Scilight

Transactions on Artificial Intelligence
https://www.sciltp.com/journals/tai



Article

Check for updates

CLOF: A Coupled Framework for Quantifying Research Impacts via A Double Iterative Power Method

Junyuan Fang [1], Zhihao Yu [1], Xiangyu Cai [2], Hao Wang [1,*] and Dapeng Oliver Wu [1]

[1] Department of Computer Science, City University of Hong Kong, Hong Kong SAR, China
[2] Computer Science and Engineering Division, University of Michigan, Ann Arbor, MI 48109, USA
* Correspondence: hao.wang@my.cityu.edu.hk

How To Cite: Fang, J.; Yu, Z.; Cai, X.; et al. CLOF: A Coupled Framework for Quantifying Research Impacts via A Double Iterative Power Method. *Transactions on Artificial Intelligence* **2026**, *2*(1), 245–265. https://doi.org/10.53941/tai.2026.100016

Received: 5 February 2026
Revised: 9 November 2026
Accepted: 9 November 2026
Published: 18 November 2026

Abstract: Research impact quantification plays a crucial role in academic assessment. Conventional citation-based metrics have several limitations in fair and comprehensive research evaluation. They often fail to recognize foundational work that receives limited early attention. They may also overlook sustained influence over time, differences across research fields, citation quality reflected by citation contexts and the expertise of citing authors, and variations in author contributions. We incorporate these aspects at the paper level through four factors: originality-oriented downstream influence, long-term influence, field-specific ranking, and citation quality. At the researcher level, we quantify credit according to author order. Based on these, we propose CLOF, a unified metric framework that quantifies the impacts of papers and researchers in an interdependent and iterative manner. To reduce reliance on manually specified weighting schemes, we employ multiple large language model (LLM)-based paper scorers. Their outputs are aggregated using an α-trimmed procedure to generate robust global signals for learning the combination weights. These learned weights further guide the coupled iterative updates of paper and researcher scores via a Double Iterative Power (DIP) method with theoretical convergence guarantees. Extensive experiments on academic datasets show that CLOF uncovers influence patterns across papers and researchers that are not captured by existing impact metrics.

Keywords: research impact quantification; citation networks; coupled metric modeling; large language models; robust aggregation

1. Introduction

Research impact quantification has consistently drawn significant attention in the academic community, industry, and governments. It characterizes not only how often research papers or researchers are cited, but also the depth and breadth of their influence in academic society. Therefore, the corresponding impact metrics are essential for a wide range of academic assessments [1–6]. For instance, they support funding and grant allocation by helping agencies identify high-impact research areas and researchers [7]. They also contribute to academic rankings, influencing how institutions are perceived globally [8–10]. They further assist career development, supporting annual reviews, tenure decisions, job applications, and related evaluations [11]. To evaluate the research impact, a variety of metrics and methods have been proposed from different perspectives, such as the metrics from Bibliometrics like raw citation counts, h-index, i10-index, etc. [12–16], and graph-based approaches such as PageRank, HITS, and so on [17–20].

As the number of published research papers continues to grow, the need for accurate and fair assessment of research impact has become increasingly important. However, prior studies [21–25] have pointed out several limitations of existing metrics. These limitations include undervalued recognition of foundational work, limited time sensitivity, poor comparability across research fields, and insufficient consideration of citation quality. In addition to these paper-level limitations, existing metrics typically evaluate researchers without distinguishing author contributions.

Copyright: © 2026 by the authors. This is an open access article under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).
Publisher's Note: Scilight stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

More concretely, highly theoretical or foundational papers may receive relatively few early citations because their contributions are not immediately recognized or widely adopted. As a result, their broader downstream influence may be only partially reflected by direct citation counts. Moreover, while the h-index metric and corresponding variants characterize the long-term influence of researchers to some extent, they do not capture paper-level temporal dynamics such as the active period of individual papers. In addition, papers from different research fields often exhibit distinct citation patterns. These differences include the average number of references and the typical number of co-authors. Therefore, direct comparison based on raw citation counts may be unfair across disciplines. Furthermore, existing metrics often treat every citation as a uniformly positive signal with equal weight, regardless of context or who is citing. Citation weights should account for both local citation contexts and the expertise of citing authors. In this way, differences in citation context and academic expertise can be reflected in the importance assigned to each citation. At the same time, researcher-level evaluation often assigns equal credit to all authors, ignoring differences in contribution and failing to characterize author roles or leadership on a paper. Addressing these issues is essential for a more comprehensive and fair evaluation of research impact.

A further limitation of existing metrics lies in the absence of a coupled evaluation framework between the paper-level and researcher-level indicators. Prevailing metrics typically assess the impact of research papers and researchers in isolation, disregarding the intrinsic interdependence between these two levels. In practice, the impact of papers and researchers influences each other. This looping relationship implies that paper-level and researcher-level metrics should not be treated independently. The lack of a mechanism to capture such iterative coupling may compromise the comprehensiveness and validity of research impact evaluation.

To address the above limitations, this work proposes a unified quantification framework that jointly assesses the impact of research papers and their corresponding researchers through a coupled iterative process.

Building on a comprehensive corpus containing full texts and rich metadata, we propose CLOF (the name CLOF is derived from the four components of the paper-level metric, arranged for readability). The proposed framework quantitatively evaluates the impact of both research papers and researchers while addressing the aforementioned limitations. The CLOF metric for research papers (CLOF-P) incorporates originality-oriented downstream influence, long-term influence, field-specific ranking, and citation quality. The researcher-level metric (CLOF-R) focuses on differences in author contribution. Moreover, we introduce a coupled mechanism based on a Double Iterative Power (DIP) method. This mechanism enables mutual updating between paper-level and researcher-level metrics. The DIP method enables efficient updates and provides theoretical convergence guarantees. In addition, we employ multiple large language model (LLM) scorers to avoid manually specifying the relative importance of the four dimensions. The resulting paper-level evaluations are aggregated through an α-trimmed procedure to obtain robust global signals. The estimated weights are then applied in the coupled updates. The main contributions are summarized as follows.

(1) We propose a unified metric framework, CLOF, which jointly evaluates the impact of research papers and researchers to address the above limitations at both the paper and researcher levels. The paper impact metric (CLOF-P) incorporates several key factors, including originality-oriented downstream influence, long-term influence, field-specific ranking, and citation quality, while the CLOF researcher impact metric (CLOF-R) accounts for the paper contribution of different authors.
(2) We incorporate multiple LLM-based paper scorers. Their outputs are aggregated through an α-trimmed procedure to reduce subjective bias when determining the relative importance of different factors. These signals enable CLOF to learn combination weights in a constrained least squares manner, and the resulting weights are applied to all papers during the subsequent score updates.
(3) We introduce a coupled mechanism based on a Double Iterative Power method to address the interdependence between paper and author evaluations. The proposed mechanism enables mutual updating of impact scores and provides theoretical convergence guarantees.
(4) We conduct extensive experiments to validate the effectiveness and efficiency of CLOF on the benchmark dataset.

The rest of this paper is organized as follows. Section 2 reviews related work, focusing on prior studies in Bibliometrics and graph-based methods. Section 3 introduces the proposed CLOF framework, detailing the metric design of both research papers and researchers, along with the iterative optimization process. Section 4 presents the construction of our experimental datasets, followed by a thorough evaluation and discussion of the proposed framework. Finally, Section 5 concludes the paper and outlines directions for future research.

2. Related Work

With the growing interest in evaluating research impact, a variety of methods and metrics have been proposed, which we broadly divide into two categories, namely Bibliometrics and graph-based methods. In the following, we briefly introduce the current efforts on them.

2.1. Bibliometrics

Bibliometrics [14,26–30] is the statistical analysis of academic publications and their citations. Among all bibliometric metrics, citation count [15] is one of the most straightforward measures. It uses the number of times a paper is cited by other papers as an indicator of impact. Several metrics were subsequently developed to evaluate the long-term impact of researchers beyond the simple citation count. For instance, the h-index [16] extends the simple citation count. A researcher has an h-index of h if the researcher has published h papers, each of which has received at least h citations. Moreover, the i10-index denotes the number of papers published by a researcher that have been cited at least 10 times. G-index [31] is another metric that gives more weight to highly cited papers by ensuring that the top g papers of a researcher have at least g^2 citations. Field-weighted citation impact, proposed by Elsevier and based on the Scopus datasets (https://www.scopus.com (accessed on 5 October 2025)), utilizes the average citations of papers belonging to the same publication year, type, and discipline to normalize the original citations. Besides the above metrics, several further metrics have been proposed, such as a-index [32], contemporary h-index [33], Hx-index [34], DS-index [35], etc.

Beyond traditional citation-based indicators, recent studies have explored alternative dimensions of scientific impact. Yang et al. [36] proposed a dual measurement framework based on disruption and consolidation. Their framework introduces metrics such as disruptive citation and consolidating citation to evaluate whether scientific contributions redirect future research away from existing knowledge structures or reinforce established paradigms. Another line of research focuses on the temporal dynamics of scientific recognition. For instance, Zhao et al. [37] investigated delayed recognition from a co-citation perspective and identified cases in which combinations of existing knowledge remained dormant for extended periods before receiving substantial attention.

While these approaches provide valuable perspectives on scientific impact, they primarily focus on individual dimensions of research evaluation. In practice, scientific impact is inherently complex and cannot always be adequately characterized from a single perspective. It may involve disruption, recognition dynamics, influence propagation, field-specific differences, citation quality, and other complementary factors.

2.2. Graph-Based Metrics

Graph centrality measures [17, 38–40], such as degree centrality, betweenness centrality, and closeness centrality, have been widely used to evaluate node importance in academic graphs. The nodes may represent research papers or researchers, while edges represent their relationships. In citation networks, the PageRank-based metric [41] evaluates the impact of a paper by analyzing the citation relationships between papers, including both the citation count and the corresponding citation weight, where the importance of a node increases if it is cited by several highly ranked nodes. Besides the original PageRank method, several variants have been proposed. For instance, time-aware PageRank [42] incorporates the temporal decay of citations by assigning larger weights to recent citations. HITS [43] iteratively updates the scores of authority nodes and hub nodes by aggregating information from incoming and outgoing links, effectively identifying authoritative papers and survey-like hub papers in citation networks. The Eigenfactor metric [44] evaluates the overall influence of a journal by considering the number of incoming citations and their sources.

