ScientoBASE: A Framework and Model for Computing Scholastic Indicators of non-local influence of Journals via Native Data Acquisition algorithms

Gouri Ginde, Snehanshu Saha, Archana Mathur, Sukrit Venkatagiri, Sujith Vadakkepat, Anand Narasimhamurthy, B. S. Daya Sagar

Abstract Defining and measuring internationality as a function of influence diffusion of scientific journals is an open problem. There exists no metric to rank journals based on the extent or scale of internationality. Measuring internationality is qualitative, vague, open to interpretation and is limited by vested interests. With the tremendous increase in the number of journals in various fields and the unflinching desire of academics across the globe to publish in ”international” journals, it has become an absolute necessity to evaluate, rank and categorize journals based on internationality. Authors, in the current work have defined internationality as a measure of influence that transcends across geographic boundaries. There are concerns raised by the authors about unethical practices reflected in the process of journal publication whereby scholarly influence of a select few are artificially boosted, primarily by resorting to editorial manoeuvres. To counter the impact of such tactics, authors have come up with a new method that defines and measures internationality by eliminating such local effects when computing the influence of journals. A new metric, **Non-Local Influence Quotient** (NLIQ) is proposed as one such parameter for internationality computation along with another novel metric, **Other-Citation Quotient** as the complement of the ratio of self-citation and total citation. In addition, SNIP and International Collaboration Ratio are used as two other parameters. As these journal parameters are not readily available in one place, algorithms to scrape these metrics are written and documented as a part of the current manuscript. Cobb-Douglas production function is utilized as a model to compute JIMI (Journal Internationality Modeling Index). Current work elucidates the metric acquisition algorithms while delivering arguments in favor of the suitability of the proposed model. Acquired data is corroborated by different supervised learning techniques. As part of future work, the authors present a bigger picture, RAGIS- Reputation And Global Influence Score, that will be computed to facilitate the formation of clusters of journals of high, moderate and low internationality.

Keywords: Journal Influence Score; Journal Internationality modeling index (JIMI); web scraping; feature extraction; Cobb-Douglas Production Function; convex optimization; supervised learning; Non-Local Influence Quotient (NLIQ), Source-Normalized Impact per Paper (SNIP).

1 Introduction

In recent times, various authors and research scholars have been exploring means to find suitable and reputed journals for publication of their research work. The drive behind this is to own appreciation or award for the quality work that they do. Also, institutional assessment and evaluation depends heavily on peer-reviewed publications whether it be for academia or research labs. Generally, the trend observed among various faculties is to publish their research in journals with an ’international’ tag attached to it. Thus, evaluating internationality is an open problem owing to the

Gouri Ginde
Department of Computer Science and Engineering, PESIT South Campus Bangalore, India e-mail: gouri.ginde@gmail.com

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

Sukrit Venkatagiri
Department of Computer Science and Engineering, PESIT South Campus Bangalore, India e-mail: 95sukrit@gmail.com

fact that such journals are vast in number; every such entity claims ”internationality” but citation and influence are a bare minimum.

Data collected from IEEE Xplore in the year 2009 showed an exponential increase of 25% in journal publications, when compared with previous years. A study conducted by Buchandiran [1] reveals an enormous increase in publication of journals between the years 2004 and 2009, whereby in the year 2009, 6,132 Indian institutions have contributed 23,745 papers out of which 15,880 were from academic institutions. This clearly shows that academic institutions contribute to the majority of such published works. Leonard Heilig and Stefan VoB cite2 observed a significant increase in the number of research publications (only in the area of cloud computing) from 2008 onwards. Elsevier’s Scopus covered 15,376 publications till 2014 and Thomson Reuters Web of Science covered 8,262 publications in the same field.

On the flip side, there exists scam open access publishers who unethically and unprofessionally exploit the open access publishing model for financial rewards. They charge authors for publication fees and publish their work without providing true editorial services as well as other types of services associated with any legitimate journal. This shady publishing practice was first noticed by Jeffrey Beall,an academic librarian and a researcher at the University of Colorado in Denver. He scrutinized and investigated further and based on his findings, published his first compiled list of predatory publishers in 2010 [19]. Continuing on the same line, Jeffrey Beall,[20] regularly updated this list of fake publishers and had put forth certain criteria for categorizing such publications in order to prevent newer scholars from falling prey to such practices.

The phenomenon of predatory publishing (also known as the dark side of open access publishing) has exploded in recent years with the number of such publications expanding from 53,000 in 2010 to 420,000 in 2014. Accepting articles quickly without peer-review, pursuing academicians to submit articles or to serve on editorial boards, notifying authors of article processing fees only after paper acceptance, improper usage of ISSN and counterfeit Impact Factor values are some of the key indicators which have emerged from the observed working pattern of fake, predatory publishers. Till date, no fool-proof method has been devised to distinguish legitimate publishers from illegitimate ones.

An abundance of work has been done to evaluate the influence or prestige of scholarly articles and journals. Citation Index, a concept defined by Eugene Garfield (founder of Science Citation Index, SCI and the Institute for Scientific Information, ISI) serves as a link between similar scientific journals and literature. Citation pattern and citation frequency used by Garfield in his foundational work for Web of Science (Thompson Reuters Web of Science) initiated a vast spectrum of research and provided fully indexed and searchable research content. Thompson Reuters then initiated publishing Journal Citation Reports (JCR) to evaluate citation frequency of journals and down-the-line, Impact Factor.

Another initiative, Elseviers Scopus has a vast collection of peer reviewed scholarly journals and citations in scientific, medical, technical and social science domain. Scopus utilizes its database to provide another type of journal metric used for ranking for its journals through the SCImago Journal and Country Rank (SJR) portal [8]. The SJR rank is a score evaluated from the past five years’ data addressing a small number of journals. It is claimed that SCI, Thompson Reuters is a little more selective than Scopus. The concept of citation index, Impact Factor and SJR ranking provide a limited respite to the above mentioned challenges of distinguishing and ranking legitimate publishers from the fake entities. This gives plenty of motivation and reason to work on proving a journal’s credibility and integrity as well as ascertaining the quality, impact and influence of the publications.

Our initiative, ScientoBASE epitomizes a new approach for evaluating journals in a ”height-weight” manner. The database, when complete, will help identify and bring adequate attention to quality journals, including industry practitioner domains, which otherwise would not be possible because certain journals namely Software Quality Porfessional (SQP) refuse to be indexed.

The remainder of the paper is organized as follows. Section 2 contains literature survey carried out on existing work related to measuring a journal’s internationality vis-a-vis non-local influence. The section also brings forth systemic lags in establishing an unbiased score for ranking of journals. Section 3 defines internationality as perceived by the authors. It presents a schematic view of the approaches used to model internationality. Section 4 presents the algorithmic overflow for calculating internationality. Section 5 discusses in detail different techniques and algorithms used to collect scholastic parameters for the model described in section 6. These parameters are programatically scraped from multiple web sources such as Google Scholar, IEEEXplore, SCImago and Aminer [25, 26]. Once parametric data is acquired, these are then fed into the Cobb Douglas production function; an econometric model that is described in detail in section 6. Section 7 sheds some light on the merits of Source-Normalized Impact per Paper (SNIP) and shows why SNIP - and not Impact Factor - is a good albeit incomplete indicator for estimating non-local influence. Further, the section introduces new metrics, Non-Local Influence Quotient (NLIQ) and Other Citation Quotient and argues in favor of the usefulness of such metrics towards computing internationality. The paper concludes with a discussion on future work embodying ranking and clustering of journals according to internationality in their respective subject areas. The future work is commensurate with the current framework and model proposed in this paper.

2 Literature Survey

Neelam Jangid Snehanshu Saha, Siddhant Gupta, Mukunda Rao J [6, 7] in their work used a lightweight approach and introduced a new metric, Journal Influence Score (JIS), which is calculated by applying principal component analysis (PCA) and multiple linear regression (MLR) on citation parameters, extracted and processed from various scholarly articles in different domains, to obtain a score that gauges a journals impact. The higher the score, the more the journal is valued and accepted. Journals ranking results are compared with ranks of SJR, which internally uses Google's PageRank algorithm to calculate ranks. The results showed minimal error and the model performed reasonably well. Seyyed Mehdi et al. [9] studied the scientific output of fifty countries in the past 12 years. In order to measure the 'quality' and 'quantity' of research output, a two-dimensional map is constructed and analyzed. Clusters are generated after analysis to represent country wise research output. There exists no ranking mechanism to rank countries with the maximum output in terms of quality and quantity of journals themselves.

