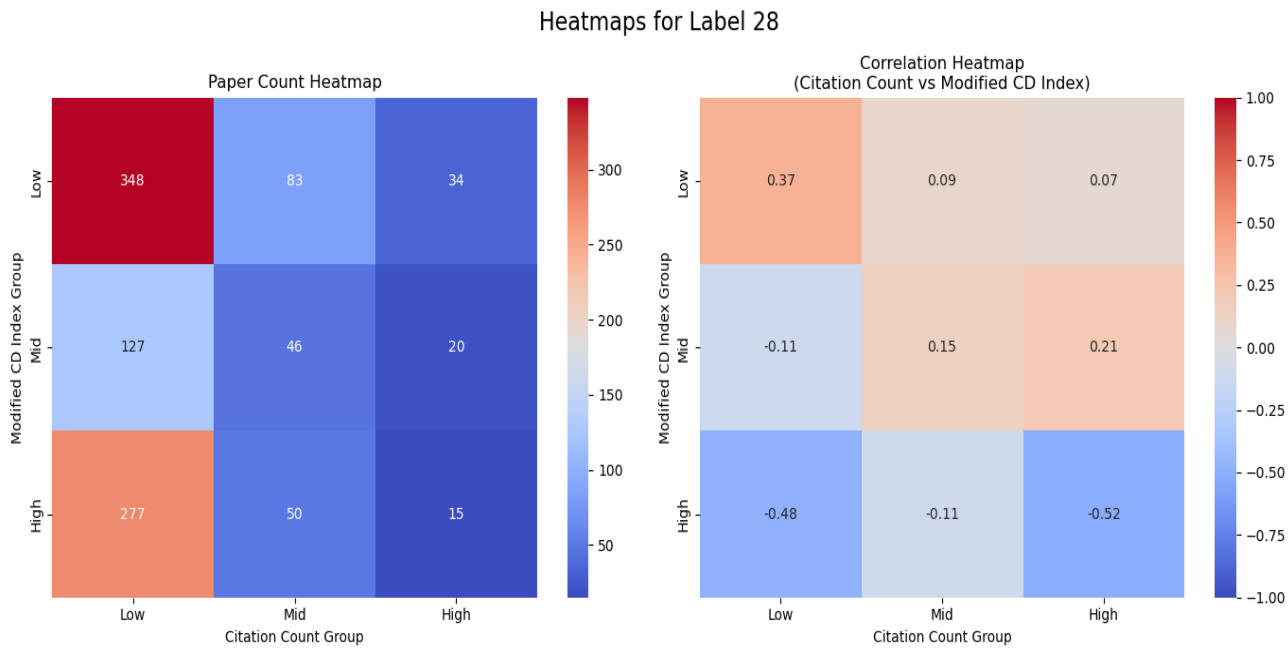


Figure 3.26: Label 28: Number of papers as per 9 categories and correlation coefficient

3.2 Conclusions

Following conclusions can be drawn based on the findings:

1. **Temporal dynamics of the CD index in a field of study:** Its seen that there is decline in the average CD index with time suggesting that the disruptiveness of papers is decreasing while consolidative and incremental work is getting dominated. The trend lines of different CDs are closely knit indicating not many successors are introduced in the citation network for a long duration. Apart from this, the trend lines for specific fields of study are highly fluctuating compared to the trend lines for the entire data set which might be an effect of smaller data size.
2. **Correlation between CD index and citation count:** Considering papers of a specific field of study, there is no correlation between the citation counts and CD index. This suggests citation counts are not sufficient for identifying the disruptive impact of paper. Citation count can help find identify papers which are frequently referred to and well recognized, but that does not imply that the paper is disruptive. Similarly, for modified CD index and citation count, the correlation is negligible except for some rare cases where they are slightly correlated. Some cases show negative correlation too, indicating that papers with high citation count may be negligibly disruptive. This re-affirms that citation count is not sufficient for measuring true scholarly impact.
3. **Disruptive potential temporal trends:** It has been observed that papers with high citation counts may initially seem non-disruptive, but with time, their disruptiveness can increase, considering the CD index trend of such papers. This emphasizes the evolving nature of research impact and the importance of reassessing scholarly work regularly, considering its effects in both the near and distant future.
4. **Incorporation of bread and depth of citation network:** Modified CD index tries to incorporate the number of overlaps between the focal paper and the successors along with the temporal difference between the focal paper and the predecessors. This metric shows that, plotting the histogram gives a positively skewed plot, indicating majority of papers have negative CD index. A peak at -1 and 1 shows that there is a large chunk of papers which are consolidative and a large

chunk has insufficient successors to predecessor citations. The correlation not only varies within the 9 categories, but also as per the field of study.

5. **Impact of the incompleteness of the citation graph:** The peak observed in the distribution of the CD indices at +1 is likely an artifact of the incompleteness of the citation graph available to us. The same may partially hold for the peak in the distribution of CD index at -1, where a few key successor papers may have been ignored. This missing successor information could be impacting the evaluated CD index of about 20% of the papers that have a CD index of 1.