Beyond direct citation-based network measures, recent studies have explored higher-order citation influence through citation chains. For example, Franceschet and Colavizza [45] proposed a higher-order citation framework for quantifying indirect scientific influence beyond direct citations. The framework models influence propagation through citation paths of arbitrary length. Such approaches demonstrate that a substantial proportion of scientific influence is transmitted through higher-order citation structures rather than direct citations alone.

However, most existing approaches focus on specific dimensions of scientific impact, such as direct citation impact, indirect influence propagation, disruption, or recognition timing. While these perspectives are valuable, research impact is inherently complex and often involves multiple complementary factors simultaneously. In particular, aspects such as long-term impact persistence, field-specific differences, and citation quality are rarely considered in a unified manner. Moreover, existing studies typically rely on predefined evaluation metrics. They provide limited discussion of how different impact dimensions should be combined. In addition, research papers and researchers are usually evaluated independently despite their intrinsic interdependence. Thus, in this work, we aim to propose a unified framework for research impact quantification by jointly modeling multiple complementary dimensions and the coupled relationship between papers and researchers.

3. Framework

In this section, we present the CLOF framework in detail, including metric definitions and updates, convergence analysis, and the overall framework.

3.1. Paper Impact Metric of CLOF

To address the limitations of existing Bibliometric-based and graph-based metrics, we propose a more comprehensive paper metric, CLOF-P. These limitations include (1) neglecting highly original or disruptive research papers; (2) underestimating the active period of publications; (3) applying uniform metrics to fields with distinct citation behaviors; and (4) failing to distinguish between supportive and critical citations while ignoring the expertise of citing authors. Specifically, the CLOF-P metric considers the originality-oriented downstream influence, long-term influence, field-specific ranking, and citation quality to solve the drawbacks of current metrics. We give the details of the above four sub-metrics in the following.

3.1.1. Originality-Oriented Downstream Influence

The originality-oriented downstream influence in CLOF reflects the extent to which a paper introduces ideas that serve as a foundation for subsequent knowledge development. It also captures the ability of these ideas to stimulate future research. Such influence may propagate through multiple generations of publications and therefore is not always fully captured by direct citation counts alone. As a result, some papers may receive less recognition than their broader downstream influence indicates when only direct citations are considered.

Motivated by this observation, the proposed originality-oriented downstream influence sub-metric S_o considers not only direct citations but also the downstream influence of a paper within the citation network. Specifically, we construct a citation tree based on citation relationships and measure the cumulative downstream citations associated with the root paper. The intuition is that papers that stimulate extensive follow-up research may exert a stronger foundational influence on the development of a research area, even though part of the downstream impact is jointly shaped by later papers in the citation tree. Such foundational influence is often associated with original research that introduces ideas capable of generating sustained downstream development. For instance, as shown in Figure 1, the raw citation count of paper A is 2, while its cumulative downstream citation count is 4. This example illustrates how the influence of a paper can extend beyond its immediate citation neighborhood through subsequent knowledge diffusion. Formally, let $\mathcal{D}(i)$ denote the set of papers that directly cite paper i. The cumulative downstream citation count is defined recursively as

$$S_o(i) = \begin{cases} 0, & \text{if } \mathcal{D}(i) = \emptyset, \\ |\mathcal{D}(i)| + \sum_{p \in \mathcal{D}(i)} S_o(p), & \text{otherwise.} \end{cases}$$

where $|\mathcal{D}(i)|$ denotes the number of papers in $\mathcal{D}(i)$. It should be noted that the proposed formulation aggregates influence along citation paths. Therefore, a descendant reachable through multiple citation paths may contribute multiple times, reflecting influence propagation through multiple independent lines of research.

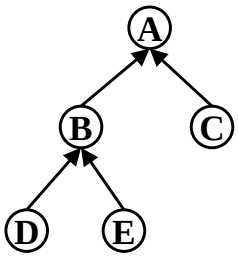


Figure 1. An example of citation tree where the root paper is A.

3.1.2. Long-Term Influence

Long-term influence of a paper refers to whether a paper can continuously attract attention over the years. Therefore, this sub-metric adopts an AUC (Area Under the Curve)-like calculation to evaluate the influence of a paper across the years after its publication. Specifically, given a series of citation thresholds $T = \{t_1, t_2, \cdots, t_m\}$ where $t_1 < t_2 < \cdots < t_m$, we first calculate how many years the citation of a paper exceeds each threshold, respectively. Then, the total years of all thresholds will be summed up to obtain the final score of this sub-metric. For example, as shown in Figure 2, the first bin indicates the number of years that the citation of a specific paper exceeds 5 times. It should be noted that these thresholds are cumulative rather than mutually exclusive. Therefore, a year counted under a higher citation threshold is also counted under all lower thresholds. The final sub-metric score is obtained by summing the y values of all bins. Formally, the sub-score of a paper i can be obtained as follows.

$$S_{lt}(i) = \sum_{j=1}^{m} y_{t_j}, \tag{1}$$

where y_{t_j} is the number of years that the citation count of paper i remains above the threshold t_j.

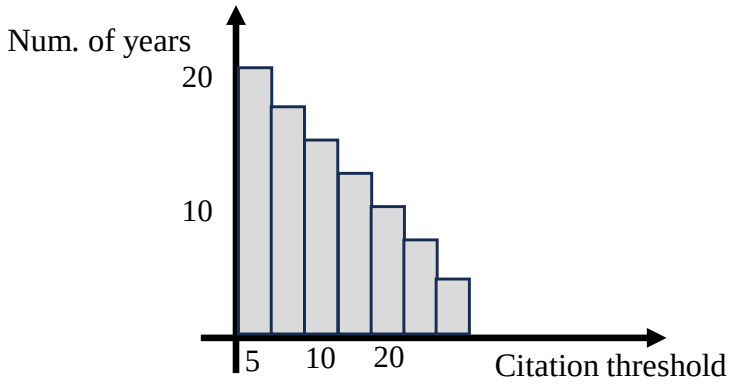


Figure 2. An example of the AUC-like sub-metric of a specific paper by using a series of thresholds.

3.1.3. Field-Specific Ranking

Research papers across different research fields, such as engineering, social science, etc., often exhibit distinct citation patterns. For example, papers from engineering fields often have shorter reference lists, while those from social science fields tend to cite more extensively. These disciplinary differences make the raw citation count an incomplete and potentially biased measure of research impact. Thus, this sub-metric is determined by the citation count of research papers in its specific research field. More importantly, the percentile ranking is adopted to mitigate the bias caused by the varying number of papers in different research fields. Formally, for a paper i published in a specific field, assuming there were n papers published in this field and its citation count ranks r-th among them, we have

$$S_{fr}(i) = (1 - \frac{r}{n}) \times 100. \tag{2}$$

In particular, if a paper involves multiple research fields, the proposed sub-metric ranks it separately within each field. The resulting percentile scores are then summed. This design aims to reflect the relative standing of a paper within individual fields. It also captures the breadth of its influence across multiple research areas. Therefore, papers that achieve strong rankings in several disciplines may receive higher scores, reflecting their potential interdisciplinary impact.

3.1.4. Citation Quality

The raw citation count usually treats all citations as positive indicators of impact. However, some papers are cited not to endorse their findings, but to critique or refute them. Therefore, it is important to analyze the motivation of citations via sentiment analysis tools [46]. Specifically, this sub-metric is determined by the impact score of researchers who cite it and the motivation of the citation. Intuitively, this sub-metric is related to two key factors: who cites you and how they judge you. If top researchers endorse a specific work, this work is likely to receive a high evaluation through this citation, and vice versa. Thus, considering a specific paper i that has been cited by l papers, the sub-score of this paper can be as follows.

$$S_{cq}(i) = \sum_{j=1}^{l} S_j \cdot CQ_j, \tag{3}$$

where j denotes a specific paper that cites paper i, and S_j denotes the total impact scores of researchers (which will be introduced in the next subsection) appearing in the author list of paper j. Furthermore, $CQ_j \in [0, 1]$ is a continuous value indicating the motivation of this citation, where 1 represents a positive citation and 0 represents a negative citation. It should be noted that this component is not intended to measure scientific influence itself. Rather, it aims to characterize endorsement-oriented citation quality as a complementary dimension of research impact. Specifically, the citation motivation of each citation can be determined by conducting sentiment analysis on the citation context using established natural language processing tools.

However, a key challenge lies in determining what should be regarded as the citation context. Existing efforts usually take either the single sentence or the entire paragraph in which the citation appears. Both choices have clear limitations. A single sentence may not contain sufficient semantic information to reveal the intent of the citation. In contrast, a full paragraph may include unrelated material that distracts from the actual motivation.

To address this limitation, we propose an adaptive context segmentation module to dynamically determine how many surrounding sentences should be included when forming the citation context. Instead of relying on a fixed sentence window, we adopt an adaptive procedure that adjusts the window size according to the semantic relevance of surrounding sentences. The process begins from the citation sentence and evaluates neighboring sentences on both sides. New sentences are included only when they strengthen the semantic alignment between the context and the title of the cited work. The expansion stops once additional content no longer contributes meaningful information. In this way, the resulting context is tailored to the specific citation rather than constrained by predefined boundaries. The complete procedure is summarized in Algorithm 1.

Algorithm 1: Citation Context Segmentation

Input: Sentence sequence $\mathcal{S} = \{s_1, s_2, \ldots, s_N\}$; index of the citation sentence anchor; embedding function $\phi(\cdot)$; similarity function $\mathrm{sim}(\cdot,\cdot)$; maximum expansion length H.
Output: Extracted citation context $\mathcal{S}[L:R]$.
$E \leftarrow \phi(\mathrm{cited_title})$;
$L \leftarrow \mathrm{anchor}$;
$R \leftarrow \mathrm{anchor}$;
$\mathrm{best} \leftarrow \mathrm{sim}\Big(E,\ \phi\big(\mathcal{S}[L:R]\big)\Big)$;
for all $\mathrm{direction} \in \{\mathrm{left},\ \mathrm{right}\}$ **do**
  for $\mathrm{step} = 1$ **to** H **do**
    if $\mathrm{direction} = \mathrm{left}$ **then**
      $j \leftarrow \mathrm{anchor} - \mathrm{step}$;
      if $j < 1$ **then**
        break
      end if
      $\mathrm{score} \leftarrow \mathrm{sim}\Big(E,\ \phi\big(\mathcal{S}[j:R]\big)\Big)$;
      if $\mathrm{score} \geq \mathrm{best}$ **then**
        $L \leftarrow j$;
        $\mathrm{best} \leftarrow \mathrm{score}$;
      else
        break
      end if
    else
      $j \leftarrow \mathrm{anchor} + \mathrm{step}$;
      if $j > N$ **then**
        break
      end if
      $\mathrm{score} \leftarrow \mathrm{sim}\Big(E,\ \phi\big(\mathcal{S}[L:j]\big)\Big)$;
      if $\mathrm{score} \geq \mathrm{best}$ **then**
        $R \leftarrow j$;
        $\mathrm{best} \leftarrow \mathrm{score}$;
      else
        break
      end if
    end if
  end for
end for
return $\mathcal{S}[L:R]$

3.1.5. Sub-Metric Combination Based on LLMs

A key challenge in combining the above four sub-metrics lies in determining their relative importance, as manually assigning weights would inevitably introduce subjective bias. Thus, we adopt a data-driven module in

which LLMs provide global evaluations based on these four sub-metrics. Prior to weight estimation, each submetric is independently normalized to the range $[0, 100]$. Since it is impractical to obtain LLM scores for all papers, we apply LLMs only to a randomly sampled subset of C papers. This subset serves as a computationally efficient supervision set for weight estimation. Specifically, the detailed learning process for each weight learning stage is given as follows.