Anup Kumar Das, Sanjaya Mishra [10] discussed how research communities are preferring article-level metrics (ALM) over Journal Impact Factor (JIF) to assess the performance of individual scientists and their contributions. Gunther K. H. Zupanc [11] also stressed on the unsuitability of using Journal Impact Factor to compare the influence of journals, especially when journals are from different areas. He claims that authors are tempted to publish their work in high-Impact Factor journals instead of journals that are best suited for their research work. A. Abrizah et al. [16] compared the coverage, ranking, impact and subject categorization of Library and Information Science journals, in which 79 titles were from Web of Science and 128 from Scopus. The prestige factor score of journals from JCR (Journal Citations Report 2010) and SJR (SCImago Journal Rank 2010) was extracted and the difference in ranks was noted. They observed a high degree of similarity in impact factor of titles in both Web of Science and Scopus. At the same time, authors also observed that the two databases differ in the number of journals covered.

Henk F. Moed [29] introduced a different indicator of journal citation. impact, Source Normalized Impact per Paper (SNIP). SNIP is defined as the ratio of the journals citation count per paper and the citation potential in its subject field. It aims to allow direct comparison of sources in different subject fields. There is no single perfect indicator of journal performance. Delimitation of a journals subject field does not depend upon some predefined categorization of journals into subject categories but is entirely based on citation relationships. It is carried out on a paper-by-paper basis, rather than on a journal-by-journal basis. SNIP is based on citations from peer-reviewed papers to other peer-reviewed papers. Ludo Waltman et al. [12] have discussed a number of modifications that were recently made to the SNIP indicator. The SNIP indicator considers a source normalized approach to correct the differences in citation practices between scientific fields. The key benefit of this approach is that it does not require the classification of subject fields, where the boundaries of fields are defined explicitly. There are some arguments around the original SNIP indicators properties that may be considered counter-intuitive. For instance, it is possible that additional citation has a negative correlation with journals SNIP value.

Gaby Haddow, Paul Genoni [13] defined a new model - Excellence for Research for Australia (ERA) to determine the efficacy of citations measures in order to determine the quality of Australian social science journals. Chiang Kao [33] investigated the contribution of different countries to international repositories of research in industrial engineering journals. After compiling journal data from ISI from 1996 to 2005, it was evident that the USA, UK and China are the top three countries to contribute articles to IE journals and six Asian countries are in the top ten. Yu Lipinga et al. [34] classified common journal evaluation indicators into three categories, namely three first-level indicators. They are, respectively, the indicators on journal impact, on timeliness, and on journal characteristics. The three categories of indicators are correlated with one another, so a structural equation may be established. Then authors calculated the value of three first-level indicators and gave subjective weights to these indicators. This approach provides a new perspective for scientific and technological evaluation, in a general sense. There are some limitations of this approach:

- The availability of basic data and the rationality of modeling bear much upon the evaluation results.
- If there are too many indicators in a scientific and technological evaluation, data availability will be relatively difficult, and the evaluation cost will increase. If indicators are too few, they cannot provide adequate information.
- If the data is inaccurate or wrong, no satisfactory results will be obtained. In scientific and technological evaluation, sometimes certain data is very difficult to gather.

Gualberto Buela-Casal et al. [14] performed a survey on existing measures of internationality and observed that a valid and quantitative internationality index should differentiate between quality and internationality. They suggested that in order to measure internationality, suitable weights should be assigned to certain identified parameters using a large-scale census of journal data. They proposed a neuro-fuzzy system to construct an unambiguous journal internationality index.

Chia-Lin Changa et al. [15] examined the issue of coercive journal citations and the practical usefulness of two recent journal performance metrics i.e. Eigenfactor Score, which may be interpreted as measuring journal influence, and the Article Influence Score, using Thomson Reuters Web of Science. Authors compared the two new bibliometric measures with the existing ISI metrics, total citations and the 5-Year Impact Factor (5Y-IF) of a journal. It is shown that the sciences and social sciences are different in terms of the strength of the relationship of journal performance metrics, although the actual relationships are very similar. Authors concluded that the Eigenfactor Score (measuring journal influence) and Article Influence performance metrics for journal are shown to be closely related empirically to the two existing ISI metrics, and hence add little in practical usefulness to what is already known, except for eliminating pressure arising from coercive journal self-citations.

Predatory publishing has earned a lot of attention (in terms of approval as well as criticism) from different sections of research communities across the globe [21]. Beall's list of predatory journals has been welcomed by many open access supporters, whereas others have raised serious doubts about it's credibility. Walt Crawford [24] in 2014 thoroughly investigated the list and called it a ”One Man's List”. He concluded that it should be ignored and suggested some steps to evaluate a journal's trustworthiness prior to submission.

Step 1: To make a pertinent decision whether ”The International Journal of A” is a good target, one must look it up in the Directory of Open Access Journals (doaj.org). If the journal is not in the directory, look for another journal in a similar subject category.

Step 2: If the journal is in DOAJ, explore its site, its APC policy, quality of English used, its editorial board members - whether they are real people. Otherwise start from step 1.

Step 3: Check whether article title over the past issues makes sense within the journal's scope or if any author show up repeatedly within the past few issues. If so, go to step 1 again.

One can escape from predatory journals utilizing this technique. Nonetheless, it needs a lot of involvement in knowing how to assess journals as there is no scientific model which will do so for us. Additionally, this algorithm is, to a greater extent, a manual investigation and hence ungainly and elaborate. Therefore, there is a pressing need to build a complete, end to end web interface that also serves as repository and information visualization toolkit for scientometric evaluation, modeling and analysis. ScientoBASE is designed to serve this purpose and cater to internationality modeling and interface estimation of peer-reviewed journals in the fields of science and technology.

3 Definition, Objective and Schematic View

This section defines internationality and presents an overview of the steps to achieve the end results. The authors would like to take this opportunity to stress that the ”internationality of a journal” is defined here as a measure of influence beyond restricted boundaries. These boundaries may be geographical or regional or even cliques or networks of journals. It was observed during the course of this research that citations occur mostly within the journal from which the original citing article was published. The authors believe that in a community which is essentially international by nature, such trends don’t bode well. A new metric which is an offspring of this realization, will be elaborated in due course. In order to remain clear about our objectives and dispel any confusion, we reiterate that internationality as defined and measured throughout this paper is a reflection of ”non-local influence” and therefore does concur with the standard definition. The basic steps taken to achieve the end results are as follows:

- Defining and measuring internationality
- Creating a suitable model
- Validating the model
- Generating granular clusters of international journals and conferences (part of future work)
- Model the diffusion of internationality (part of future work)

Definition: Internationality of a journal, as proposed by the authors, is a holistic parameterization, of the international aspects of a journal’s dimensions. These dimensions include - but are not limited to - quality of publications and measure international span of subscribing readers, authors and reviewers. These additionally evaluate the geographic source of a journals citations and the impact it spreads across nations. Authors explore these dimensions in succession and refer to International Collaboration Ratio; a parameter that indicates the ratio of articles whose author affiliations are from distinct nation, has the potential to be a suitable candidate for evaluating a journal’s prestige. Likewise, extensive self-citation is a self promoting strategy which is unfairly used by authors to artificially boost their scientific influence, and thus indirectly inflate the publishing journal’s impact factor. If used skillfully, this self-citation parameter can provide a good insight to judge a journal’s credibility. It must be stated at this juncture, that some of the most common attributes of internationality such as ISSN number, constitution of editorial boards, country of publication and reputation of publishers as input factors are not considered. This is precisely because the authors do not view these attributes as entirely sufficient measures for internationality. Rather, use of such attributes as yardsticks in judging internationality is viewed as impediments towards objectively classifying journals.

There are well-accepted influence measurement parameters used by various web portals. Source-Normalized Impact per Paper (SNIP), allows comparison of sources across and within the same subject field by calculating their citation potential and normalizing their citation impact by dividing their RIP’s (Raw Impact per Paper) with the calculated database citation potential. Integrity of academic publications would be at risk if editors coerce authors to cite their journals for enhancing their impact factor. With the intention to weaken the effects of this strategy, authors have introduced a new metric Non-Local Influence Quotient (NLIQ) which is a ratio of ”non-local” citations of an article to the total number of citations. Larger the value of NLIQ, more ”international” a journal is.

Taking all these factors into account, authors proposes a high-level design and methodology to model internationality index that would scrape and assemble the above mentioned parameters and generate journal clusters of high, moderate and low internationality. As already indicated, after computing ”internationality”, generating granular clusters of journals is a future plan of action. The current work embodies various algorithms for parameter acquisition and discusses the suitability of the new metric for influence calculation.

