Pitfalls of Publish or Perish: A novel framework for Modeling and Ranking Internationality of Scholarly Publications

Abhishek Bansal, Snehanshu Saha, Archana Mathur, Gouri Ginde, Sriparna Saha, Saroj K. Meher, Sandra Anil, Gambhire Swati Sampatrao, Sudeepa Roy Dey & Suryoday Basak

Abstract Receiving citations has become a means of support for academicians who aspire to have active research profile and wish to secure tenure. Publications tend to be associated with citations and the impact factor of the journals where articles of academicians got published in. This triggers a cycle where journals strive to attain higher impact factor and authors wish to achieve fat citation count and H-indices. The practices adopted to achieve such milestones have been under scrutiny and suspicion. Standard metrics are abused and manipulated. The paper presents a greed-survival deadlock mitigation policy that plagues scientific publishing and proposes a few novel metrics. The premise is built on the assumption that reputation of journals and authors must not be confined to local circles. A true measure of impact and influence of journals and authors must transcend familiar borders. International recognition of scholarly publications may not be solely dependent upon Impact factor of journals. The paper puts forward mathematical models to track and quantify the metrics of influence and internationality of journals and authors by proposing a novel, difficult to manipulate and quantifiable metrics. The paper proposes, for the first time, a machine learning based classification system of journal internationality.

Keywords: Non Local Influence Quotient (NLIQ); Copious citation quotient (CCQ); Cognizant Citations (CogCQ); Other-Citation-Quotient (OCQ); Stochastic Frontier Analysis (SFA); Unified Granular Neural Networks (UGNN); internationality classification; Sparse Principal Component Analysis(SPCA);Equivalence Class; Genealogy Citation Model;

1 Introduction

Recently, the number of new journals in various fields has been increased noticeably. While this offers a number of newer avenues for authors to publish their research, the danger of proliferation of spurious journals cannot be ignored. Given that the evaluation of faculty members in various academic and research institutions depends heavily on the peer-reviewed publications, an inclination is seen among authors to publish research in International Journals as a mode of increasing the number of publica-

Abhishek Bansal
Indian Institute of Technology, Patna, e-mail: abhibansal28@gmail.com

Snehanshu Saha
Department of Computer Science and Engineering, PESIT South Campus Bangalore, India, e-mail: snehanshusaha@pes.edu

Archana Mathur
Indian Statistical Institute, Bangalore, e-mail: archana_plp@isibang.ac.in

Gouri Ginde
Department of Computer Science and Information Systems, University of Calgary, Canada e-mail: gouri.deshpande@ucalgary.ca

tions. The "international" tag attached to various journals is considered as possessing more credibility thereby, enticing authors to publish their research largely in "International" Journals than in National. Likewise, comparable with the quality of work an author does, he may have difficulties in evaluating a journal's suitability for submitting his/her work. There is thus a need to introduce new methods of evaluation that measures the influence of journals that also take into account the internationality aspect.

India has witnessed a significant increase in scientific publication, particularly in the year 2009, when the increase of 25% in publications is observed in comparison to previous years. Bibliometric information for most of the research papers is documented, but according to [16], the information is rarely been used for analysis. Buchandiran [20] organized an exploratory analysis of the scientific output of science and technology publications, and expressed that an Indian author is more interested in publishing his/her scientific paper, in an internationally reviewed journal. The study also reveals that leaving aside a few, most of the Indian Science and Technology journals are low in quality, are less internationally recognized and only a limited number is covered in ISI and Scopus. While there exist well-known methodologies for journal ranking (examples include the Web of Science by Thomson Reuters and SCImago Journal and Country Rank), these indices cover only a small fraction of the overall published journals. The focus of this work is to propose and validate an objective methodology for scoring journals. The number of journals in publication is growing rapidly. While on one hand, this is partly due to the proliferation of journals of questionable credibility, on the other hand, this is also due to increased research output and new journals may indicate emerging niche areas in some cases.

The motivation of this work is two-fold, namely to have a methodology which is applicable to a large number of journals (either already indexed or not) and is lightweight i.e. does not involve extensive data compilation and computation. The direct impact would be to enable evaluating newer journals (especially those that are between three to five years old), more importantly, it could also provide a basis for authors to assess where to publish their work. The indirect impacts include the establishment of publication appraisal policy and laying guidelines for funding agencies and accreditation bodies across country towards measuring the research output. These are especially important in the Indian context. Predatory publishing is on the rise in India which is governed by the need and greed of thousands of Indian researcher to get published and earn tenure. According to Jeffrey Beall [30], every week new predatory journals are emerging in India. Under these circumstances, the proposed methodology could aid appraisal of publications at individual and institutional level. While the one who is evaluating a journal would be the best judge of how to use these scores, it is fair to say that providing the necessary metrics aids decision-making. Recognizing that it is nearly impossible to cover all aspects with a single metric, the proposed methodology provides multiple scores, each covering different quality aspects. The manuscript is motivated towards the quantitative evaluation of internationality, quality, and influence of journals and authors which are the vital cogs in scholarly publications.

1.1 Issues in Scientometric Evaluation

Evident from the discussion in the preceding section, it might appear an overwhelmingly bright thing to have journals and conferences proliferate. This is certainly a "shot in the arm" measure for the policy makers and management in academia but not a reason for unbridled elation for the academicians. The reasons are manifold.

- Scientometric indicators of Internationality are flawed [1], [19]
- A single indicator or metric shouldn't be used for evaluation
- Raw Impact factor or H-index is crowned with undue importance [18]
- Strong speculations exist regarding manipulations or gaming the metrics
- Compliance to ethical and scientific practices in evaluation is not beyond reasonable doubt.

A section of academia is not very pleased with scientometric indicators being used as the primary yardstick for faculty evaluations. There are "theories" that discredit the entire methodology and blame "Publish or Perish" doctrine. The authors of the paper have been to discourses/symposia where Eugene Garfield, the father of Impact Factor has been criticized, unfairly and relentlessly. There is some element of truth in the claims that the evaluation scheme is not fair; however, the premise that the field of Scientometric analysis should not be taken seriously, is a bit far-fetched. Everybody in academia and in the business of publishing scholastic articles and new research findings loves citations. Therefore, it has become a tool for survival and a weapon to attain glory. It is not the metric or the study of metrics, rather survival instinct and human greed that are responsible for importing uncertainty. So, incremental improvements and sophisticated modeling and evaluation methods are required in the Scientometric analysis, not the extremist recommendation of ignoring it altogether.

2 Problem Definition

The Internationality of journals is a recent and relevant topic simply on the merit of evaluation mechanism of scholarly output being built around the doctrine of the ability of publishing in "International Journals". The singular problem of evaluating Internationality of journals is the absence of metrics and models of internationality. Surely, considering the claims of publishers and vested interests would be naive, to say the least. For reasons not understood by the authors, relevant and meaningful studies on this topic are rare. Unavoidably, there arises a need to define additional metrics for computing scholastic influence, as a single metric does not suffice. We postulate internationality on the tenets of scholastic influence transcending Geo-physical boundaries. This is in stark contrast with what many scientometricians believe! The problem may be dissected into the following sub-categories:

- Current literature lacks viable and scientific quantification paradigm of Journal Internationality. It lacks any model to quantify the concept (except by [1]).
- There is no theory or practice guideline to rank Journals based on Internationality, not even by the leaders in Scientometric research.
- There exists no baseline documentation on the classification of journals based on Internationality.
- There is no metric or model to measure "Author Internationality". Author Internationality implies the scholastic influence of authors, in terms of citations, readership etc. outside her/his peer group, advisors and collaborators. Raw citations will fail to disregard the local influence of authors and therefore is not advised as a useful metric.
- Therefore, models are required to be defined, built and used to compute internationality score of an author or a journal. Since, some of the "popular yet prone to gaming" metrics can’t be used, the problem needs a few novel metrics. These metrics are termed suitably as "greed-aware" metrics.

Authors believe that these greed-aware metrics would disregard any local influence within a journal. They ensure that the proposed Internationality-model and novel metrics will invalidate any artificial boost of journals influence via coercive citation, extensive self-citations or copious citations (sections 5).

3 Our Contribution

We summarize our contribution to ensure smooth reading before presenting the scholarly merit of the manuscript in detail.

1. Scientometric contribution

- Quantitative definition of internationality is absent until recently [1].A Popular perception is that journals are either national or international. We have shown, by theoretical and empirical calculation that such binary classification of internationality does not capture the true picture.
- We have also shown that rankings based on internationality of journals may not match rankings based on impact factor. Our work paves the way for a fresh, new ranking system of journals based on internationality diffusion. May we say, this is a path-breaking idea.
- Journal internationality upon quantification turns out to be finely graded and granular and therefore there should be more classes and not just two. Internationality requires multiple class discrimination and is therefore, an evolving concept, not orthogonal to journal impact and quality as many position papers would like us to believe.
  This is one of the resounding contributions of the paper where we established journal internationality as a quantitative parameter, solidly grounded on theory and proceeded to create a more reasonable and accurate classification paradigm, unavailable in existing literature.
- We defined a set of novel metrics for internationality quantification. One such metric is genealogy and non-genealogy citations applicable to author/scholar internationality modeling.
- Internationality should be free from local influence. By local influence, we do not imply being tied to Geo-physical locations and proximal regions. Rather, local influence is equivalent to journals or authors, receiving citations from friendly network. In the case of an author, it could be his/her colleagues, Ph.D. students, advisor etc.
- We defined, derived and quantified very important metrics, NGC, CGR etc which discounts local influence. This shall help establish the notion that internationality diffusion must be impersonal.
- We provided a framework for author internationality which may reflect international influence and diffusion of scholarly impact of individuals. This is a novel concept and detailed methodologies have been discussed to realize the framework.

2. Technical contribution

- Stochastic Cobb-Douglas (SCD) model (ref. section 6): Since global optimization principles have been used to compute the maximum internationality of journals/scholars, we require smoothness of functional properties, to ensure global optima. Deterministic CD model [4] suffers from curvature violation which in turn affects the smoothness of the internationality curve. This makes locating the global maxima computationally expensive. We mitigate the problem by proposing the SCD model and addressed the issue of curvature violation, such that the global maximum internationality score of journals or scholars may be computed efficiently and accurately.
- We have done extensive stochastic frontier analysis and applied sparse principal component analysis (SPCA) to ensure that key features used for Internationality computation are estimated reasonably well.
- Equivalence Classes (ref. section 7.3.1): This is a novel concept where the citation corpus could be split into a set of equivalence class partitions by exploiting and recognizing citation patterns and collaborations among scholars. This helps track copious citations efficiently.
- Non-genealogy citation (ref. section 9): Algorithmic interpretation of genealogy tree is accomplished for the first time. The novel algorithm helps quantify genealogy and non-genealogy citations by using an elegant tree data structure. This is a key factor toward computing penalty contribution in the scholar internationality computation.
- Unified granular neural networks (UGNN), (ref. section 8.1) was proposed to classify journals into three layers of internationality- high, medium and low. This is a new postulate in the literature of journal internationality and UGNN, a machine learning approach, was the suitable choice.
- Graph theoretic models (section 7.3) are used to compute key features (metrics) defined and utilized in our model.

The paper intends to define novel journal-level metrics that are free from any kind of bias or influence via unreasonable practices. An esoteric model of journal's internationality is proposed exploiting novel metrics like OCQ (other-citation-quotient), ICR (international collaboration ratio), NLIQ (non lo-

cal influence quotient), Copious citation metric, non-genealogy citations etc. These metrics are used to compute a discriminating score that quantifies journal's "internationality" and consequently, on the basis of score, journal's "internationality" is granulized on a scale of low, medium and high. We present a few metrics in Scientometry and Bibliometry, commonly used and widely known. The flaws and limitations are discussed and remedial and novel metric definitions are proposed.

The remainder of the paper is organized as follows. We start by looking at the standard metrics and their drawbacks. The next section defines and elaborates the remedial metrics and proposes mathematical models to synchronize quantification. Sections 6 explains Cobb-Douglas model and its usage in computing internationality score and further explores on using Stochastic Frontier Analysis and Sparse PCA for parameter computation and dimensionality reduction. Section 7 calculates internationality scores for journals and authors with penalty imposed for unfair practices. The section also showcases different models and algorithms to detect fraudulent practices that are performed at author and journal level. Section 8 dwells on classifying journals on the basis of internationality score and section 9 elaborates on genealogy citation model for computing author internationality score. The driver algorithm, embodying the aggregate evaluation model is presented in the next section. We conclude with the merits and pitfalls of our approach. Algorithms and summary of scraped data are presented in appendices.

4 Standard Metrics

The section defines some standard matrics used frequently to indicate the impact of scholarly research. The section also explores pitfall of their usage.

- Raw citations: Number of citations an article receives without considering the article's field/subject area.
- I.F: The impact factor (IF) of a journal is a ratio of the average number of citations a journal receives to the recent articles published in that journal.
- i-10: Count of the number of articles cited at least ten times.
- Total Docs./Total Documents: Output of the selected period. All types of documents are considered, including citable and non citable documents.
- H-index [6]: The h-index expresses the number of articles (h) a journal publishes that have received at least h citations.
- SJR (SCImago Journal Rank) indicator [6]: It represents the average number of weighted citations received by a journal in a particular year by documents published by the same journal in three previous years.

Some other metrics may include Total Docs (3years) , Total References, Total Cites (3years), Citable Documents, Cites per Documents (2 years), Cites per Doc (3 years) [6], Cites per Doc (4 years), Ref/Doc, Self-Cites, Non-citable documents (Available in the graphics), Cited Documents (Cited Doc.) and Uncited Documents (Uncited Doc.) [6].

International Collaboration [6] is defined as the document ratio whose affiliation includes more than one country address. It is a critical component toward measuring internationality of journals. A journal's International Collaboration Ratio (ICR) reflects the contribution of an article in terms of dissimilarity of affiliating countries. Every article of a journal is inspected and author's affiliation of the article is matched with the journal's originating country. Different weights are assigned to various combinations of authors' affiliations and journal's origin. The description of computing methodology used for International Collaboration Ratio is borrowed from our previous work [1], in which exclusive algorithms are written for accumulating journal's country information, for collecting author's affiliations and finally to compute ICR based on the weights assigned to the different coalition of contributing authors. The primary reason for this approach is the occurrence of dummy affiliations in middle-east countries [21].

It is not odd to have some authors being affiliated to more than one institute within the same country. In the western world, land grant institutions affiliated to public or private universities are set up. Typically, such institutes are small and don't demand additional intellectual resources. Therefore, this results in faculty having multiple affiliations. In such cases, multiple affiliations do not imply different institutions.

4.1 Pitfalls

The mechanism by which Thomson Reuters (ISI Web of Science) calculates IF and sources of their citation database is known only to them. This indicates that journals that are not existing in their database are not accredited an IF value. Since not all journals are indexed, authors are deterred from comparing non-indexed but otherwise competent journals. I.F is not useful and fairly weighted until it is normalized across subject areas. Raw Impact factor therefore, should not be considered as an absolute metric for journal influence.

Bhattacharjee [22] reported that some Saudi Arabian universities collaborate with highly cited authors by incentivizing them to add their institutions in publication to artificially inflate their university rankings. In response to the favor, the Universities compensate the authors with attractive salary and positions of adjunct Professorship. Gingras [23], [24] calls such type of affiliations as dummy affiliations since the research activity carried out by such institutes is not noteworthy. Those institutes receive an artificial boost in ranking positions, without having to carry out meaningful and rigorous scientific investigation.

Editorial and other kinds of Nexus: Any of the standard metrics mentioned above, conceptualized as good indicators of scholarly impact is subject to to manipulation. Some of the malpractices could be attributed to human survival instincts and some could be pure greed. Regardless of how unethical these practices are, some of these practices that include citing friends and expecting citations in return, coercing colleagues and fellow authors by exploiting Editorial power, collaborating with the editor of another journal and spike boosts in journal citations, proof of such activities is difficult to establish. Algorithmic and mathematical intervention is required to track and catch such instances among a pool of extremely large data. Next section discusses some of the novel metrics defined by the authors of this paper, with the specific goals of addressing these practices.

Note: *The proposed model to track this trend, cognizant citations, is discussed in section 7.*

5 Remedial Metric definitions

There are possibly several parameters which are not considered in commonly used Scientometric practices. We propose and list a few of those which could embellish the metric and internationality score functions and hopefully counter/mitigate artificial boosting arising out of survival or greedy inclinations.

- Non Local Influence Quotient (NLIQ): According to [1], considering a journal A, NLIQ of A is the ratio of citations received from articles outside journal A to the total citations received by A. Higher NLIQ indicates the number of external citations is high, which implies less localized influence.

$$NLIQ = \frac{x}{y}$$

where
x is the number of citations received from articles published in Journal other than A.
y is the total citations received by A.

We observed a common tendency where articles of a journal cite articles belonging to the same journal. As a result, the journal's prestige is increased artificially because of increased citations from within. It has been noted that even if the SNIP of a journal is high, [1] the quality of work may not have diffused significantly, and the journal may possess low NLIQ (citation contribution from outside of the journal). This is the indication of the fact that although SNIP is strong indicator of journal's influence, it may not be considered as a comprehensive measure of its internationality.
PS: The authors retain the copyright to the definition of NLIQ and encourages peers to use this definition while acknowledging the source. Please note Scopus has a similar metric defined, external cites for journals. This appeared on their site after publication of [1]. This may be coincidental, we are not sure!

- Journal Effusion Index (JEI): This index is a computation of citations which originate from a journal (say Journal X) and are made to articles in journals other than journal X.

$$JEI = \frac{a}{b}$$

where
a is the number of citations given to articles published in Journal outside X.
y is the total citations made by X.

The algorithm for computing JEI is in Appendix A. Inherently, it evaluates the integrity of journals, whether a journal indulges in any kind of self-citation to unnaturally increase its prestige. A high value of JEI would indicate that journal promotes research activity in a legitimate way.

- Copious Citation Quotient: The Copious citation is defined as a condition in which two authors (say A and B) cite each other's work. This means A cites all published papers of B and vice-versa. The intention behind such practices is to unfairly boost citations. Eliminating the effect of copious citation is possible by including a penalty in the main score function whenever an instance of copious citation emerges. The main score model can be formulated as a profit function, whereas an author's internationality score is the revenue function and penalty due to copious citation becomes the cost function.
- SNIP: Computation of Source-Normalized Impact per Paper (SNIP) involves characteristics and citation potential of source's subject domain [26]. If the citation potential of the subject field is high, it is biased to receive citations many times. SNIP is average citation count per paper normalized by the citation potential of the relevant subject field. Hence, it allows direct comparison of sources in various subject domains.
- Cognizant Citations (CG): Cognizant citations are citations built under the influence of strategic cognizance between two Editor-in-Chiefs (EIC). At times, (EIC) of reputed journals, in an effort to improve citations and their ranks, insist authors to cite papers of their journal. Assume there are two journals, A and B, EiC of A aspires under the influence of cognizant citations to raise citations for B and vice versa. "The International Journal of Nonlinear Science and Numerical Simulation (IJNSNS), started publishing in 2000, by Freund Publishing House, has gained attention because of the increased impact factor in the category of "Mathematics, Applied". The editor-in-chief of IJNSNS and a member of the editorial board of IJNSNS published and cited papers of their journals and also cited each other generously so much so that the Journal almost topped in impact factor ranking chart. The journal charged USD 90 per page up to six pages, USD 50 each additional page thereafter. During the computation of journal's internationality score, such cases should be penalized.
- Other-Citation-Quotient (OCQ): OCQ [1] is the ratio of 1 - (self-citation/Total Citations) for a journal. If a citing article and cited article of an author belonging to the same journal, then we term it as self-citation for that journal. OCQ reflects a journals integrity owing to the fact that no legitimate journal will promote authors and allow them to indiscreetly cite their own work. OCQ is a subset of NLIQ.

- Ancestor-inheritance citations: The citations which an author gets from its advisor (i.e ancestor) are called ancestor inheritance citations. There may be some possibility of that advisor cite his/her advisee frequently in his/her work for advisee's benefit. This type of citation is part of genealogy citation. We discuss genealogy citation in detail in section 7.1.

5.1 Framework of the remainder of the paper

Figure 1. shows the flowchart of sequence of processes and techniques covered in the manuscript. Table 1 shows the list of various remedial metrics that are used in the current research work. According to authors, there are other important metrics which can be explored and used in computation of journal's influence. These metrics which are part of future work, are listed in table 1 with definitions in Appendix .

	Metrics
Current Work	NLIQ SNIP Copious citations Quotient Cognizant-citations Other-Citation-Quotient Ancestor inheritance citations
Future work	Journal Effusion Index Weighted NLIQ Differentiating citations Turnaround Time Time Window Readership Profile Volumetric Information

Table 1: **Use of various Metrics**

Internationality score mirrors the genuine picture of execution of any journal or author. It truly tells about the influential capacity of any journal or author. In our work, we try to propose a technique through which we can ascertain internationality's score. We gather author and journal data through web scraping in the Data collection stage. Further, we extract metrics and derive the necessary features such as, Non-Local Influence Quotient, which is computed using Self-citation count etc. and normalize these features. Normalization eases comparison of subject fields such as Computer Science, the Social Sciences and Mathematics where collaboration and citation trends differ remarkably. In the next stage, we have used sparse principal component analysis (SPCA) to obtain the minimal feature set which explains the maximal variance of the original dataset. The new feature set is fed to Stochastic Cobb-Douglas Model, which is used in economics as a production function. In order to estimate the coefficients of Stochastic Cobb-Douglas model, we proceed to compute SFA (stochastic frontier analysis) on it. SFA yields technical efficiency, which is defined as the ratio of observed output to maximum possible output. If this efficiency is insignificant then Ordinary Least Squares (OLS) method is used to compute the coefficients. In the next stage, Score is computed using the Stochastic Cobb-Douglas

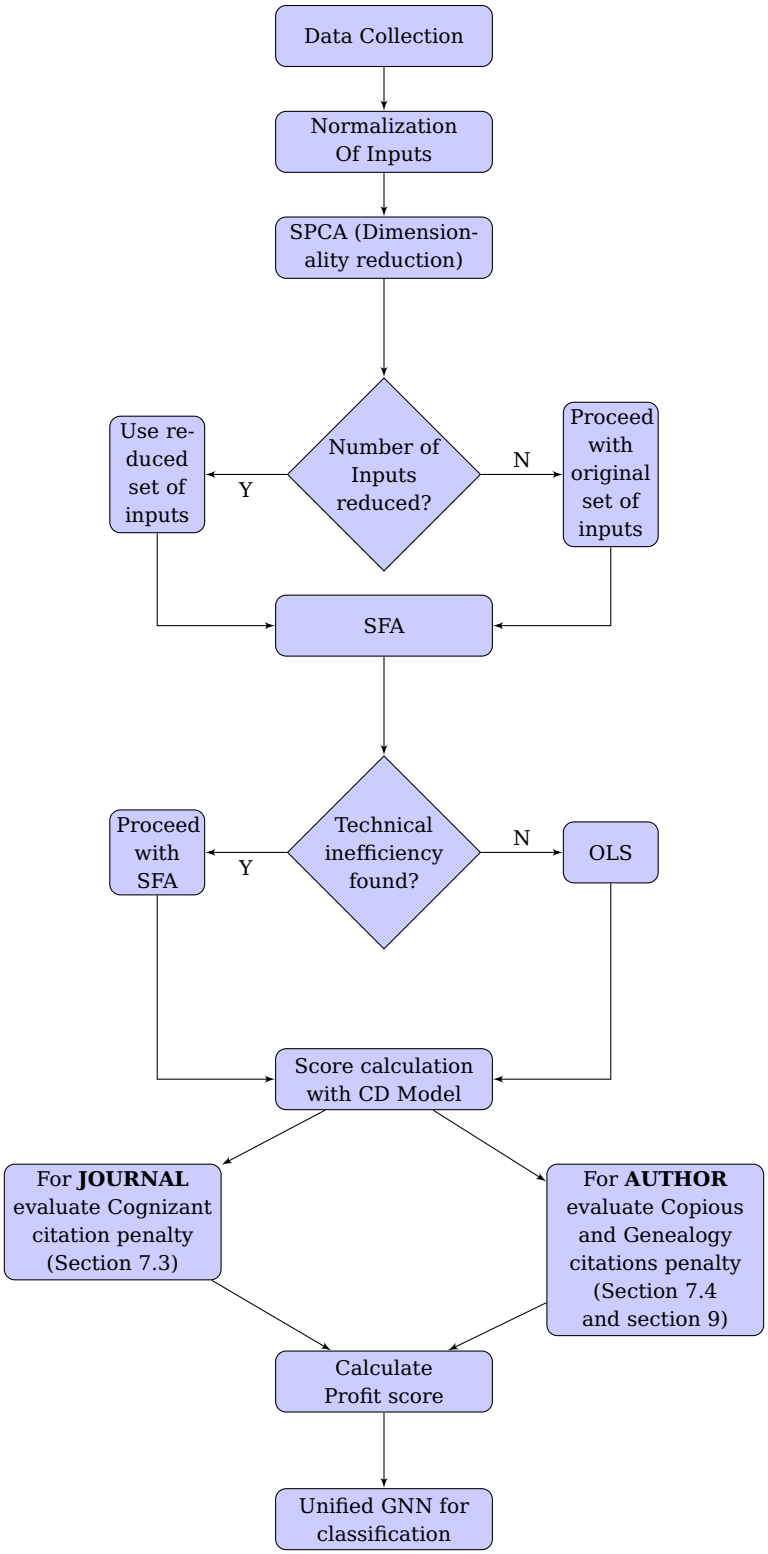


Fig. 1: Research Framework: Profit Score is the effective Internationality score of journals or scholars after penalty, if at all there is one.

model. The Journals, as well as authors, are penalized for practicing unfair means. Following are the two independent scenarios to impose penalty while computing Journal's and author's internationality score.