Score Aggregation

For each sampled paper i, let $\{y^{(j)}(i)\}_{j=1}^{K}$ denote the valid paper-level scores returned by the K LLM scorers. Each score is given under a carefully designed prompt that reflects the definitions and intended roles of the four sub-metrics. Particularly, in our implementation, the LLM scorers are instantiated by a mixture-of-experts (MoE) model, which activates only a subset of experts at each inference step. This structure not only reduces inference cost compared with dense models of similar size, but also provides natural diversity across scorers so that the aggregated judgment reflects a broader range of expert perspectives. Even with this efficiency gain, evaluating all papers remains costly at scale, so only a subset of C papers is queried to LLMs, and the resulting weights generalize to the remaining set.

Specifically, we use a lightweight instruction template summarized below.

Template of Prompt (T):

> *You are an expert in research impact evaluation. Given the semantic definitions of sub-metrics of Originality, Long-term Influence, Field Ranking, and Citation Quality, provide an overall 0–100 impact score for each paper. Your judgment should be based on the conceptual meanings of the sub-metrics rather than their numerical magnitudes alone. Provide a concise explanation summarizing your holistic reasoning.*

After obtaining the K scores, we first remove responses that cannot be successfully parsed into the required output format. We then apply an α-trimmed mean operation to mitigate the influence of extreme scores and improve the stability of the supervision signal. Thus, after sorting these values, we discard $\lfloor \alpha K \rfloor$ scores from each tail and average the remaining to obtain a robust target $\hat{y}(i)$.

Weight Estimation

We determine the combination weights of four sub-metrics by fitting them to the LLM-derived paper scores through a simple constrained least squares approach. Specifically, we construct the matrix $X \in \mathbb{R}^{C\times 4}$ with rows

$$x_i^\top = \big[S_o(i),\, S_{lt}(i),\, S_{fr}(i),\, S_{cq}(i)\big],$$

where S_o, S_{lt}, and S_{fr} remain fixed, and $S_{cq}(i)$ is the citation quality induced by $M_{pr}\boldsymbol{S_R}$ at the beginning of current stage. Let

$$\hat{y} = [\,\hat{y}(1), \ldots, \hat{y}(C)\,]^\top.$$

These C samples constitute the supervision set for weight estimation.

We estimate the weights $\hat{w} = [\hat{w}_1, \hat{w}_2, \hat{w}_3, \hat{w}_4]^\top$ via the following constrained least squares step, i.e.,

$$\min_{\hat{w}\in\mathbb{R}^4} \frac{1}{C}\big\|\hat{y} - X\hat{w}\big\|_2^2 \quad \text{s.t.} \quad \hat{w}_k > 0, \sum_{k=1}^{4} \hat{w}_k = 1. \tag{4}$$

Sub-Score Combination

With the fitted weights w, the score for paper i within current stage is computed as

$$S_P(i) = w_1 S_o(i) + w_2 S_{lt}(i) + w_3 S_{fr}(i) + w_4 S_{cq}(i). \tag{5}$$

These weights are fixed throughout the stage to ensure that the subsequent coupled updates of papers and researchers proceed in a stable manner.

3.2. Researcher Impact Metric of CLOF

In addition to the score of research papers, we propose the CLOF-R metric to evaluate the impact of corresponding researchers. A simple and straightforward idea is to distribute the credit of a paper to all of its co-authors. In particular, different authors should have different levels of contribution to a paper. For instance, the first author

and corresponding author may have higher contributions than other authors. Thus, we adopt a linearly decreasing mechanism to allocate the score of a paper. Specifically, the first author and corresponding author receive the two highest contribution scores. The contribution scores then decrease as the author order increases. Assuming a paper has k authors, the first author receives the highest weight of k, the corresponding author receives $k-1$, and so on, until the last author receives a score of 1. To ensure the total weights of all authors normalize to 1, we divide the original weight of each author by the total score $k(k+1)/2$. An example of the above score allocation strategy is given in Figure 3. Finally, the total score of a specific researcher i, $S_R(i)$, can be obtained by summing up all of the paper contributions of i's published papers. In particular, for single-author papers, all credit is assigned to the sole author. When the first and corresponding author are the same person, the corresponding contribution is accumulated accordingly. Assuming researcher i has published β papers, we have

$$S_R(i) = \sum_{j=1}^{\beta} p_j, \tag{6}$$

where p_j denotes the paper contribution of researcher i in paper j.

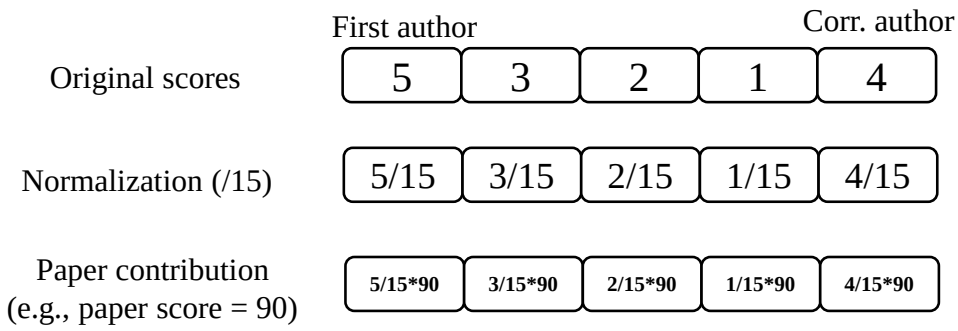


Figure 3. An illustration of the score allocation based on the author order. Here we take the number of authors to be 5, i.e., $k = 5$, and the paper score to be 90 as examples.

3.3. Metric Updates Based on a Double Iterative Power Method

From the above definitions, the citation quality sub-metric relies on researcher scores. These researcher scores, in turn, depend on paper scores. As a result, a coupled and iterative update process naturally arises. To effectively model such recursive influence propagation of scores, we draw inspiration from the power method [20,47], which has proven highly effective in evaluating node importance within networks. The power method is an iterative algorithm designed to compute the dominant eigenvector of a matrix. This eigenvector corresponds to the largest eigenvalue in magnitude and captures the most influential components in a network system. Specifically, the power method ensures fast convergence, scalability, and mathematical stability, making it particularly suitable for application on large-scale, sparse academic networks. Therefore, motivated by the above advantages, we adopt a power method-based strategy to update the impact scores.

Given a non-negative, column-stochastic matrix $P \in \mathbb{R}^{n\times n}$, the traditional power method updates a score vector $\boldsymbol{r} \in \mathbb{R}^n$ iteratively as follows.

$$\boldsymbol{r}^{(t+1)} = P\boldsymbol{r}^{(t)}, \tag{7}$$

where $\boldsymbol{r}$ converges to the dominant eigenvector of P under mild conditions. This eigenvector captures the steady-state distribution or relative importance of entities in a network system and serves as a foundational tool in many influence propagation models.

However, unlike the traditional power method, which focuses on a single entity, the CLOF framework consists of a coupled structure where the scores of researchers and research papers need to be updated based on each other. Therefore, we further propose a novel double iterative power method (DIP) to achieve the coupled update of metrics. Specifically, assuming we have two score vectors, namely $\boldsymbol{S_P} \in \mathbb{R}^n$ for scores of research papers and $\boldsymbol{S_R} \in \mathbb{R}^m$ for scores for researchers, where n and m denote the number of research papers and the number of researchers, respectively. The update of scores of research papers and researchers are given as follows.

3.3.1. Updates of Research Papers

From (5), we know the scores of research papers are related to four sub-metrics, including originality-oriented downstream influence, long-term influence, field-specific ranking, and citation quality. Among them, the sub-metric of citation quality S_{cq} of a paper is related to the score of the corresponding researcher who cites the specific paper. Thus, we can further transform the updates of S_{cq} based on the variant of the power method, as follows.

$$\boldsymbol{S}_{\boldsymbol{cq}}^{(t)} = M_{pr}\boldsymbol{S}_{\boldsymbol{R}}^{(t)}, \tag{8}$$

where $M_{pr} \in \mathbb{R}^{n\times m}$ is the transition matrix of researcher-to-paper. Specifically, based on the formulation of (5), we can easily obtain that a specific value of this transition matrix, $M_{pr}(i,j)$, denotes the citation quality CQ that researcher j gave to paper i. It is worth noting that a researcher may cite a paper several times in different published papers, thus $M_{pr}(i,j)$ indicates the sum of all these citations.

Then, combining (5) and (8), we have the following.

$$\boldsymbol{S}_{\boldsymbol{P}}^{(t)} = w_1 \cdot \boldsymbol{S}_{\boldsymbol{o}} + w_2 \cdot \boldsymbol{S}_{\boldsymbol{lt}} + w_3 \cdot \boldsymbol{S}_{\boldsymbol{fr}} + w_4 \cdot \boldsymbol{S}_{\boldsymbol{cq}}^{(t)}, \tag{9}$$

where $\boldsymbol{S}_{\boldsymbol{o}}$, $\boldsymbol{S}_{\boldsymbol{lt}}$, and $\boldsymbol{S}_{\boldsymbol{fr}}$ denote the score vectors of the first three sub-metrics. In particular, the combination weights w are fixed within the same stage.