Empowered by data acquisition techniques, two approaches are put forth for modeling (Fig. 1). The first approach takes data from the Scopus and SJR portals and calculates a journal’s score (JIS) [6] using a multiple linear regression model on the scraped scientific indicators. Second approach, uses non-indexed, non-Scopus/non-Web of Science databases to acquire scientific parameters and evaluate a journal’s internationality score generated from a Modeling Index (JIMI, Journal Internationalty Modeling Index) [30, 31]. The approach uses Cobb-Douglas [35, 36] and Log Production Model on the parameters scraped from web. The algorithms and procedures are described in section 5 and 6.

The prestige/internationality of a journal is a convex combination of JIS [please refer additional files on GitHub, **32**] and Internationality Score, represented as- Internationality of a journal,

$$YI = \alpha \text{ JIS} + (1 - \alpha) \text{ JIMI} ;$$
$$0 < \alpha < 1$$

where YI refers to the internationality score as response variable(to be sorted in decreasing order), JIS is the influence score obtained from metric JIS, JIMI is the score evaluated from work done using two parameters (JIMI) and α is a weight deduced from the cross correlation. For JIS, refer [17] and Appendix I in[32].

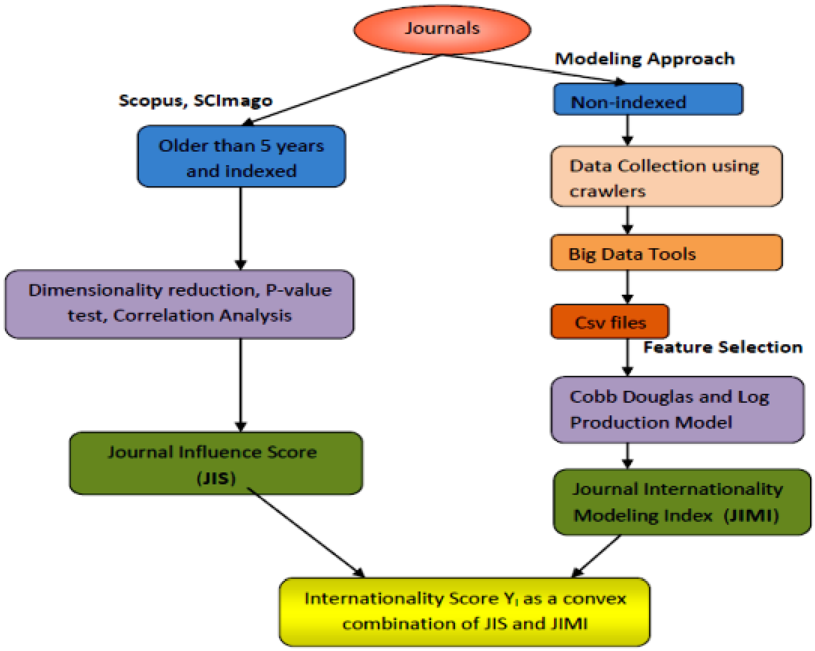

Journals
Modeling Approach
Scopus, SCImago
Non-indexed
Older than 5 years and indexed
Data Collection using crawlers
Big Data Tools
Dimensionality reduction, P-value test, Correlation Analysis
Csv files
Feature Selection
Cobb Douglas and Log Production Model
Journal Influence Score (JIS)
Journal Internationality Modeling Index (JIMI)
Internationality Score Y_I as a convex combination of JIS and JIMI

Fig. 1: Block diagram of Methodology.

3.1 *Journal Internationality Modeling Index(JIMI)*

The prestige of an academic journal is derived from quantifiable as well as non-quantifiable factors. Some commonly accepted factors that reflect a journals prestige are Impact Factor (IF), Eigenfactor, SCImago Journal Rank (SJR), Source-Normalized Impact per Paper (SNIP), Impact per Publication (IPP), internationality etc. Impact Factor, as per Thomson Reuter's definition [22] is a measure of the frequency with which an article of a journal has been cited in a particular duration. The IPP (Impact per Publication) measures the ratio of citations in a year to scholarly papers published in the three previous years divided by the number of scholarly papers published in those same years. When normalized for the citations in the subject field, the Impact per Publication becomes the Source-Normalized Impact per Paper (SNIP). The SJR or SCImago Journal Rank is a measure of the scientific prestige of scholarly sources. SJR assigns relative scores to all of the sources in a citation network.

In this section, authors discuss a technique [30] to quantify internationality by exploiting a mathematical model, which determines the internationality of a journal by using two major metrics - Source-Normalized Impact per Paper (SNIP) and International Collaboration. Here author stresses on the efficacy of such a model and confirm the model theoretically. We prove that the model has a global maxima where a particular value of the inputs (SNIP and

International Collaboration) would ensure some maximum value of internationality, subject to a constraint or set of constraints.

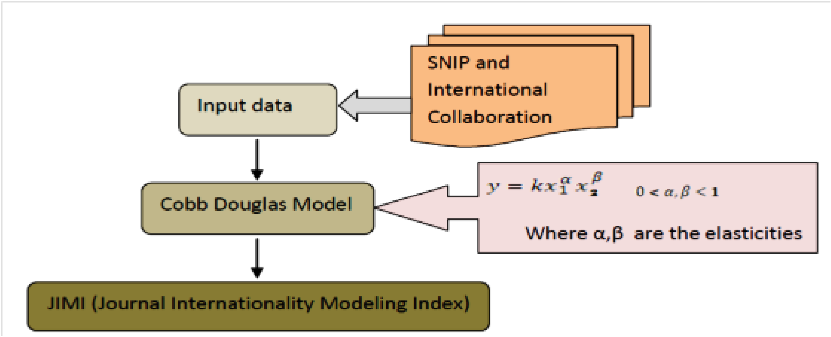

SNIP and International Collaboration
Input data
Cobb Douglas Model
$y = kx_1^{\alpha} x_2^{\beta}$
$0 < \alpha, \beta < 1$
Where α,β are the elasticities
JIMI (Journal Internationality Modeling Index)

Fig. 2: Computation model of Journal Internationality Modelling Index (JIMI).

As shown in Fig. 2, the ,modeling approach uses web scraping technique to extract required features of various journals to generate CSV data files. All these features from various data sources are processed further for only desired features of a journal which will contribute to the evaluation of internationality index. These features are specifically,

- International Collaboration Ratio
- Source-Normalized Impact per Paper (SNIP)
- Other-Citation Quotient
- Non-Local Influence Quotient (NLIQ)

Using the above features as input parameters to Cobb-Douglas production function [35, 36], the authors intend to measure the internationality index, proposing a score to gauge the influence of peer-reviewed journals.

Sample Data: This modeling approach uses a consolidated database generated by crawling the web (using software tools) to gather all recent (non-indexed) journals (that are older than 3 years and younger than 5 years) from Google Scholar.

4 Algorithmic Overview:

This section discusses the basic steps taken to compute internationality and to generate granular clusters.
Step 1: Collect data (algorithms to extract data are shown in section 5.3)
Step 2: Pose internationality as a score: ”y” as response variable.
Step 3: Model y = $f(x_1, x_2, x_3, ...x_i)$; i= 1,2,.....n where $x_1, x_2....x_i$ are the input variables as will be discussed in secton 6.
Step 4: (i)Perform down-selection, in case there are too many input variables, some of which could be highly correlated. Otherwise, go back to Step 3.
(ii) For simulation and visualization aid use a 3-D down-selection model.

$$y = A \prod_{i=1}^{2} {x_i}^{\alpha_i}$$

, obtain ”best” estimate of α_i ; use the best fit values.
Step 5: Compute ”y” for each category.
Step 6: Observe the density and histogram plot.
Step 7: Decide on the granularity of internationality into several classes.
Step 8: Predict/visualize ”variations” in ”y” based on small perturbations in $x_1 and x_2$.

The next section details the procedures and algorithms critical for data acquisition from the public domain. These include extracting data for the input parameters required for the model. Additional information such as journal name, country name etc is scraped for building a public repository.

5 Data Acquisition from Google Scholar

5.1 Collection

There are many advantages of using Google Scholar as a data source because it is free to access, easy to use and quick and comprehensive in its coverage. Various studies [5] have also shown that Google Scholar is a serious alternative data source for various reasons.