1. Cognizant citations penalty is computed based on the journal level data. For a journal, the Stochastic Cobb-Douglas score is manipulated to reflect this penalty, and the net score is reduced to cumulative profit score for a journal in question. A threshold is set to omit a nominal/minuscule penalty score. The new profit score is then used for classification of journals using Unified Granular Neural Network (UGNN) algorithm.
2. Copious and Genealogy citation penalty are computed based on the author level data. For an author, the Stochastic Cobb-Douglas score is manipulated to reflect the penalty and the net score is reduced to cumulative profit score for an author in question. A threshold is set to omit the nominal/minuscule penalty score from the computation. The new profit score, **a measure of effective internationality of a scholar**, is then used for classification of scholars using Unified Granular Neural Network (UGNN) algorithm. All the technical terms and the implications of various algorithms are explained in the following sections.

The next section examines the suitability of using Cobb-Douglas model and proves, with the help of suitable theorems and lemmas that the function is concave and attains global maxima. The section investigates the necessity of SFA and explains how the model parameters can be estimated. The classic method of dimensionality reduction, PCA, is associated with the disadvantage that the principal components are a linear combination of the original feature set. The authors explore on the usage of Sparse PCA for obtaining the reduced feature set and still, achieving the maximum variance in the data.

6 Aggregate Model: Econometrics at play

Cobb-Douglas Production function, an econometric function, is extensively used to model a relationship between output and input [1]. A production function is used for the first time, to compute internationality score for ranking journals. The parameters are algorithmically obtained from various sources. The output y, Internationality score, varies over time and depends on scholastic parameters, subject to evaluations, constant scrutiny, and ever-changing patterns. Depending upon the elasticity values, the function exhibits convexity/concavity and attain global maxima/minima that can be employed to model influence or internationality.
Cobb-Douglas function is given by

$$y = A\prod_{i=1}^{n} {x_i}^{\alpha_i}$$

where y is the internationality score, x_i are the predictor variables/input parameters and α_i are the elasticity coefficients.

6.1 Proof of Concept

The section proves the efficacy of the model for n number of variables/inputs, n being countably finite. To begin with, authors have done the simulation using 2 variables (this facilitates visualization, fig. 25 in Appendix B) and extended the exercise to 3 variables. This can be extended to n variables as shown below. Consider the following production function:

$$y = \prod_{i=1}^{n} k x_i^{\alpha_i}$$

$n = 4$, x_1 to x_4 are the input parameters, namely Other-Citations Quotient, International Collaboration / 100, SNIP value / maximum SNIP value and Non-Local Influence Quotient respectively. Please note, n could be indexed to accommodate more input parameters and dimensionality reduction eventually prunes the final set of parameters.

Lemma I:
Optimality of journal internationality score is accomplished at decreasing returns to scale i.e.

$$\sum_{i=1}^{n} \alpha_i < 1$$

where α_i is the elasticity/exponent of the input variable x_i. Let us consider the following production function:

$$y = \prod_{i=1}^{n} k x_i^{\alpha_i}$$

Need to prove:

$$\sum_{i=1}^{n} \alpha_i < 1$$

Consider the profit function i.e the difference between the revenue and cost functions as the following:

$$\pi_n = \prod_{i=1}^{n} k x_i^{\alpha_i} - \sum_{i=1}^{n} w_i x_i$$

where w_i is Unit cost of inputs.
Profit maximization is achieved when: $p\frac{\partial f}{\partial x_i} = w_i$. *The lemma determines the choice of optimal elasticity under constraints such that the prestige/influence of a journal is maximized under suitable penalty for malpractice.*

Lemma II:
$f \in C, U \subset R; U$ is a convex, open set, $f : R \to R, f$ is a concave iff

$$f(x+\theta) \leq f(x) + \nabla f(x)\theta; \quad \forall \ \theta \in R^N; x+\theta \in A;$$

C: Class of continuous and first order differential functions,
NOTE: The lemma is required for proving theorem I, necessary and sufficient condition for global optima of journal/author internationality.

Theorem 1: $f \in C^2; x \in R; f : R^2 \to R$ is concave iff the Hessian Matrix, $H \equiv D^2 f(x)$ is negative semi-definite $\forall x \in U$. [necessary and sufficient condition for concavity]

Implications of Theorem 1: The internationality function is concave for certain conditions on elasticity which make the Hessian Matrix of the function negative semi-definite (NSD). This is a pre-requisite for concavity and ensures global maxima. Now consider, the internationality function, $f(x_1, x_2) = k x_1^{\alpha} x_2^{\beta}$ with $k, \alpha, \beta > 0$ for the region $x_1 > 0$ and $x_2 > 0$. The hessian matrix is computed as follows:

$$H = \begin{bmatrix} \alpha(\alpha-1)k x_1^{\alpha-2} x_2^{\beta} & \alpha\beta k x_1^{\alpha-1} x_2^{\beta-1} \\ \alpha\beta k x_1^{\alpha-1} x_2^{\beta-1} & \beta(\beta-1)k x_1^{\alpha} x_2^{\beta-2} \end{bmatrix}$$

First order principal minors [2] of H are:

$$M_1 = \alpha(\alpha-1)k {x_1}^{\alpha-2} {x_2}^{\beta}; \quad M_1' = \beta(\beta-1)k {x_1}^{\alpha} {x_2}^{\beta-2}$$

and second order principal minor is:

$$M_2 = k\alpha\beta x_1^{2\alpha-2} x_2{}^{2\beta-2}[1-(\alpha+\beta)]$$

H must be negative semi-definite, this implies $f(x_1,x_2)$ is concave. This will happen if $M_1 \leq 0,\ M_1' \leq 0\ \ and\ \ M_2 \geq 0$. Conditions for decreasing and constant returns to scale are satisfied by : $\alpha+\beta \leq 1$, therefore

$$\begin{aligned}
&\alpha \leq 1, \beta < 1 \\
&\Rightarrow (\alpha-1) \leq 0 \\
&\Rightarrow M_1 \leq 0 \\
&(1-(\alpha+\beta)) \geq 0 \\
&\Rightarrow M_2 \geq 0
\end{aligned}$$

Both conditions for concave function are satisfied by decreasing and constant returns to scale. Therefore, $f(x_1,x_2)$ is concave, if and only if

$$\alpha \geq 0, \beta \geq 0, \alpha+\beta \leq 1$$

NOTE: The extrema of the internationality function, an analytical property is imperative to theorize and compute global maxima of internationality" score. The modeling exercise is founded on the principle of existence of maximum internationality score and the score in the neighborhood may be classified as shades of internationality. Therefore, the theoretical exploration of optima is consequential.

Theorem 2: on global maxima:

Let $f(x_1,x_2) = kx_1^{\alpha}x_2^{\beta} : U \subset R^2 \to R$ be concave function on U; U is an open convex set; the critical point, x^* is a global maximum.

Note:

1. The functional modeling, $f(x_1,x_2) = kx_1^{\alpha}x_2^{\beta}$ may be extended to $f(x_1,x_2\ldots x_n) = k\prod_{i=1}^{n} x_i^{\alpha_i}$; in which case $f : U \subset R^n \to R$ & the global maxima holds.
2. U may be open or closed since the search for optima is allowed on the boundary of the set.
3. the proposed function is quasi concave.
4. The values of elasticity are computed by using **fmincon** command in Matlab. These elasticity values are the exponents in the expression, $f(x_1,x_2\ldots x_n) = k\prod_{i=1}^{n} x_i^{\alpha_i}$; the function **fmincon** is a built-in convex optimization tool in MATLAB and is in agreement with **Lemma II**.
5. $n = 4$ may be extended to any number of input variables that could include all the remedial metrics mentioned in previous sections.

Initially, seven predictor variables are fed as input to the model. Theoretically, the model can scale up to any number of inputs but practically, as the number of predictor variables/input grows, the complexity increases and determining elasticity may become difficult because of curvature violation of the internationality production function. Stochastic frontier analysis, however, may resolve this problem [4], [17]. Authors intend to investigate other econometric models that are unaffected by curvature violations. The method suffers from the curse of dimensionality and as a result feature reduction becomes unavoidable. The dimensionality reduction methods help to recognize the crucial features that may produce a major impact when used in the model.

By appropriately choosing the exponent of NLIQ in Cobb-Douglas model, its effect on the internationality score can be modulated. Through the process of elasticity boosting, one can choose elasticity in such a way that internationality score of the journal remains unaffected. This may be needed in situations when one parameter is low but other is high, for example, NLIQ for a journal is low but SNIP is high. The boosting of elasticity can be carried out through Design of Experiments (DoE). The process deals with conducting experiments to assess the contribution of every input factor in our model. If a

certain factor is contributing minimally, the elasticity is adjusted adhering to the constraints of the optimization problem solved during the process. (Theorem 2)

Estimation of the constant of proportionality, k, in the proposed Model: We have assumed the value of k as 1 for simplicity. However, k in the model formulation may be estimated from data by using sophisticated fitting models and constrained optimization techniques. Once k is suitably estimated, elasticity may then be predicted/fitted accordingly.

6.2 Need for Stochastic Frontier Analysis

Cobb Douglas model can scale up to any number of inputs in theory. However, the increase in the numbers of inputs leads to exponential increase in the complexity of this model. This increase in complexity may cause curvature violation of the Cobb-Douglas model to a great extent which would cause erroneous elasticity coefficients estimations. In order to mull this ill effect we have used stochastic frontier analysis(SFA), a method used in economics for modeling, to estimate the values of elastic coefficients α_i's and k. The following section elaborates on SFA, evaluates the Cobb-Douglas Stochastic Frontier Model and elucidates the usage of maximum likelihood estimation approach to compute the parameters of CD-SFA.

6.2.1 Stochastic Frontier Analysis(SFA)

Econometric estimation techniques witness observed choices deviate from optimal ones due to two factors:

1. Failure to optimize i.e.inefficiency
2. Random noise.

Stochastic Frontier Analysis incorporates these factors seamlessly. It is one of the best techniques to model input behavior, produce individual estimates and produce individual scores that have greater accuracy. The basic idea of SFA lies in the introduction of an additive error term consisting of a noise and an inefficiency term. Thus, SFA can help to identify the predictor variables which need corrective measures. Hence, we have used SFA to produce efficiency estimates or efficiency scores of a journal. We used these estimates to identify the predictor variables which need intervention and corrective measures. It is important to note that the efficiency score varies across journals as it is dependent on journal's characteristics. This relationship can be expressed in terms of a function of single dependent variable(output) with one or more explanatory variables(inputs). Mathematically we can express it as following.

$$y = f(x_1, x_2, x_3, ..., x_n)$$

where parameters are described below:
y: *Dependent variable*
x: *Explanatory variables* $\quad 1 \leq j \leq n$
$f()$: Mathematical function

$f()$ can have different algebraic form depending upon relationship between dependent variable and explanatory variable. One such function is Cobb-Douglas Function

$$y = \prod_{i=1}^{n} k x_i^{\alpha_i}$$

then, the ***Cobb Douglas Frontier is:***

$$lny_j = x_j^T \alpha - u_j$$

y_j: Output of jth journal
x_j: K*1 vector containing log of input

$$x_j^T = \left[1 \ \ln x_{j1} \ \ln x_{j2} \ \ln x_{j3} \ \ln x_{j4} \ . \ . \ . \ . \ \ln x_{j(K-1)}\right]$$

α : K*1 vector of unknown parameter

$$\alpha^T = \left[\ln k \ \alpha_1 \ \alpha_2 \ \alpha_3 \ \alpha_4 \ . \ . \ . \ . \ \alpha_{K-1}\right]$$

u_j: non negative random variable associated due to technical efficiency.
The frontier does not bring about inclusion of measurement errors or other statistical noise, so another random variable, v_j, is included which represents noise.

$$\ln y_j = x_j^T \alpha + v_j - u_j$$

This can be written as

$$\ln y_j = \ln k + \alpha_1 \ln x_{j1} + v_j - u_j$$

This form is called *Cobb Douglas Stochastic Frontier Model*

$$y_j = \exp(\alpha_0 + \alpha_1 \ln x_{j1}) \exp(v_j) \exp(-u_j)$$

where α_0=$\ln k$
$\exp(\ln k + \alpha_1 \ln x_{j1})$-*deterministic component*
$\exp(v_j)$-*noise*
$\exp(-u_j)$-*inefficiency*

Estimation of parameters:

Estimation in SFA is more complicated due to inclusion of two random terms v_j and u_j. The following assumptions are made for further computation:

- E(v_j)= 0 (zero mean) , E(v_j^2)=σ_v^2 , E($v_j v_k$) = 0 for all $j \neq k$ (uncorrelated)
- E(u_j^2) = constant , E($u_j u_k$) = 0 for all $j \neq k$ (uncorrelated) , E(u_j)$\neq 0$

To estimate the slope coefficients α_j and intercept, Maximum Likelihood Estimation (MLE) approach is chosen which makes assumption of a certain inefficiency distribution and normal noise distribution based on maximum likelihood. The approach is better than most of the other, as MLE has many desirable large sample (asymptotic) properties. Further, half-normal model as suggested by Aigner et.al. [32] is used to obtain MLE with the following assumptions:
$v_j \sim iidN(0, \sigma_v^2)$
and $u_j \sim iidN^+(0, \sigma_u^2)$
Here, positive half-normal distribution is taken for u due to which expectation of random variable u is zero i.e. $E(u) = 0$

The probability density function of $f(v)$ is

$$f(v_j) = \frac{1}{\sqrt{2\pi}\sigma_v} \exp(\frac{-{v_j}^2}{2\sigma_v^2})$$

The probability density function of f(u) is

$$f(u_j) = \frac{2}{\sqrt{2\pi}\sigma_u}\exp(\frac{-{u_j}^2}{2\sigma_u^2}) \quad for\ u \geqslant 0$$
$$= \quad 0 \qquad\qquad\qquad for\ u < 0$$

Here, composite error, $e_j = v_j - u_j = \ln y_j - x_j^T\alpha$
The distribution of e_j is the convolution of the distribution of v_j and $-u_j$ i.e.

$$f(e) = \int_{-\infty}^{\infty} f(u)f(e+u)du$$
$$f(e) = \int_{0}^{\infty} f(u)f(e+u)du$$
$$f(e_j) = \frac{\sqrt{2}}{\sqrt{\pi}\sigma}\phi(\frac{-\lambda e_j}{\sqrt{\sigma^2}})\exp(\frac{-{e_j}^2}{2\sigma^2})$$

where $\sigma^2 = \sigma_v^2 + \sigma_u^2$ and $\lambda^2 = \frac{\sigma_u^2}{\sigma_v^2}$
ϕ is the distribution function of standard normal with mean zero and variance one.
Likelihood function of f(e):

$$L(y \mid \alpha, \sigma, \lambda) = \prod_{j=1}^{J} f(e_j)$$
$$L(y \mid \alpha, \sigma, \lambda) = \prod_{j=1}^{J} \frac{\sqrt{2}}{\sqrt{\pi}\sigma}\phi(\frac{-\lambda e_j}{\sqrt{\sigma^2}})\exp(\frac{-{e_j}^2}{2\sigma^2})$$

Log likelihood function of f(e):

$$L(y \mid \alpha, \sigma, \lambda) = \frac{-J}{2}\ln(\frac{\pi\sigma^2}{2}) + \sum_{j=1}^{J}\ln(\phi(\frac{-\lambda e_j}{\sqrt{\sigma^2}})) - \frac{1}{2\sigma^2}\sum_{j=1}^{J} e_j^2$$

To maximize log likelihood function of $f(e)$, the first derivative is taken with respect to unknown parameters and equalize it with zero. The equations thus obtained are nonlinear and analytically non solvable. An iterative optimization procedure is used in which some initial values for unknown parameters are assumed and iteratively updated, until the assumed values, maximize the log likelihood function. Battese and Corra, [16] claims that it is simpler to parameterize the log-likelihood in terms of σ^2 and $\gamma = \sigma_u^2/\sigma^2$. It is known that γ lies between zero and one. Thus, if γ = 0, all deviations from the frontier happen because of noise, and γ =1 implies technical inefficiency causes the deviations. This property is exploited to facilitate iterative optimization.

6.3 Dimensionality Reduction using Sparse Principal Component Analysis (SPCA)

Principal Component Analysis (PCA) is a technique that transforms a set of correlated/dependent variables into a smaller set of uncorrelated/independent variables. These features (Principal Components) are orthogonal to each other (and therefore linearly independent) effecting largest possible variance in the entire data set. Though PCA is a popular choice for dimensionality reduction, the technique is disadvantaged by the fact that new set of variables are linear combinations of input factors of original data and therefore span the original data space and not the transformed one. Hui Zou, Trevor Hastie and Robert Tibshirani [7] proposed that dimensionality reduction and reduction in number of explic-

itly used features plays significant role in improving efficiency of statistical model. The lasso [8] is an efficient machine learning technique for variable selection that uses a variety of statistical models for regularization of selected variables. Zou and Hastie [9] proposed the elastic net, a generalization of the lasso, which has some advantages. Trevor Hastie and Robert Tibshirani [10] introduced a new approach for estimating PCś with sparse loadings, known as sparse principal component analysis(SPCA). SPCA is is an improvemnt over PCA in the sense that a regression-type optimization problem endowed with a quadratic penalty may be factored into the original PCA formualtion where the regression criterion is syntesized with the lasso penalty (via the elastic net). Thus PCA is recreated with sparse loadings. Following is the SPCA algorithm which explains this mathematically.

General SPCA Algorithm

1. Initialize A, the array of elasticity at V[, 1 : k]. Thus, the first k ordinary principal components are loaded.
2. Define A = [α_1,.....,α_k]; solve the following elastic net problem for j = 1, 2,...,k

$$\beta_j = argmin(\alpha_j - \beta)^T X^T X(\alpha_j - \beta) + \lambda \|\beta\|^2 + \lambda_{1,j}\|\beta\|_1$$

3. For a fixed B = [β_1,.....,β_k], compute the singular value decomposition (SVD) of $X^T XB = UDV^T$, update A = UV^T.
4. Repeat Steps 23, until convergence.
5. Normalize the loadings: $V j = \frac{\beta_j}{\beta_j}$, j=1,...,k. 6. Exit

6.4 Result of SPCA:

The data set we used for our analysis consists of 86 journals from Computer Networks domain. We start with seven predictor variables as inputs. The variables are included from standard and remedial metrics defined in section 4 and 5. These are OCQ (other-citation-quotient), IC (international collaboration ratio), NLIQ (non local influence quotient), IF (impact factor), SNIP, HINDEX and TotalCit/TotalDocs. Firstly we normalized seven input parameters. Normalization allows a fair comparison between different subject fields such as Computer Science, the Social Sciences and Mathematics where collaboration and citation trends differ remarkably. After normalization we performed sparse PCA on the normalized input data set with seven input variables. We used Elastic Net package in language R to investigate sparse PCA with three different combination of number of nonzero loadings. Here we discuss this results.

Case	No Of Non zero Loadings	Explained Variance(in percentage)	Inputs(without nonzero loading)
1	(6,6,6,6,6,6)	84.8	OCQ
2	(5,5,5,5,5,5)	83.1	TotalCites/TotalDoc
3	(5,4,4,4,4,4)	83.5	OCQ and TotalCites/TotalDoc

Table 2: **Different Loading result of sparse PCA**

Note: In Table 2, we considered the three cases, where we considered first six PCs and explained variance(in percentage). Cases 1, 2 and 3 explain the variance by first three PCs together and inputs (without nonzero loading). It shows that those inputs for first three PCs did not assign non-zero loading. Table 3 clarifies these in detail. case 1 for OCQ all the three PCs did not assign nonzero loading. Similarly, in case 2 for TotalCit/TotalDocs, all the three PCs did not assign nonzero loading and in case 3 for TotalCit/TotalDocs and OCQ all the three PCs did not assign nonzero loading.

Observation: The results in Table 3 show that all the three cases from Table 2 explain approximately equal percentage of variance. In the first two cases, first three PCś did not assign nonzero loadings to only one input variable i.e OCQ in the first case and TotalCite/ToatalDocs in the second case but in case 3, first three PCś did not assign nonzero loading for two input variables OCQ and TotalCit/TotalDocs. Case 3 indicates that those two inputs OCQ and TotalCit/TotalDocs are irrelevant for further explanation of our model. Itś very intuitive to choose case 3 because dimension reduction is more evident as compared to case 1 and case 2. We show the result of sparse PCA in case 3 in table 5.

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**
OCQ	0.000	0.000	0.000	0.000	0.028	0.000
IC	0.301	0.947	0.000	0.108	0.000	0.000
HINDEX	0.322	0.000	0.289	-0.888	0.000	0.154
Impact Factor	0.504	-0.160	-0.098	0.000	0.000	-0.843
SNIP	0.722	-0.260	-0.279	0.261	0.000	0.515
TotalCites/TotalDocs	0.000	0.000	0.000	0.000	1.000	0.000
NLIQ	0.174	-0.097	0.910	0.362	0.000	0.014

6 sparse PCs
Percentage of explained variance : 44.5 24.9 14.1 7.3 4.6 2.8

where FVs:Feature Vectors

Table 3: **Result of sparse PCA in case 3**

6.5 Result of SFA

After sparse PCA, stochastic frontier analysis (SFA) is applied on the reduced set of input variables. In order to estimate values of elastic coefficients α_i and slope k through SFA, we use output of Cobb-Douglas model with three conditions: increasing rate of scale, decreasing rate of scale and constant rate of scale one at a time. **fmincon**, a function which is included in optimization toolbox in **Matlab** is used for calculation of output from Cobb-Douglas Function. Fmincon yields the values of estimated elastic coefficients through SFA for all three cases. Table 4 include the values of gamma and gammmaVar, which are obtained after stochastic frontier analysis (SFA) for three cases. The value of γ stipulates the role of stochastic error (v) and technical inefficiency (u) in explaining the deviations from the production function.*The parameter γ lies between zero and one, where zero shows that one can proceed with results of Ordinary Least Square and u (technical inefficiency) can be discarded. Likewise when γ is one, it can be stated that all divergence from the production frontier is due to technical inefficiency and noise term (v) can be ignored. Parameter gammaVar explains variance by inefficiency.* A package **frontier** from language **R** is used for the stochastic frontier analysis(SFA). We can infer following results based on the values of gamma and gammaVar in Table 4:

- The estimate of γ is close to zero for all three cases. Also, the likelihood values of stochastic frontier analysis (SFA) and ordinary least squares (OLS) are equal. This spells out that the inefficiency is not important for explaining deviations from the production function in our case.
- gammaVar explains 0.00085342% variance in CRS, 0.00073068% in IRS and % variance in DRS. These variance percentages are negligible which indicates that the inefficiency term is irrelevant.
- The variance explained by noise is 99.99968% in CRS, 99.99974% in IRS and 99.99226% in DRS. This indicates a significant role of noise in our data set.

ROS	**gamma**	**gammaVar**	**Explained variance By Noise(in %)**	**LV(SFA)**	**LV(OLS)**
CRS	8.5342e-06	3.1012e-06	99.99968	102.82	102.82
IRS	7.3068e-06	2.6552e-06	99.99974	87.248	87.248
DRS	2.1280e-04	7.7339e-05	99.99226	719.7	719.7

where ROS=Return Of Scale LV=Likelihood Value

Table 4: **Result Of SFA With Five Input Variables**

After observing the results of SFA with five input variables, the method is applied to the original data set with seven input variables. This checks whether the technical inefficiency explains any variation at all. The results are in Table 5. We can infer following results based on the values of gamma and gammaVar of Table 5:

- Since the estimate of γ is close to zero for all three cases and, the likelihood values of stochastic frontier analysis (SFA) and ordinary least squares (OLS) are equal, one can conclude that inefficiency is not important for explaining deviations from the production function.
- gammaVar indicates 0.007276% variance in CRS, 0.041578% in IRS and 0.012843% in DRS. Overall, it is a negligible variance percentage, which also indicates about the irrelevancy of inefficiency term.
- The variance indicated by noise is 99.97355% in CRS, 99.84849% in IRS and 99.91849% in DRS which justifies the role of noise in our data set.
- This SFA computation on the original set of seven input variables is optional. We computed this to analyze the effect of technical inefficiency with seven input variables.