3.3.2. Updates of Researchers

For researchers, a similar update process to that of research papers can be obtained based on the author order. Formally, the updates of researcher scores can be as follows.

$$\boldsymbol{S}_{\boldsymbol{R}}^{(t+1)} = M_{rp}\boldsymbol{S}_{\boldsymbol{P}}^{(t)}, \tag{10}$$

where $M_{rp} \in \mathbb{R}^{m\times n}$ is the transition matrix of paper-to-researcher. Specifically, a specific value of this transition matrix, $M_{rp}(i,j)$, denotes the paper contribution of researcher i on paper j based on i's author order. Since the combination weights are fixed within a stage, the only time-varying component in (10) is the citation-quality sub-score inherited from $\boldsymbol{S}_{\boldsymbol{cq}}^{(t)}$. This design ensures that the researcher-paper interaction evolves consistently during the inner iteration.

3.3.3. Iteration and Stopping Criteria

Given the fitted weights within the current stage, the proposed DIP method alternately updates paper scores and researcher scores according to (9) and (10). To monitor the convergence of the inner-loop iterations, we define the iteration loss as

$$L^{(t)} = \left\| \boldsymbol{S}_{\boldsymbol{P}}^{(t)} - \boldsymbol{S}_{\boldsymbol{P}}^{(t-1)} \right\|_2, \tag{11}$$

which measures the difference between two consecutive paper-score vectors. The inner-loop iteration terminates when

$$L^{(t)} < 10^{-6}.$$

After the inner-loop convergence is reached, the combination weights are re-estimated through (4) at the stage level. The stage-wise updates also terminate when the difference between the fitted weights of two consecutive stages becomes smaller than 10^{-6}.

3.4. Convergence Analysis

In this part, we provide a convergence analysis of the DIP method. For both transition matrices, column-wise normalization is applied during construction so that each column sums to one. Therefore, both M_{pr} and M_{rp} are column-stochastic matrices. Formally, within a specific stage where the combination weights are fixed, we have the following.

Proposition 1. *Let $M_{rp} \in \mathbb{R}^{m\times n}$ and $M_{pr} \in \mathbb{R}^{n\times m}$ be two non-negative, column-stochastic matrices representing paper-to-researcher and researcher-to-paper influence, respectively. Define the composite matrix*

$$M = M_{rp}M_{pr}.$$

Under the update rule

$$\boldsymbol{S}_{\boldsymbol{R}}^{(t+1)} = w_4 M \boldsymbol{S}_{\boldsymbol{R}}^{(t)} + \boldsymbol{b}, \tag{12}$$

where

$$\boldsymbol{b} = M_{rp}(w_1 \boldsymbol{S}_{\boldsymbol{o}} + w_2 \boldsymbol{S}_{\boldsymbol{lt}} + w_3 \boldsymbol{S}_{\boldsymbol{fr}}),$$

and

$$0 < w_4 < 1,$$

the iteration converges to a unique stationary solution $\boldsymbol{S}_{\boldsymbol{R}}^{*}$. Moreover, the corresponding paper scores $\boldsymbol{S}_{\boldsymbol{P}}^{*}$ also converge.

Proof.

(1) Non-negativity. Since M_{rp} and M_{pr} are non-negative, their product $M = M_{rp}M_{pr}$ is also non-negative. The vector $\boldsymbol{b}$ is a non-negative linear combination of non-negative sub-metric vectors and is therefore non-negative.

(2) Spectral radius bound. Since both M_{rp} and M_{pr} are column-stochastic,

$$\|M_{rp}\|_1 = \|M_{pr}\|_1 = 1.$$

Thus,

$$\|M\|_1 = \|M_{rp}M_{pr}\|_1 \leq \|M_{rp}\|_1 \|M_{pr}\|_1 = 1.$$

Since

$$\rho(M) \leq \|M\|_1,$$

we obtain

$$\rho(M) \leq 1.$$

(3) Contraction property. Since $0 < w_4 < 1$,

$$\rho(w_4 M) = w_4 \rho(M) < 1.$$

Therefore the mapping

$$F(\boldsymbol{S}_{\boldsymbol{R}}) = w_4 M \boldsymbol{S}_{\boldsymbol{R}} + \boldsymbol{b}$$

is a contraction mapping.

(4) Convergence. By the Banach fixed-point theorem [48], the iteration

$$\boldsymbol{S}_{\boldsymbol{R}}^{(t+1)} = w_4 M \boldsymbol{S}_{\boldsymbol{R}}^{(t)} + \boldsymbol{b}$$

converges to a unique fixed point $\boldsymbol{S}_{\boldsymbol{R}}^{*}$. Because

$$\boldsymbol{S}_{P}^{(t)} = w_1 \boldsymbol{S}_o + w_2 \boldsymbol{S}_{lt} + w_3 \boldsymbol{S}_{fr} + w_4 M_{pr} \boldsymbol{S}_{R}^{(t)},$$

the paper scores are a linear transformation of a convergent sequence. Hence $\boldsymbol{S}_{P}^{(t)}$ also converges.

□

From Proposition 1, we conclude that the proposed DIP method converges to a unique stationary solution within each stage when the transition matrices are non-negative and column-stochastic and the citation-quality weight satisfies $0 < w_4 < 1$.

When weights are estimated per stage via (4), we fix w within each stage so that the contraction condition in Proposition 1 remains valid. Allowing the DIP inner loop to converge before updating the weights ensures that the estimation of them is based on stable influence scores rather than transient intermediate states, which prevents oscillation in the outer stages. Moreover, since the LLM-derived scoring is costly to obtain and does not change rapidly, it is unnecessary to refit the weights before the inner dynamics have stabilized. Stages are therefore repeated until the change of the fitted weights between two consecutive stages becomes smaller than 10^{-6}, giving practical stability to the overall procedure.

3.5. Overall Procedure

We summarize the entire framework in a stage–iteration structure. Each stage refreshes the paper-level LLM scores, performs robust aggregation, fits the combination weights, and then carries out an inner loop of coupled

updates. The detailed procedure is as follows.

- Step 1: Initialization. Randomly initialize the score vector of researchers $\boldsymbol{S}_{\boldsymbol{R}}^{(0)}$. Select a fixed subset of C papers for the subsequent stages.
- Step 2: Pre-calculation. Pre-calculate the three sub-metrics for research papers, namely the originality-oriented downstream influence, long-term influence, and field-specific ranking, as they do not need to be updated during the stages.
- Step 3: Stage-level LLM score. At the beginning of current stage, obtain multiple paper scores from the LLM scorers for each paper in the fixed subset of C papers.
- Step 4: α-trim aggregation. Aggregate the valid LLM outputs of the sampled papers through the α-trim procedure to obtain the robust paper scores.
- Step 5: Weight fitting. Estimate the combination weights through the constrained least squares approach based on (4). The fitted weights w remain fixed within the current stage and are applied to all papers.
- Step 6: Inner iterations. With the weights fixed, update the paper scores based on (8) and (9), and researcher scores based on (10) in alternating iterations. Stop the inner loop when the change of paper scores between two consecutive iterations becomes sufficiently small based on (11).
- Step 7: Stage convergence check. After the inner iterations converge, compare the current fitted weights with those from the previous stage. If the change is below the threshold, i.e., 10^{-6}, proceed to Step 8. Otherwise start the next stage and return to Step 3.
- Step 8: Convergence and Outputs. Output the latest impact scores $\boldsymbol{S_P}$ for research papers and $\boldsymbol{S_R}$ for researchers.

The overall procedure of the proposed CLOF framework is given in Figure 4.

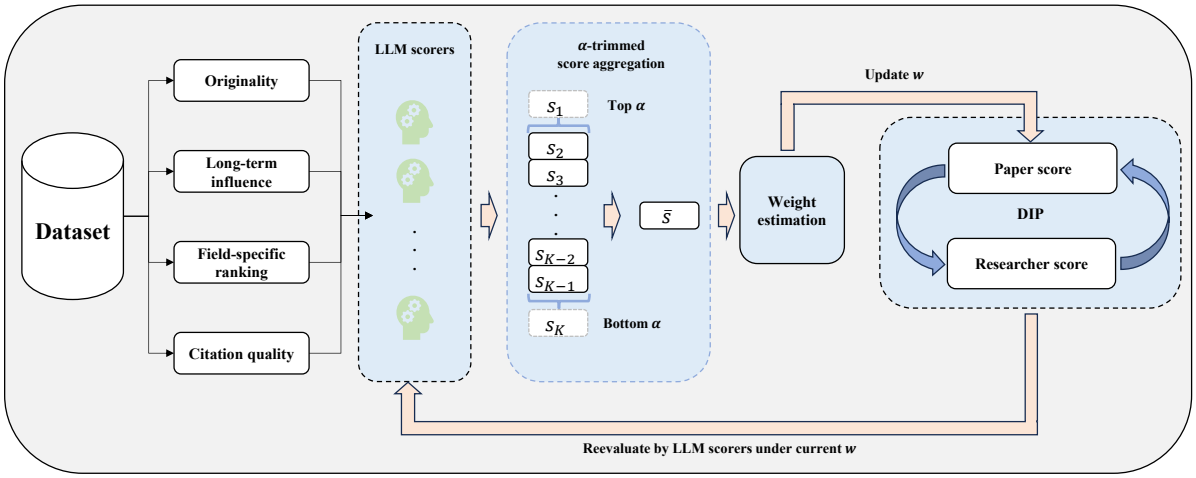


Figure 4. Overview of the proposed CLOF framework. The four paper-level sub-metrics are first extracted from the dataset and combined through LLM-guided weight estimation with α-trimmed score aggregation. The learned weights are then used in the Double Iterative Power (DIP) process to iteratively update paper and researcher scores until convergence.

Compared with traditional Bibliometric and graph-based metrics, CLOF provides several advantages. First, it jointly considers multiple complementary factors, including originality-oriented downstream influence, long-term influence, field-specific ranking, and citation quality, rather than relying solely on raw citation counts or graph centrality. Second, the relative importance of these factors is learned through LLM-guided weight estimation instead of manually predefined weighting schemes. Third, CLOF explicitly accounts for differences in author contributions when evaluating researchers. Finally, unlike existing approaches that assess papers and researchers independently, CLOF models their mutual influence through the proposed DIP mechanism with theoretical convergence guarantees.

4. Experiments

In this section, we conduct experiments to verify the effectiveness of the proposed CLOF framework. We first give the details of the dataset construction process, followed by experimental setups. Then, detailed experimental results and discussions are presented.

4.1. Dataset

4.1.1. Data Collections

We construct the experimental dataset based on the AMiner platform [49]. The platform provides comprehensive metadata on academic publications, including titles, authors, publication years, venues, and citation information. These metadata are essential for constructing reliable citation networks. However, AMiner does not provide the full text of papers, which is necessary for performing sentiment analysis on citation contexts in our research impact evaluation. To address this limitation and manage the high computational cost of retrieving and processing the full text of papers from external sources, we construct a focused subset centered around a single researcher.