1. Not everything published on the internet is counted in Google Scholar:
   Google Scholar indexes only scholarly publications. As their website indicates ”we work with publishers of scholarly information to index peer-reviewed papers, theses, preprints, abstracts and technical reports from all disciplines of research”. Some not scholarly citations, such as student handbooks, library guides or editorial notes slip through. There might be some overestimation of the number of non-scholarly citations in Google Scholar, for many disciplines this is preferable to very significant and systematic under-estimation of scholarly citations in ISI or SCOPUS.
2. Non-ISI publications can be high-quality publications:
   There is a misconception that ISI listing is a stamp of quality and one should ignore non-ISI listed publication and citations.
   However, there are a few problems with this assumption. a) ISI has a bias towards Science and English language, b) ISI ignores the majority of publications in the social sciences and humanities as well as engineering and computer science fields.
3. Google Scholar flaws dont impact citation analysis much:
   There is no doubt that the Google Scholars automatic parsing occasionally provides us with nonsensical results. However, these errors do not appear to be frequent or important. They do not generally do not impact the results of author or journal queries much, if at all.
   What is more important is that these errors are random than systematic. In contrast, the commercial databases such as ISI and Scopus have systematic errors that do not include many journals, nor have good coverage of conference proceedings, books or book chapters. Therefore, although it is always a good idea to use multiple data sources, rejecting Google Scholar out of hand because of presumed parsing errors is not rational.

In spite of the fact that Google Scholar is an incomprehensible storehouse with uninhibited accessible, it does not provide an API. Likewise, Google entirely obstructs any computerized web crawling. Subsequently we turned to web scraping. With included occasional time delays in the script, we could gather the required data from Google Scholar.

5.2 Organization

Web scraping is the procedure of consequently gathering data from the World Wide Web. Under this, we plan to develop completely robotized frameworks that can change over whole web website into organized data for further handling. Fig. 4 demonstrates the essential segments of our methodology of web scraping Google Scholar. DOM Parsing is the philosophy which assists the system with retrieving element content created by client-side scripts utilizing undeniable web program controls, for example, the Internet Explorer browser or the Mozilla browser control. These program controls likewise parse web pages into a DOM tree, in light of which program can recover parts of the pages.

1. Create Scraping Template: Inspect Element is a developer tool that allows to view the HTML, CSS, and JavaScript that is currently on a web page. On nearly any web page, one can right click and select inspect element. This will pull up the developers console to view will the HTML and CSS of the web page. Using this tool we explored the Document Object Model (DOM) tree of the Google Scholar Engineering and Computer Science Section. Fig. 4 depicts the screen capture of the Inspect Element tool in use.

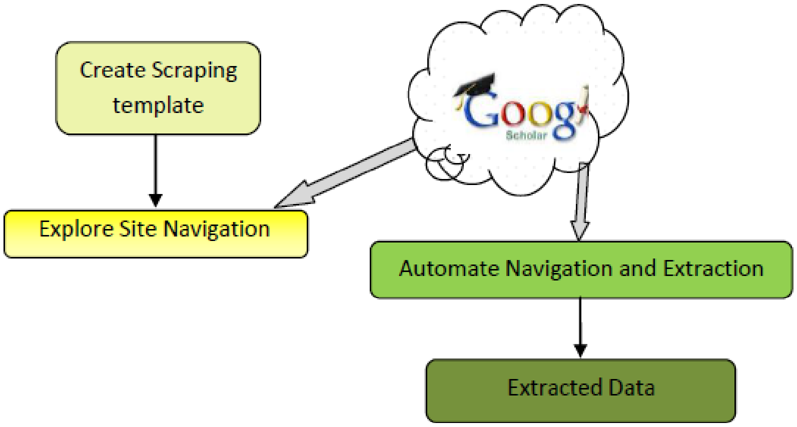
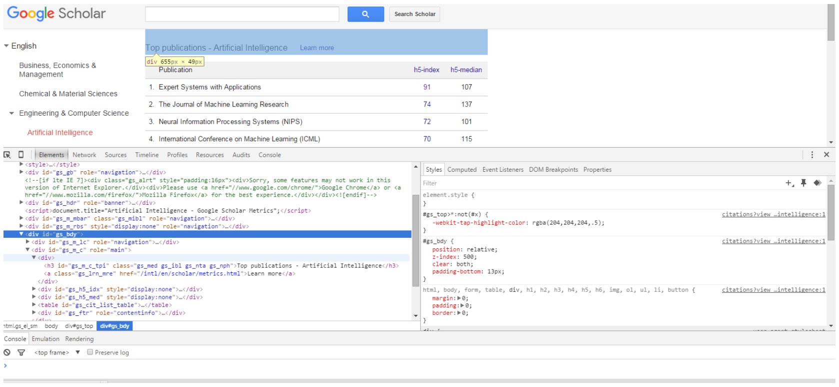


Fig. 3: Block diagram of Web Scraping Methodology

Left lower panel in the Fig. 4 is the Elements panel used to inspect all elements in the Google Scholar web page (top pane) in one DOM tree. Using this Elements panel one can select any element and inspect the styles applied to it. The right lower panel is Styles pane. It shows the CSS rules that apply to the selected element, from highest priority to lowest. Styles pane is used to view and change the CSS rules applied to any selected element in Elements panel. Following are the various other information access provided in Inspect Element tool.

- Elements: Shows the HTML for the current page
- Network: Shows all the GET and POST requests that are made while the developers console is open. One can also identify the requests that are taking the longest to process.
- Sources: Allows to view the JavaScript files (and other files) associated with the page. This is most used for debugging as a web page is being developed, but can be helpful for coding your own JavaScript in Qualtrics as well.
- Timeline: The timeline shows where time is invested when a web page is loaded/refreshed. It logs GETs, PUTs, calculations, parsing JavaScript, etc.
- Profiles: Also helps see where time is being spent on a page. One can record time spent by function, by JavaScript Object, and by script
- Resources: Allows to inspect the resources that are loaded onto a page. (i.e. cookies)
- Audits: Analyzes a page as it is loading and then gives suggestions to decrease the load time
- Console: This is a JavaScript console where one can try out code as if he/she were coding it for the web page. One can use it to log information about debugging, to test out code snippets, etc.



	Publication	h5-index	h5-median
1.	Expert Systems with Applications	91	107
2.	The Journal of Machine Learning Research	74	137
3.	Neural Information Processing Systems (NIPS)	72	101
4.	International Conference on Machine Learning (ICML)	70	115

```
▶<style>…</style>
▶<div id="gs_gb" role="navigation">…</div>
 <!--[if lte IE 7]><div class="gs_alrt" style="padding:16px"><div>Sorry, some features may not work in this
 version of Internet Explorer.</div><div>Please use <a href="//www.google.com/chrome/">Google Chrome</a> or <a
 href="//www.mozilla.com/firefox/">Mozilla Firefox</a> for the best experience.</div></div><![endif]-->
▶<div id="gs_hdr" role="banner">…</div>
 <script>document.title="Artificial Intelligence - Google Scholar Metrics";</script>
▶<div id="gs_m_mbar" class="gs_mibl" role="navigation">…</div>
▶<div id="gs_m_rbs" style="display:none" role="navigation">…</div>
▼<div id="gs_bdy">
  ▶<div id="gs_m_lc" role="navigation">…</div>
  ▼<div id="gs_m_c" role="main">
    ▼<div>
       <h3 id="gs_m_c_tpi" class="gs_med gs_ibl gs_nta gs_nph">Top publications - Artificial Intelligence</h3>
       <a class="gs_lrn_mre" href="/intl/en/scholar/metrics.html">Learn more</a>
     </div>
    ▶<div id="gs_h5_idx" style="display:none">…</div>
    ▶<div id="gs_h5_med" style="display:none">…</div>
    ▶<table id="gs_cit_list_table">…</table>
    ▶<div id="gs_ftr" role="contentinfo">…</div>
```

```
element.style {
}
#gs_top>*:not(#x) {                                              citations?view …intelligence:1
   -webkit-tap-highlight-color: rgba(204,204,204,.5);
}
#gs_bdy {                                                        citations?view …intelligence:1
   position: relative;
   z-index: 500;
   clear: both;
   padding-bottom: 13px;
}
html, body, form, table, div, h1, h2, h3, h4, h5, h6, img, ol, ul, li, button {   citations?view …intelligence:1
   margin:▶0;
   padding:▶0;
   border:▶0;
}
```

Fig. 4: Inspect Element tool usage to explore Document Object Model(DOM).

2. Explore Site Navigation: To further explore and understand the Google scholar site navigation for dynamic URL formulation, we used Beautiful Soul Parser. BeautifulSoup parser is also called Elixir and Tonic The Screen-Scrapers Friend [18] It uses a pluggable XML or HTML parser to parse a (possibly invalid) document into a tree representation. Beautiful Soup provides methods and pythonic idioms that make it easy to navigate, search, and modify the parse tree.