ROS	**gamma**	**gammaVar**	**Explained Variance By Noise(in %)**	**LV(SFA)**	**LV(OLS)**
CRS	7.2762e-04	2.6453e-04	99.97355	291.17	291.17
IRS	4.1578e-03	1.5149e-03	99.84849	266.35	266.35
DRS	1.6234e-03	3.1235e-03	99.91849	412.83	412.83

where ROS=Return Of Scale LV=Likelihood Value

Table 5: **Result Of SFA With Seven Input Variables**

Note: *We investigate our model with stochastic frontier analysis in two cases. One, with original set of inputs i.e. seven inputs and other, with reduced set of inputs i.e. five inputs. In both the cases we obtain same results. Thus, we can conclude that inefficiency term is not relevant in explaining deviations from the production function in our model and deviations are only due to the stochastic random*

error component i.e. v. Also, the only advantage of SFA over ordinary least squares (OLS) is that SFA is able to explain the role of technical inefficiency. When role of technical inefficiency is negligible then we choose OLS. So to estimate the value of constant k and elastic coefficients α_i we use Ordinary Least Squares method.

Table 6 shows the results of ordinary least squares method. From OLS results we can conclude following:

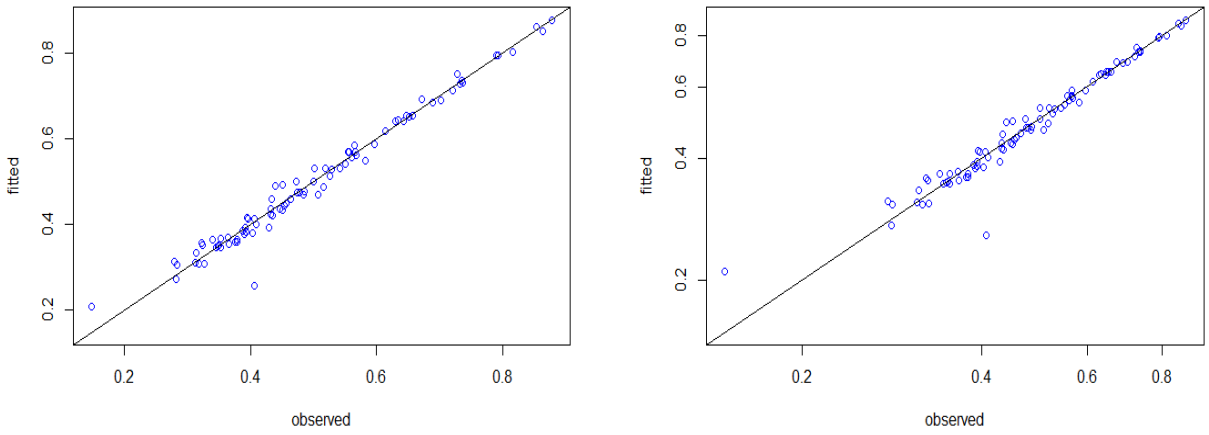
1. Since the coefficients of all inputs and correspondingly, the output variables are positive, the conceptually important condition of SFA-the monotonicity, is fulfilled globally. However, the coefficient of the NLIQ has highest value among the estimated value of predictor variables.
2. The internationality function proposed by us is impaired by strict essentiality in the sense that the output (internationality of a journal) becomes zero, as soon as any of the input quantities becomes zero. The predicted output quantity is non-negative as long as k and the input quantities are non-negative, where k = exp(α_0) is always positive. Predicted (fitted) output quantities may be computed manually by considering the internationality function, input quantities (obsrved) and the estimated parameters. However, predicting values of the dependent variable from an estimated model is more convenient.

	Estimate	**Std. Error**	**t value**	**Pr(> | t |)**
Intercept	0.01743	0.02402	0.726	0.47
log(IC)	0.12805	0.01254	10.215	3.70e-16
log(HINDEX)	0.10539	0.01307	8.062	6.12e-12
log(Impact Factor)	0.11091	0.01960	5.659	2.29e-07
log(SNIP)	0.08975	0.02029	4.424	3.02e-05
log(NLIQ)	0.32528	0.02500	13.012	< 2e-16

Residual standard error: 0.0759 on 80 degrees of freedom
Multiple R-squared:0.9462 *Adjusted R-squared:0.9428*
F-statistic: 281.2 on 5 and 80 DF *p-value:< 2.2e-16*

Table 6: **Result Of OLS**

3. Figure 2(a) uses a linear scale for the axes. Figure 2(b) uses a logarithmic scale for both axes. Hence, the deviations from the 45° line captures absolute deviations in the left panel and relative deviations in the right panel.
4. The elasticity of scale is the sum of all output elasticities, Hence, we can calculate it by summing up all coefficients except for the intercept. The sum of coefficients is 0.75938, which implies that the data set has decreasing returns to scale.



(a) linear scale (b) logarithmic scale

Fig. 2: Stochastic Cobb-Douglas production function: fit of the model

So far, we have defined various remedial metrics which can be transformed into numerical values to compute author's and journal's internationality. These metrics pose a promising insight to mitigate citation boosting practices. To begin with, we have explored the nuances of the econometric production function called Cobb-Douglas model, a concave function, which provides global maxima when used along with a convex optimization tool. However, with an increase in the number of input variables, the complexity of this model increases. This increase in complexity may cause curvature violation of the Cobb-Douglas model to a great extent which would cause erroneous elasticity coefficient estimations. Hence, we have designed a modified Cobb-Douglas model called Stochastic Cobb-Douglas model. This model uses stochastic frontier analysis(SFA), a method used in economics to estimate the values of elastic coefficients. The next section uses this model to compute the journal's and author's internationality score, derives a penalty for unethical practices at journal and author level and later, classify journals into low, medium and high international using UGNN classifier. Classification of authors is a part of future work and it can be inferred that since the author's score can be highly granular, fuzzy neural networks can be used for author's classification.

7 Model for computation of journal's and author's profit score (effective internationality score after penalty)

This section deals with computing the effective internationality score of journals and authors and using the score to classify journals into layers of internationality- **high, medium and low**. To the best of our knowledge, such classification technique, which is based on the understanding of the granularity of internationality, has never been attempted. This is a *cornerstone* of our work. As already stated, classifying authors is a part of future work.

7.1 Profit score computation for a journal

We define journal's profit score on the scale of 0 to 1 as a reflection of the effective international prestige of any journal, computed using stochastic Cobb-Douglas model and penalty function. In order to compute this score, we have explored the unethical methods by the journal's publishers and the authors, who might have boosted the internationality score remarkably. In this section, we propose methodologies to penalize such unethical practices. The next subsection section dwells on computing the effective internationality score of journals and devises scheme to use those scores in classifying journals to different layers of internationality, a novel postulate.
As defined in section 5, cognizant citation implies unethical means of boosting the citation count of the journal by misusing the acquaintance among the high authority members of the editorial board of two different journals. Penalizing this behavior should reflect in the overall score from the Stochastic Cobb-Douglas Model, which is handled as below

Consider the profit function:

$$\pi = \prod_{i=1}^{n} k x_i^{\alpha_i} - \sum_{i=1}^{n} w_i z_i$$

π:*Profit score*
z_i:*Penalty input variables*
w_i:*Cost of penalty input variables*

Net score after this penalty corresponds to the actual internationality score of that journal. For example, Let's consider a journal J which is involved in cognizant citations nexus with journals J_1,J_2 and J_3 that tells us all the articles of journal J have cited one or more articles of journal J_1,J_2,J_3 and vice versa. Assume, total number citations from journal J to journals J_1,J_2 and J_3 are 15,20 and 18 respectively.

Steps of penalization in case of cognizant citations:
1. Normalize the cognizant citations for a journal.
2. Sum all the normalized cognizant citations.
3. Subtract the value obtained through step 2 with the Stochastic Cobb-Douglas output score of that journal.

So, to solve, we have normalized values as, 0.75, 1, 0.9 respectively and their equals 2.65. Assuming, Stochastic Cobb-Douglas output score for J as 0.87562. On substituting the above values in the equation, the profit score for J is -1.77438. Similarly we can compute cognizant citation for J_1, J_2 and J_3.

Note: We consider weight w_i as unity in profit function in case of cognizant citations because we believe that occurrence and existence of cognizant citation demonstrates a high level of unethical practice by the journal. Thus it is liable to 100% penalty in such scenarios. In brief, the increase in the degree of cognizant citations increases penalty in the profit score. The profit score falls abruptly below zero in such a case. In the crux, it is evident that the net profit score, with penalty correction, reflects the true worth of a journal on the scale of 'internationality'. Hence journals, which belong to diverse fields, can be compared based on this new profit score. Evidence of cognizant citations, which reflects a higher negative value of profit score, showcases evidence of journal's tactical malpractice to boost its credibility.

We are still working on the data gathering part to verify this proposed solution for cognizant citations.

7.2 Profit score computation of scholar/authors

In this sub-section, we propose a methodology for calculation of profit score for an author. We considered six predictor variables as inputs in Stochastic Cobb-Douglas function, namely, OCQ (Other-Citation-Quotient), Author-SNIP, h-index, Author-IC (International Collaboration Ratio), Author-NLIQ (Non-Local Influence Quotient) and i-10 index. Copious and genealogy citations are used as a penalty.

- **OCQ:** Other-citation-quotient for an author is equal to:

$$1 - \frac{SelfCitations}{TotalCitations}$$

  Self-citation for an author A is a case when he cites his own work.
- **Author-SNIP:** Author-SNIP is equal to the cumulative of Article-SNIP for every article authored/co-authored by him/her, where

$$ArticleSNIP = \frac{JournalSNIP}{TotalJournalArticles}$$

- **h-index:** A scholar with h-index of 'h' has published 'h' papers, each of which has been cited at least 'h' times. [28]
- **Author-IC:** Author's International Collaboration Ratio is equal to the ratio of the number of collaborators from other nations and total number of collaborators.
- **Author-NLIQ:** The total count of citations for an article received by an author that originated from journals other than publishing journal plus citations originating outside co-authorship network or genealogy network. This is complex to evaluate.
- **i-10 index:** Count of the number of articles cited at least ten times.
- **Citation Genealogy Ratio (CGR):** The ratio of total number of an author's citations from his/her genealogy network to the total number of that author's complete citation network (excluding self-citations) is defined as citation genealogy ratio

To calculate Stochastic Cobb-Douglas score for an author, normalize the input variables, then Sparse PCA is run to reduce the dimensionality of data set, if possible, and finally, SFA to generate an estimated value of elasticity coefficients α_i's is applied. If SFA results fail to justify the role of technical inefficiency then we resort to ordinary least squares method.

From the Stochastic Cobb-Douglas output score for an author, the net score is computed by penalizing authors, if any signs of using suspicious practice are detected. Section 7.3 elaborates on these practices and suggests models/algorithms to identify such attempts. Two scrupulous mechanisms widely used by scholars to boost their citation counts are identified as:
1) Genealogy Citations (this is elaborately covered in separate section, section 9, Genealogy Citation Model)
2) Copious Citations (section 7.3.2).
Consider the profit function,

$$\pi = \prod_{i=1}^{n} k x_i^{\alpha_i} - \sum_{i=1}^{n} w_i z_i$$

π:*Profit score*
z_i:*Penalty input variables*
w_i:*Cost of penalty input variables*

As defined in section 5, genealogy citations can act as a catalyst that can significantly boost a scholar's citation count. Generally, a scholar's successors, who are his students, and their successors, are seen frequently citing each other irrespective of any contribution. Hence, we inflict a penalty on author if the number of citations from scholar's genealogy tree crosses a threshold. This penalty is computed as below:

$$w_{GC} = \frac{(GC - \gamma_1 * TC)}{GC} \quad if \quad \frac{GC}{TC} \geqslant \gamma_1 \;\; or \;\; CGR \geqslant \gamma_1$$

where
γ_1:*Threshold Level*
GC:*Genealogy Citations for an author*
TC:*Total Citations for an author*

Note: All the terminologies and genealogy tree are elaborated in section 9

The copious citation is another staggeringly used mechanism to boost citation by a set of scholars for their mutual benefit. For example, two scholars cite each other in all of their published work with other author's minimal or insignificant contribution. Such scholars are liable to penalty in the form of some reduction in the Stochastic Cobb-Douglas score. It should be noted that, among the mutually benefiting scholars, none of them appear in other scholar's genealogy tree. They are mutually exclusive in terms of successor citations. Example: Let's consider a scholar A is involved in copious citations with authors A_1,A_2 and A_3. That is in all of the A's published work, A cited authors A_1,A_2 and A_3 and vice versa. Also, genealogy network of A does not include author A_1,A_2 and A in it. Suppose total number of times author A cited authors A_1,A_2 and A_3 is 10,12 and 8 respectively.

Steps of penalization in case of copious citation:
1. Normalize the copious citations for an author.
2. Take the sum of normalized copious citations.
3. Subtract the value obtained through step 2 with the Stochastic Cobb-Douglas output score of that author obtained after penalizing for genealogy citations, if that author is liable for the penalty.

Solving the example based on these above stated steps we get normalized values as 0.8333,1 and 0.667 respectively. Sum is 2.5. Assuming Stochastic Cobb-Douglas output score for A is 0.78412 then, the net profit score for A is -1.71588. Similarly we can penalize scholars A_1,A_2 and A_3. We consider weight w_i is unity in profit function in case of copious citations, because we believed that copious citations is deliberate unrighteous method. Hence, all the scholars involved in it are liable to 100 % penalty.

For an author the total penalty score is:

$$\sum_{i=1}^{2} w_i z_i = w_{GC} * normalized\ GC + Copius\ citation\ penalty$$

The net score derived after the penalty is a fair measure of one's performance. Also, this score helps in legitimate comparison of the scholars, who might also belong to different domains. In conclusion, when a scholar deliberately indulges in unfair mechanisms of citation boosting such as genealogy citations or copious citations or both increases then, his/her profit score, computed with the additional penalty, falls abruptly below zero. This higher negativity of profit score becomes evidence of grave involvement of that scholar in different unethical practices of citation boosting.

We are trying to build a large corpus of author level data through web scraping, which is work in progress. Threshold value i.e γ_1 is decided by observing the results of a simulation which should be performed on a very large data-set. However, due to unavailability of data set we couldn't set the values of threshold by simulation.

This subsection exhibits a mechanism to compute a net score of journal's and author's internationality. The next subsection will unveil various mathematical models and algorithms to detect fraudulent practices prevailing in the system. The first part of the section demonstrates graph-based approach to identify cognizant citations followed by models to detect copious citation. Genealogy citations are detected and computed using graph theory approach, the details of which are covered in a separate section, 9.

7.3 Greed-aware Mathematical models: Metric Quantification

Classification of journals based on the profit score needs quantification of penalty parameters – copious and cognizant citations. This section deals with different terminologies and their associated usage in tracking copious citations and self- citation. In order to structure the data for processing using algorithms, we start with definitions of the adjacency matrix, directed graph, equivalence classes and bipartite graphs and increment count of 1's in $i-i$ and $i-j$ intersection of the adjacency matrix.

Definitions:

- Directed Graph is a graph in which edges connect vertices in a specified direction. Figure 3 shows a graph to specify citation relationship between set of authors, where nodes q_0, q_1 etc. denote authors, a directed edge q_0 to q_1 conveys q_0 cites q_1 and loop signifies self-citation.
- Adjacency matrix is a matrix structure that represents a graph in which an entry (0 or 1) of a matrix indicates whether a vertex is adjacent or not. Here, it represents author's citation pattern such that if an entry is '0', there exists no citation relationship between authors, whereas a '1' indicates one cites the other.
- Equivalence class is a class that has a set of elements which satisfies equivalence relation, a (binary) relationship that satisfies properties of reflexivity, symmetry and transitivity.

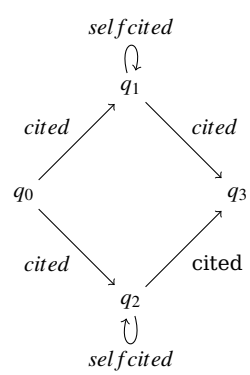


Fig. 3: digraph for citations

7.3.1 The models, schemes and Algorithms to Detect Cognizant Citations in Journals

Equivalence Class Algorithm 1: The algorithm uses an adjacency matrix to reckon a number of self-cites and community citations within journals. As already explained, the adjacency matrix is a

```
Input: An adjacency matrix a[i][j] representing citation information
Output: An equivalence class array contains self citation and copious citation information
Initialize all elements of an array representing number of self citation of an author to zero
Equi_class[] ← 0
for every author: selfcitation[i] do
    selfcitation[i] ← 0
end for
for every journal: JNames[i] do
    A digraph represents citation relationship between set of authors
    Convert the digraph to an adjacency matrix a[i][j], where
    a[i][j] ← 0 implies no citation
    a[i][j] ← 1 implies there exists citation
    Check Reflexive condition by
    if a[i][i] ← 1 then
        Increment self citation count of an author by
        selfcitation[i] ← selfcitation[i] + 1
        Check Symmetric condition by
        if a[i][j] ← 1 and a[j][i] ← 1 then
            Check Transitive condition by
            if a[i][j] ← 1 and a[j][k] ← 1 and a[i][k] ← 1 then
                Add [i,j,k] to set of equivalence classes
                Equi_class[i] ← Equi_class[i] + [i, j, k]
            end if
        end if
    end if
end for
```

Algorithm 1: $Equivalence_class(a[i][j])$

$$\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{pmatrix} \end{array}$$

Fig. 4: Matrix for Describing Citations

representation of the citation pattern of authors for a particular journal. A '1' at the intersection of the ith row and jth column indicates author i has cited author j and '0' indicates no citation. Similarly, if the intersection of the ith row and ith column is '1' it denotes the self-citation of an author in the current journal. Repeating the search in adjacency matrices of every journal, total self-citation of an author across all journals can be obtained. In order to compute the Copious citation coefficient, the equivalence classes present in the matrix of a journal must be determined.

The algorithm first determines if an author has imparted journal self-citations by computing all reflexive relationships. Figure 10 shows sample author network in which each arrow indicates a citation. The arrow-head indicates the author who has been cited and the tail end of arrow indicates the author who has cited. Nodes a,b,c,d,f and h have arrows pointing to itself indicating that these authors have cited themselves i.e self-citations. Node pairs (a,b),(a,c)(b,d),(c,d),(d,f),(f,g),(e,h) have mutually cited each other. In the example matrix Fig. 11 (which is derived from Fig 10), authors a,b,c,d,f and h have self-cited. Then it proceeds to determine the symmetric relations i.e if the two authors a_i and a_j have cited each other from the same journal. In any matrix, if intersections of ith row and j^{th} column and j^{th} row and i^{th} column are 1, it points at a symmetric relation between i and j. Looking at the matrix, (a,b), (a,c) and (c,d) have a symmetric relation. Finally, the transitive relations are determined by searching for an intermediate author who has cited two inter-related

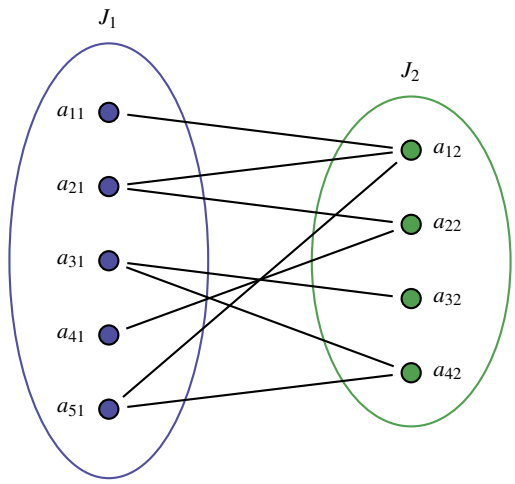


Fig. 5: Bipartite graph

authors. In the matrix considered, a has cited c, c has cited h and a is citing h. From the example under consideration, it is observed that [a,c,h] form an equivalence class. The main objective of the algorithm is to segregate similar equivalence classes. The final equivalence classes are later used to determine the Copious Citation coefficient. Please note that minor modifications can be made in recognizing author citation patterns belonging to different journals. In that case, such behavior will be used to determine a penalty for individual authors and not for journals.

Bipartite Graph: To inspect Cognizant Citations- A bipartite graph (Fig. 5) consists of vertices split into two independent components or sets such that every edge connects vertices from the two different sets but never from the same set. The graph is used to represents two journals J_1 and J_2. a_{j1} and a_{j2} represent articles of journal J_1 and J_2 respectively. All articles of journal J_1 have cited one or more articles of journal J_2 and vice versa. This mutual relationship between two independent journals, represented by a bipartite graph, is called Cognizant Citations.

Computation Of Cognizant Citations-Algorithm 3: This algorithm extracts all the journal names except input journal name in an array $JNames[]$: line 3 by calling function $Article_name()$ and collecting all article names of input journal in an array $arr[]$: line 4. Then, for every journal present in $JNames[]$, the algorithm checks the condition essential for cognizant citations. If both journals, the input journal and the journal under consideration satisfy the condition, then algorithm extracts

1: **Input:** Journal name
2: **Output:** All article names in the input journal
3: $k \leftarrow 1$
4: **for** every author name in the graph **do**
5: **if** $V(graph)[i].journal == input\,journalname$ **then** ▷ compare each vertices(articles) of the graph having attribute journal name with the input journal name
6: $arr[k] \leftarrow$ *V(graph)[i].name* ▷ put name of vertices(articles) in the arr[]
7: $k \leftarrow k+1$
8: **end if**
9: **end for**
10: **return** $(arr[])$

Algorithm 2: $Article_name(journalname)$:Algorithm to extract all the article names from a graph for a particular journal

all article names of the journal, $JNames[i]$ and merges those article names with the articles of input journal and eventually collects them in an array $arr2[]$. As a result, a sub-graph is generated. The sub-graph contains all the articles present in $arr2[]$ as nodes and edges represent citations given by those articles to each other. We count all those edges which are between articles of different journals and omitted those edges which are between articles of same journals. We repeat this process for every journal present in $JNames[]$: line 5 to 18. Thus, a tracker is obtained which counts cognizant citations between input journal and every other journal present in our graph. *Clearly, elements of $arr2[]$ are represented as **bipartite graphs** and articles present in $arr2[]$ can be divided into two disjoint sets U and V (that is, U and V are each independent sets) and we have to count every edge which is from element of U to element of V.*

```
1:  Input: Journal name
2:  Output: Cognizant Citations between input journal and all other journals in the graph
3:  JNames[] ← extract all journal names except input journal name from graph
4:  arr[] ← Article_name(input journal name)
5:  for every journal name i in the JNames[] do
6:      Cog_Cit ← 0
7:      if one or more edges are present in between all articles of input journal and journal JNames[i] and vice versa then
8:          arr1[] ← Article_name(JNames[i])
9:          arr2[] ← merge(arr[],arr1[])
10:         induce a sub graph from original graph having all the elements of arr2[] as nodes and all the edges between
                nodes
11:         for every edge in the sub graph do
12:             if both nodes(articles) of that edge from different journal then
13:                 Cog_Cit ← Cog_Cit + 1
14:             end if
15:         end for
16:     end if
17:     print (Cog_Cit)
18: end for
```

Algorithm 3: Algorithm to calculate cognizant citations for a journal

Trace Algorithm 3: Cognizant citation detection For sake of clarity, the algorithm is traced through an example. Table 7 shows the Journal list and the articles published in Journals. Figure 6,7,8,9 are subgraphs to compute Cognizant citations between Journals using Algorithm 2 and 3:

Journal's Name	J_1	J_1	J_2	J_2	J_3	J_3
Article's Name	a	c	b	d	e	f

Table 7: **sample data**

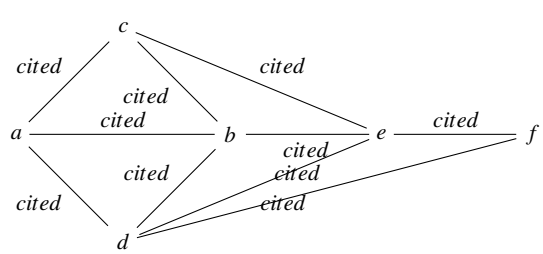
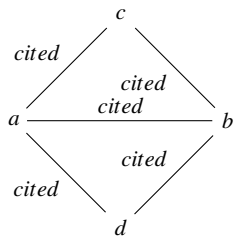
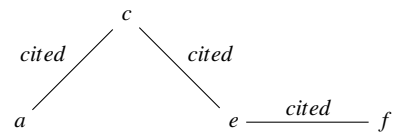


Fig. 6: Graph Of Sample Data



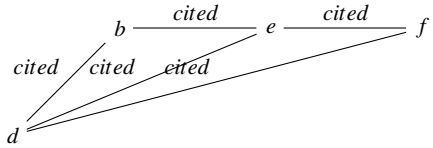
Cognizant citations between journal J_1 and J_2=3

Fig. 7: subgraph of articles of journal J_1 and J_2



cognizant citation criteria between J_1 and J_2 unsatisfied

Fig. 8: subgraph of articles of journal J_1 and J_3



Cognizant citations between journal J_2 and J_3=3

Fig. 9: subgraph of articles of journal J_2 and J_3

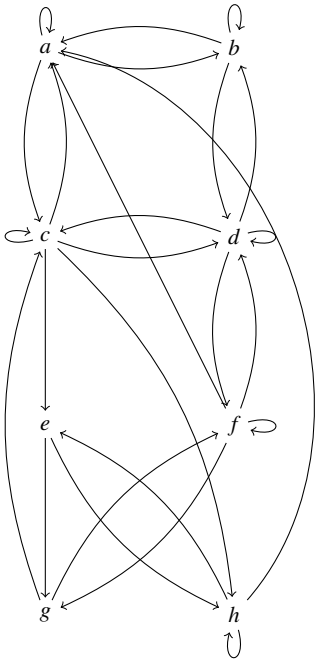


Fig. 10: Sample author network to determine citations

$$
\begin{array}{c|cccccccc}
 & a & b & c & d & e & f & g & h \\
\hline
a & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
b & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
c & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 1 \\
d & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 \\
e & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
f & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
g & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
h & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1
\end{array}
$$

Fig. 11: Adjacency matrix representation of author network. 1 indicates citation and 0 indicates no citation, If A[i,j]=1 then it means that author i has cited author j. If A[i,i]=1 then it indicates self citation by author i.