Specifically, we first extract the corresponding published papers of the selected researcher as the central nodes. Then, we extract a local citation subgraph from the AMiner dataset, including the research papers that cited or were cited by these central papers. The citation relationships among the extracted papers are preserved during extraction.

After the prior step, we have extracted a subset of the AMiner dataset beginning from a selected central researcher. For citation-context sentiment analysis, we collect full texts only when openly accessible versions are available. We use arXiv (https://arxiv.org/ (accessed on 5 October 2025)) as the primary source and Semantic Scholar (https://www.semanticscholar.org/ (accessed on 5 October 2025)) solely to locate open-access versions. Then, we perform sentiment analysis on the corresponding citation context to obtain the scores of citation quality of citations based on Algorithm 1. Only the sentiment scores of citations that can be matched to references in the AMiner dataset will be retained. Unmatched citations in the networks will be assigned the average score of these matched citations to ensure consistency and completeness of the dataset.

Besides the citation quality, the proposed paper metric also incorporates the research fields of each paper. Specifically, in this work, we adopt the four broad disciplinary categories defined by the Scopus database, namely Life Science, Physical Sciences, Health Sciences, and Social Sciences, as the types of research fields. Due to inconsistencies and noise in the original venue data, each paper is assigned to one or more disciplinary categories based on the most similar Scopus-indexed venue. The assignment is determined through a venue-similarity matching strategy.

In summary, we build a researcher-centered local citation network by integrating structured data from AMiner with supplementary full text information, which provides a solid foundation for the subsequent design and analysis of research impact metrics.

4.1.2. Dataset Statistics

Table 1 presents the key statistics of the constructed datasets before and after the removal of self-citations. Specifically, the constructed dataset contains a total of 11,682 papers authored by 18,723 unique authors. While the number of papers and authors remains unchanged, the number of links decreases from 81,284 to 74,544 after removing self-citations. As a result, the average number of citations for each paper slightly drops from 6.96 to 6.38. The removal of self-citations helps improve the reliability of the citation information for subsequent analysis.

Table 1. Dataset statistics.

Statistics	Before	After
# Nodes (Papers)	11,682	11,682
# Links (Citations)	81,284	74,544
# Authors	18,723	18,723
# Citations per Paper	6.96	6.38

4.2. Experimental Setups

For the experimental setting, we first exclude self-citations to improve evaluation fairness. For the long-term influence sub-metric, we set the citation-threshold group to $T = \{10, 20, \cdots\}$ with an interval of 10. Citation matching adopts a similarity threshold of 0.9 to ensure accurate alignment between citation contexts and reference entries in the AMiner dataset. For venue matching, papers with a venue similarity above 0.9 are assigned the corresponding Scopus disciplinary category. For semantic embedding ϕ in Algorithm 1, we employ SciBERT [50] and use cosine similarity as the similarity function. The maximum expansion length is set to $H = 10$. For sentiment analysis, we adopt DistilBERT [51], which has been fine-tuned on SST-2 for binary sentiment classification.

The default combination weights of four sub-metrics are initialized to 0.25. For weight estimation, we sample $C = 200$ representative papers using citation-stratified sampling. The LLM evaluation is implemented using the Mixtral-8x7B-Instruct model [52]. For each sampled paper, the model generates impact scores through $K = 10$

independent stochastic generations under an identical prompt template (temperature = 0.4). The generated scores are aggregated using an α-trimmed mean ($\alpha = 0.1$), where the highest and lowest 10% of scores are discarded before averaging. The resulting aggregated scores are used as supervision signals for estimating the weights of the four sub-metrics. In cases where the corresponding author is not explicitly indicated, we adopt the convention of treating the last author as the corresponding author to ensure reproducibility. In particular, we employ a sparse initialization strategy. A small number of entries in $\boldsymbol{S}_{\boldsymbol{R}}^{(0)}$ are randomly initialized as non-zero values, while most entries remain zero. This strategy improves convergence speed. All experiments are conducted on a single NVIDIA A800 GPU with 80-GB GPU memory and 14-vCPU Intel Xeon Gold 6348 CPUs @ 2.60-GHz. The source code and processed datasets will be publicly released upon publication.

4.3. Baselines

We compare the CLOF framework against widely recognized metrics adopted by established academic platforms and ranking systems. For the impact of research papers, we utilize Google Scholar Citation Count (Google Scholar) [15], a foundational and universally accepted indicator of paper influence, which serves as the basis for numerous Bibliometric analyses and is supported by major databases such as Scopus and Web of Science. For the impact of researchers, we adopt the ScholarGPS ranking [53], which integrates productivity (i.e., number of published papers), citation impact (i.e., citation count), and quality (i.e., h-index) into a composite score.

4.4. Results and Discussions

4.4.1. Correlations with Existing Metrics

We conduct the correlation analysis between our two CLOF metrics with Google Scholar and ScholarGPS metrics, respectively. Specifically, the Pearson coefficient measures the linear relationship between actual values, while Spearman correlation assesses the monotonic relationship between ranked values. For paper-impact evaluation, Table 2 shows that CLOF-P exhibits a positive correlation with the Google Scholar metric. The correlation is particularly strong in terms of Spearman coefficients, suggesting that CLOF-P captures information that is broadly consistent with existing indicators.

Table 2 also presents the corresponding correlation coefficients of CLOF-R and ScholarGPS. Since the Spearman coefficient measures the consistency of rankings rather than absolute score values, a lower Spearman correlation indicates larger ranking differences between the two researcher evaluation metrics. This result is expected because CLOF-R considers author contributions according to author order and is jointly influenced by the coupled paper–researcher update process, whereas ScholarGPS primarily relies on publication counts, citation counts, and h-index related indicators. These methodological differences naturally lead to larger variations in researcher rankings while maintaining a positive overall correlation.

Table 2. Correlation coefficients between CLOF scores and existing metrics.

Metric Pair	Correlation Type	Value
CLOF-P vs. Google Scholar	Pearson	0.3319
	Spearman	0.8958
CLOF-R vs. ScholarGPS	Pearson	0.4956
	Spearman	0.4601

4.4.2. Illustrative Comparison with Existing Metrics

Based on previous experiments, we know that CLOF metrics exhibit a certain degree of consistency with existing metrics. In this part, we shift our focus to the differences, as some discrepancies have also been observed. We conduct the following experiments from a micro perspective by selecting several representative papers/researchers. Specifically, we selected two papers/researchers where the first has a high CLOF score while a low Google Scholar/ScholarGPS score, and the second is the contrary. The detailed findings are as follows.

- Originality-oriented downstream influence. For the originality-oriented downstream influence, we adopt the Sankey diagram to visualize the citation flow of papers, highlighting the influence propagates from the rooted paper to downstream layers. We first select a representative paper that has a high originality-oriented downstream influence score but a low citation count. As shown in Figure 5, though this paper has only 18 raw citation count, it has a high number of higher-order citations (i.e., 8292 citation count). If we directly

adopt the simple Google Scholar citation count, we may ignore the specific credit it deserves. However, the situation differs for another selected paper, as shown in Figure 6. Although this paper received a relatively high citation count (i.e., 239), it accumulated only 634 higher-order citations. This result suggests that its influence on subsequent research may be more limited and that it generated less downstream attention.

- Long-term Influence. For the long-term influence, we select another two representative papers based on a similar strategy. Paper 248813, identified by Google Scholar, obtained a better Google Scholar rank, while Paper 479658 received a better CLOF-P rank. We plot the annual citation distribution of these two papers, as shown in Figure 7. The figure shows that the paper identified by Google Scholar failed to attract continuous attention over time as its citations were largely concentrated within a few years after publication. In contrast, the paper identified by CLOF-P demonstrates a more persistent citation trajectory, suggesting that our metric is better suited to uncover works with enduring academic influence and deeper integration into the research landscape.
- Field-specific Ranking. Then, for the field-specific ranking, we present the details of two representative papers in Table 3. Specifically, the selected two papers, which are from the fields of Health Sciences and Social Sciences, received 36 and 160 citations, respectively. As a result, the second paper ranks significantly higher than the first when evaluated solely by Google Scholar metric. However, once field-specific citation patterns are considered, the ranking is reversed: the first paper rises to 1st, while the second drops to 40th. This highlights the importance of incorporating disciplinary context into impact assessment, as citation behaviors vary substantially across research fields. The proposed CLOF-P metric better captures the relative impact within each research field, enabling fairer and more meaningful comparisons across disciplines.
- Citation Quality. The detailed citations of the two selected papers in terms of citation quality are given in Figures 8 and 9, respectively. Although the second paper received more citations than the first paper, i.e., 10 vs. 3, the majority of its citing authors have relatively low researcher scores. By contrast, the first paper was cited by high-impact researchers, and the citation was strongly positive, substantially enhancing its citation quality score. The above results suggest that traditional Google Scholar metric alone may not reliably reflect the impact of papers. What truly matters is who cites the work and in what context. A positive citation from highly influential researchers carries far more weight than numerous superficial or even negative citations. Our CLOF-P metric incorporates both the credibility of citing authors and the motivation of their citations, offering a more meaningful assessment of research impact.
- Author Position. Last but not least, we selected two representative researchers based on the assessment of CLOF-R and ScholarGPS metrics, respectively. We plot the author position distribution in their published papers of these two researchers, as shown in Figure 10. We observe that the researcher identified by ScholarGPS ranks highly according to that metric. However, most publications place this researcher in middle-author positions, suggesting a less dominant contribution to many of the papers. In contrast, the researcher identified by our CLOF-R frequently appears as either the first author or corresponding author, indicating more central contributions. Our method accounts for author position through a linearly decreasing credit allocation, thereby offering a more accurate reflection of individual scholarly leadership rather than assigning equal credit to all authors.