3. Automate Navigation and Extraction: Python is a scripting language which is easy to learn, powerful programming language. We have used the python interpreter library which is freely available in source and binary form for all major platforms.

5.3 Extraction

5.3.1 Algorithm for Feature Extraction - Algorithm 1

Algorithm 1 is for features extraction and internationality index computation for all the listed journals from Google Scholar (data source) under engineering and computer science field. Features such as total citations, other-citation, International Collaboration Ratio and SNIP are computed for each one of these journals later on. Also, additional data was obtained from Aminer Citation Network data set[26] based on a paper by Jie Tang et al. [25]. Using web scraping we first extract all the journal names from the source: line 1. Then extract Total Citations count and all the Articles published in each one of these journals: line 3 and 4. Further on we compute the cumulative/averaged parameter values for that journal from the various values extracted for each article: line 5 to 8. The various function calls in these lines are explained ahead in the report under respective algorithms. the average value for the International Collaboration is computed: line 11. Finally, line 12 and 13 invoke the functions to compute the SNIP and Internationality Index.

Algorithm 1 Algorithm to extract various features and to compute Internationality Index of Journals

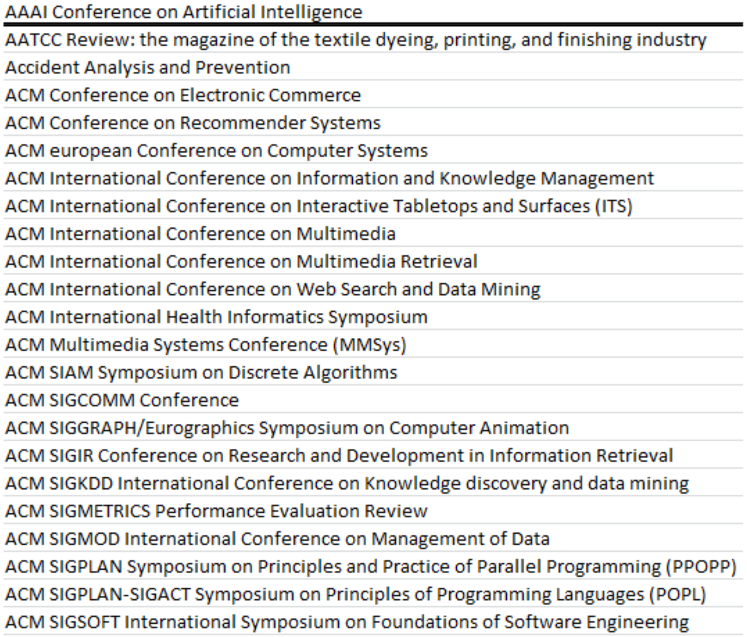
```
 1: Input: URL link of Google Scholar
 2: Output: Features such as International Collaboration Ratio, SNIP, Other-Citations and Internationality Index
 3: $JNames[]$ = $Fetch\_Journal\_Names\_from\_Google$(Engineering and Computer Science)
 4: for every journal: $JNames[i]$ do
 5:     TotalCites = Get the totalcites value
 6:     Get all the published articles/papers: $X[]$
 7:     for every article: $X[i]$ do
 8:         $JNames[i]$.Selfcites += compute_SelfCitations($X[i]$)
 9:         $JNames[i]$.Intl_Collaboration_Ratio += compute_Intl_Collaboration_Ratio($X[i]$)
10:     end for
11:     $averaged\_Intl\_Collaboration\_Ratio(JNames[i])$
12:     $compute\_SNIP(JNames[i])$
13:     $compute\_NonLocalIQ(JNames[i])$
14:     $compute\_Internationlity(JNames[i])$
15: end for
```

5.3.2 Journal Name Extraction - Algorithm 2

This algorithm is to extract the journal names for Algorithm 1 to work upon. For the given source's URL, we perform web scraping to first extract all the subcategories of the Engineering and Computer Science field: line 1 and then in turn scrape the 20 journal names listed in each of the web links for these subcategories by dynamically generating the URL addresses using subcategory names: line 2 Then on we accumulate these scraped journal names in the spreadsheet: line 3 We successfully extracted about 1160 journal names from all the subcategories listed under Engineering and Computer Science category in Google Scholar. Fig. 5 shows the sample capture of the journal names extracted.

Algorithm 2 $Fetch\_Journal\_Names\_from\_Google()$: Algorithm to Extract Journal Names from Google Scholar

```
Input: A html file of Google Scholar web page: $HLINK$
Output: List of Journal Names
for every sub category link in $HLINK$ : $SUBLINK$ do
    for every hyperlink in the $SUBLINK$ : $JLINK$ do
        Print $JLINK.gs\_title$ from $< td >$ tag to spreadsheets
    end for
end for
```



AAAI Conference on Artificial Intelligence
AATCC Review: the magazine of the textile dyeing, printing, and finishing industry
Accident Analysis and Prevention
ACM Conference on Electronic Commerce
ACM Conference on Recommender Systems
ACM european Conference on Computer Systems
ACM International Conference on Information and Knowledge Management
ACM International Conference on Interactive Tabletops and Surfaces (ITS)
ACM International Conference on Multimedia
ACM International Conference on Multimedia Retrieval
ACM International Conference on Web Search and Data Mining
ACM International Health Informatics Symposium
ACM Multimedia Systems Conference (MMSys)
ACM SIAM Symposium on Discrete Algorithms
ACM SIGCOMM Conference
ACM SIGGRAPH/Eurographics Symposium on Computer Animation
ACM SIGIR Conference on Research and Development in Information Retrieval
ACM SIGKDD International Conference on Knowledge discovery and data mining
ACM SIGMETRICS Performance Evaluation Review
ACM SIGMOD International Conference on Management of Data
ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPOPP)
ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL)
ACM SIGSOFT International Symposium on Foundations of Software Engineering

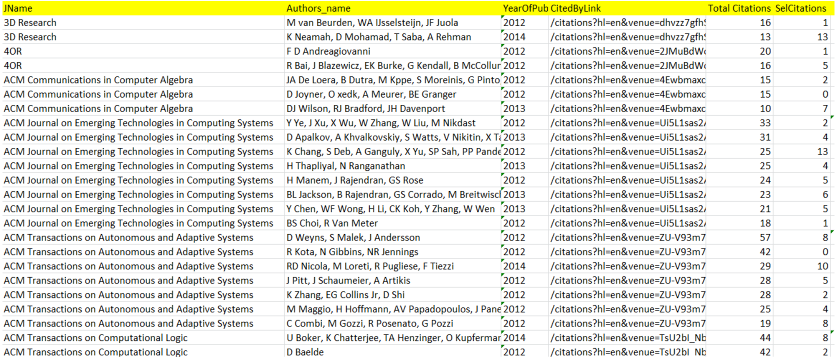
Fig. 5: Sample list of Journals extracted into spreadsheet.

5.3.3 Other-Citation Quotient Computation - Algorithm 3

Self-citation is defined as a citation where the citing and the cited paper share at least one author. Other-Citation is the complement of self-citation/total citations, i.e $1-$ self_citation/total citations. Algorithm 3 provides the skeleton of self-citation computation for an article in a journal. The denominator, total citations, is already computed by parsing web sources. The key to computing Other-Citations Quotient is to calculate self-citations. For this, we first scrape all the cited papers for the input article name (line 1). Then for each one of these cited papers check if it shares at least one common author name with the input article. If true then we increment the self-citation count (lines 3 and 4). Google Scholar lists a maximum of 1000 cited papers for any listed article. By adding all the individual self-citation counts for every article in a journal, we will get the total self-citations count for a journal (line 5 in Algorithm 1). Fig. 6 shows the output from this algorithm which is a raw data extracted to spreadsheet. Fig. 7 shows the processed raw data which provided the total self-citations for every listed journal.