7.3.2 Models and Algorithm to detect Copious Citations for author network

We proceed to compute the greed-aware metrics needed to design the author internationality score model. *These metrics are defined already in section 5.*

The objective of our study is to develop algorithms to trace authors who might have fabricated their citations in order to increase citation count. The most common method is self citation. Authors may cite their own work indiscreetly to boost their citations. There are other ways to accomplish higher citation count, mutual and copious citation between pair of authors being the notable ones. Eventually, the network of authors increases disproportionately, leading to the formation of a community where all authors part of the network are benefited. Here, authors lay stress on the thought process that neither self nor mutual citation should be considered candidates for penalty. It is only when these citation practices cross certain threshold, the authors involved are liable to be penalized. Algorithms 4, 5 and 6 help finding all types of community-bound citation practices.

```
procedure calc_self_cites(adj_matrix, size)
    authors ← {}
    i ← 0
    while i < size do
        if adj_matrix[i,i] = 1 then
            authors ← authors ∪ {i}
        end if
        i ← i + 1
    end while
    return authors
end procedure
```

Algorithm 4: Get Self-Citations

Algorithm **4** is used to compute the self-citations of author. It takes the citation information of authors as an adjacency matrix as input and returns the authors who have cited themselves.The output given is according to the sample graph considered in Fig. 10 and Fig. 11. Consider the set of authors who have **Self-Cited:** a,b,c,d,f and h. We have, in the adjacency matrix representation

A[1,1] = 1
A[2,2] = 1
A[3,3] = 1
A[4,4] = 1
A[6,6] = 1
A[8,8] = 1

which implies that these authors have cited themselves.

```
procedure CALC_REFLEXIVE_CITES(adj_matrix, size)
    authors ← {}
    i ← 0
    while i < size do
        j ← i + 1
        while j < size do
            if adj_matrix[i, j] = 1  AND  adj_matrix[j, i] = 1 then
                authors ← authors ∪ {{i, j}}
            end if
            j ← j + 1
        end while
        i ← i + 1
    end while
    return authors
end procedure
```

Algorithm 5: Get mutual citation (copious citation) count

Algorithm 5 returns the pairs of authors who have collaborated by citing each other. Similar to the algorithm to compute Self-Citations, the input is an adjacency matrix representation and the results given are according to the sample graph in Figure 10 and the corresponding matrix representation Figure 11 considered. Let us consider the following example: **Mutual Citations:** (a,b),(b,d),(a,c),(d,f), (c,d) and (e,h). The collaboration between these pair of authors is reflected in the adjacency matrix as

A[1,2] = A[2,1] = 1
A[2,4] = A[4,2] = 1
A[1,3] = A[3,1] = 1
A[4,6] = A[6,4] = 1
A[3,4] = A[4,3] = 1
A[5,8] = A[8,5] = 1

These authors form a citation cartel i.e. they mutually form a network by citing the others part of the network. By this arrangement the citations of all the authors in the network will increase. Further evidence of doubtful collaboration can be demonstrated from the example of **Equivalence class:** [a,b,c,d]. The adjacency matrix represents this as:

A[1,2]=A[2,4]=A[4,3]=A[3,1]=1

Here, a, b, c and d form a group of authors who have mutually collaborated and formed a separate set. The Equivalence class algorithm is shown in previous subsection as algorithm 1.

As already stated, self-citations and mutual citations are not liable for any penalty until it becomes evident that their count has exceeded certain boundary values. The adjacency matrix only detects the community citation pattern, and to detect and penalize copious citations, we need a representation that reflects the citation pattern as well as its count. We make a small addition in adjacency matrix representation, in which, a matrix entry will be:

$$A[i,j] = 1^{\frac{p}{n}}$$

where
i and j are two citing authors
A[i,j] as '1' indicates i cites j
'p' represents number of times i has cited j
'n' represents total articles j has published.
Technically, the ratio p/n reflects a number of citations received by author j from i, when he has a total of 'n' publications. Conversely, A[j,i] will indicate citation count of author i cited by j alone. If the ratios exceed certain minimum threshold, i and j are liable for penalty in their Cobb-Douglas internationality score for indulging in Copious Citation.

Within a set of authors, a class of related authors must be established. In order to achieve this, the existence of a direct relation or indirect relation involving other intermediate authors is determined. If such relation exists between an author and one or more other authors, they are considered to belong to the same class. Warshall's algorithm is used to find the shortest paths in the network. Using Warshall's algorithm, the reachability to a node from another node is determined. Each author is considered a node and if an author is reachable from another, there exists a citation relationship between them and they are grouped under the same class.

```
Input:Author adjacency matrix A[n][n]
Ouptut:Classes of related authors
for k from 0 to n-1 do
    for i from 0 to n-1 do
        for j from 0 to n-1 do
            if A[i][j]==1 and A[j][k]==1 then
                A[i][k] = 1
            end if
        end for
    end for
end for
for i from 0 to n-1 do
    for j from 0 to n-1 do
        if A[i][j]==1 then
            Append j to class[i]
        end if
    end for
end for
```

Algorithm 6: Establish author relations

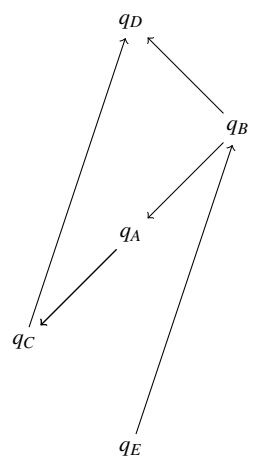


Fig. 12: Sample author network to determine reachability of each author. Each arrow indicates a citation and thereby a relation between two authors. Author A has cited C and C has cited author D. Hence the class of author A comprises of (A,C,D). Similarly, a class is formed for each node by tracing the relations of each node adjacent to it.

$$\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

Fig. 13: Author adjacency Matrix for reachability graph. A 1 indicates that a relationship exists between two nodes and 0 indicates no relationship exists.(A,C,D) form the class of Node A. Hence in the matrix A A[A,C]=1 and A[C,D]=1.

The adjacency matrix in Figure 13 is a representation of Author network in Figure 12. Consider the algorithm above for matrix in Figure 13. Class[A]=[A,C,D] Class[B]=[A,C,D] Class[C]=[A,C,D] Class[D]=[A,C,D] Class[E]=[A,B,C,D]

The greed-aware metrics described above are used in the penalty estimation when computing the profit score (effective internationality score of journal and author). These metrics together are used in the final model of computing the profit score. Once the score is obtained, we classify journals into three classes of internationality using Unified Granular Neural Networks, based on the granularity of the profit score for a set of journals. The next section describes the pseudo code of the classifier and demonstrates a high accuracy of classification when compared with other models.

8 Classification Of Journals

Now, we propose a framework for binary classification of journals based of their profit score. One way of classification is the division of journals into two classes, good and not so good. In order to explore a strong discrimination factor for this classification, we performed Shapiro Wilk test on the dataset. The test results revealed frequency distribution of journal's profit score to be a Gaussian distribution. That is, a few data points lie on the lower tail, a few lie on higher tail and rest of the scores points lie around the mean. Intuitively, we decide to classify journals into three classes. Namely, **Low**, **Medium** and **High**. $0 < c \leqslant 0.4$ signifies class **Low**,$0.4 < c \leqslant 0.7$ signifies class, **Medium** $0.7 < c \leqslant 1$ signifies class **High**. Note that normalization of input variables lay profit scores in the range of 0 and 1. Due to this unit width range, profit scores of journals fall very close to each other. This close proximity makes journal comparison and distinct classification challenging. For this task, we have used the classification model, i.e., developed based on Unified Granular Neural Networks (UGNNs) [11]. Performance of UGNNs is compared with k-nearest neighbor (kNN with k= 3, 5, 7) and conventional neural network (NN). UGNN outperforms kNN and NN. The following section briefly describes the motivation in the use of UGNNs over NN and its advantages.

Firstly, we perform KNN classification to build a basic standard from where unified GNN's performance can be compared. Next, we apply NN then GNN and finally unified GNN. We will discuss the results of classification in further sections. The following section provides brief information on UGNN.

8.1 Unified Granular Neural Networks (UGNNs)

Artificial neural networks (ANNs) are the computational systems mostly used for classification data sets with highly overlapping class boundaries. ANNs have the abilities to learn and generalize, similar to the abilities of the human brain and they can also classify. The weight parameters between the layers of ANN are updated using back error propagation learning algorithm [33]. The weight parameters between the layers of ANN are updated using back error propagation learning algorithm. However, ANNs are often treated as black-boxes due to their inabilities in explaining the complexity, contribution of weights and their connections. To overcome this shortcoming and acquire the operational transparency of ANNs, granular neural networks (GNNs) are used for classification tasks. Intuitively, gaining more knowledge by looking inside into the network architecture enlighten the developer for further improvement in the final decision. GNNs work with the principle of granular

computing [12], where the basic building blocks called granules play the key role in the granulation of information at different granularity. An architecture of GNNs is built based on the *if then* rules which are extracted using Kasabov Rule Extraction (KRE) method from [13] granulated data. There are two broad methods of information granulation, which are used to granulate the input data:

(i) Class non supportive granulation
(ii) Class supportive granulation.

CNS granulation, uses three fuzzy granules, *low*, *medium* and *high* to granulate the information along each feature axis. CNS granulation of input data does not consider the class-wise belonging features. CS granulation of input data improves the performance of classifier compared to CNS method, because it considers class-wise belonging of each feature to the classes present in the dataset.
Each of the set of rules develop GNN and the process leads to the cumbersome task of choosing the best GNN among them. One possible solution is to consider all the GNNS developed through different sets of rules and combine them in the framework of multiple decision systems, to form unified granular neural networks (UGNN). UGNN provides the collective opinion of individual GNN through the decision combination operations, such as mean, median, maximum, sum, minimum and product etc.

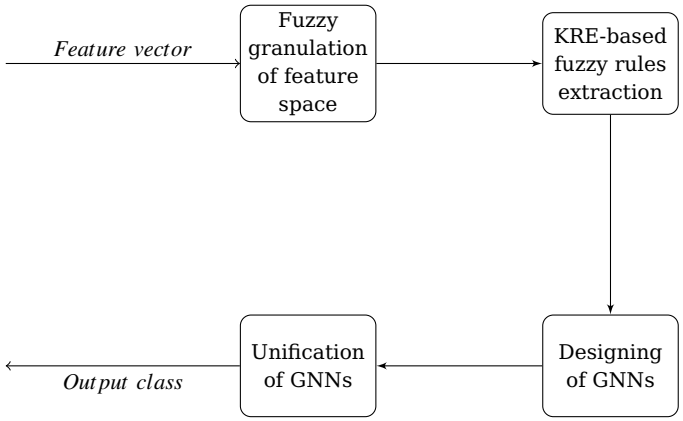


Fig. 14: Block diagram of model of Unified GNN.

Pseudo code of unified GNN classification model

Input: Training and test samples
Looping for T iterations,where T = number of training samples
1.Fuzzy granulation of features of training samples using CS(Class Supportive) and CNS(Class Non Supportive) methods
2.Process these granulated features through a fuzzy neural network and extract knowledge in terms of granulated rules(four rules) using KRE(Kasabov rule extraction) method
3.Use these rules to build four GNNs,where one set of rules leads to one GNN
4.Training these GNNs using back propagation learning algorithm
5.Unifies all four GNNs to develop a unified GNN classification system
Return the trained UGNN classifier

Training Sample(in %)	Model 1	Model 2
60	79.14729%	80.57142%
80	82.4444%	83.88888%

Model 1:Ungranulated FVs + KNN Algorithm
Model 2:Class Supportive Granulated FVs + KNN Algorithm

where FVs:Feature Vectors and K=5

Table 8: **Result Of Classification Using KNN**

Training Sample(in %)	Model 3	Model 4	Model 5	Model 6
60	83.00001%	85.71426%	87.28526%	94.36122%
80	86.7385%	89.44442%	91.8887%	96.10785%

Model 3:Ungranulated FVs + Back Propagation Algorithm(NN)
Model 4:Class Supportive Granulated FVs + Back Propagation Algorithm(NN)
Model 5:Class Supportive Granulated FVs + Granular Neural Network(4 rules per class)
Model 6:Class Supportive Granulated FVs + Unified Granular Neural Network(4 GNNs)

where FVs:Feature Vectors

Table 9: **Result Of Classification Using Different NN Models**

Results of using UGNN to predict the labels of test samples

1. Table 8 and Table 9 shows the accuracy of classification through different models. Among the four models, classification accuracy is lowest in Model 1 and highest in Model 6, which is very intuitive.
2. All the results in Table 8 and Table 9 are obtained through **10 fold cross validation**. In 10-fold cross validation (k fold cross-validation [31] is one of the cross-validation techniques for optimizing parameters to fit model as best as possible), the data set is randomly divided into 10 equal sized subsets. Every time, one of the k subset is used as the test data and the leftover 9 subsets are utilized for training the model. Each of the 10 subsets is used for validation exactly once and the results of all 10 validations are averaged for the final estimation.
3. In unified GNN, model is constructed using four different GNNs because we have extracted four rules. We take the decision of each GNN into consideration and combine them using five different methods such as mean, median, product, minimum and maximum. With this data set, the highest accuracy is produced when we combine individual GNN's decision through the decision combination operation called median. The results obtained through model 4 in Table 8 are generated by median decision combination criteria.
4. Our data set has few number of journals, increase in the number of journals will reflect good variation in the results.
5. As described in [11], the improvement in performance was not significant with the cost of additional computation if the number of rules increased beyond four. We tried with five rules and the improvement in performance was not very fascinating with the cost of additional computation so we continued with four rules.
6. We used class supportive granulation because as proved in [11], models with CS granulation provided better results than models with CNS granulation. This is because the class supportive granulation gives more emphasis on class-wise belonging of features to classes in the process of granulation.
7. We used π type membership function to represent the granules because the fuzzifier parameter can be tuned according to the requirement.

This section describes a mechanism to classify a set of journals exploiting a novel methodology called UGNN based on the granularity of the profit scores for a set of journals. Also, we proved the nuances of the model through various supporting analysis and respective results. Next section gives a detailed explanation of Genealogy Citation Model. The model supports in obtaining genealogy citations to induce penalty to authors who have exorbitantly high genealogy citations.

9 Genealogy Citation Model

Genealogy is the study of members of families and discovering the sequence of members, each of which is considered to have generated from its ancestor and their history. Also, Genealogy tree is a tree-like structure in which nodes represent a member of the family and edges represent the relationship between ancestors, descendant and all other members of the family and of the other genealogical group. In the context of scholarly articles and journals data, nodes represent author and edges between nodes represent adviser/advisee relationship. For example, assume, we have three authors A, B and C as shown in figure 15. A is an academic adviser of B and B is the adviser of C and D. The tree represents the ancestor-descendant relationship between authors A, B, C and D. It is called Author Genealogy tree. In this section we put forward some definitions related to Genealogy Citations (GC), Non Genealogy Citations (NGC) and Citation Genealogy ratio (CGR), and discuss different cases one may face while calculating GC,NGC AND CGR. Algorithms for calculation of GC,NGC and CGR for all possible cases are also proposed in the next sub-section. This author's

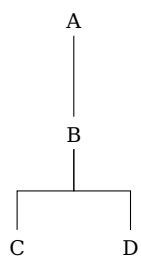


Fig. 15: Author Genealogy Tree Example

genealogy tree is used to calculate various metrics such as Genealogy citations(GC), Non-genealogy citations(NGC) and CGR (citation genealogy ratio). Following are the definitions for these metrics.

1. **Author Genealogy Network**:-An author's/scholar's genealogy network includes all those authors, who are one or two hops away (up or down) from his/her position in the tree, and authors who are his/her siblings. For example, consider Fig 15, the advisor of any author A (say Adv(A)) is one hop above him, the advisor of Adv(A) is two hops upward in the tree. Likewise, one hop below is the advisee of author A (say Ads(A)) and two hop below is Ads(A)'s advisee and so on.
2. **Genealogy citations(GC)**: Genealogy citations are citations which an author gets only from his/her Genealogy Network.
3. **Non-Genealogy Network citations(NGNC)**: Let X be the total citations for an author A from his/her Genealogy Network. Let Y be the citations that he receives from authors of the complete citation network (excluding self-citations). Then Non-Genealogy Network Citations can be defined as

$$\textbf{Non-Genealogy Network Citations} = Y - X$$

4. **Citation Genealogy Ratio(CGR):** Let X be the total of number of citations an author receives from his/her Genealogy Network. Let Y be the number of citations the author receives from his/her complete citation network (excluding self-citations). CGR for the author:

$$\textbf{Citation Genealogy Ratio(CGR)} = \frac{X}{Y}$$

Note: We use the word genealogy tree interchangeably with genealogy network.

9.1 Genealogy Tree Representation

Genealogy is a study of a researcher's line of descent in the academic world that includes his academic advisers, advisee their ancestors and descendants. Figure 16 is a sample graph comprising of nodes depicted as researchers and edges as adviser-advisee relationships. Table 10 shows a list of authors/researchers and their id's which will be used as a reference in explanation of different algorithms. However, the structure shown in Figure 16 does not have inherent tree properties. There are nodes like H, Z, L, M, K etc which have two advisers, which shows that these nodes have two parents. Hence, there arise needs to re-configure the whole structure into a representation which has a hierarchy and which does not disrupt tree property by ensuring there is no node in the structure which has two parents. Figure 17 is an equivalent representation of the sample graph in Figure 16. Nodes which have two parents (authors with two advisers) are forked into two sub-nodes that have the same name, same id but has different weights. Taking the case of author H, his two advisers are C and D. Node H is bifurcated into H1 (weight as 1) and H2 (weight as 2). H1 remains associated with C1 and H2 is merged with D1 along with its descendants. Likewise, all authors that have two advisers are branched into two sub-nodes differentiated by their weights.

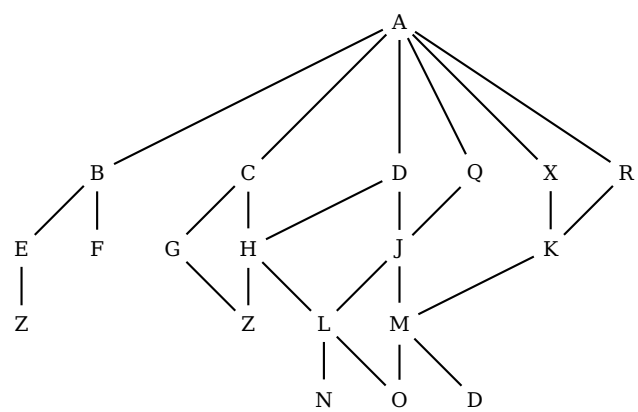
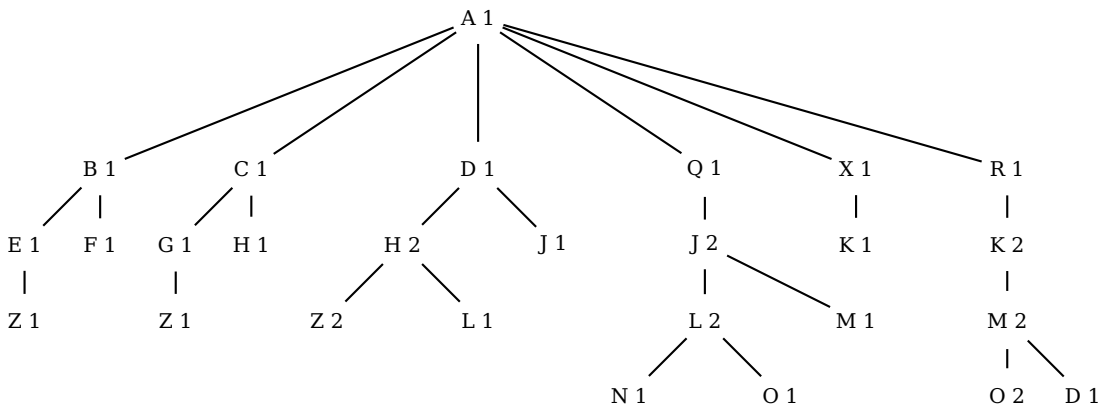


Fig. 16: **Original Representation Of Sample Data**

Author's Name	A	B	C	D	X	Q	R	E	F	G	H	J	K	L	M	N	O	D	Z	Z
Author's Id	1	2	3	4	5	17	18	6	7	8	9	10	11	12	13	14	15	16	20	19

Table 10: Sample Data



Note:Number associated with each author's name is its weight.