3.3 Future work

(a) **Similarity score between the focal paper and predecessor:** Assigning a similarity score between the focal paper and the predecessor will help to understand the whether the work proposed in the focal paper is novel or incremental. This score can be based on the vector embedding of the entire paper.

(b) **Reference list of successors:** Number of papers in the reference list and the percentage of papers in that reference list having common predecessors can be considered as factors to penalize the focal paper via the successors.

(c) **Evaluating region, author and institute specific metric:** Analyzing the disruptive scores for a specific regions can help understand the citation pattern variation as per regions. Author and institute level metric can help to rank them for fund allocations , resource allocation and research grants.

(d) **Relation with other existing metrics:** Examining the correlation between the existing metrics like h-index and g-index can help to understand if these metrics can be more fruitful and disruptive than the CD index and the improved CD index.

Bibliography

[1] Eugene Garfield. "Citation analysis as a tool in journal evaluation". In: *Science* 178.4060 (1972), pp. 471–479. URL: https://ethw.org/w/images/3/33/Ref21-Garfield1972-CitationAnalysis.pdf.

[2] Eugene Garfield. "The History and Meaning of the Journal Impact Factor". In: *JAMA* 295.1 (Jan. 2006), pp. 90–93. ISSN: 0098-7484. DOI: 10.1001/jama.295.1.90. eprint: https://jamanetwork.com/journals/jama/articlepdf/202114/jco50055.pdf. URL: https://doi.org/10.1001/jama.295.1.90.

[3] Per O. Seglen. "Why the impact factor of journals should not be used for evaluating research". In: *BMJ* 314.7079 (1997), pp. 498–502. DOI: https://doi.org/10.1136/bmj.314.7079.497.

[4] Ronald N. Kostoff. "The use and misuse of citation analysis in research evaluation". In: *Scientometrics* 43.1 (1998), pp. 27–43. DOI: 10.1007/BF02458392. URL: https://doi.org/10.1007/BF02458392.

[5] Diana Hicks et al. "Bibliometrics: The Leiden Manifesto for research metrics". In: *Nature* 520.7548 (2015), pp. 429–431. DOI: 10.1038/520429a. URL: https://doi.org/10.1038/520429a.

[6] Loet Leydesdorff and Caroline S. Wagner. "International Collaboration in Science and the Formation of a Core Group". In: *Journal of Informetrics* 2.4 (2008), pp. 317–325. DOI: 10.1016/j.joi.2008.07.003. URL: https://www.sciencedirect.com/science/article/pii/S1751157708000448.

[7] Vincent Larivière and Yves Gingras. "The Impact Factor's Matthew Effect: A Natural Experiment in Bibliometrics". In: *Journal of the American Society for Information Science and Technology* 61.2 (2010), pp. 424–427. DOI: 10.1002/asi.21232. URL: https://doi.org/10.1002/asi.21232.

[8] James Wilsdon et al. *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management*. Tech. rep. ISBN 1902369273. Higher Education Funding Council for England (HEFCE), July 2015. DOI: 10.13140/RG.2.1.4929.1363. URL: https://www.ukri.org/wp-content/uploads/2021/12/RE-151221-TheMetricTideFullReport2015.pdf.

[9] Leo Egghe. "Theory and practice of the g-index". In: *Scientometrics* 69.1 (2006), pp. 131–152. DOI: 10.1007/s11192-006-0144-7. URL: https://link.springer.com/article/10.1007/s11192-006-0144-7.

[10] Lutz Bornmann, Rüdiger Mutz, and Hans-Dieter Daniel. "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". In: *Journal of the American Society for Information Science and Technology* 59.5 (2008), pp. 830–837. DOI: 10.1002/asi.20806. URL: https://doi.org/10.1002/asi.20806.

[11] Raj Kumar Pan and Santo Fortunato. "Author Impact Factor: Tracking the Dynamics of Individual Scientific Impact". In: *Scientific Reports* 4 (2014), p. 4880. DOI: 10.1038/srep04880. URL: https://www.nature.com/articles/srep04880.

[12] András Schubert and Tibor Braun. "Cross-field normalization of scientometric indicators". In: *Scientometrics* 36.3 (1996), pp. 311–324. DOI: 10.1007/BF02129597. URL: https://link.springer.com/article/10.1007/BF02129597.

[13] Loet Leydesdorff and Tobias Opthof. "Scopus's source normalized impact per paper (SNIP) versus a journal impact factor based on fractional counting of citations". In: *Journal of the American Society for Information Science and Technology* 61.11 (2010), pp. 2365–2369. DOI: 10.1002/asi.21371. URL: https://onlinelibrary.wiley.com/doi/full/10.1002/asi.21371.

[14] Wolfgang Glänzel and András Schubert. "Analyzing Scientific Networks Through Co-Authorship". In: *Handbook of Quantitative Science and Technology Research*. Ed. by Henk F. Moed, Wolfgang Glänzel, and Ulrich Schmoch. Springer, 2004, pp. 257–276. DOI: 10.1007/1-4020-2755-9_12. URL: https://link.springer.com/chapter/10.1007/1-4020-2755-9_12.