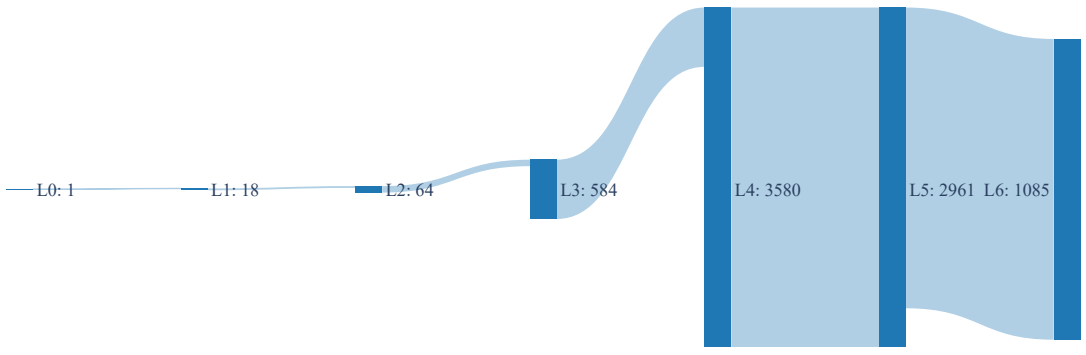


Figure 5. Citation Sankey diagram rooted at the representative paper (i.e., Paper 836134) selected by CLOF-P metric, which ranks 1st by CLOF-P (Originality-oriented downstream influence) and 860th by Google Scholar.



Figure 6. Citation Sankey diagram rooted at the representative paper (i.e., Paper 926881) selected by CLOF-P metric, which ranks 31st by Google Scholar and 607th by CLOF-P (Originality-oriented downstream influence).

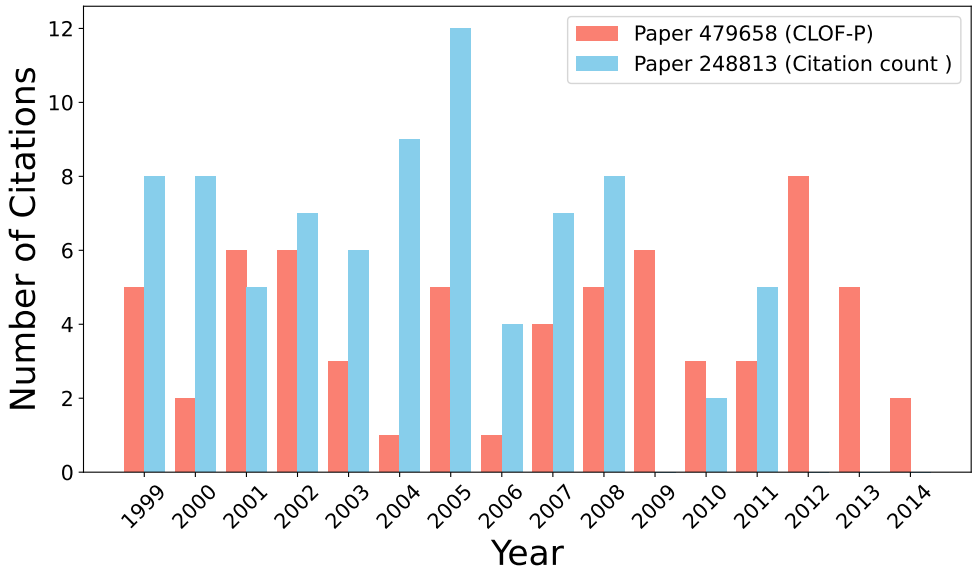


Figure 7. Comparison of annual citation distributions for representative papers selected by CLOF-P (Long-term Influence) and Google Scholar metrics. Specifically, Paper 248813 ranks 140th by Google Scholar and 265th by CLOF-P (Long-term Influence), and Paper 479658 ranks 164th by CLOF-P (Long-term Influence) and 187th by Google Scholar.

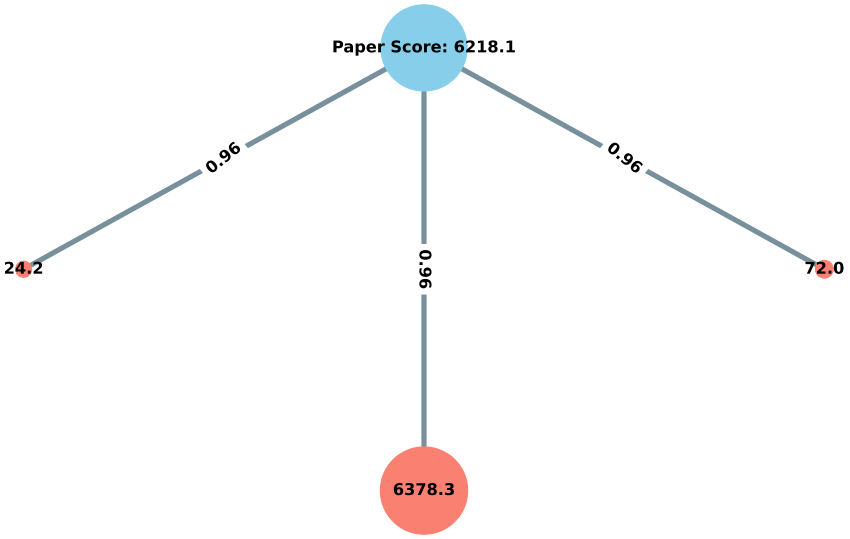


Figure 8. Citation quality visualization of the representative paper (Paper 939087) selected by CLOF-P (Citation Quality), which ranks 1356th by CLOF-P (Citation Quality) and 2986th by Google Scholar. Orange nodes represent the center papers, and blue nodes are their citing neighbors. Node size reflects the total author influence score (for neighbors) or the final score of paper (for the center). Edge thickness encodes the sentiment score of the citation context.

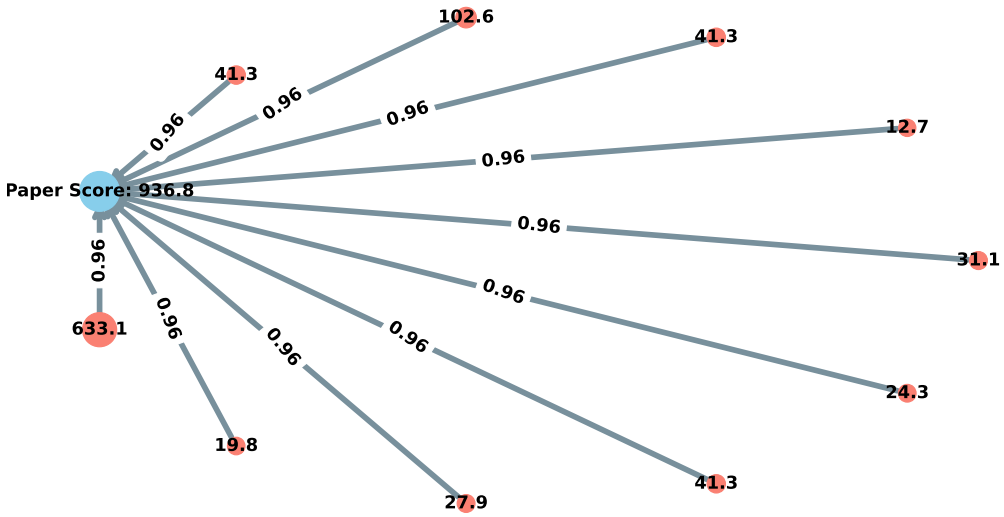


Figure 9. Citation quality visualization of the representative paper (Paper 863385) selected by CLOF-P (Citation Quality), which ranks 1459th by Google Scholar and 2504th by CLOF-P (Citation Quality). Orange nodes represent the center papers, and blue nodes are their citing neighbors. Node size reflects the total author influence score (for neighbors) or the final score of paper (for the center). Edge thickness encodes the sentiment score of the citation context.

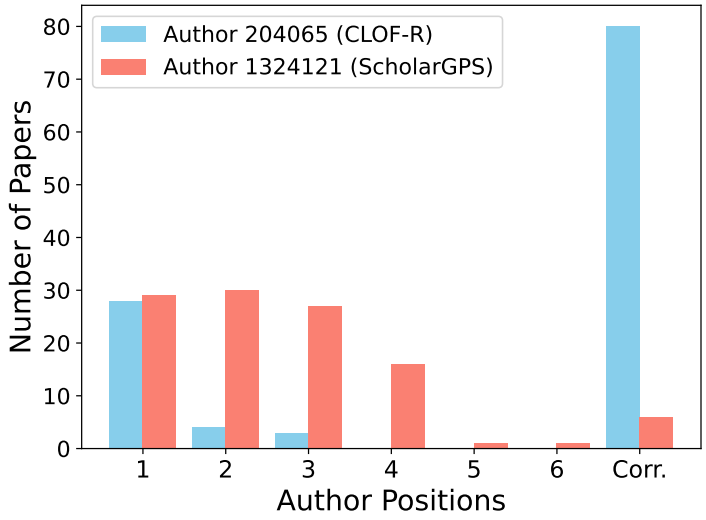


Figure 10. Comparison of author position distributions for representative researchers selected by CLOF-R and ScholarGPS metrics. In particular, 'Corr.' refers to the corresponding author. Author 204065 ranks 287th by CLOF-R and 779th by ScholarGPS, and Author 1324121 ranks 311st by ScholarGPS and 649th by CLOF-R.

Table 3. Comparison of two papers across Google Scholar and CLOF-P Metrics (by Removing Self-citations).

Paper ID	Field	Citation Count	Google Scholar Rank	CLOF-P (Field-specific) Rank
430746	Health Sciences	36	388	1
310500	Social Sciences	160	52	40

4.4.3. Component-Wise Analysis of Top Papers

We further visualize the detailed sub-scores of the top papers identified by CLOF-P. Figure 11 presents the sub-scores of the top 10 papers identified by the CLOF-P metric. This observation is consistent with the weight analysis in Section 4.4.4, where the field-specific ranking metric receives relatively larger learned weights than the other sub-metrics. As a result, papers with strong field-specific ranking scores are more likely to obtain higher CLOF-P scores and consequently appear among the top-ranked papers shown in Figure 11.

At the same time, most of these papers also exhibit strong performance in one or more additional factors, including originality-oriented downstream influence, long-term influence, and citation quality. This suggests that while field-specific ranking serves as an important indicator of research impact of papers, the remaining sub-metrics provide complementary information for distinguishing papers with similarly strong field-specific rankings. Therefore, CLOF-P enables a more comprehensive evaluation of research impact by considering multiple aspects simultaneously rather than relying on a single indicator.