Algorithm 3 $compute\_Self\_Citations()$: Algorithm to compute Self-Citation Count

```
**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 forreturn** $SelfCitationCount$
```



JName	Authors_name	YearOfPub	CitedByLink	Total Citations	SelCitations
3D Research	M van Beurden, WA IJsselsteijn, JF Juola	2012	/citations?hl=en&venue=dhvzz7gfh	16	1
3D Research	K Neamah, D Mohamad, T Saba, A Rehman	2014	/citations?hl=en&venue=dhvzz7gfh	13	13
4OR	F D Andreagiovanni	2012	/citations?hl=en&venue=2JMuBdW	20	1
4OR	R Bai, J Blazewicz, EK Burke, G Kendall, B McCollur	2012	/citations?hl=en&venue=2JMuBdW	16	5
ACM Communications in Computer Algebra	JA De Loera, B Dutra, M Kppe, S Moreinis, G Pinto	2012	/citations?hl=en&venue=4Ewbmaxc	15	2
ACM Communications in Computer Algebra	D Joyner, O xedk, A Meurer, BE Granger	2012	/citations?hl=en&venue=4Ewbmaxc	15	0
ACM Communications in Computer Algebra	DJ Wilson, RJ Bradford, JH Davenport	2013	/citations?hl=en&venue=4Ewbmaxc	10	7
ACM Journal on Emerging Technologies in Computing Systems	Y Ye, J Xu, X Wu, W Zhang, W Liu, M Nikdast	2012	/citations?hl=en&venue=Ui5L1sas2	33	2
ACM Journal on Emerging Technologies in Computing Systems	D Apalkov, A Khvalkovskiy, S Watts, V Nikitin, X Ta	2013	/citations?hl=en&venue=Ui5L1sas2	31	4
ACM Journal on Emerging Technologies in Computing Systems	K Chang, S Deb, A Ganguly, X Yu, SP Sah, PP Pande	2012	/citations?hl=en&venue=Ui5L1sas2	25	13
ACM Journal on Emerging Technologies in Computing Systems	H Thapliyal, N Ranganathan	2013	/citations?hl=en&venue=Ui5L1sas2	25	4
ACM Journal on Emerging Technologies in Computing Systems	H Manem, J Rajendran, GS Rose	2012	/citations?hl=en&venue=Ui5L1sas2	24	5
ACM Journal on Emerging Technologies in Computing Systems	BL Jackson, B Rajendran, GS Corrado, M Breitwiscl	2013	/citations?hl=en&venue=Ui5L1sas2	23	6
ACM Journal on Emerging Technologies in Computing Systems	Y Chen, WF Wong, H Li, CK Koh, Y Zhang, W Wen	2013	/citations?hl=en&venue=Ui5L1sas2	21	5
ACM Journal on Emerging Technologies in Computing Systems	BS Choi, R Van Meter	2012	/citations?hl=en&venue=Ui5L1sas2	18	1
ACM Transactions on Autonomous and Adaptive Systems	D Weyns, S Malek, J Andersson	2012	/citations?hl=en&venue=ZU-V93m7	57	8
ACM Transactions on Autonomous and Adaptive Systems	R Kota, N Gibbins, NR Jennings	2012	/citations?hl=en&venue=ZU-V93m7	42	0
ACM Transactions on Autonomous and Adaptive Systems	RD Nicola, M Loreti, R Pugliese, F Tiezzi	2014	/citations?hl=en&venue=ZU-V93m7	29	10
ACM Transactions on Autonomous and Adaptive Systems	J Pitt, J Schaumeier, A Artikis	2012	/citations?hl=en&venue=ZU-V93m7	28	5
ACM Transactions on Autonomous and Adaptive Systems	K Zhang, EG Collins Jr, D Shi	2012	/citations?hl=en&venue=ZU-V93m7	28	2
ACM Transactions on Autonomous and Adaptive Systems	M Maggio, H Hoffmann, AV Papadopoulos, J Pane	2012	/citations?hl=en&venue=ZU-V93m7	25	4
ACM Transactions on Autonomous and Adaptive Systems	C Combi, M Gozzi, R Posenato, G Pozzi	2012	/citations?hl=en&venue=ZU-V93m7	19	8
ACM Transactions on Computational Logic	U Boker, K Chatterjee, TA Henzinger, O Kupfermar	2014	/citations?hl=en&venue=TsU2bl_Nb	44	8
ACM Transactions on Computational Logic	D Baelde	2012	/citations?hl=en&venue=TsU2bl_Nb	42	2

Fig. 6: Sample data extracted into spreadsheet.

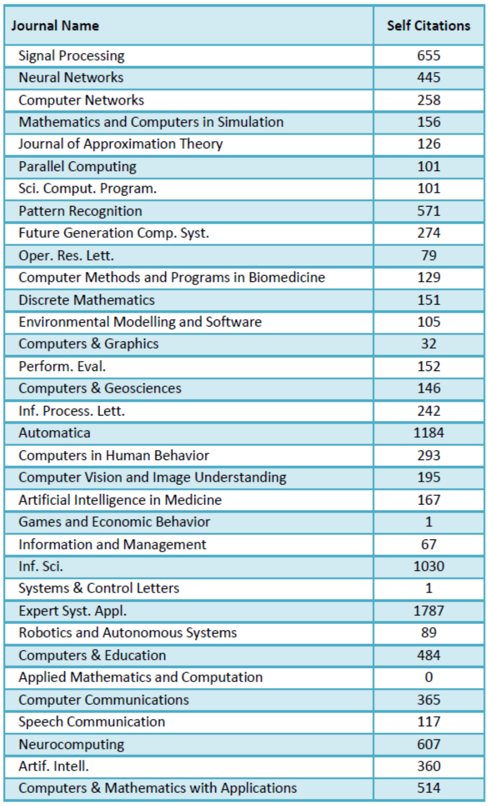
5.3.4 Non-Local Influence Quotient (NLIQ) - Algorithm 4

Influence is termed as a factor which causes a paper to be cited by other papers. Non-local refers to the fact that some citations originate from different journals; that is, not from the same journal in which the cited paper is published in. Thus, Non-Local Influence Quotient (NLIQ) is defined as follows,

Let **A** be the number of citations made from articles in one journal X to articles belonging to a number of different journals. Let **B** be the number of citations from articles in journal X to articles in the same journal, X. Then, for a given journal, we have:

$$\textbf{Non-Local Influence Quotient} = \frac{A}{A+B}$$

It must be stressed that **other-citations** are uniquely different from **non-local influence**. Namely, an other-citation occurs when a paper cites another paper where no authors are in common. On the other hand, non-local influence is the number of citations made from one paper in a given journal, to a number of different journals - divided by the total number of citations. Section 7 outlines the implications of collaboration and NLIQ towards ”internationality” and shows that SNIP values of journals are weakly correlated to their Non-Local Influence Quotient (NLIQ).



Journal Name	Self Citations
Signal Processing	655
Neural Networks	445
Computer Networks	258
Mathematics and Computers in Simulation	156
Journal of Approximation Theory	126
Parallel Computing	101
Sci. Comput. Program.	101
Pattern Recognition	571
Future Generation Comp. Syst.	274
Oper. Res. Lett.	79
Computer Methods and Programs in Biomedicine	129
Discrete Mathematics	151
Environmental Modelling and Software	105
Computers & Graphics	32
Perform. Eval.	152
Computers & Geosciences	146
Inf. Process. Lett.	242
Automatica	1184
Computers in Human Behavior	293
Computer Vision and Image Understanding	195
Artificial Intelligence in Medicine	167
Games and Economic Behavior	1
Information and Management	67
Inf. Sci.	1030
Systems & Control Letters	1
Expert Syst. Appl.	1787
Robotics and Autonomous Systems	89
Computers & Education	484
Applied Mathematics and Computation	0
Computer Communications	365
Speech Communication	117
Neurocomputing	607
Artif. Intell.	360
Computers & Mathematics with Applications	514

Fig. 7: Journals with corresponding self-citation counts.