Fig. 17: **Equivalent Representation Of Sample Data**

9.2 Data Structures

- **All-Author matrix**:-It is a matrix of size $n \times n$ where n denotes the total number of authors in the genealogy tree. Each element of the matrix indicates the number of times an author has cited another author. For example, each element a_{ij} of the matrix represents the number of times an author j has cited author i. Figure 18 shows the All Author matrix for 7 Authors.

$$
\begin{array}{c}
\\ A \\ B \\ C \\ D \\ X \\ E \\ F \\ G \\ H \\ J \\ K \\ L \\ M \\ N \\ O \\ D \\ Q \\ R \\ Z \\ Z
\end{array}
\begin{array}{c}
\begin{array}{ccccccccccccccccccccc}
A & B & C & D & X & E & F & G & H & J & K & L & M & N & O & D & Q & R & Z & Z
\end{array} \\
\begin{pmatrix}
1 & 0 & 5 & 6 & 7 & 1 & 4 & 6 & 2 & 4 & 6 & 8 & 7 & 1 & 2 & 8 & 1 & 0 & 9 & 1 \\
3 & 1 & 1 & 6 & 5 & 3 & 4 & 5 & 1 & 5 & 0 & 0 & 7 & 1 & 7 & 9 & 6 & 4 & 0 & 1 \\
5 & 11 & 8 & 6 & 4 & 1 & 5 & 2 & 2 & 9 & 1 & 8 & 1 & 0 & 6 & 1 & 0 & 4 & 4 & 1 \\
6 & 12 & 9 & 6 & 2 & 0 & 4 & 7 & 0 & 8 & 6 & 4 & 2 & 9 & 5 & 2 & 9 & 1 & 0 & 0 \\
8 & 13 & 1 & 6 & 3 & 1 & 6 & 1 & 1 & 7 & 5 & 3 & 3 & 7 & 4 & 3 & 9 & 6 & 7 & 6 \\
9 & 5 & 7 & 6 & 1 & 8 & 0 & 0 & 9 & 8 & 6 & 9 & 4 & 8 & 2 & 5 & 6 & 7 & 5 & 7 \\
12 & 6 & 0 & 6 & 0 & 4 & 8 & 9 & 5 & 6 & 6 & 1 & 1 & 7 & 8 & 3 & 1 & 3 & 3 & 2 \\
11 & 8 & 2 & 6 & 4 & 0 & 6 & 6 & 3 & 5 & 6 & 8 & 5 & 3 & 9 & 6 & 2 & 2 & 1 & 1 \\
12 & 9 & 7 & 6 & 8 & 0 & 4 & 1 & 2 & 9 & 6 & 8 & 6 & 4 & 2 & 7 & 3 & 1 & 0 & 10 \\
0 & 1 & 1 & 6 & 9 & 1 & 5 & 6 & 6 & 1 & 6 & 9 & 7 & 1 & 6 & 6 & 1 & 9 & 0 & 1 \\
4 & 7 & 5 & 6 & 2 & 4 & 1 & 9 & 8 & 0 & 6 & 1 & 8 & 5 & 2 & 2 & 1 & 9 & 9 & 5 \\
5 & 3 & 4 & 6 & 4 & 2 & 2 & 4 & 9 & 2 & 6 & 9 & 9 & 6 & 5 & 3 & 2 & 8 & 8 & 4 \\
7 & 2 & 11 & 6 & 5 & 5 & 3 & 5 & 2 & 9 & 6 & 0 & 3 & 2 & 4 & 5 & 3 & 7 & 7 & 9 \\
1 & 2 & 3 & 6 & 9 & 3 & 4 & 3 & 0 & 7 & 6 & 3 & 3 & 7 & 8 & 9 & 4 & 6 & 6 & 7 \\
9 & 3 & 5 & 6 & 1 & 4 & 9 & 2 & 1 & 5 & 6 & 4 & 1 & 8 & 7 & 1 & 3 & 4 & 5 & 5 \\
14 & 5 & 2 & 6 & 0 & 7 & 0 & 6 & 8 & 1 & 6 & 5 & 2 & 9 & 4 & 2 & 6 & 5 & 4 & 3 \\
10 & 7 & 1 & 6 & 3 & 8 & 7 & 1 & 4 & 2 & 6 & 6 & 3 & 5 & 3 & 1 & 7 & 3 & 3 & 2 \\
1 & 9 & 1 & 6 & 7 & 9 & 5 & 0 & 6 & 3 & 6 & 7 & 9 & 1 & 1 & 2 & 7 & 2 & 2 & 0 \\
1 & 8 & 9 & 6 & 5 & 1 & 4 & 6 & 2 & 10 & 6 & 8 & 7 & 2 & 2 & 3 & 8 & 1 & 1 & 1 \\
8 & 1 & 1 & 5 & 6 & 9 & 0 & 3 & 8 & 10 & 6 & 2 & 7 & 4 & 0 & 3 & 9 & 0 & 10 & 1
\end{pmatrix}
\end{array}
$$

  - **Note:** each element a_{ij} of the matrix represents the number of times an author j has cited author i.

Fig. 18: All Author Matrix

- **Two Advisor Matrix**: It is a matrix of size $n \times 2$ where n denotes the number of authors which have two advisors. The two columns store ids of author's two advisers. Figure 19 shows two Adviser Matrix defined for the sample graph. Here author H has two advisors, C and D, their ids are placed in two columns of the matrix. **Note:** In Advisee Matrix 0 represents that author has no advisee.

$$\begin{array}{c} \begin{array}{cc} \textbf{Advisor Id} & \textbf{Advisor Id} \end{array} \\ \begin{pmatrix} 3 & 4 \\ 4 & 17 \\ 5 & 18 \\ 9 & 10 \\ 10 & 11 \\ 12 & 13 \\ 8 & 9 \end{pmatrix} \begin{array}{l} \text{Advisors of author H} \\ \text{Advisors of author J} \\ \text{Advisors of author K} \\ \text{Advisors of author L} \\ \text{Advisors of author M} \\ \text{Advisors of author O} \\ \text{Advisors of author Z} \end{array} \\ \\ TwoAdvisorMatrix \end{array}$$

$$\begin{array}{c} \begin{array}{cc} \textbf{Advisee Id} & \textbf{Advisee Id} \end{array} \\ \begin{pmatrix} 12 & 20 \\ 12 & 13 \\ 13 & 0 \\ 14 & 15 \\ 15 & 16 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{array}{l} \text{Advisees of author H} \\ \text{Advisees of author J} \\ \text{Advisees of author K} \\ \text{Advisees of author L} \\ \text{Advisees of author M} \\ \text{Advisees of author O} \\ \text{Advisees of author Z} \end{array} \\ \\ AdviseeMatrix \end{array}$$

Fig. 19: Matrices

- **Advisee Matrix:**It is a matrix of size $n \times m$ where n denotes the total number of authors who have two advisors and m denotes a maximum number of advisee of the authors from genealogy tree. In figure 19, since there are 7 authors with 2 advisers, each of which has two advisees, the size of Advisee matrix is 7×2.

9.3 Overall Procedure to Compute GC, NCNG and CGR

The overall sequence of steps taken to compute CGR, GC and NGC are as follows. For an author (say A), we accumulate ids of authors which are in author A's Genealogy Network. From the $All-Author$ matrix, we extract citation count of A received from each of these authors. On adding these citations, we get Genealogy Citations (X) for A. X represents total citation that A has received from his/her Genealogy Network. Then, with the help of $All-Author$ matrix, we extract the number of citations received by author A from the entire genealogy tree. Summing all these citations, we get total citations(Y). Non-Genealogy Network Citations for A is computed from $Y-X$. Citations Genealogy Ratio is derived from X/Y. The algorithms discussed in the next section gives a detail computation of GC, NCNG and CGR.
During the course of execution, there may arise four scenarios, depending upon the number of times author's name appears and the number of advisers an author may have in his genealogy network. In most of the universities, a researcher can be mentored by at the most two advisers. The four scenarios are:-
1. An author's name is unique in the entire genealogy tree (unique name case). Authors A, B, C are all such examples in the sample graph of Figure 16
2. Different authors may have same names in genealogy tree (multiple name case). Author D as an example.
3. An author may have two advisers regardless of whether he falls in unique name or multiple name case. (two adviser case) Author M as an example
4. Authors may have the same name in genealogy tree and they may also have two advisers (multiple names and two adviser case). Author Z (with id as 19 and 20) as an example.
Algorithms sufficiently bring about all cases for the computation of CGR, NGNC and GC. However, as a result of tree-restructuring, there are two cases which require special attention when extracting citations especially for advisers and advisees of authors who have two advisers in the genealogy network. Section 9.4.3 discusses the two cases with details.

9.4 Algorithms:-

Main Algorithm-Algorithm 7: We assume that a genealogy tree consisting of authors and their adviser/advisee relationship is already built. If there exists a two-adviser edge for any node, its equivalent representation (depicted in fig.17) is taken into consideration for further processing. An All-Author matrix is also stored for creating arrays and other supporting data structures. The array $AName[]$ and $AId[]$ store the names and ids of an author. The author whose NCR and CGR is to be computed is taken as input. Let's say, the user enters a name Z. Algorithm 7, the Main Algorithm, uses $Name_occurence$ variable to find the frequency of occurrence of Z's name in the genealogy tree: line 3 to 12. If the frequency of occurrence is nonzero four cases can exist (as mentioned in the previous section).
1. *two adviser case* - Which means Z may have two advisers.
2. *multiple name case* - There may be two authors with name 'Z'
3. both *two adviser and multiple name case* - There are two authors with name 'Z' and one of the authors has two advisers.
4. *unique name case* - 'Z' has a distinct name in the Genealogy Tree and has only one adviser.

To figure out, which case prevails for author Z, the algorithm constructs and use various arrays and matrices. Array $Id_collect[]$ collects all author ids with name Z. In parallel, $two_advisor_author_arr[]$ store ids of all authors which have two advisers. The array helps in building $two_advisor_matrix$ and $advisee_matrix$: line 26 to 28. These matrices keep track of all authors who have two advisers and their respective advisee for the entire genealogy tree. arr1[] and arr2[] arrays are constructed and length of the arrays resolve whether Z belongs to 'two adviser case', 'multiple name case' or 'unique name case'. Here are the details.
1. If length of $arr1[]$ is nonzero and length of $arr2[]$ is zero, it indicates Z has two advisers and $two_advisor$ function(i.e algorithm 15) is called.
2. If length of $arr2[]$ is one and length of $arr1[]$ is zero, it means *unique name case* exists and $unique_name$ function(i.e algorithm) is called.
3. If length of $arr2[]$ is more than one and length of $arr1[]$ is zero, *multiple name case* exists and $unique_name$ algorithm runs for each id present in $arr2[]$.
4. If length of $arr1[]$ and $arr2[]$ both are nonzero, then it is a *two adviser + multiple name case*. Hence both $two_advisor$ algorithm and $unique_name$ algorithm runs and length of $arr2[]$ decides the number of times unique name algorithm should run: line 29 to 40.

9.4.1 Adviser Id extraction-Algorithm 8:

This algorithm extracts the adviser ids up to a specified level from the current node *(algorithm ascends the tree and pick ids of intermediate nodes from the specified node up to given level l)*. We used **ancestor** search which is a search option in **R** language under the inbuilt package **data.tree**. This is a non-standard traversal mode that does not traverse the entire tree. Instead, the ancestor mode starts from a node, then walks the tree along the path from parent to parent, up to the root:line 3. The required adviser-ids are extracted from search results:line 4 to 6.

```
Input: author name
Output: CGR
Take an author name as input from user
AName[] ← name of all the author in the tree
AId[] ← id of all the author in the tree
Name_occurence ← 0
k ← 0
for every author name in AName[i] do
    if AName[i] == Input_author_name then
        Name_occurence = Name_occurence + 1;
        Id_collect[k] ← AId[i]
        k ← k + 1
    end if
end for
if Name_occurence == 0 then
    Input author is not present
end if
two_advisor_author_arr[] ← collect those ids from AId[] which occur twice in the genealogy tree   ▷ twice occurence
            of author id indicates that those authors have two advisor
two_advisor_ matrix(two_advisor_author_arr[])   ▷ creation of two advisor matrix
advisee_matrix(two_advisor_author_arr[])   ▷ creation of advisee matrix
arr1[] ← extract all the ids which occur twice in Id_collect[]   ▷ it contains id of input author if he/she has two advisor
arr2[] ← extract all the ids which occur once in Id_collect[]   ▷ it contains id of input author if its name is unique or
            same as any other author's name(i.e multiple name case)
if length_of_arr1[] > 0 then
    for i from 0 to length_of_arr1[] -1 do
        two_advisor(tree, arr1[]i)   ▷ algorithm 15
    end for
end if
if length_of_arr2[] > 0 then   ▷ In unique name case length of arr2[] = 1 and In multiple name case length of arr2[] > 1
    for i from 0 to length_of_arr2[] -1 do
        unique_name(tree, arr2[i])   ▷ algorithm 13
    end for
end if
```

Algorithm 7: Main Algorithm: Driver function to integrate all subroutines and print the final result

```
Input: Tree,Level and Input author Id
Output: Advisers id up to required level
advisor[] ← advisors id from input node id upto root id   ▷ ancestor search for input author
for i from 0 to level-1 do
    required_advisor[i] ← advisor[i]
end for
return (required_advisor[])
```

Algorithm 8: $advisor_search(Tree, Level, Input_author_id)$ ▷ generate adviser ids up to a specified level

9.4.2 Advisee Id extraction-Algorithm 9:

This algorithm extracts the advisee ids up to specified level from the current node *(it descends the tree and collects author ids of all intermediate nodes up to given level l)*. Basically, to compute Genealogy Citations, it searches an author with its id and find advisees that are one or two hop down in Genealogy Network: line3 to 4. *Since the algorithm finds specified author in genealogy tree before descending the tree for advisees, the worst case time complexity of algorithm comes to* $O(N)$ *where N is total number of authors in genealogy tree.*

1: **Input:** tree having all the authors as nodes,level up to which advisees required and input author id
2: **Output:** advisees id up to given level
3: *required_advisee*[] ← *extract the immediate advisees id from tree upto level L for input author*
4: **return** (*required_advisee*[])

Algorithm 9: *advisee_search*(*Tree*, *Level*, *Input_author_id*) ▷ generate advisee ids up to specified level

9.4.3 Extracting Ids for a node which is either adviser or advisee of author who has two advisers-Algorithm 4,5 and 6:

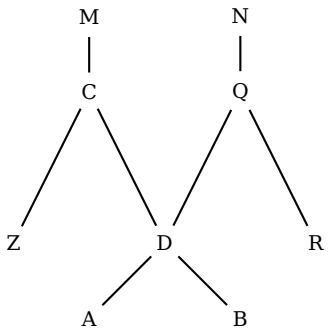


Fig. 20: Original Representation

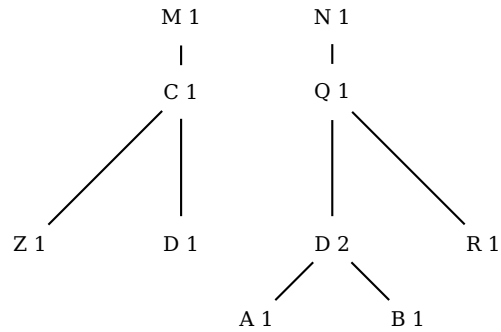


Fig. 21: Equivalent Representation

Two Advisor Case: The *Adviser_search* and *advisee_search* algorithms dig up an author's genealogy network from the whole tree. Fig 20 shows a small section of the tree, that describes the relationship of an individual with its advisers/advisees. But for computation of NGNC and CGR, an equivalent representation (Fig 21) is built. As explained in the previous section, this reformed presentation is utilized in order to deal with the tree property violation (a node in a tree should not have two parents) whereby the alternate presentation bifurcates node into two sub-nodes and assigns weights 1 and 2 to each. All descendants of the forked node are attached to sub node of weight 2. As a result, there may arise some discord while acquiring ids of advisers and advisees of authors who have two advisers. Two conditions may arise while handling such cases. For clarity, we discuss these two cases and propose remedy for that:-

- **First case***(To find NGNC and CGR of a researcher who is one of the advisers of an author who has two advisers)* Looking at the original representation in Figure 13, consider node C as an adviser of D (but D has two advisers, C and Q). Genealogy citations for C require computing citation from adviser M, advisee Z, D, A, B(one or two hop away from C) and from Q (since the distance of Q is two hops from author C). Algorithm 9 fetches M and algorithm 10 picks only Z and D. Nodes A, B and Q which are a part of C's Genealogy Network remains unreachable. This is in effect since equivalent representation (Fig 21) does not have a path connecting C and A, B, Q. Evidently, these nodes are not picked up by *advisee_search* algorithm, this is when the two matrices come to rescue - **Two Advisor Matrix**, **Advisee Matrix**. For the graph shown in Figure 20, these matrices contains,

$$TwoAdvisorMatrix = \left[C's\ ID\ Q's\ ID\right]$$

$$AdviseeMatrix = \left[A's\ ID\ B's\ ID\right]$$

If an author, V, who has two advisers and three advisees, is added to the tree, the advisee matrix will be of dimensions 2×3 (number of rows is 2 as there are two authors D and V who have two advisers

and no of columns are three because max(2,3)=3). The maximum number of advisee is taken to make it a square matrix.
Algorithm 10 and 11 shows the construction of these matrices. Algorithm 12 includes mechanism of extracting the otherwise inaccessible nodes (A, B and Q for C's citation count) by using TwoAdvisorMatrix. Every time such node is extracted for authors with two advisers, the id_collect algorithm checks for author id (C's id) in every row of the matrix. If it is found in a particular row, the id of its corresponding column (which is Q's id) is copied in a separate array. Likewise, the algorithm also captures ids (A and B) of the corresponding row of the advisee matrix to pull out descendants which get isolated due to tree bifurcation. For example, if C's id is found in the first row of TwoAdviserMatrix, then same row in AdviseeMatrix fetches the remaining advisees for author C. With the help of these two matrices along with Collect_Id, adviser_search and advisee_search algorithms, all the ids required for calculation of genealogy citation of C are gathered(id's of M,Z,D,A,B and Q). To retrieve Genealogy Citations for C, citations received from all these authors is summed up by using All Author Matrix. The same matrix is again used to compute C's total citations by summing up citations received from every other author in the tree. From Genealogy Citations and Total Citations, NGNC and CGR are computed.

- **Second case***(When we want to find NGNC and CGR for authors who are one of the advisees of author with two advisers)*-:Referring to the same Figure, author A is the advisee of D and D has two advisors C and Q. Figure 20 indicates genealogy network of A to have authors B, D, C and Q. One can easily extract ids of A's immediate adviser (i.e. D and Q) by algorithm 8 and sibling (i.e B) by algorithm 9. But, equivalent representation does not have a connected path between A and C which incorrectly implies that node C is independent of A's genealogy network. To include C and all such disconnected nodes, the algorithm checks for A's id in advisee matrix and the specific row in which it is found, its corresponding row in TwoAdviserMatrix is retrieved. The first element of the row is the node-id that remained secluded because of graph-restructuring. Let's say, A's id is found in the first row of advisee matrix, so the first element of the first row of TwoAdvisorMatrix gives the required node. Once all the nodes are collected via different algorithms, the computation of GC, Total Citations, NGNC and CGR would be same as explained in section "First Case".
- *Algorithm 10 and 11 describes the procedure of generation of two advisor matrix and advisee matrix. Algorithm 12 generate required ids of authors when any case among those two cases arises.*
- *Complexity of algorithm 10,11 and 12 is $O(n)$ where n is the number of authors in two advisor author array.*

```
Input: Tree
Output: Two advisor matrix
for i from 0 to count-1 do                          ▷ count is the length of two_advisor_author_arr[]
    for j from 0 to 1 do                            ▷ loop run two times because number of advisor is two
        if node.id == two_advisor_author_array[i]  &&  node.weight == j then
            arr[j] ← advisor_search(Tree,1,two_advisor_author_array[i])
        end if
        two_advisor_matrix[i][j] ← arr[j]
    end for
end for
return (two_advisor_matrix[][])
```

Algorithm 10: $two_advisor_matrix(two_advisor_author_arr[])$ ▷ create two advisor matrix

```
Input: Tree
Output: Advisee matrix
for i from 0 to count-1 do                                   ▷ count is the length of two_advisor_author_arr[]
    if node.id == two_advisor_author_array[i]  &&  node.weight == 2 then   ▷ advisee always with the author having
            weight=2
        arr[] ← advisee_search(Tree,1,two_advisor_author_arr[i])
    end if
    advisee_matrix[i][] ← arr[]
end for
return (advisee_matrix[][])
```

Algorithm 11: $advisee_matrix(two_advisor_author_arr[])$ ▷ create advisee matrix

```
Input: Tree,Level and Input Author Id
Output: Authors Id whenever any case among the two case arises
for i from 0 to count-1 do                                   ▷ count equal to length of two_advisor_author_arr[]
    if Input_author_id == two_advisor_matrix[i][0] then
        temp[] ← two_advisor_matrix[i][1]
        total_temp[] ← merge(temp[],advisee_matrix[i][]
    else
        if Input_author_id == two_advisor_matrix[i][1] then
            total_temp[] ← two_advisor_matrix[i][0]
        end if
    end if
end for
for i from 0 to count-1 do                                   ▷ count equal to length of two_advisor_author_arr[]
    for j from 0 to count1-1 do   ▷ count1 equal to maximum among the number of advisee of authors having two
            advisor
        if Input_author_id == advisee_matrix[i][j] then
            total_temp[] ← merge(total_temp[],two_advisor_matrix[i][0]
        end if
    end for
end for
return total_temp[]
```

Algorithm 12: $id_collect(Tree, Input_author_id)$ ▷ create an array having id of authors if they have any association with any case mentioned above

9.4.4 Computation Of NGNC (Non-Genealogy Network Citations) and CGR(citation genealogy ratio) in unique name case-Algorithm 13:

Algorithm 13 computes NGNC and CGR when the author has a unique name in entire genealogy tree. $Input_Id$ variable stores the author id: line 3. Algorithm 14 collects and stores Id of those authors who are in $Input_Id's$ Genealogy Network in $arr1[]$: line 4. Algorithm 12 collects all those node-ids which could not be retrieved because of the disconnected path between related authors in equivalent representation. If those node-ids exist, $arr2[]$ stores them:line 5. $arr1[]$ and $arr2[]$ are merged in a single array $final[]$. $Unique_final[]$ picks up the unique ids from $final[]$:line 6 to 7. All Author Matrix algorithm computes the total genealogy citations and total citations received from all authors(excluding self-citations) for an input author: line 8 to 16. From these values, NGNC and CGR is computed: line 17 to 19.*Complexity of this algorithm is $O(N)$ in the worst case where N is the total number of authors in genealogy tree.*

```
1: Input: tree having all the authors as nodes,level and input author id
2: Output: CGR
3: Input_Id ← Input_author_id
4: arr1[] ← collect_genealogy_network_id(Tree,Input_id)          ▷ Algorithm 14
5: arr2[] ← id_collect(Tree,Input_author_id)                     ▷ Algorithm 12
6: final[] ← merge(arr1[],arr2[])
7: Take unique author Id from final[] into unique_final[]
8: X ← 0
9: for every author id in unique_final[j] do
10:     elememt ← unique_final[j]
11:     X = X + All_Author_Matrix[Input_Id][element];
12: end for
13: Y ← 0
14: for every author id k in the tree do
15:     Y = Y + All_Author_Matrix[Input_Id][k];
16: end for
17: NGC = (Y − X);
18: CGR = X/Y;
19: return (CGR)
```

Algorithm 13: $unique_name(Tree, Input_author_id)$ ▷ compute CGR if author is related to unique name case

9.4.5 Extracting adviser's advisee's and siblings node-ids from author's genealogy network -Algorithm 14:

By running algorithm 8 and 9, advisers and advisees of an author are extracted in $ances[]$ and $child[]$ arrays :line 3 to 5. The next step takes adviser-id from array $ances[]$. The purpose here is to extract author's siblings. So by running advisee_search algorithm on author's adviser, his siblings are pulled out in array $sib[]$:line 6 to 7. All three arrays are finally merged into a single array $final[]$:line8. *Complexity of this algorithm is in worst case O(N) where N is the total number of authors in genealogy tree.*

```
1: Input: tree having all the authors as nodes,level and input author id
2: Output: ids of advisor,advisee upto two hop and siblings
3: if node.id == Input_author_id then
4:     ances[] ← advisor_search(Tree,2,Input_author_id)          ▷ Algorithm 8
5:     child[] ← advisee_search(Tree,2,Input_author_id)          ▷ Algorithm 9
6:     ances1← ances[1]
7:     sib[] ← advisee_search(Tree,1,ances1)                     ▷ Algorithm 9
8:     final[] ← merge(ances[],child[],child1[])
9: end if
10: return (final[])
```

Algorithm 14: $collect_genealogy_network_id(Tree, Input_author_id)$ ▷ generate ids of all authors present in genealogy network of any author

9.4.6 Computation Of NGNC(Non-Genealogy Network citations) and CGR(citation genealogy ratio) in two adviser case-Algorithm 15:

In the case when authors have two advisers, genealogy tree represents such nodes as two sub nodes with the same name, ids but different weights (weight=1, 2). For a node with weight=1,

collect_genealogy_network_tree and id_collect algorithms takes all nodes which are in author's genealogy network and stores in $final1[]$ array. The same step is performed for a node with weight=2: line 4 to 14. With the help of all_author_matrix, the total citations received from genealogy network and from the entire tree (excluding self-citations) is computed for the input author: line 17 to 26. Then, NGNC and CGR is calculated from formulas defined in definition section: line 27 to 29

```
1: Input: Tree,Level and Input Author Id
2: Output: CGR
3: Input_Id ← Input_author_id
4: for i from 0 to 1 do
5:     if Input_author_weight == i then
6:         temp_arr[] ← collect_genealogy_network_id(Tree,Input_author_id)        ▷ algorithm 14
7:         temp_arr1[] ← id_collect(Tree,Input_author_id)                          ▷ algorithm 12
8:         if i == 1 then
9:             final1[] ← merge(temp_arr[],temp_arr1[])
10:        else
11:            final2[] ← merge(temp_arr[],temp_arr1[])
12:        end if
13:    end if
14: end for
15: final[] ← merge(final1[],final2[])
16: Take unique author Id from final[] into unique_final[]
17: q ← Input_author_id
18: X ← 0
19: for every author id in unique_final[j] do
20:     elememt ← unique_final[j]
21:     X = X + all_author_matrix[q][element];
22: end for
23: Y ← 0
24: for every author id k in the tree do
25:     Y = Y + all_author_matrix[q][k];
26: end for
27: NGC = (Y − X);
28: CGR = X/Y;
29: return (CGR)
```

Algorithm 15: $two_advisor(Tree, Input_author_id)$:id_collect() algorithm which produce output only when the input author is advisor or advisee of an author having two advisors in the genealogy tree.