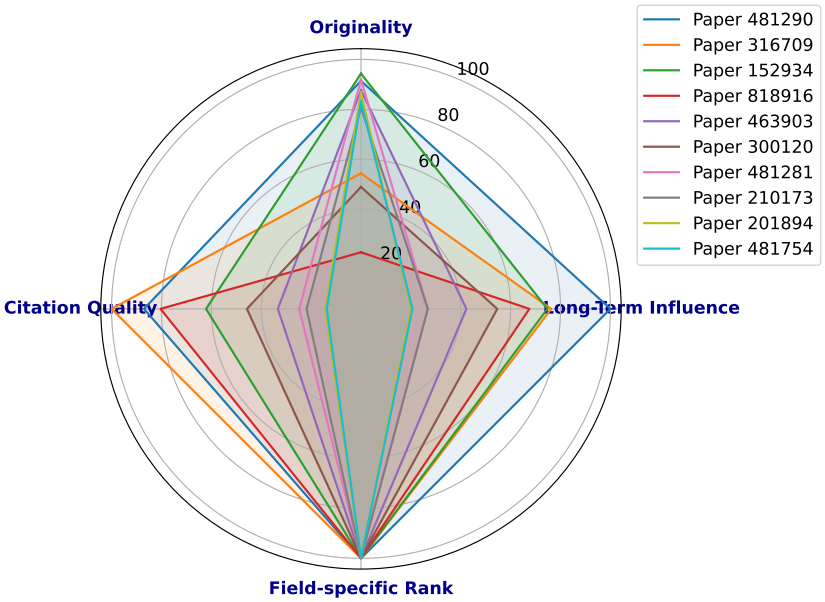


Figure 11. Radar chart visualization of the top 10 papers ranked by CLOF-P, showing individual scores across the four sub-metrics. Each axis represents one sub-component of the CLOF-P.

4.4.4. Impact of the Size of Sampled Papers

Based on the overall pipeline of CLOF, the sampled papers play a critical role in determining the combination weights of four sub-metrics. We therefore examine how different sample sizes influence the resulting weights. We consider three sampling-size settings of 100, 200, and 300 papers. Figure 12 shows the evolution of the learned weights across stages under these settings. Within each setting, the weights remain relatively stable throughout the stage-wise updates. Among all sub-metrics, the field-specific ranking metric contributes more consistently to the final paper score. Moreover, the differences across the three sampling sizes are small, indicating that the CLOF framework is not sensitive to the size of sampled papers.

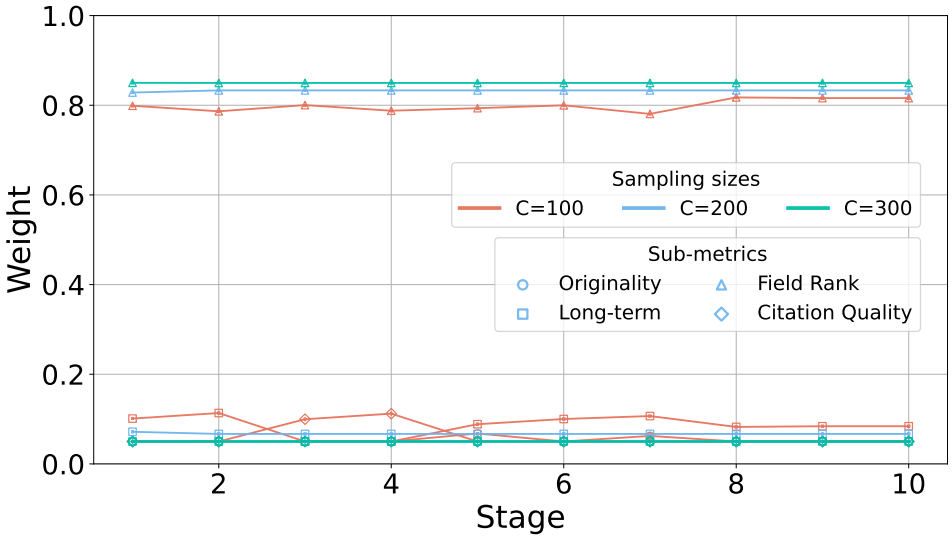


Figure 12. Evolution of sub-metric weights across stage-wise refinements.

4.4.5. Convergence Analysis

Finally, we conduct a convergence analysis of the CLOF framework. Figure 12 illustrates the stage-wise convergence behavior of CLOF. The learned weights stabilize within only a few stages, indicating that the update process rapidly reaches a consistent weight distribution. Moreover, Figure 13 presents the change of loss with regard to running iteration in the final stage. It is evident that our CLOF achieves a rapid convergence within 5 iterations, demonstrating its computational efficiency and stability. This rapid convergence not only ensures practical feasibility for iterative updates but also highlights the potential scalability of the CLOF framework for applications in large-scale academic data.

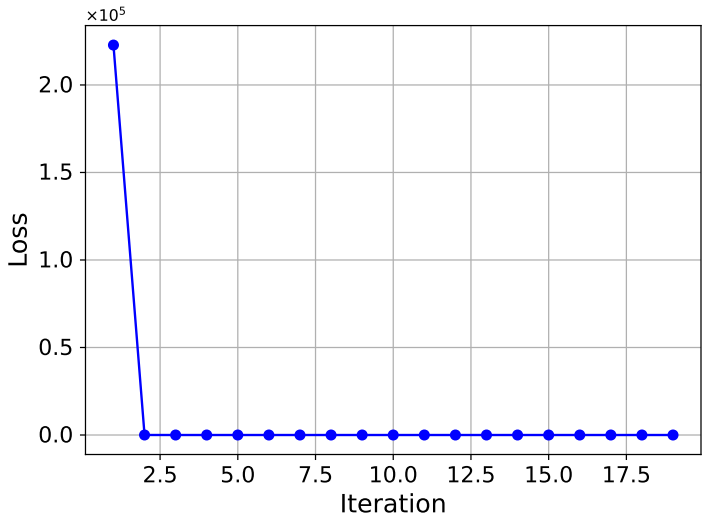


Figure 13. Loss change within the inner-loop iterations of the final stage.

5. Conclusions

In this study, we propose a coupled framework, CLOF, to quantify the research impact of both research papers and researchers. Specifically, the paper metric incorporates several key factors, including originality-oriented downstream influence, long-term influence, field-specific ranking, and citation quality. The researcher metric further accounts for the contributions of different authors. Moreover, the citation quality component employs an adaptive semantic segmentation strategy to extract more relevant citation contexts for sentiment evaluation. In particular, unlike existing methods that evaluate research papers and researchers separately, we adopt a coupled strategy to iteratively update their impact scores. This strategy is implemented through the proposed Double Iterative Power (DIP) method and enables a more comprehensive assessment of research impact. The framework further incorporates an LLM-enhanced module to learn the combination weights of paper-level sub-metrics. Multiple LLM scorers provide stage-wise evaluation signals, which are aggregated through an α-trimmed mean. The resulting scores are then fitted using a constrained least squares approach. We conduct extensive experiments to illustrate the behavior, interpretability, and practical applicability of the proposed CLOF framework on a benchmark dataset.

Despite the encouraging results, the proposed CLOF framework still has several limitations. First, some components rely on practical approximations and simplifying assumptions. For example, the originality-oriented downstream influence, long-term influence, citation quality, and author-contribution mechanisms are designed to capture specific aspects of research impact. They are not intended to provide complete characterizations of these concepts. Their effectiveness may be influenced by factors such as disciplinary conventions, temporal patterns, and the availability of metadata. Second, several modules depend on auxiliary models and external resources. For instance, citation quality is estimated through sentiment analysis and adaptive context extraction, while submetric weights are learned through an LLM-assisted procedure. These designs provide a practical way to incorporate information beyond citation counts. However, their reliability has not yet been systematically validated against large-scale expert annotations or dedicated benchmark datasets. Third, the current evaluation is conducted on a researcher-centered benchmark dataset with specific disciplinary coverage. This setting provides a well-defined testbed for studying coupled paper-researcher impact quantification. Nevertheless, further validation on broader academic corpora, more diverse research communities, and larger citation networks would provide stronger evidence regarding the generalizability and robustness of the proposed framework. Future work will focus on addressing these limitations, exploring alternative modeling and aggregation strategies, and extending CLOF to larger-scale and more diverse academic datasets.

Author Contributions

J.F.: Conceptualization, investigation, writing, and revision; Z.Y.: Conceptualization, investigation, writing, and revision; X.C.: Conceptualization, investigation, writing, and revision; H.W.: Conceptualization, investigation, writing, and revision; D.O.W.: Conceptualization, investigation, writing, and revision. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The source code and processed datasets will be made publicly available through an online repository upon publication.

Conflicts of Interest

Given the role as Editor-in-Chief of Transactions on Artificial Intelligence, D.O.W. had no involvement in the peer review of this paper and had no access to information regarding its peer-review process. Full responsibility for the editorial process of this paper was delegated to another editor of the journal.

Use of AI and AI-Assisted Technologies

During the preparation of this work, the authors used ChatGPT for language polishing and grammatical refinement. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

1. Kostoff, R.N. Research Impact Quantification. *R D Manag.* **1994**, *24*, 207–218. https://doi.org/10.1111/j.1467-9310.1994.tb00874.x.
2. Lane, J. Let's Make Science Metrics More Scientific. *Nature* **2010**, *464*, 488–489. https://doi.org/10.1038/464488a.
3. Mustafa, G.; Rauf, A.; Ahmed, B.; et al. Comprehensive Evaluation of Publication and Citation Metrics for Quantifying Scholarly Influence. *IEEE Access* **2023**, *11*, 65759–65774. https://doi.org/10.1109/access.2023.3290917.
4. Klebel, T.; Traag, V.; Grypari, I.; et al. The Academic Impact of Open Science: A Scoping Review. *R. Soc. Open Sci.* **2025**, *12*, 241248. https://doi.org/10.1098/rsos.241248.
5. Calò, L.N. Impact Metrics and Science Evaluation. *Rev. Peru. Med. Exp. Salud Publica* **2022**, *39*, 236–240. https://doi.org/10.17843/rpmesp.2022.392.11171.
6. Wei, Q.; Mingkun, W.; Danyang, L.; et al. Identification and Portraits of Open Access Journals Based on Open Impact Metrics Extracted from Social Activities. *J. Sch. Publ.* **2023**, *54*, 597–623. https://doi.org/10.3138/jsp-2022-0081.
7. Ou, G.; Zhao, K.; Zuo, R.; et al. Effects of Research Funding on the Academic Impact and Societal Visibility of Scientific Research. *J. Inf.* **2024**, *18*, 101592. https://doi.org/10.1016/j.joi.2024.101592.
8. Faghri, A.; Bergman, T.L. Ranking Academic Institutions Based on the Productivity, Impact, and Quality of Institutional Scholars. *J. Data Inf. Sci.* **2024**, *9*, 116–154. https://doi.org/10.2478/jdis-2024-0017.
9. Wiechetek, Ł.; Pastuszak, Z. Academic Social Networks Metrics: An Effective Indicator for University Performance? *Scientometrics* **2022**, *127*, 1381–1401. https://doi.org/10.1007/s11192-021-04258-6.
10. Leal Filho, W.; Viera Trevisan, L.; Wahaj, Z.; et al. University Rankings and Sustainable Development: The State of the Art. *Int. J. Sustain. High. Educ.* **2025**, *26*, 1615–1636. https://doi.org/10.1108/ijshe-11-2023-0530.
11. Rushforth, A.; De Rijcke, S. Practicing Responsible Research Assessment: Qualitative Study of Faculty Hiring, Promotion, and Tenure Assessments in the United States. *Res. Eval.* **2024**, *33*, rvae007. https://doi.org/10.1093/reseval/rvae007.
12. Agarwal, A.; Durairajanayagam, D.; Tatagari, S.; et al. Bibliometrics: Tracking Research Impact by Selecting the Appropriate Metrics. *Asian J. Androl.* **2016**, *18*, 296–309. https://doi.org/10.4103/1008-682x.171582.
13. Lyu, P.H.; Liu, X.; Yao, T. A Bibliometric Analysis of Literature on Bibliometrics in Recent Half-Century. *J. Inf. Sci.* **2023**, *52*, 324–344. https://doi.org/10.1177/01655515231191233.