Algorithm 4 *calculate_NLIQ*(): Algorithm to calculate Non-Local Influence Quotient

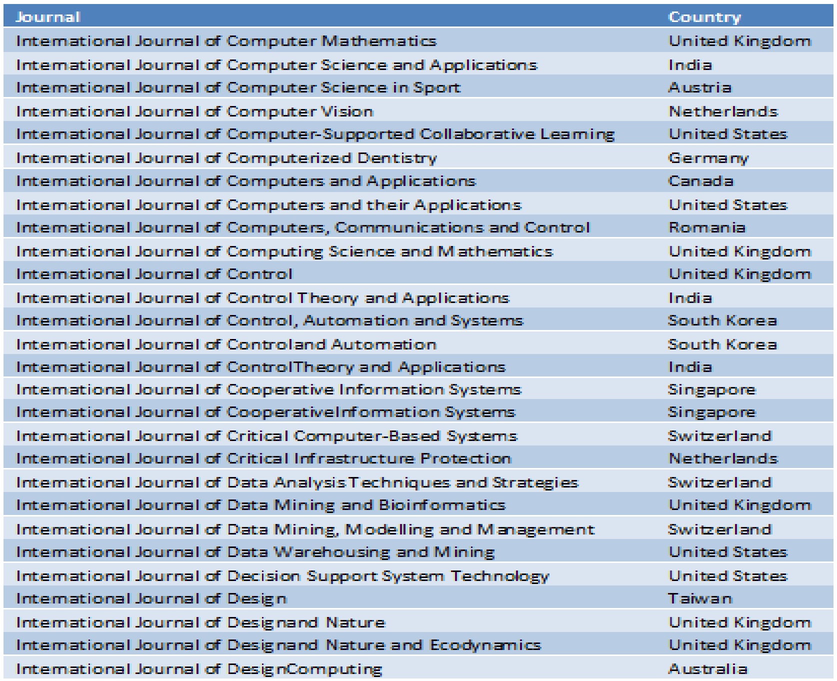
```
Input: $journal\_name, citation\_database$
Output: $NLIQ\ \ of\ journal\_name$
$A \leftarrow 0$                                              ▷ external citation count
$B \leftarrow 0$                                              ▷ internal citation count
$J\_articles \leftarrow [\ ]$                                 ▷ used to store articles in a journal
$count \leftarrow 0$
for each $article \in citation\_database$ do                  ▷ get all articles in a journal
    if $article[journal] = journal\_name$ then
        $J\_articles[count++] \leftarrow article$
    end if
end for
for each $article \in J\_articles$ do                         ▷ get count of internal, external cites
    for each $reference \in article[references]$ do
        if $reference \in ARTICLE\_TYPE$ then                 ▷ reference is an article
            if $reference[journal]$ != $journal\_name$ then
                $A \leftarrow A + 1$
            else
                $B \leftarrow B + 1$
            end if
        end if
    end for
end for
$NLIQ \leftarrow A\ /\ (A + B)$
$return\ \ NLIQ$
```

5.3.5 International Collaboration Ratio - Algorithms 5, 6, 7

International collaboration accounts for the articles that have been produced by researchers from several countries. In order to compute this parameter, we first extracted the country information of the journal and then the author affiliations for each one of the published articles in that journal. Every author's country is matched with the country of publishing journal. Ratio is calculated on the basis of weights assigned to different combination of authors affiliation and origin of the publishing journal. Algorithm 5 is for collecting the country information of the journals. Algorithm 6 shows computation of International Collaboration Ratio of a journal. Fig. 8 shows the sample of country names of a few listed journals.

Algorithm 5 $Country\_info(Journal\_name)$: Algorithm to fetch country information

```
Input: List of Journal names from Algorithm 1: $J$
Output: Country information of the Journals
for Every journal_name: $JNames[i]$ do
    $FetchURLhttp://www.scimagojr.com/journalsearch.php?q = " + journal\_name + "\&tip = jou"$
    $write\_to\_spreadsheet(forall('div', 'id' :' derecha\_contenido'))$
end for
```



Journal	Country
International Journal of Computer Mathematics	United Kingdom
International Journal of Computer Science and Applications	India
International Journal of Computer Science in Sport	Austria
International Journal of Computer Vision	Netherlands
International Journal of Computer-Supported Collaborative Learning	United States
International Journal of Computerized Dentistry	Germany
International Journal of Computers and Applications	Canada
International Journal of Computers and their Applications	United States
International Journal of Computers, Communications and Control	Romania
International Journal of Computing Science and Mathematics	United Kingdom
International Journal of Control	United Kingdom
International Journal of Control Theory and Applications	India
International Journal of Control, Automation and Systems	South Korea
International Journal of Controland Automation	South Korea
International Journal of ControlTheory and Applications	India
International Journal of Cooperative Information Systems	Singapore
International Journal of CooperativeInformation Systems	Singapore
International Journal of Critical Computer-Based Systems	Switzerland
International Journal of Critical Infrastructure Protection	Netherlands
International Journal of Data AnalysisTechniques and Strategies	Switzerland
International Journal of Data Mining and Bioinformatics	United Kingdom
International Journal of Data Mining, Modelling and Management	Switzerland
International Journal of Data Warehousing and Mining	United States
International Journal of Decision Support System Technology	United States
International Journal of Design	Taiwan
International Journal of Designand Nature	United Kingdom
International Journal of Designand Nature and Ecodynamics	United Kingdom
International Journal of DesignComputing	Australia

Fig. 8: Sample list of the Journals with country name

Algorithm 6 $Intl\_Collaboration\_Ratio(JNames[i])$: Algorithm to compute international collaboration ratio of a Journal

```
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: For all authors from different countries weight=1, For all authors from same country weight = 0, For n/2 authors from one country and n/2 from others weight=0.5
for Every article: $J.all\_articles\_url[i]$ do
    $iNtrNationality\_wt[i] = compute\_wt(article)$
end for    ▷ Generate the affiliations matrix of i:author names, j:country names
$authAffs[][] = get\_author\_affiliation(J.all\_articles\_url[])$    ▷ Add all author information to the internationality matrix of a journal i:country names, j:author names
$J.iNtrNational[][]$
for every authorname: $authAffs[i][]$ do
    for every country of author i: $authAffs[i][j]$ do
        if $authAff[i][j] == J[i]$ then    ▷ if the author's country and the country of Journal are same then make entry = 0
            J.iNtrNational[i][j] = 0
        else
            $J.iNtrNational[i][j] = 1$
        end if
    end for
end for
x = Ratio of(Number of 0's and Number of 1's in $J.iNtrNational[][]$)
y = cumulative weights($iNtrNationality\_wt[i]$)
%international_collaboration = x + (1-a) y
```

Algorithm 7 illustrates steps to fetch author affiliations of an article. An article URL is obtained by searching the article in IEEE and ACM websites. If found, the author affiliations is scraped and stored.

Algorithm 7 $Fetch\_Author\_Affiliations(article)$: Algorithm to fetch author affiliations information for the article

```
Input: Link to the article from algorithm 5: $article\_URL$
Output: Author names and respective Affiliations
$Fetchthecontentsfromarticle\_URL$
if $article\_URL$ IN List[IEEE, ACM] then
    SCRAPER($List$)    ▷ fuction to fetch the author affiliations from article's web page
end if
```

6 Cobb Douglas Model: Internationality Score function

In economics, Cobb-Douglas production function [35, 37] is widely used to represent relationship of outputs to inputs. This is a technical relation which describes the Laws of Proportion, i.e., the transformation of factor inputs into outputs at any particular time period. This production function is used for the first time, to compute the internationality of a journal where the predictor/independent variables, $x_i, i = 1, 2, ...n$ are algorithmically extracted from different sources as explained in the preceeding section. Internationality, y is defined as a multivariate function of $x_i, i = 1, 2, ...n$. Internationality score varies over time and depends on scholastic parameters, subject to evaluations, constant scrutiny and ever changing patterns.
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. The function has extremely useful

properties such as convexity/concavity depending upon the elasticity's. The properties yield global extrema which are intended to be exploited in the computation of internationality.
A sample Cobb-Douglas production function for two inputs, x_1 and x_2 and internationality of journal as output, y, is written as -

$$y = A{x_1}^{\alpha}{x_2}^{\beta}$$

where:

- $0 < \alpha, \beta < 1$
- y: Internationality of journal
- x_1: International Collaboration (percentage)
- x_2: SNIP (Source-Normalized Impact per Paper)

As explained in the subsequent sections, the sample model is easily extended to accommodate all relevant input/predictor variables extracted during the acquisition process [Please refer Section 5].

6.1 Functional Form

Here, x_1 and x_2 values which are modified values of international collaboration and SNIP respectively, are taken into consideration for different journals and using these optimal values for α and β is computed. In order to have data lying between 0-1, transformations on the data set is performed, which give the final input values for Cobb-Douglas production function. Finally, the two variables along with the elasticity values of α and β are placed in Cobb-Douglas equation to compute y. Following is the algorithm used:
Algorithm to find optimum values of α and β:
1. Input values of x_1, x_2
2. Vary α for x_1 such that the corresponding y reaches its maximum value. Similarly compute β values by varying x_2.
In the figure (Fig. 9) below it can be seen that y is maximum for α=0.1 and β=0.1

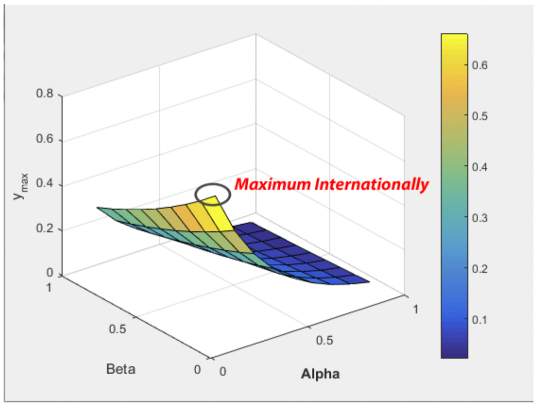


Fig. 9: Optimum values of α and β

Hence, the optimum values of α and β are 0.1. Using these values of α and β in the Cobb-Douglas production function

$$\mathrm{y} = A{x_1}^{\alpha}{x_2}^{\beta}$$

y is computed which represents internationality of journal. In the next subsection, regression method is used to compute α and β. The general model is endowed to accomodate any number of factors as proved in section 6.2. However in section 6 and 6.1, two factors are considered as illustration for the 3-D plot. As observed, α and β are varied to capture the maximum value of "y", internationality of the journal. If the number of input parameters are increased, visualization becomes untenable.