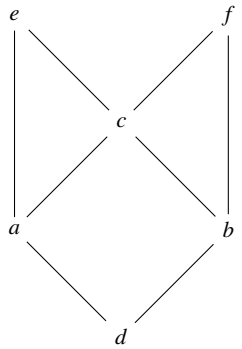


Fig. 22: Example graph structure

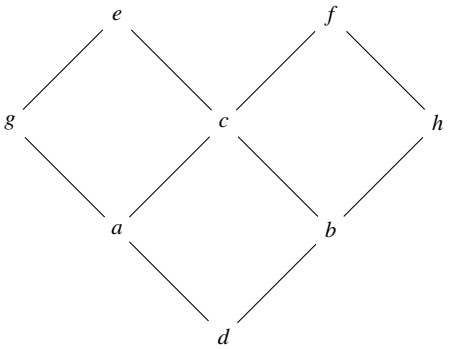


Fig. 23: Complex structure

We see two structures wherein authors may have two advisers in various complex combinations. It is claimed that algorithms proposed in this section to compute CGR, NGNC and GC efficiently deal with all such complex structures.

Case	Input	Algorithm 14 output	Algorithm 12 Output	Final[]	Unique Final[]	X	Y	NGC	CGR
Unique Name	Q(id=17)	1,10,12,13 2,3,4,5,18	4	1,10,12 13,2,3 4,5,18,4	1,10,12 13,2,3 4,5,18	41	81	40	0.5062
Two advisor	M(id=13)	10,17,12,11 18,15,16	4,5	10,17,12 11,18,15 16,4,5	10,17,12 11,18,15 16,4,5	45	98	53	0.4592
Multiple name and Two advisor	Z(id=19)	6,2	None	6,2	6,2	9	90	81	0.1
	Z(id=20)	8,3,9,4,12	3	8,3,9 4,12,3	8,3,9 4,12	19	92	73	0.2065
Multiple name	D(id=4)	1,9,12,20 10 2,3,5 17,18	17,12,13 3	1,9,12,20 10,2,3,5 17 18,17,12 13,3	1,9,12 20 10,2 3,5,17 18,13	53	86	33	0.6163
	D(id=16)	13,11,15	10	13,11,15 10	13,11,15 10	13	93	80	0.1398

Table 11: **Output In Different Case of authors depending on their location in the genealogy tree and the citation genealogy ratio (CGR) of authors**

- Algorithm 14 collect_genealogy_network_id(): Output of this algorithm consist of Ids of authors in genealogy network of an input author.
- Algorithm 12 id_collect(): algorithm which produce output only when the input author is advisor or advisee of an author which has two advisor in genealogy tree.

We present the summary of the exhaustive algorithms presented in this section. Algorithm 8 and 9 extract the adviser and advisee ids up to a specified level from the current node. The output of algorithm 8 and 9 help create two matrices named as advisor matrix and advisee matrix in algorithm 10 and 11 respectively and those two matrices will help in the creation of an array in algorithm 12 (if input author has any association with any case mentioned above). Next, we discriminate our problem on the basis of occurrence of author names(whether one or multiple) and on the number of advisers he may have(one or two), in genealogy tree and compute CGR in the following manner:

- If an author falls in unique-name case then, algorithm 13 is used for the computation of CGR.
- If input author has two advisors, algorithm 15 computes CGR.
- In case, if input author has multiple name occurrences in Genealogy tree then, after collecting the required ids in an array, algorithm 13 computes the value of CGR.
- In case, when an author has multiple names and also has two advisors, both algorithm 13 and 15 computes CGR.

The inherent complexity of the data structure is understood and exploited to compute the greed-aware metric, CGR and NGC and could be used as a penalty in the author internationality score model. The layered algorithmic approach shall help us achieve this complicated task.

10 DISCUSSION

A ranking system on internationality on the basis of internationality score of journals obtained by our model is evolved. The ranking is shown in Tables 17-20 in Appendix D. Impact factor is one predictor variable which is fed to the Stochastic Cobb-Douglas model for score calculation. If we rank journals (present in our data set) on the basis of impact factor only and compare with the ranking system which was generated on the basis of journal's score (Sections 6 and 7), then we can infer following:

- Only 5 journals among 86 have identical rank in both ranking systems. This indicates a strong mismatch between two ranking systems. However, this was expected and we remind the readership that this was one of the motivations behind considering the research problem.
- Further inspection reveals that there are some journals which have earned good rank according to our ranking system based on internationality score but their ranks in impact factor based ranking system is "off the mark". Let us consider an example, **Journal of System and Computer Sciences** is ranked 28 in our ranking system but occupies rank 56 in the impact factor based ranking system. We have found many other journals whose ranking degraded abruptly.
- We have observed reverse trend as well. Rank of some journals got some up-gradation in impact factor based ranking system as compared to the ranking system based on internationality score. **IEEE transactions on network and service management** ranked 11 in internationality score based ranking system but in Impact Factor (I.F.) based ranking system, it is ranked 2.
- Although impact factor plays a significant role in the calculation of internationality score (if we see 11 journals which belong to class HIGH, among those 9 journals are present in top 11 journals according to impact factor based ranking system). It suggests that impact factor alone cannot reflect journal's internationality and we have to consider some other features like NLIQ, SNIP, h-index, IC also as inputs. These features perform an essential role in the calculation of internationality score. However, it does drive home a very significant point. **Impact factor is not orthogonal to internationality as we are made to believe.**

This paper attempts to serve as the quantifier to the philosophical and abstract problem of *"greed-survival deadlock"*. The complex and esoteric nature of the problem demands *"out of the box"* perspective using novel metric definitions. These new metrics coupled with various mathematical models have been proposed to track and quantify the influence and internationality of journals as well as of authors. We like to call the manuscript as *"an expedition in new directions"*. The fundamental reason for this expedition is the exponential increase in the number of journals in publication and their tall claims of internationality, originality, higher visibility and transparency. However, this could be also because of the emerging niche areas of research. The concoction of these two factors has a tremendous impact on the credibility evaluation of new journals. The problem intensity manifolds as evaluation not only provides a basis for authors to assess where to publish their work but also helps the formulation of publication appraisal policy of institutions across the country (Since Institutions insist so much on international journal publications). Also, evaluation serves as a useful guideline for funding agencies and accreditation bodies towards measuring the research output.
It is impossible to cover all aspects, such as coercive citations, extensive self-citations, copious citations etc. with a single baseline model. Hence, we have proposed methodologies that provide multiple scores each covering different quality aspects. In this bid, we have proposed new metrics such as Cognizant citations, Other-Citation-Quotient, genealogy citations, Ancestor inheritance citations and NLIQ. These metrics are defined, quantified and computed from web sources and used in our model so that greed-driven and survival-oriented performance metrics can be measured effectively. As explained in detail in sections 6 and 7, these metrics are defined to quantify penalty accounting for unethical practices.

The complete process internationality computation and classification can be separated into 3 parts. In the first part, we gather all the relevant dataset from various websites through web scraping for a large set of journals and the authors. Next, we process this data, extract metrics and derive the newly defined metrics. Finally, sparse principal component analysis(SPCA) is used to obtain the minimal fea-

ture set and the new feature set is fed to Stochastic Cobb-Douglas Model. Further, in order to estimate the coefficients of Stochastic Cobb-Douglas model, we computed SFA (stochastic frontier analysis) on it. SFA yields technical efficiency, which is defined as the ratio of observed output to maximum feasible output. If this inefficiency is insignificant, then Ordinary Least Squares(OLS) method is employed to estimate the coefficients. Next, Score is computed using the Stochastic Cobb-Douglas model. We have used Stochastic Cobb-Douglas model because this model provides global maximum internationality score, hence the score/values in the neighborhood can be classified as the levels of internationality. We employ two independent computations of Stochastic Cobb-Douglas Production score. Features, processes vary and depend on the input data. The computation method can be broadly categorized as follows.

- Journal input data:
  Our approach: If the analysis is based on the journal input data, then this score represents the measure of Cognizant Citation penalty with respect to a Journal. The new profit score, a penalized output score from Stochastic Cobb-Douglas model can be zero or negative due to heavy cognizant citations. In such case, we consider zero profit score. On the other extreme, if the penalty score is below a certain level we ignore the weight function and proceed with the net score from Stochastic Cobb-Douglas function. Eventually, this net profit score is used for classification of journals using Unified Granular Neural Network (UGNN) methodology.
  Benefits: It is evident that the net profit score, with penalty correction, reflects the effective international prestige of a journal. Hence journals, which belong to diverse fields can be compared based on this new profit score. Evidence of cognizant citations, which reflects the higher negative value of profit score, showcases evidence of journals tactical malpractice to boost its credibility.
  Data structure used: In order to compute citations, we first represent the complete citation information in the form of an adjacency matrix. The matrix represents citation pattern among all the authors of a journal. Next, we analyze equivalence classes present in the matrix of a journal to compute the copious citation coefficient. For cognizant citations, we use bipartite graph representation, where the citations among 2 journals are traced at article level and then we compute the count of mutual citations which amounts to cognizant citation.
  We have developed various algorithms to trace authors who have fabricated their citations in order to increase their total cited works. Apart from self-citations, authors also indulge in copious citations, which is a collaboration among a group of authors to mutually cite each other.
- Author input data:
  Our approach: If the analysis is based on the author level data then the net score represents a measure of copious and genealogy citation penalty. Penalty score is considered only if it is above a specified threshold. The net profit score can then, be used for classification of scholar's/author's. Classification is part of future work.
  Benefits: The author level profit score will aid comparison of the scholars from different domains. Smaller or negative value of the scholar's profit score implies clear unethical practices such as genealogy citations and copious citations for citation boosting. In both these scenarios, we have used UGNN system. Since the profit scores are in the range of 0 and 1, a unit width range, profit scores of journals lie very close to each other. This close proximity makes journal comparison and distinct classification challenging. For this task, we have used the classification model, i.e., developed based on Unified Granular Neural Networks (UGNN) [11].
  Data structure used: In this paper we have proposed genealogy citation model, which is based on the genealogy tree principles. This model is novel in this context since it acts as an input to data structure construction for actual metric computation. Genealogy in our context is a study of a researchers line of descent in the academic world that includes her/his academic advisers, advisee, their ancestors and descendants. Scholarly articles and journals data are nodes representing authors and edges between nodes represent adviser/advisee relationships.
  While computing the journal's internationality score, we analyze the citation behavior of all authors of all the articles of a particular journal, mainly through graph-theoretic representation and algo-

rithmic learning. Author's genealogy tree is used to compute author-level metrics only. This is the prominent differentiator from the data analysis perspective for two mutually exclusive methodologies mentioned above. Author and journal-level data are used to construct authors genealogy tree. It was crucial to transform the text information into a tree structure to extract meaningful information for further computation of metrics. As elaborated in section 9, we have proposed various algorithms, such as, advisee name search, advisee matrix, two adviser matrix, etc. to extract a scholar's genealogy information for computation from an otherwise vast genealogy tree structure. The genealogy tree structure is exploited to generate the adjacency matrix, which is then used to derive various metrics such as Genealogy citations (GC), Non-genealogy citations (NGC) and CGR (citation genealogy ratio). We have explained algorithms to compute these novel metrics, which have potential to filter all the contribution towards unethical citation boosting.
The manuscript is a major contribution in the field of scientometrics encompassing novel mechanisms in data analysis, novel metric definitions, new data structures, methods and interpretation. The natural outcome of such an elaborate exercise are the convincing results which have not been explored so far including the multi-class discrimination of journal internationality doctrine.

11 CONCLUSION AND FUTURE WORK

Defining "Internationality" with regard to Journals and authors has never been done before. There has been no consensual, undisputed definition of internationality till now. Researchers have agreed to call a journal as "international" solely, on the basis of the constituency of Editorial board, ISSN number, publication language etc. and have completely ignored the fact that internationality, in the true sense, measures the extent of diffusion of influence across nation's (and domain's) boundary. In consistent with the definition given in [1], the manuscript quantifies journal's as well as author's influence by building a model that rewards the deserving and penalizes all attempts to build non-genuine influence and reputation by artificial means. The manuscript first explores the already existing scholastic indicators and investigates how easily the system can be manipulated by academicians who use local influence for their personal benefit. Authors of the manuscript, then, propose a quintessential model of "internationality" (Stochastic Cobb-Douglas model) at the journal and author level and endorse the usage of fair indicators like NLIQ, International Collaboration Ratio, SNIP, OCQ to capture the non-local facet of journal's (and author's) influence. Stochastic Frontier Analysis used in combination with Cobb-Douglas Model avoids curvature violation and ensures accurate estimation of elasticity coefficients. Sparse PCA is performed on seven initial parameters to check if the number of input parameters can be reduced and simultaneously generate the largest variance in the data set. After Sparse PCA, the final set of parameters are fed into Stochastic Cobb-Douglas Model for profit score computation of Journals and authors which enforces penalty calculation for spurious practices. At Journal level, Cognizant citations are penalized by inducing 100% penalty in the net profit score. Likewise, authors are liable for penalty if they indulge in copious or genealogy citations.

Computation of cognizant, copious, genealogy or extensive self-citations requires data set with information on journals and their countries, articles published by them, contributing authors with their affiliations and citations. The computation of NLIQ, OCQ, International Collaboration Ratio and SNIP also requires a similar database. This was built through web scraping, a process in which dedicated computer scripts scrape and store the required data from IEEE and ACM website in JSON format. Data scraping, cleaning and the output JSON-sample are shown in Appendix A. Using this data, exclusive algorithms and models are written to compute aforesaid parameters and citations. These algorithms, too, are explained in Appendix A. A journal's country information together with authors' affiliation computes International Collaboration Ratio, and citation database of the journal is used to compute its NLIQ and SNIP. Using approaches explored in section 7.3, the adjacency matrix and bipartite graphs are built to identify cognizant and copious citations for penalty correction in profit scores of journal and author.

The genealogy citations are derived from a different model altogether, explored in section 9, in which internal matrices (*two_advisor*, *advisee*) are put up and algorithms are developed to compute an author's genealogy citations. The Genealogy Citation model takes care of every possible scenario of genealogy network in terms of authors of the same name and/or authors with a different number of advisers.

The Cobb-Douglas Stochastic Frontier Model is used to calculate the internationality score of journals and authors with score values ranging from 0 to 1. The econometric model uses Stochastic Frontier Analysis approach to measure the level of technical inefficiency in the system. The model converges to attain global maxima indicating a maximum internationality score for journals (or authors). The model and the selection of input parameters (NLIQ, SNIP, ICQ, h-index and IF) ensure a fair policy by imposing a penalty to eliminate all unexplained citations originating from a local peer group. The manuscript is geared to breed an impartial policy while recruiting new faculty or evaluating/extending/promoting tenure to the existing ones.
Some authors may succumb to the pressures of the *publish or perish* doctrine. Our framework is to sympathize with those who did and carve out a fair policy that benefits the deserving. Some may argue that since their work is restricted to niche areas, it may be unavoidable not to cite within a particular peer group. Having stated that, we humbly present the following argument: even if that's the case, the reward model for influence/internationality/impact is nonlinear and the penalty for local citations is linear. Therefore, the net score may not penalize the authors working in such niche areas to a great extent but will do so to authors whose subject areas appeal to a much broader audience offering no reason to justify a bulk of their citations originating from local peer groups (copious/genealogy citations etc). The threshold for the penalty is 50% which is hardly stringent. For such cases not justifiable being niche, penalty needs to be imposed so that a fair framework across different domains is created and authors do not feel short-changed for contributing to niche areas as compared to the ones who don't! This breeds an impartial policy while recruiting new faculty or evaluating/extending/promoting tenure to the existing ones. Greed-aware mathematical models proposed here are interpreted in that spirit.

We combined metrics for measuring internationality of authors and journals since they complement each other and reputation of one may help build the reputation of the other. Additionally, the model is similar and there is no reason to articulate identical methods in two different manuscripts. Journal or author internationality is a measure of prestige. The manuscript defines internationality as a metric to evaluate journals/authors based on non-local diffusion and impact of publications. Our work bridges the gap that the discourse of internationality suffers from i.e. a clear definition accompanied by benchmarks and baseline. It is perplexing that more than two decades of discussion could not bring out plausible parameters other than the editorial board composition, ISSN # etc. The fact that these are not quantitative or qualitative measures are tested on a set of 500 journals satisfying these criterion. It is no surprise to the authors that all of these journals could barely make it to the lowest rung of internationality class (high, medium and low are the three classes defined in the manuscript). In fact, the entire discourse on internationality, according to the authors, is a measure of the degree or extent of it. The discourse is much beyond if a journal is international or not!

Some degree of convergence with I.F. establishes our supposition that the journal internationality is indeed a measure of impact and prestige, computed in a different manner from I.F. Moreover, internationality is beyond just I.F. Therefore, ranking and classifying journals and authors in future is a meaningful exercise. The motivation for creating author internationality evaluation framework is to match the results of our author internationality class scheme with ISI-Thompson Reuters' list of highly cited researchers in future. We will propose a composite ranking of journals based on I.F. and internationality in future work. This will be a natural sequel to the existing work.

References

1. Gouri, G., Saha, S., Mathur, A., Venkatagiri, S., Vadakkepat, S., Narasimhamurthy, A., Daya Sagar, B.S. (June 2016). ScientoBASE: A Framework and Model for Computing Scholastic Indicators of non local influence of Journals via Native Data Acquisition algorithms. Journal Of Scientometrics. 1-51. doi:10.1007/s11192-016-2006-2 http://link.springer.com/article/10.1007/s11192-016-2006-2
2. Ginde, G., Saha, S., Balasubramaniam, Chitra., R.S, Harsha., Mathur, A., Daya Sagar, B. S., Narsimhamurthy, A. (August 2015). Mining massive databases for computation of scholastic indices - Model and Quantify internationality and influence diffusion of peer-reviewed journals. Proceedings of the Fourth National Conference of Institute of Scientometrics, SIoT.
3. Ginde, G. (2016). Visualisation of massive data from scholarly Article and Journal Database A Novel Scheme. CoRR abs/1611.01152
4. Bora K., Saha S., Agrawal S., Safonova M., Routh S., Narasimhamurthy A. M. (2016). CD-HPF: New Habitability Score Via Data Analytic Modeling. Journal of Astronomy and Computing. Preprint arxiv:1604.01722v1
5. Webpage: https://journalmetrics.scopus.com/; accessed on 04/08/2017.
6. Webpage: http://www.scimagojr.com/journalrank.php/: accessed on 25/08/2017.
7. Zou, Hui and Hastie, Trevor and Tibshirani, Robert, ”Sparse principal component analysis”*Journal of computational and graphical statistics*,vol. 15,no. 2,pp. 265–286, 2006
8. Tibshirani, Robert, "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society. Series B (Methodological), 267-288, (1996)
9. Zou, Hui, and Trevor Hastie, "Regularization and variable selection via the elastic net." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67.2, 301-320, 2005.
10. Hastie, Tibshirani and Wainwright, Statistical Learning with Sparsity: The Lasso and Generalizations, Monographs on Statistics and Applied Probability 143, CRC press
11. D. Arun Kumar, Saroj K. Meher, Debananda Kanhar, K. Padma Kumari, ”Unified granular neural networks for pattern classification”*Neurocomputing*,vol. 216,pp. 109–125,2016
12. W. Pedrycz, A. Skowron, V. Kreinovich, Handbook of Granular Computing, Wiley-Interscience, 2008
13. N. K. Kasabov, Learning fuzzy rules and approximate reasoning in fuzzy neural networks and hybrid systems, Fuzzy Sets Syst., vol. 82, pp. 135149, 1996.
14. S. K. Pal, S. K. Meher, and S. Dutta, Class-dependent rough-fuzzy granular space, dispersion index and classification, Pattern Recogn., vol. 45, pp. 26902707, 2012
15. Henningsen, Arne and Cadavez, Vasco and Rodrigues, Orlando, “Introduction to Econometric Production Analysis with R” *Leanpub*,2015
16. Battese, G. E. and Corra, G. S., Estimation Of a Production Frontier Model: With Application To The Pastoral Zone Of Eastern Australia. Australian Journal of Agricultural Economics, 21: 169179, 1977: doi:10.1111/j.1467-8489.1977.tb00204.x
17. Coelli, Timothy J and Rao, Dodla Sai Prasada and O’Donnell, Christopher J and Battese, George Edward, “An introduction to efficiency and productivity analysis” *Springer Science & Business Media*, 2005.
18. L Waltman, NJ van Eck, "The inconsistency of the h-index", Journal of the American Society for Information Science and Technology
19. Nees Jan van Eck, Ludo Waltman, " Generalizing the h and g indices", Journal of Informetrics
20. G. Buchandiran, *An Exploratory Study of Indian Science and Technology Publication Output*, Department of Library and Information Science, Loyola Institute of Technology Chennai http://www.webpages.uidaho.edu/ mbolin/buchandiran.htm.
21. Lutz Bornmann http://arxiv.org/pdf/1407.2037.pdf
22. Bhattacharjee, Y. (2011). Saudi universities offer cash in exchange for academic prestige. Science, 334(6061), 1344-1345. doi:10.1126/science.334.6061.1344
23. Gingras, Y. (2014a). *The abuses of research evaluation, University World News.* Retrieved from http://www.universityworldnews.com/article.php?story=20140204141307557
24. *How to boost your university up the rankings, University World News. Retrieved* from http://www.universityworldnews.com/article.php?story=20140715142345754
25. Henk F. Moed *Measuring contextual citation impact of scientific journals,* Journal of Informetrics, Volume 4, Issue 3, July 2010
26. Ludo Waltman, Nees Jan van Eck, Thed N. van Leeuwen, Martijn S. Visser, *Some modifications to the SNIP journal impact indicator*, Journal of Informetrics 7 (2013) 272 285
27. https://aminer.org/billboard/citation *As accessed on 21/1/2016.*
28. Hirsch, J. E. *An index to quantify an individual’s scientific research output*, Proceedings of the National Academy of Sciences (2005) 16569–16572
29. http://sahascibase.org/
30. Jeffrey Beall *Predatory publishers are corrupting open access* Nature, Volume 489, 179–179, 2012, doi:10.1038/489179a

31. Kohavi, Ron *A study of cross-validation and bootstrap for accuracy estimation and model selection*, Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (1995) San Mateo, CA: Morgan Kaufmann. 2 (12), pages: 1137-1143.
32. Aigner, Lovell, Schmidt *Formulation And Estimation Of Stochastic Frontier Production Function Models*, Journal of Econometrics(1977), pages:21-37.
33. David L. Bailey, Donna Thompson *Developing neural-network applications*, AI Expert(1990), Pages: 34-41.

APPENDIX A

Data Scraping for ACM journals

Data accumulation and pre-processing are tenacious but crucial parts of our research. We required corpus of a huge and reliable data source to work with. Data of scientific journals and articles may be acquired from websites like SciMago. We preferred to write scripts which retrieves article and journal data by mining authentic websites. Downloading data from 3rd party websites may not guarantee data consistency or completeness. Even though it's an easier option, it is a better practice to fetch the data from the publisher websites by writing scripts. This process is known as Web-scraping and took us approximately 7 months! We wrote a web scraping script in Python to fetch information regarding scientific journals and articles directly from their digital library. This decision, even though provided the surety of accurate and complete data, provided several other hurdles along the way. A web-scraping script works by getting the page source code of a web page which for an average website contains the data presented by the browser. In such a case, obtaining required data from a page becomes the task of merely identifying the path of HTML tags (code components). This process lets the script crawl through data. Publishers make this increasingly complicated every day, if not horrifyingly difficult, as it was discovered that the data was being dynamically fetched when the initial web-page had loaded. Whenever a web crawling script requests for a web page, the primitive code is fetched and not the completely processed ones generated by a browser. Automating this task was excruciatingly painful but rewarding. After successfully coding a script to scrape the website, we decided to let it run from a dedicated server to scrape the entire website. However, most sites possess IP load balancers with a stringent request limit. Only after a couple of issues of a journal were scraped, the IP was blocked for 12 hours. The task of scraping one journal consumed a significant amount of time. Eight independent nodes had to be deployed to run the script simultaneously in order to complete the task. Another issue discovered while scraping was there were many inconsistencies and variations in the way the data was presented by the website. This forced us to update the script to handle the anomalies on a regular basis. The task was completed and information for all the journals was obtained in top-down fashion. The script started from the highest level, i.e. the journal and moved to the next level, discovering all the volumes and issues. Finally, all required data of all issues and volumes of the journals were obtained. Algorithm 16 shows the scraping of the journals and Fig 24. shows the output in the JSON structure.