[15] Jarno Hoekman, Koen Frenken, and Robert J.W. Tijssen. "Research Collaboration at a Distance: Changing Spatial Patterns of Scientific Collaboration within Europe". In: *Research Policy* 39.5 (2010), pp. 662–673. DOI: 10.1016/j.respol.2010.01.012. URL: https://doi.org/10.1016/j.respol.2010.01.012.

[16] David A. King. "The Scientific Impact of Nations". In: *Nature* 430.6997 (2004), pp. 311–316. DOI: 10.1038/430311a. URL: https://www.nature.com/articles/430311a.

[17] Anthony F. J. van Raan. "Measuring Science: Capita Selecta of Current Main Issues". In: *Handbook of Quantitative Science and Technology Research*. Ed. by Henk F. Moed, Wolfgang Glänzel, and Ulrich Schmoch. Springer, 2004, pp. 19–50. DOI: 10.1007/1-4020-2755-9_2. URL: https://link.springer.com/chapter/10.1007/1-4020-2755-9_2.

[18] Mike Thelwall and Kayvan Kousha. "ResearchGate: Disseminating, Communicating, and Measuring Scholarship?" In: *Journal of the Association for Information Science and Technology* 66.5 (2015), pp. 876–889. DOI: 10.1002/asi.23236. URL: https://doi.org/10.1002/asi.23236.

[19] Mike Thelwall and Paul Wilson. "Mendeley Readership Altmetrics for Medical Articles: An Analysis of 45 Fields". In: *Journal of the Association for Information Science and Technology* 67.8 (2016), pp. 1962–1972. DOI: 10.1002/asi.23501. URL: https://doi.org/10.1002/asi.23501.

[20] Lutz Bornmann, Félix de Moya Anegón, and Loet Leydesdorff. "Do Scientific Advancements Lean on the Shoulders of Giants? A Bibliometric Investigation of the Ortega Hypothesis". In: *PLoS ONE* 5.10 (2010), e13327. DOI: 10.1371/journal.pone.0013327. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0013327.

[21] Russell J. Funk and Jason Owen-Smith. "A Dynamic Network Measure of Technological Change". In: *Management Science* 63.3 (2017), pp. 791–817. DOI: 10.1287/mnsc.2015.2366. URL: https://doi.org/10.1287/mnsc.2015.2366.

[22] Michael Park, Erin Leahey, and Russell J. Funk. "Papers and Patents Are Becoming Less Disruptive Over Time". In: *Nature* 613.7942 (2023), pp. 138–

144. DOI: 10.1038/s41586-022-05543-x. URL: https://www.nature.com/articles/s41586-022-05543-x.

[23] Nicholas Bloom et al. "Are Ideas Getting Harder to Find?" In: *American Economic Review* 110.4 (Apr. 2020), pp. 1104–44. DOI: 10.1257/aer.20180338. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20180338.

[24] Benjamin F. Jones. "The Burden of Knowledge and the 'Death of the Renaissance Man': Is Innovation Getting Harder?" In: *Review of Economic Studies* 76.1 (2009), pp. 283–317. DOI: 10.1111/j.1467-937X.2008.00531.x. URL: https://academic.oup.com/restud/article/76/1/283/1577537.

[25] Patrick Collison and Michael Nielsen. *Science is Getting Less Bang for Its Buck*. *The Atlantic*. 2018. URL: https://www.theatlantic.com/science/archive/2018/11/diminishing-returns-science/575665/.

[26] George Basalla. "The Spread of Western Science". In: *Science* 156.3775 (1967), pp. 611–622. URL: https://warwick.ac.uk/fac/arts/history/students/modules/hi997/2018-week10/science.156.3775.611.pdf.

[27] Filippo Radicchi, Santo Fortunato, and Claudio Castellano. "Universality of Citation Distributions: Toward an Objective Measure of Scientific Impact". In: *Proceedings of the National Academy of Sciences* 105.45 (2008), pp. 17268–17272. DOI: 10.1073/pnas.0806977105. URL: https://www.pnas.org/doi/10.1073/pnas.0806977105.

[28] Henk F. Moed. *Citation Analysis in Research Evaluation*. Vol. 9. Information Science and Knowledge Management. Springer, 2005. ISBN: 978-1-4020-3713-9. DOI: 10.1007/1-4020-3714-7. URL: https://link.springer.com/book/10.1007/1-4020-3714-7.

[29] Christian Leibel and Lutz Bornmann. "What do we know about the disruption index in scientometrics? An overview of the literature". In: *Scientometrics* 129.1 (2024), pp. 601–639. DOI: 10.1007/s11192-023-04873-5. URL: https://doi.org/10.1007/s11192-023-04873-5.

[30] G. Naga Narasimha Rao and Murugesan Venkatapathi. "Reliability of citation count in predicting Scientific Impact : an estimation using higher order citation indices". In: (2023).