14. Ninkov, A.; Frank, J.R.; Maggio, L.A. Bibliometrics: Methods for Studying Academic Publishing. *Perspect. Med. Educ.* **2021**, *11*, 173–176. https://doi.org/10.1007/s40037-021-00695-4.
15. Garfield, E. Citation Indexes for Science. *Science* **1955**, *122*, 108–111. https://doi.org/10.1126/science.122.3159.108.
16. Hirsch, J.E. An Index to Quantify an Individual's Scientific Research Output. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 16569–16572. https://doi.org/10.1073/pnas.0507655102.
17. Manghi, P.; Mannocci, A.; Osborne, F.; et al. New Trends in Scientific Knowledge Graphs and Research Impact Assessment. *Quant. Sci. Stud.* **2021**, *2*, 1296–1300. https://doi.org/10.1162/qss_e_00160.
18. Życzkowski, K. Citation Graph, Weighted Impact Factors and Performance Indices. *Scientometrics* **2010**, *85*, 301–315. https://doi.org/10.1007/s11192-010-0208-6.
19. Giatsidis, C.; Nikolentzos, G.; Zhang, C.; et al. Rooted Citation Graphs Density Metrics for Research Papers Influence Evaluation. *J. Inf.* **2019**, *13*, 757–768. https://doi.org/10.1016/j.joi.2019.03.006.
20. Page, L.M.; Brin, S.; Motwani, R.; et al. *The PageRank Citation Ranking: Bringing Order to the Web*; Stanford InfoLab: Stanford, CA, USA, 1999.
21. Costas, R.; Bordons, M. The H-Index: Advantages, Limitations and Its Relation with Other Bibliometric Indicators at the Micro Level. *J. Inf.* **2007**, *1*, 193–203. https://doi.org/10.1016/j.joi.2007.02.001.
22. Schreiber, M. Self-Citation Corrections for the Hirsch Index. *Europhys. Lett. EPL* **2007**, *78*, 30002. https://doi.org/10.1209/0295-5075/78/30002.
23. Maabreh, M.; Alsmadi, I.M. A Survey of Impact and Citation Indices: Limitations and Issues. *Int. J. Adv. Sci. Technol.* **2012**, *40*, 35–54.
24. Haustein, S.; Larivière, V. The Use of Bibliometrics for Assessing Research: Possibilities, Limitations and Adverse Effects. In *Incentives and Performance: Governance of Research Organizations*; Springer: Cham, Switzerland, 2015; pp. 121–139. https://doi.org/10.1007/978-3-319-09785-5_8.
25. Mesiar, R.; Gagolewski, M. H-Index and Other Sugeno Integrals: Some Defects and Their Compensation. *IEEE Trans. Fuzzy Syst.* **2016**, *24*, 1668–1672. https://doi.org/10.1109/tfuzz.2016.2516579.
26. Manoj Kumar, L.; George, R.J.; Anisha, P.S. Bibliometric Analysis for Medical Research. *Indian J. Psychol. Med.* **2023**, *45*, 277–282. https://doi.org/10.1177/02537176221103617.
27. Tomaszewski, R. Visibility, Impact, and Applications of Bibliometric Software Tools through Citation Analysis. *Scientometrics* **2023**, *128*, 4007–4028. https://doi.org/10.1007/s11192-023-04725-2.
28. Hassan, W.; Duarte, A.E. Bibliometric Analysis: A Few Suggestions. *Curr. Probl. Cardiol.* **2024**, *49*, 102640. https://doi.org/10.1016/j.cpcardiol.2024.102640.
29. Matorevhu, A. Bibliometrics: Application Opportunities and Limitations. In *Bibliometrics - An Essential Methodological Tool for Research Projects*; IntechOpen: London, UK, 2024. https://doi.org/10.5772/intechopen.1005292.
30. Alter, N.; Daiem, M.; Pontell, M.E.; et al. Limitations of Academic Bibliometric Indices: The Need for More Comprehensive Metrics. *Ann. Plast. Surg.* **2025**, *95*, 603–606. https://doi.org/10.1097/sap.0000000000004484.
31. Egghe, L. Theory and Practise of the G-Index. *Scientometrics* **2006**, *69*, 131–152. https://doi.org/10.1007/s11192-006-0144-7.
32. Burrell, Q.L. On the H-Index, the Size of the Hirsch Core and Jin's a-Index. *J. Inf.* **2007**, *1*, 170–177. https://doi.org/10.1016/j.joi.2007.01.003.
33. Khan, N.R.; Thompson, C.J.; Taylor, D.R.; et al. Part II: Should the H-Index Be Modified? An Analysis of the M-Quotient, Contemporary H-Index, Authorship Value, and Impact Factor. *World Neurosurg.* **2013**, *80*, 766–774. https://doi.org/10.1016/j.wneu.2013.07.011.
34. Ibrahim, N.; Habacha Chaibi, A.; Ben Ahmed, M. New Scientometric Indicator for the Qualitative Evaluation of Scientific Production. *New Libr. World* **2015**, *116*, 661–676. https://doi.org/10.1108/nlw-01-2015-0002.
35. Farooq, M.; Khan, H.U.; Iqbal, S.; et al. DS-Index: Ranking Authors Distinctively in an Academic Network. *IEEE Access* **2017**, *5*, 19588–19596. https://doi.org/10.1109/access.2017.2744798.
36. Yang, A.J.; Hu, H.; Zhao, Y.; et al. From Consolidation to Disruption: A Novel Way to Measure the Impact of Scientists and Identify Laureates. *Inf. Process. Manag.* **2023**, *60*, 103420. https://doi.org/10.1016/j.ipm.2023.103420.
37. Zhao, W.; Korobskiy, D.; Chacko, G. Delayed Recognition: A Co-Citation Perspective. *Front. Res. Metrics Anal.* **2021**, *5*, 577131. https://doi.org/10.3389/frma.2020.577131.
38. Albert, R.; Barabási, A.L. Statistical Mechanics of Complex Networks. *Rev. Mod. Phys.* **2002**, *74*, 47. https://doi.org/10.1103/revmodphys.74.47-97.
39. Bloch, F.; Jackson, M.O.; Tebaldi, P. Centrality Measures in Networks. *Soc. Choice Welf.* **2023**, *61*, 413–453. https://doi.org/10.1007/s00355-023-01456-4.
40. Ugurlu, O. Comparative Analysis of Centrality Measures for Identifying Critical Nodes in Complex Networks. *J. Comput. Sci.* **2022**, *62*, 101738. https://doi.org/10.1016/j.jocs.2022.101738.
41. Ding, Y.; Yan, E.; Frazho, A.; et al. PageRank for Ranking Authors in Co-Citation Networks. *J. Am. Soc. Inf. Sci. Technol.* **2009**, *60*, 2229–2243. https://doi.org/10.1002/asi.21171.
42. Fiala, D. Time-Aware PageRank for Bibliographic Networks. *J. Inf.* **2012**, *6*, 370–388. https://doi.org/10.1016/j.joi.2012.02.002.
43. Kleinberg, J.M. Authoritative Sources in a Hyperlinked Environment. *J. ACM* **1999**, *46*, 604–632. https://doi.org/10.

1145/324133.324140.

44. Bergstrom, C. Eigenfactor: Measuring the Value and Prestige of Scholarly Journals. *Coll. Res. Libr. News* **2007**, *68*, 314–316. https://doi.org/10.5860/crln.68.5.7804.
45. Franceschet, M.; Colavizza, G. Quantifying the Higher-Order Influence of Scientific Publications. *Scientometrics* **2020**, *125*, 951–963. https://doi.org/10.1007/s11192-020-03580-9.
46. Wankhade, M.; Rao, A.C.S.; Kulkarni, C. A Survey on Sentiment Analysis Methods, Applications, and Challenges. *Artif. Intell. Rev.* **2022**, *55*, 5731–5780. https://doi.org/10.1007/s10462-022-10144-1.
47. Brin, S.; Page, L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Comput. Netw. ISDN Syst.* **1998**, *30*, 107–117. https://doi.org/10.1016/s0169-7552(98)00110-x.
48. Jachymski, J.; Jóźwik, I.; Terepeta, M. The Banach Fixed Point Theorem: Selected Topics from Its Hundred-Year History. *Rev. Real Acad. Cienc. Exactas Fis. Nat. Ser. A-Mat.* **2024**, *118*, 140. https://doi.org/10.1007/s13398-024-01636-6.
49. Wan, H.; Zhang, Y.; Zhang, J.; et al. AMiner: Search and Mining of Academic Social Networks. *Data Intell.* **2019**, *1*, 58–76. https://doi.org/10.1162/dint_a_00006.
50. Beltagy, I.; Lo, K.; Cohan, A. SciBERT: A Pretrained Language Model for Scientific Text. *arXiv* **2019**, arXiv:1903.10676.
51. Sanh, V.; Debut, L.; Chaumond, J.; et al. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *arXiv* **2019**, arXiv:1910.01108. http://arxiv.org/abs/1910.01108v4.
52. Jiang, A.Q.; Sablayrolles, A.; Roux, A.; et al. Mixtral of Experts. *arXiv* **2024**, arXiv:2401.04088.
53. ScholarGPS: Quantitative Rankings and Profiles of Scholars and Institutions. Available online: https://www.scholargps.com (accessed on 23 July 2025).