Data cleansing and pre-processing algorithms

Once the data is gathered via scraping, algorithms are written to pre-process it for further usage. Here is an example shown to find similarity between two strings using Cosine similarity technique. The

```
procedure Scrape(journal, result)                                   ▷
    volumes ← journal[volumes]
    for v + olume ∈ volumes do
        issues ← volumes[issues]
        for issue ∈ issues do
            articles ← issues[articles]
            for article ∈ articles do
                result[volume][issue][article][title] ← article[title]
                result[volume][issue][article][authors] ← article[authors]
                result[volume][issue][article][citations] ← article[citations]
            end for
        end for
    end for
    return result
end procedure
```

Algorithm 16: Scrape article data

Journal Name	**Number of Articles**	**Number of Authors**	**Publication Years**	**Size of Stored Data**
ACM Computing Surveys (CSUR)	1554	323	1969-2016	29.7 MB
Journal of the ACM (JACM)	2798	907	1954-1026	26.9 MB
Journal of Data and Information Quality (JDIQ)	107	15	2009-2016	876 KB
Journal of Experimental Algorithms (JEA)	283	90	1996-2016	1.7 MB
ACM Journal of Emerging Technologies in Computing Systems (JETC)	307	114	2005-2016	2.8 MB
Journal on Computing and Cultural Heritage (JOCCH)	151	22	2008-2016	1.4 MB
ACM Transactions on Autonomous and Adaptive Systems (TAAS)	249	74	2006-2016	2.5 MB
ACM Transactions on Accessible Computing (TACCESS)	115	30	2008-2016	1.2 MB
ACM Transactions on Architecture and Code Optimization (TACO)	480	239	2004-2016	5.5 MB
ACM Transactions on Algorithms (TALG)	538	257	2005-2016	3.8 MB
ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)	290	122	2002-2016	2.6 MB
ACM Transactions on Applied Perception (TAP)	332	145	2004-2016	2.8 MB
ACM Transactions on Economics and Computation (TEAC)	95	40	2013-2016	588 KB
ACM Transactions on Embedded Computing Systems (TECS)	896	239	2002-2016	8.3 MB
ACM Transactions on Interactive Intelligent Systems (TiiS)	132	33	2011-2016	1.5 MB
ACM Transactions on Information and System Security (TISSEC)	333	129	1998-2016	4.5 MB
ACM Transactions on Intelligent Systems and Technology (TIST)	419	145	2010-2016	05 MB
ACM Transactions on Knowledge Discovery from Data (TKDD)	258	94	2017-2016	2.9 MB
ACM Transactions on Management Information Systems (TMIS)	122	26	2010-2016	1.4 MB
ACM Transactions on Computing Education (TOCE)	292	69	2001-2016	2.4 MB
ACM Transactions on Computer-Human Interaction (TOCHI)	470	183	1994-2016	7.4 MB
ACM Transactions on Computer Systems (TOCS)	444	192	1983-2016	8.6 MB
ACM Transactions on Computation Theory (TOCT)	104	44	2009-2016	784 KB
ACM Transactions on Design Automation of Electronic Systems (TODAES)	874	413	1996-2016	7.4 MB
ACM Transactions on Database Systems (TODS)	926	399	1976-2016	12.7 MB
ACM Transactions on Graphics (TOG)	2759	1573	1982-2016	31.3 MB
ACM Transactions on Information Systems (TOIS)	645	228	1983-2016	10.0 MB
ACM Transactions on Internet Technology (TOIT)	275	75	2001-2016	3.2 MB
ACM Transactions on Modeling and Computer Simulation (TOMACS)	497	138	2991-2016	4.3 MB
ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)	477	171	2005-2016	4.3 MB
IEEE/ACM Transactions on Networking (TON)	2555	1235	1993-2016	22.8 MB
ACM Transactions on Parallel Computing (TOPC)	58	183	2014-2016	383 KB
ACM Transactions on Programming Languages and Systems (TOPLAS)	1031	400	1979-2016	13.5 MB
ACM Transactions on Storage (TOS)	204	100	2005-2016	2.2 MB
ACM Transactions on Software Engineering and Methodology (TOSEM)	431	159	1992-2016	6.4 MB
ACM Transactions on Sensor Networks (TOSN)	446	217	2005-2016	5.1 MB
ACM Transactions on Spatial Algorithms and Systems (TSAS)	20	65	2015-2016	139 KB
ACM Transactions on the Web (TWEB)	194	45	2007-2016	2.8 MB

Table 12: **Summary of Scraped ACM Data**

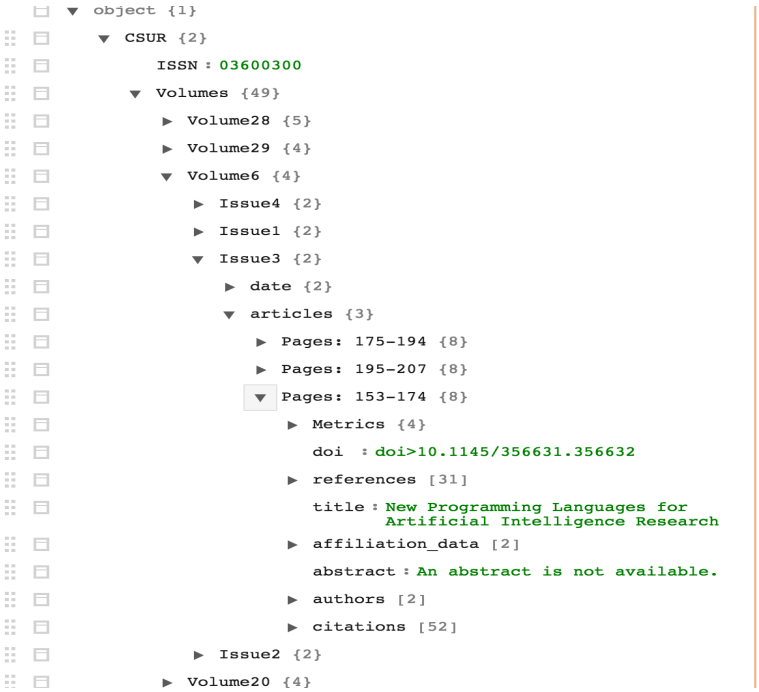


Fig. 24: JSON Structure obtained after data processing

example finds the relative similarity between the citing and cited journals to see how contextual the citations are. The sample output is shown with the cosine similarity values.

Cosine Similarity Metric:

Similarity metrics quantify similarity or dissimilarity (distance) between two text strings. For example, similarity between the strings orange and range can be considered to be much more than the strings apple and orange. Cosine similarity is a vector based similarity measure. Cosine of two vectors a, b can be derived by using the inner product formula.

$$a.b = |a||b|cos\theta$$

Where, θ represents the angle between a and b.

Sample output

Journal name of Article: Plant Molecular Biology
Journal name of Cited Article: Plant Science 247, 1-12
cosine similarity value: 0.258198889747

Journal name of Article: Theoretical and Applied Genetics
Journal name of Cited Article: Theoretical and Applied Genetics 129 (3), 469-484
cosine similarity value: 0.707106781187

Journal name of Article: Agricultural and Forest Meteorology
Journal name of Cited Article: American Society of Agricultural and Biological Engineers 59 (2), 555-560
cosine similarity value: 0.301511344578

Input: Scraped data repository from ACM
Output: Features such as International Collaboration Ratio, SNIP, Other-Citations and Internationality Index
$JNames[]$ = $Fetch_Journal_Names_from_Scraped_Repository(ACM)$
for every journal: $JNames[i]$ **do**
    TotalCites = Get the totalcites value
    Get all the published articles/papers: $X[]$
    for every article: $X[i]$ **do**
        $JNames[i]$.Selfcites+=compute_SelfCitations($X[i]$)
    end for
    $x_1 = 1 - JNames[i].Selfcites/TotalCites$; **compute other-citation-quotient**
    $x_2 = compute_Intl_Collaboration_Ratio(JNames[i])/100$; **compute International Collaboration Ratio**
    $x_3 = compute_SNIP(JNames[i])/MaxSNIP$; **compute SNIP**
    $x_4 = compute_NonLocalIQ(JNames[i])$; **compute NLIQ**
    $Internationlity_index = StochasticCobbDouglasModel(JNames[i], x_1, x_2, x_3, x_4)$;**compute JIMI** ▷ refer section 6.2 for Stochastic Cobb-Douglas Model
end for

Algorithm 17: Driver Algo [1]: Algorithm to extract various features and to compute Internationality Index of Journals: collect_genealogy_network_id().Output of this algorithm consist of Ids of authors in genealogy network of an input author.

Computation Of Derived Parameters

This section gives a brief explanation of various algorithms written and utilized to extract features such as NLIQ, other-citation quotient, International Collaboration Ratio and SNIP of every journal from the repository. These algorithms are part and parcel of author's previous work on Internationality index computation [1]. Algorithm 17 is the main algorithm that internally calls separate functions to compute the essential features of journals. Algorithm 18 computes self-citations (which is later used in computation of Other-Citation-Quotient) and algorithms 19, 20 and 21 compute NLIQ, International Collaboration Ratio and SNIP respectively. Finally, Stochastic Cobb-Douglas model is used to calculate the internationality index of every journal.

Input: article/paper name (P) from Google Scholar
Output: self-citation count for article / paper (P)
Get all citedPapers for article/paper(P): $citedBy[]$
for Every cited paper: $citedBy[i]$ **do**
    if $P.author_name$ **IN** $citedBy[i].author_names$ **then**
        $IncrByOne(P.SelfCitationCount)$
    end if
end for
return $SelfCitationCount$

Algorithm 18: $compute_Self_Citations()$ [1]: Algorithm to compute Self-Citation Count

[0] The algorithm 17 is used from our previous work [1]

```
Input: journal_name, citation_database
Output: NLIQ of journal_name
A ← 0                                          ▷ external citation count
B ← 0                                          ▷ internal citation count
J_articles ← [ ]                               ▷ used to store articles of journal_name
O_articles ← [ ]                               ▷ used to store articles of other journal
count ← 0
count1 ← 0
for each article ∈ citation_database do
    if article[journal] = journal_name then    ▷ get all articles in journal_name
        J_articles[count++] ← article
    else if article[journal] != journal_name then   ▷ get all articles of other journals
        O_articles[count++] ← article
    end if
end for
for each article ∈ O_articles do               ▷ get count of citation from outside Journal_name
    for each reference ∈ article[references] do
        if reference ∈ ARTICLE_TYPE then       ▷ reference is an article
            if reference[journal] == journal_name then
                A ← A + 1
            end if
        end if
    end for
end for
for each article ∈ J_articles do               ▷ get count of citation from outside Journal_name
    for each reference ∈ article[references] do
        if reference ∈ ARTICLE_TYPE then       ▷ reference is an article
            if reference[journal] == journal_name then
                B ← B + 1
            end if
        end if
    end for
end for
NLIQ ← A / (A + B)
return NLIQ
```

Algorithm 19: *compute_NonLocalIQ*() [1]: Algorithm to calculate Non-Local Influence Quotient

Sample Output for computation of International Collaboration Ratio [1]: Algorithm 20

Input: Author name and Country name from affiliation information
Intermediate Sets: [u'lei yang', u'pedro v sander', u'jason lawrence'] [u'Hong Kong', u'Honduras']
Mod of country set: 2 Mod of author set: 3
IC = 0.666666666667
Intermediate Sets: [u'jaewon kim', u'roarke horstmeyer'] [u'New Zealand']
Mod of country set: 1 Mod of author set: 2
IC = 0
Intermediate Sets:[u'jing dong', u'reza curtmola', u'cristina nita-rotaru']
[u'Mali', u'United States', u'Iceland']
Mod of country set: 3 Mod of author set: 3
IC = 1.0

Algorithm 20: INTERNATIONAL COLLABORATION RATIO [1] An authors affiliation of an article is fetched using algorithm 21. In order to extract the country and city names from affiliation, we have scraped the article URL and extracted author names and their complete information. In case there are multiple affiliations of an author, the algorithm extracts all the affiliation information. This data is utilized in Algorithm 20 to compute International Collaboration Ratio.

[0] Algorithms 18 and 19 are from our previous work [1]

```
Input: Journal Name: J
URL to all the articles in that Journal : J.all_articles_url[]
Country information of the Journal: J.contryName
Output: %international collaboration ratio of Journal: J        ▷ Compute the internationality weight of an article
        Based on the combination (Eg: out of 5 authors 2 are from same rest from other) deduce the weight of
        the article from a predefined values for a given combination, Eg: mod(Set of different countries) =
        3+1=4, mod(Set of all the authors) = 5 wt(article)=4/5, if all the authors belong to the same
        country i.e. mod(Set of different countries)=1 then wt(article)=0
authAffs = []
for Every article in J.all_articles_url[i] do
    Authors_Affiliation ← Fetch_Author_Affiliations(article)                                  ▷ Algorithm 21
    authAffs.append(read_auth_name&_1st_affiliation
        info(Author_Affiliation))                  ▷ Generate 2D array of i:author name, j:country name
    iNtrNationality_wt[i] = compute_wt(article)
end for
J.iNtrNational[][]                  ▷ Create one big matrix for a journal where i:country names, j:author names
for every i in authAffs do
    if Country_of(i['Affiliation']) == J.countryName) then    ▷ if author's country same as Journal's country then make
        entry = 0
        J.iNtrNational[Country_of(i['Affiliation'])][i['Author']] = 0
    else
        J.iNtrNational[Country_of(i['Affiliation'])][i['Author']] = 1
    end if
end for
x = Ratio of(Number of 1's and Size of matrix J.iNtrNational[][])
y = cumulative weights(iNtrNationality_wt[i])
return (%international_collaboration = αx + (1 − α)y)          ▷ α is a weight deduced from cross correlation
```

Algorithm 20: $Intl_Collaboration_Ratio(JNames[i])$ [1]: Algorithm to compute international collaboration ratio of a Journal

```
Input: Link to the article from algorithm 5: article_URL
Output: Author names and respective Affiliations
authors[] ⟵ scraped_author_names(article_URL)
list = []                                                                  ▷ list of dictionaries
for every author in authors[] do
    dictionary_element = {'Author' : author}
    count = 0
    affiliations = []
    for every affiliation of author do
        count = count + 1                                   ▷ First, Second, Third Affiliations
        affiliations.append(affiliation)
    end for
    dictionary_element.update{'affiliations' :affiliations}
    dictionary_element.update{'count' :count}
    list.append(dictionary_element)
end for
return list
```

Algorithm 21: $Fetch_Author_Affiliations(article)$ [1]: Algorithm to fetch author affiliations information for the article

Algorithm 22: SOURCE NORMALIZED IMPACT PER PAPER [1] This section computes SNIP value of a Journal in a particular year.

[0] Algorithms 20, 21 and 22 are from our previous work [1]

```
Input: Database of cites (cites[][]) made to publications of journal J (Jpub[] with Jsize publications) in year X to all
       documents (article, conference paper or review) in the three years preceding X
Output: SNIP value for journal J in year X
journal ← Jname
year ← read year_to_be_computed_for
citation_count ← 0
for all paper in Jpub do
    citation_count+ = num papers published last 3 years
    num_papers ← num_papers + 1
end for
RIP ← citation_count / num_papers
DCP ← Average number of 1-3 year old cited references contained in papers in the dataset citing the target journal
median ← median DCP of all journals
RDCP ← DCP / median
SNIP ← RIP / RDCP
return SNIP
```

Algorithm 22: $compute_SNIP(cites[][], Jpub[], Jsize)$: Algorithm to calculate SNIP

Algorithm 23: JOURNAL EFFUSION INDEX This algorithm computes the number of outgoing citations from a journal. Journal Effusion Index is a ratio of citations delineated to various journals and the total citations. It can be used as a measure of journal's integrity and reflects the fairness with which it publishes and promotes research.

```
Input: journal_name, citation_database
Output: EffusionIndex of journal_name
A ← 0                                                    ▷ external citation count
B ← 0                                                    ▷ internal citation count
J_articles ← [ ]                                         ▷ used to store articles in a journal
count ← 0
for each article ∈ citation_database do                  ▷ get all articles in a journal
    if article[journal] = journal_name then
        J_articles[count++] ← article
    end if
end for
for each article ∈ J_articles do                         ▷ get count of internal, external cites
    for each reference ∈ article[references] do
        if reference ∈ ARTICLE_TYPE then                 ▷ reference is an article
            if reference[journal] != journal_name then
                A ← A + 1
            else
                B ← B + 1
            end if
        end if
    end for
end for
EffusionIndex ← A / (A + B)
return NLIQ
```

Algorithm 23: $compute_EffusionIndex()$: Algorithm to calculate Journal Effusion Index

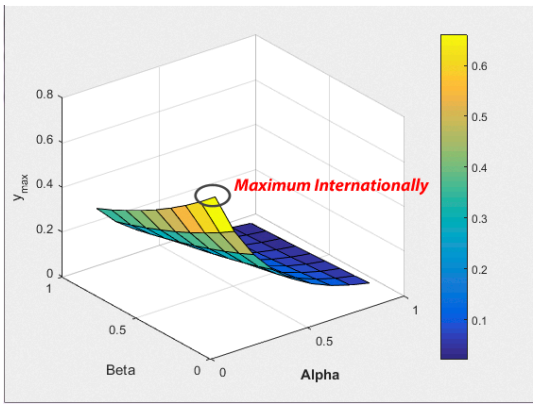


Fig. 25: Journal's Maximum Internationality for optimum values of α and β

Table 13: Journals and the computed values

Journal name	**NLIQ Algo:18**	**ICR Algo:19**	**OCQ Algo:17**	**SNIP Algo:21**
ACM Computing Surveys	0.98	0.15	0.98	1.71
Journal of the ACM	0.96	0.23	0.97	0.78
Journal of Data and Information Quality	0.83	0.14	0.89	0.17
Journal of Experimental Algorithmics	0.90	0.18	0.83	0.20
ACM Journal on Emerging Technologies in Computing Systems	0.80	0.19	0.77	0.19
Journal on Computing and Cultural Heritage	0.78	0.16	0.80	0.22
ACM Transactions on Autonomous and Adaptive Systems	0.93	0.19	0.82	0.37
ACM Transactions on Accessible Computing	0.85	0.14	0.87	0.41
ACM Transactions on Architecture and Code Optimization	0.90	0.17	0.87	0.56
ACM Transactions on Algorithms	0.95	0.28	0.82	0.50
ACM Transactions on Asian and Low-Resource Language Information Processing	1.00	0.14	0.83	0.15
ACM Transactions on Applied Perception	0.90	0.17	0.88	0.24
ACM Transactions on Economics and Computation	1.00	0.17	0.64	0.45
ACM Transactions on Embedded Computing Systems	0.93	0.18	0.87	0.33
ACM Transactions on Interactive Intelligent Systems	1.00	0.25	0.73	0.49
ACM Transactions on Information and System Security	0.96	0.17	0.94	0.45
ACM Transactions on Intelligent Systems and Technology	0.97	0.19	0.88	0.76
ACM Transactions on Knowledge Discovery from Data	0.97	0.17	0.90	0.43
ACM Transactions on Management Information Systems	0.64	0.20	0.93	0.33
ACM Transactions on Computing Education	0.90	0.13	0.91	1.02
ACM Transactions on Computer-Human Interaction	0.96	0.17	0.93	1.37
ACM Transactions on Computer Systems	0.97	0.22	0.97	1.00
ACM Transactions on Computation Theory	0.93	0.13	0.76	0.21
ACM Transactions on Design Automation of Electronic Systems	0.90	0.18	0.87	0.33
ACM Transactions on Database Systems	0.93	0.21	0.95	0.39
ACM Transactions on Graphics	0.73	0.12	0.92	2.09
ACM Transactions on Information Systems	0.96	0.16	0.95	0.65
ACM Transactions on Internet Technology	0.98	0.19	0.92	0.31
ACM Transactions on Modelling and Computer Simulations	0.90	0.17	0.91	0.29
ACM Transactions on Mathematical Software	0.96	0.18	0.88	0.52
IEEE/ACM Transactions on Networking	0.90	0.21	0.96	0.42
ACM Transactions on Storage	1.00	0.12	0.73	0.14
ACM Transactions on Programming Languages and Systems	0.95	0.17	0.95	0.65
ACM Transactions on Storage	0.89	0.14	0.92	0.53
ACM Transactions on Software Engineering and Methodology	0.96	0.20	0.80	0.91
ACM Transactions on Sensor Networks	0.90	0.20	0.93	0.48
ACM Transactions on Spatial Algorithms and Systems	1.00	0.25	0.00	0.05
ACM Transactions on the Web	0.97	0.18	0.89	0.50

APPENDIX B

SciBase site is developed in concurrence with the scientific investigation reported in our work. The site contains the information scraped (in an algorithmic manner) from different sources on the web. Please visit the site at [29]

Graph Database and Data Model

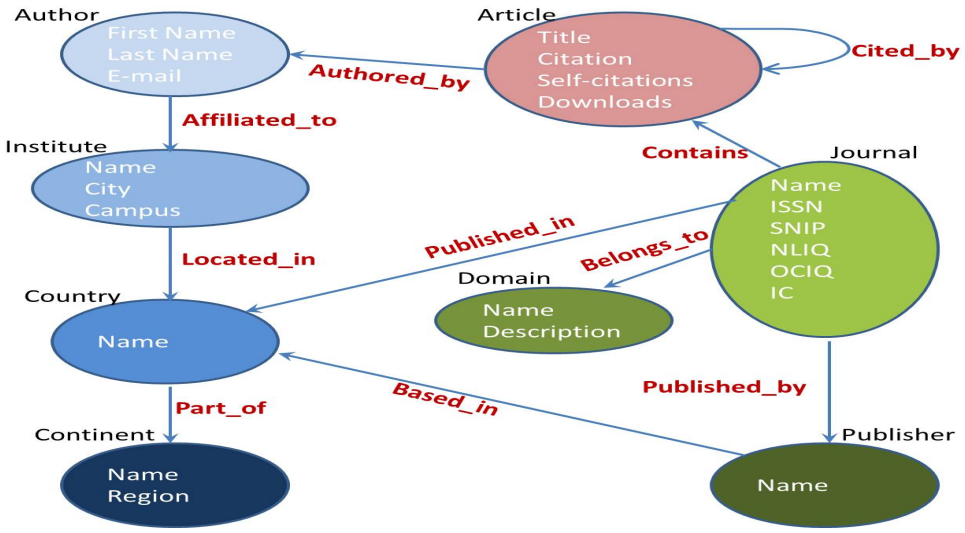


Fig. 26: Graph database data model

Data accumulated is rich, very well connected and has a lot of hidden information within it. We chose to visualize this data using graph database to extract the patterns within this massive data. A graph database is a graph-oriented database, which is a type of NoSQL database that uses graph theory to store, map and query relationships. It is basically a collection of nodes and edges.

The data model is the outcome of a thorough research on the data repository. Data modeling is based on selecting appropriate features as nodes, properties and relationships. Fig. 26 is the data model where oval shapes represent Nodes. Namely, Author, Article, Journal, Publisher, Continent, Country and Institute. The features described within the oval shapes represent the properties of the respective nodes. For example, Author has First Name, Last Name and E-Mail as the properties. One or more properties can be configured as unique properties to identify and index the nodes at the time of node creation in the database. The arrows connecting two nodes represent the relationships. For example, "Affiliated to" is an example of a relationship between Author and Institute nodes. These relationships will help in forming the cypher queries which represent the question we want to ask to the database.

Cypher is a declarative graph query language that allows for expressive and efficient querying and updating of the graph store. Cypher is a relatively simple but still very powerful language. Very complicated database queries can easily be expressed through Cypher. Cypher query's complexity can increase with the increase in the number of relationships involved in the question. The question can be as simple as "How many authors, who belong to a particular Institute, have published articles in a Journal XYZ?" Or "List of all the countries who have maximum publications to a Journal?" to as complex as "List of Journals who have least international collaboration" etc. Effective Data analysis and visualization is solely based on the formulation of these complex queries. Using such cipher queries on the database following visualizations can be generated

- Author network

- Institute network
- Country network
- Spread of a domain in a country
- Collaboration network of an institute
- Extract year-wise publication trend of an author

Data Visualization using built-in D3.js library Fig 27 shows how a particular Journal will map to Journal to Author and in turn to the Country to which his/her Institution belongs to. The blue color node represents Journal displaying the Journal name, the yellow color node represents Article with the article title, the purple color node is Author node with Author name, the red color node is Country with Country name.

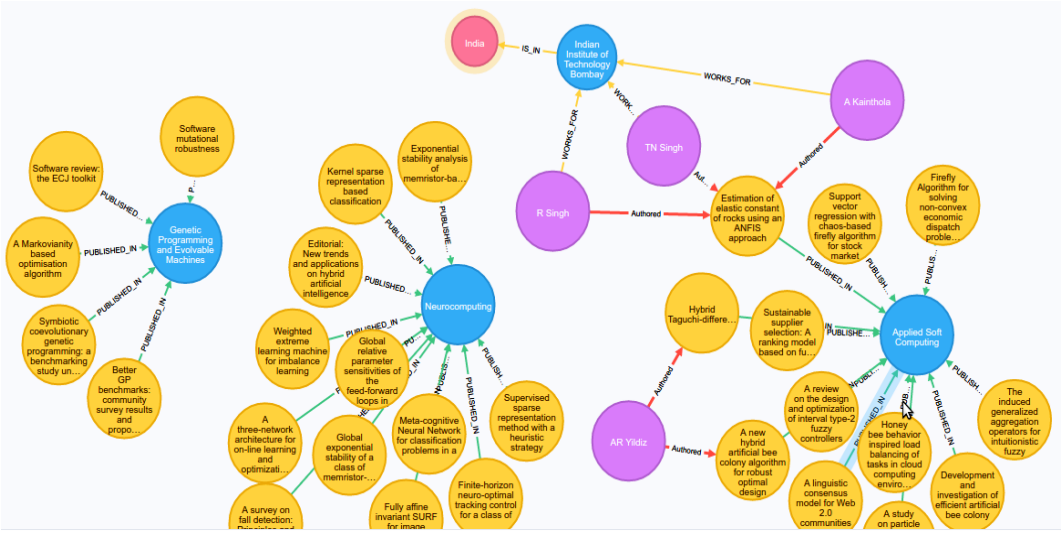


Fig. 27: Journal to Author to Country mapping

Fig 28 shows all the Author's contribution to an article through "Authored" relationship. This relationship can be used in query formulation to visualize year-wise contributions made by authors to a Journal.

Fig 29 is the data for all of the institutes that belong to a country, and countries belong to a Region. The relationships "located_in" and "Part_of" will be used in a query to extract data which can be further used to visualize various countries of a region that are contributing to the growth of a technical domain.

Fig 30 shows the mapping of an Author to an Institute to a Country. This will help in identifying the contribution trends pertaining to a particular Institute and Country in particular.

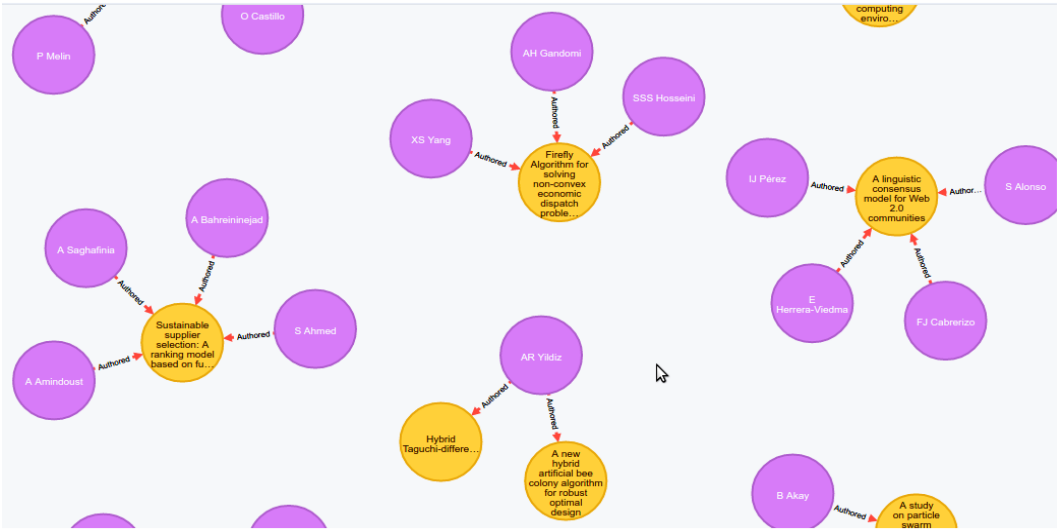


Fig. 28: Articles to Author mapping

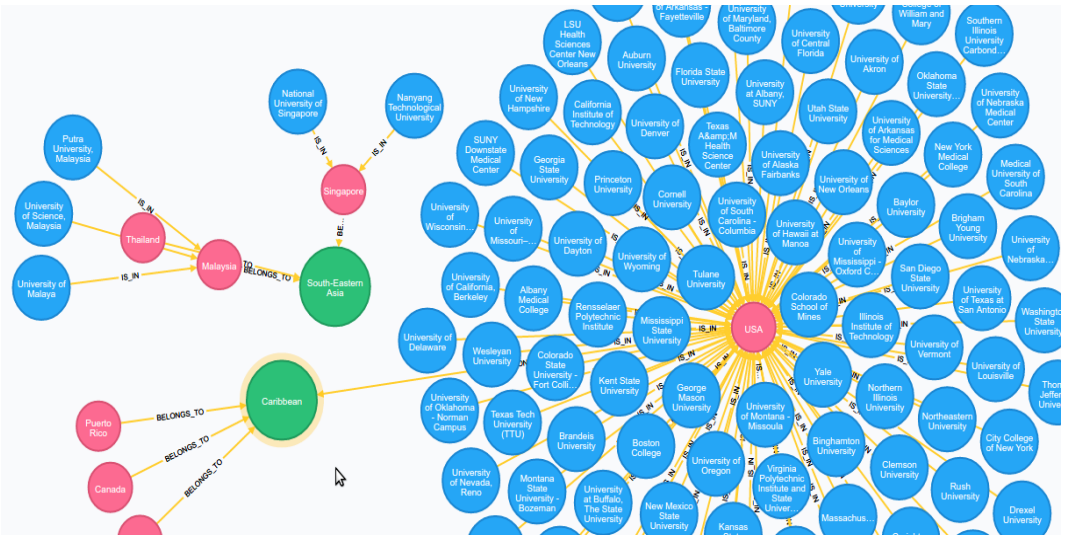


Fig. 29: Institute to Country to Region mapping

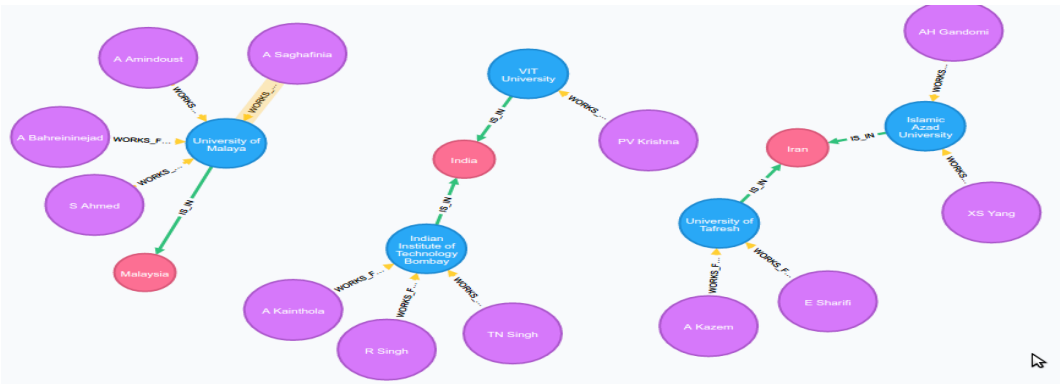


Fig. 30: Author to Institute to Country mapping

APPENDIX C:

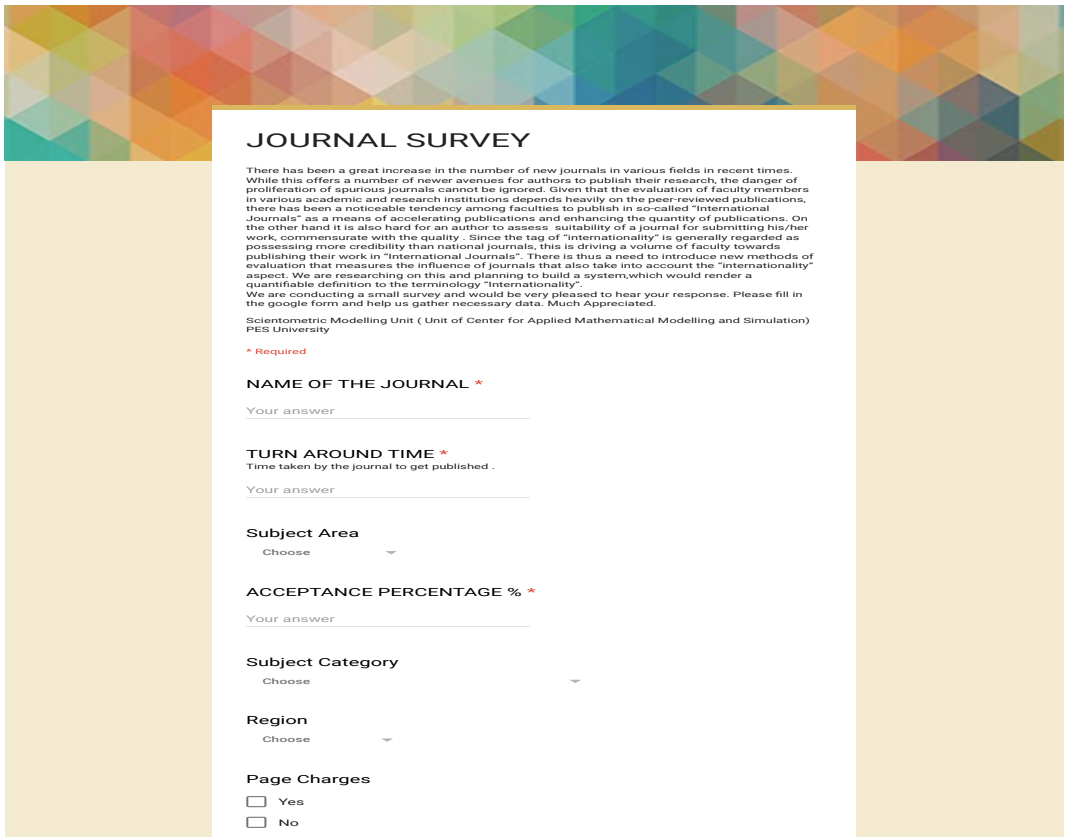
The section of Appendix covers those novel metrics which are yet to be used in the current model. These will be utilized as a part of future work.

- **Weighted NLIQ**: We defined another metric here which represents a more fair picture of citations made by articles published in different journals.

$$Weighted\ NLIQ = \frac{\alpha N_H + (1-\alpha)N_L}{N}$$

  N_H=number of citations coming from articles published in higher impact factor journals
  N_L=number of citations coming from articles published in lower impact factor journals
  N=*Total number of citations made by articles in a journal*
  α is the ratio between N_H and N which is very intuitive to give more weight to the role of citations coming from journals having high impact factor in deciding the value of NLIQ.
- **Turnaround time(amount of time from the time of submission to publication) and Acceptance Ratio ((Number Of Accepted Papers) / (Number Of Submitted Papers):** are two parameters which should be additional measures in computing Internationality Index. However, barring a few journals, this information is not available for scraping. As a future endeavor, we have created a survey to reach out to the editors and hope that this information would be available to us in six to eight months time. The Fig. 31 shows the screen-shot of the survey. Assuming lukewarm response from the Editor-in-Chief's, this task will have to be accomplished programmatically. The Stochastic Cobb-Douglas model of scoring is endowed with handling these parameters as long as numerical values could be scraped and computed. **We have not used this in our model yet.**
- **Time window**: Elsevier considers a 3-year window for SNIP mainly due to the difference in the rates at which subject fields mature, whereas Thomson Reuters has a 2-year and a 5-year window for Impact Factor (IF). As noted in section 7.1, one unmistakable advantage of SNIP over IF is that SNIPs 3-year citation window allows fields that move at a slower pace to be compared with those that advance fairly rapidly, in as fair a manner as possible. Whereas the 2-year IF and 5-year IF only favor one or the other. Thus, authors have considered a window of 3-5 years in order to cater to journals in both categories. Another reason is that many Journals shutdown due to various reasons in a very short span of time. Hence any journal needs minimum incubation time up to 3 years to prove its half-life.
- **Readership Profile**: An important additional factor, readership profile has to be included as an input parameter to the internationality index/influence computation model. It is interesting to know the source of article downloads country wise. This could be a challenging task as IP addresses are often masked. Finding a workaround is challenging and the current exercise has no provisions to build on this. This is one weakness that needs to be resolved. We may use Research Gate to gather this information on a limited scale.
- **Volumetric information**: NLIQ may vary widely across domains and this may hurt some journals more than others. Normalization, not implemented yet in the computation of NLIQ, is a pertinent landmark to accomplish. In order to obtain normalized NLIQ, we could divide the NLIQ of a Journal with the total number of Journals belonging to a domain. The challenge lies in correctly identifying the classification of journals for categorization and count as there is always some overlap across domains. SCOPUS, WOS and GS all have their own logic for segregating the Journals. ACM subject classification is useful and clear enough and might be used for this purpose. The good part of the overlap mentioned above is that journals in niche domains don't get isolated and subsequently NLIQ computation doesn't suffer abruptly. This should alleviate the concern of decent and good Journals having lower NLIQ.
- **Differentiating citations**: Based on the type of article, it is possible to differentiate between number of survey and original research article citations of a journal. Surveys or review articles, written tutorials and technical reports tend to receive a large number of citations. It is necessary to distin-

guish between journals which publish original research articles only, a mix of research and review articles and review articles only (ACM Computing Survey). Therefore, internationality score and parameter quantification need to be normalized accordingly to ensure a fair comparison between those journals.



JOURNAL SURVEY

There has been a great increase in the number of new journals in various fields in recent times. While this offers a number of newer avenues for authors to publish their research, the danger of proliferation of spurious journals cannot be ignored. Given that the evaluation of faculty members in various academic and research institutions depends heavily on the peer-reviewed publications, there has been a noticeable tendency among faculties to publish in so-called "International Journals" as a means of accelerating publications and enhancing the quantity of publications. On the other hand it is also hard for an author to assess suitability of a journal for submitting his/her work, commensurate with the quality . Since the tag of "internationality" is generally regarded as possessing more credibility than national journals, this is driving a volume of faculty towards publishing their work in "International Journals". There is thus a need to introduce new methods of evaluation that measures the influence of journals that also take into account the "internationality" aspect. We are researching on this and planning to build a system,which would render a quantifiable definition to the terminology "Internationality".
We are conducting a small survey and would be very pleased to hear your response. Please fill in the google form and help us gather necessary data. Much Appreciated.

Scientometric Modelling Unit (Unit of Center for Applied Mathematical Modelling and Simulation) PES University

* Required

NAME OF THE JOURNAL *

Your answer

TURN AROUND TIME *
Time taken by the journal to get published .

Your answer

Subject Area

Choose

ACCEPTANCE PERCENTAGE % *

Your answer

Subject Category

Choose

Region

Choose

Page Charges

Yes

No

Fig. 31: Screenshot of the survey form

APPENDIX D

	Estimate	**Std.Error**	**z value**	**Pr(> | z |)**
Intercept	1.2409e-06	1.7857e-05	0.0695	0.94460
log(IC)	0.099995	9.4970e-06	10529.1331	< 2.2e-16
log(HINDEX)	0.10000	9.5781e-06	10440.5624	< 2.2e-16
log(Impact Factor)	0.10001	1.4339e-05	6974.5700	< 2.2e-16
log(SNIP)	0.099989	1.4927e-05	6698.3695	< 2.2e-16
log(NLIQ)	0.099991	1.8623e-05	5369.3074	< 2.2e-16

Table 14: **Result of SFA with DRS**

	Estimate	**Std.Error**	**z value**	**Pr(> | z |)**
Intercept	0.021560	0.11759	0.1834	0.8545198
log(IC)	0.13360	0.014054	9.5066	< 2.2e-16
log(HINDEX)	0.10659	0.014637	7.2819	3.290e-13
log(Impact Factor)	0.11322	0.022540	5.0232	5.082e-07
log(SNIP)	0.087938	0.023204	3.7897	0.0001508
log(NLIQ)	0.36985	0.028053	13.1838	< 2.2e-16

Table 15: **Result of SFA with IRS**

	Estimate	**Std.Error**	**z value**	**Pr(> | z |)**
Intercept	0.017606	0.096683	0.1821	0.8555
log(IC)	0.12805	0.011925	10.7384	< 2.2e-16
log(HINDEX)	0.10539	0.012631	8.3440	< 2.2e-16
log(Impact Factor)	0.11091	0.018694	5.9332	2.970e-09
log(SNIP)	0.089746	0.019383	4.6301	3.654e-06
log(NLIQ)	0.32528	0.023998	13.5543	< 2.2e-16

Table 16: **Result of SFA with CRS**

Journal Name	Journal Internation-ality Score	Normalized Impact Factor(IF)	Class	Rank Based On Score	Rank Based On IF
ieee internet computing	0.877355594	1	HIGH	1	1
ieee communications magazine	0.860482353	0.53454307	HIGH	2	6
ieee journal on selected areas in communications	0.850307023	0.675994753	HIGH	3	3
ieee transactions on information forensics and security	0.80236338	0.644075208	HIGH	4	4
ieee intelligent systems	0.795182462	0.36270223	HIGH	5	12
ieee security and privacy	0.79452798	0.526235243	HIGH	6	7
information systems journal	0.75030428	0.563620463	HIGH	7	5
computer graphics forum	0.736308761	0.503279405	HIGH	8	8
computer networks	0.729782365	0.279186707	HIGH	9	17
ieee transactions on services computing	0.72940446	0.429602099	HIGH	10	11
ieee transactions on network and service management	0.712797841	0.718408395	HIGH	11	2
computer communications	0.692723997	0.316571928	MEDIUM	12	15
ieee transactions on mobile computing	0.690139691	0.327503279	MEDIUM	13	14
data mining and knowledge discovery	0.684616943	0.330126804	MEDIUM	14	13
parallel computing	0.654680851	0.267380848	MEDIUM	15	19
ad hoc networks	0.653621343	0.25994753	MEDIUM	16	20
distributed computing	0.652240431	0.206821163	MEDIUM	17	26
wireless networks	0.643580455	0.155225186	MEDIUM	18	32
networks	0.642035324	0.22540446	MEDIUM	19	25
multimedia systems	0.641056515	0.190642763	MEDIUM	20	29
information systems research	0.619051639	0.15347617	MEDIUM	21	33
acta informatica	0.586660159	0.143419327	MEDIUM	22	37
information systems frontiers	0.585802505	0.471141233	MEDIUM	23	10
neural processing letters	0.568273621	0.152601662	MEDIUM	24	34
real-time systems	0.568191481	0.197420201	MEDIUM	25	28
world wide web	0.568068949	0.164188894	MEDIUM	26	30
performance evaluation	0.562469907	0.232619152	MEDIUM	27	23
journal of computer and system sciences	0.556014401	0.066462615	MEDIUM	28	56
journal of computer security	0.549110242	0.447748142	MEDIUM	29	10

Table 17: **Ranking based on computed Internationality Score and Impact Factor: Only 5 journals among 86 journals match ranks. This indicates a total mismatch between two ranking system hypothesizing that Internationality may not be depend on Journal Impact Factor alone.**

Journal's Name	Calculated Score	Normalized Impact Factor(IF)	Class	Rank Based On Score	Rank Based On IF
cluster computing	0.540714872	0.148010494	MEDIUM	30	35
telematics and informatics	0.531407513	0.113248798	MEDIUM	31	43
new generation computing	0.530560129	0.093353739	MEDIUM	32	49
multimedia tools and applications	0.530323023	0.122868387	MEDIUM	33	39
journal of technology in human services	0.528141033	0.240926979	MEDIUM	34	21
mobile networks and applications	0.514086212	0.066243988	MEDIUM	35	57
international journal of telemedicine and applications	0.499847179	0.157411456	MEDIUM	36	31
security and communication networks	0.499678144	0.097726279	MEDIUM	37	46
journal of grid computing	0.491846319	0.302579799	MEDIUM	38	17
mobile information systems	0.491061168	0.239615216	MEDIUM	39	22
international journal of neural systems	0.488592191	0.070397901	MEDIUM	40	54
journal of parallel and distributed computing	0.477355158	0.066899869	MEDIUM	41	55
international journal of security and networks	0.474940881	0.065150853	MEDIUM	42	58
journal of heuristics	0.474629156	0.050065588	MEDIUM	43	64
international journal of ad hoc and ubiquitous computing	0.470028521	0.200043725	MEDIUM	44	27
universal access in the information society	0.469053228	0.093353739	MEDIUM	45	49
journal of computer-mediated communication	0.45918763	0.026016616	MEDIUM	46	81
journal of network and systems management	0.458760731	0.056405772	MEDIUM	47	64
international journal of distributed sensor networks	0.449149541	0.06165282	MEDIUM	48	61
international journal of network management	0.443216582	0.124617403	MEDIUM	49	38
systems engineering	0.436685918	0.11674683	MEDIUM	50	42
journal of network and computer applications	0.43639752	0.118058592	MEDIUM	51	41
electronic commerce research and applications	0.43297394	0.028202886	MEDIUM	52	78
international journal of internet technology and secured transactions	0.422533758	0.146917359	MEDIUM	53	36
international journal of information and computer security	0.420078698	0.236335811	MEDIUM	54	23

Table 18: **Ranking based on computed Internationality Score and Impact Factor: Only 5 journals among 86 journals match ranks. This indicates a total mismatch between two ranking system hypothesizing that Internationality may not be depend on Journal Impact Factor alone.**

Journal's Name	Calculated Score	Normalized Impact Factor(IF)	Class	Rank Based On Score	Rank Based On IF
international journal of web services research	0.4165797	0.059029296	MEDIUM	55	63
problems of information transmission	0.413855084	0.087888063	MEDIUM	56	51
international journal of web based communities	0.413088385	0.060996939	MEDIUM	57	62
journal of intelligent information systems	0.400489685	0.04110188	MEDIUM	58	67
photonic network communications	0.391909615	0.08810669	LOW	59	50
international journal of mobile network design and innovation	0.391709708	0.084827285	LOW	60	52
international journal of information security	0.384216366	0.035417578	LOW	61	73
international journal of communication systems	0.382962752	0.060996939	LOW	62	62
international journal of distance education technologies	0.380717994	0.064494972	LOW	63	59
international journal of mobile computing and multimedia communications	0.376743644	0.03301268	LOW	64	74
journal of high speed networks	0.369266275	0.031263664	LOW	65	75
international journal of web and grid services	0.365812183	0.071053782	LOW	66	53
optical switching and networking	0.365423732	0.094228247	LOW	67	48
international journal of networking and virtual organisations	0.364896337	0.229558373	LOW	68	24
journal of location based services	0.359237823	0.02864014	LOW	69	77
international journal of grid and high performance computing	0.358626002	0.039352864	LOW	70	69
international journal of communication netwoks and distributed systems	0.355853317	0.098600787	LOW	71	45
international journal of web engineering and technology	0.353184531	0.02645387	LOW	72	80
wseas transactions on communications	0.351893809	0.046348929	LOW	73	65
international journal of sensor networks	0.351024986	0.107564495	LOW	74	44
international journal of information and communication technology	0.34823629	0.039571491	LOW	75	68
wseas transactions on signal processing	0.346503467	0.038041102	LOW	76	72
journal of web engineering	0.346311551	0.039353	LOW	77	19

Table 19: **Ranking based on computed Internationality Score and Impact Factor: Only 5 journals among 86 journals match ranks. This indicates a total mismatch between two ranking system hypothesizing that Internationality may not be depend on Journal Impact Factor alone.**

Journal's Name	Calculated Score	Normalized Impact Factor(IF)	Class	Rank Based On Score	Rank Based On IF
international journal of information technology and management	0.333080343	0.027328378	LOW	78	79
journal of electronic commerce in organizations	0.313417874	0.028858767	LOW	79	76
international journal of electronic government research	0.31141811	0.045037167	LOW	80	66
international journal of high performance computing and networking	0.308124672	0.025579362	LOW	81	84
international journal of electronic security and digital forensics	0.307287226	0.038696983	LOW	82	71
international journal of computer applications in technology	0.306806973	0.038478356	LOW	83	72
international journal of applied cryptography	0.272246728	0.025797989	LOW	84	83
journal of computer and systems sciences international	0.257704031	0.118495846	LOW	85	40
international journal of e-collaboration	0.209401439	0.063401836	LOW	86	60

Table 20: **Ranking based on computed Internationality Score and Impact Factor: Only 5 journals among 86 journals match ranks. This indicates a total mismatch between two ranking system hypothesizing that Internationality may not be depend on Journal Impact Factor alone.**