6.2 *Proof of Concept*

The section proves the efficacy of the model for n number of variables/inputs, n being countably finite
As a first exercise, authors have done the simulation for 2 variables and extended to 3. This can be extended to n variables as shown below.
Consider the following production function:

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

$n = 4$, x_1 to x_4 are the input parameters as described below:

- x_1 : Other-Citations Quotient = 1 - (self-citations / total citations)
- x_2 : International Collaboration / 100
- x_3 : SNIP value / maximum SNIP value
- x_4 : Non-Local Influence Quotient

Note: Due to the fact that these input parameters are not just raw numbers but instead defined as quotients having values between 0 and 1 - that is, they are normalized - the Cobb-Douglas model allows for a fair comparison between different subject fields where collaboration and citation trends differ dramatically (such as in Computer Science, the Social Sciences and Mathematics).

We have discussed the parameters, x_1, x_2 in section **5 - Data Acquisition**. This section following the analytical explanation of the Cobb-Douglas model, section 7 contains discussion on x_3 and x_4, since both are not merely acquisition oriented but deserve discussion by their own merit.

Lemma I: Maximum internationality score can be obtained at decreasing returns to scale which is true when -

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

where α_i is the i^{th} elasticity of the input variable x_i. Consider the following production function:

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

To prove:

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

Consider the profit function:

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

w_i: Unit cost of inputs
Profit maximization is achieved when: $p\frac{\partial f}{\partial x_i} = w_i$. Deriving the condition for optimization:

$$pk\frac{\alpha_1}{x_1}\prod_{i=1}^{n}x_i^{\alpha_i} = w_1 \tag{1}$$

$$pk\frac{\alpha_2}{x_2}\prod_{i=1}^{n}x_i^{\alpha_i} = w_2 \tag{2}$$

$$.$$
$$.$$
$$.$$
$$.$$

$$pk\frac{\alpha_n}{x_n}\prod_{i=1}^{n}x_i^{\alpha_i} = w_n \tag{3}$$

Multiplying these equations with x_i, respectively-

$$p\alpha_1\prod_{i=1}^{n}kx_i^{\alpha_i} = w_1x_1 \Rightarrow p\alpha_1 y = w_1x_1 \tag{4}$$

$$p\alpha_2\prod_{i=1}^{n}kx_i^{\alpha_i} = w_2x_2 \Rightarrow p\alpha_2 y = w_2x_2 \tag{5}$$

$$.$$
$$.$$
$$.$$
$$.$$

$$p\alpha_n\prod_{i=1}^{n}kx_i^{\alpha_i} = w_nx_n \Rightarrow p\alpha_n y = w_nx_n \tag{6}$$

Dividing equations (62) to (63) by (61), following equations are obtained:

$$x_2 = \frac{\alpha_2}{\alpha_1}\frac{w_1}{w_2}x_1$$
$$x_3 = \frac{\alpha_3}{\alpha_1}\frac{w_1}{w_3}x_1$$
$$.$$
$$.$$
$$x_{n-1} = \frac{\alpha_{n-1}}{\alpha_1}\frac{w_1}{w_{n-1}}x_1$$
$$x_n = \frac{\alpha_n}{\alpha_1}\frac{w_1}{w_n}x_1$$

Substituting these values ofx_i in equation (58),

$$pk\frac{\alpha_1}{x_1}\prod_{i=1}^{n}x_i^{\alpha_i} = w_1$$
$$\Rightarrow pk\alpha_1 x_1^{\alpha_1-1}\left(\frac{\alpha_2}{\alpha_1}\frac{w_1}{w_2}x_1\right)^{\alpha_2}\left(\frac{\alpha_3}{\alpha_1}\frac{w_1}{w_3}x_1\right)^{\alpha_3}....\left(\frac{\alpha_{n-1}}{\alpha_1}\frac{w_1}{w_{n-1}}x_1\right)^{\alpha_{n-1}}\left(\frac{\alpha_n}{\alpha_1}\frac{w_1}{w_n}x_1\right)^{\alpha_n} = w_1$$
$$\Rightarrow pk{x_1}^{(\alpha_1+\alpha_2+....+\alpha_n)-1}\alpha_1^{1-(\alpha_2+\alpha_3+....+\alpha_n)}\alpha_2^{\alpha_2}....\alpha_n^{\alpha_n}w_1^{-1+(\alpha_2+\alpha_3+....+\alpha_n)}w_2^{-\alpha_2}....w_n^{-\alpha_n} = 1$$
$$\Rightarrow x_1 = \left(pk\alpha_1^{1-(\alpha_2+\alpha_3+....+\alpha_n)}\alpha_2^{\alpha_2}....\alpha_n^{\alpha_n}w_1^{-1+(\alpha_2+\alpha_3+....+\alpha_n)}w_2^{-\alpha_2}....w_n^{-\alpha_n}\right)^{\frac{1}{1-(\alpha_1+\alpha_2+....+\alpha_n)}}$$

Performing similar calculations following values of $x_i, (i >= 2)$ are obtained,

$$x_2 = \left(pk\alpha_2^{1-(\alpha_1+\alpha_3+....+\alpha_n)}\alpha_1^{\alpha_1}....\alpha_n^{\alpha_n}w_2^{-1+(\alpha_1+\alpha_3+....+\alpha_n)}w_1^{-\alpha_1}....w_n^{-\alpha_n}\right)^{\frac{1}{1-(\alpha_1+\alpha_2+....+\alpha_n)}}$$

$$.$$

$$.$$

$$x_n = \left(pk\alpha_n^{1-(\alpha_1+\alpha_2+....+\alpha_{n-1})}\alpha_1^{\alpha_1}....\alpha_{n-1}^{\alpha_{n-1}}w_n^{-1+(\alpha_1+\alpha_2+....+\alpha_n)}w_2^{-\alpha_2}....w_{n-1}^{-\alpha_{n-1}}\right)^{\frac{1}{1-(\alpha_1+\alpha_2+....+\alpha_n)}}$$

Substituting values of x_i in production function,

$$y = \left(kp^{(\alpha_1+\alpha_2+....+\alpha_n)}\alpha_1^{\alpha_1}\alpha_2^{\alpha_2}....\alpha_n^{\alpha_n}w_1^{-\alpha_1}w_2^{-\alpha_2}....w_n^{-\alpha_n}\right)^{\frac{1}{1-(\alpha_1+\alpha_2+....+\alpha_n)}}$$

y increases in price of its output and decreases in price of its inputs iff:

$$1-\sum_{i=1}^{n}\alpha_i > 0$$
$$\sum_{i=1}^{n}\alpha_i < 1$$

Therefore decreasing returns to scale, is validated.

6.3 Proof of Concavity of Cobb-Douglas function using Hessian Matrix

This section proves that the Cobb-Douglas [36, 37] production model is concave in nature and hence a maximum internationality score can be found at a particular value of input factors which in this case are international collaboration and SNIP value. The concavity of the function is proved by showing that the Hessian Matrix of the function is negative semi-definite.

Definition:

1. Suppose $f \in C^2$, U is an open curve set, then $f: U \subset R^n \to R$ is concave/strictly concave iff the Hessian Matrix $D^2f(x) = H$ is negative semi-definite/ negative definite $\forall x \in U$.
   C^2: Class of continuous and second order differential functions[11].
2. Let S be a convex set [12]; x_1, x_2 be any two points in S; then a function $f: S \subset R_n \to R$ is concave if,

$$(1-\lambda)f(x_1)+\lambda f(x_2) \le f((1-\lambda)x_1+\lambda x_2); \qquad \lambda \in [0,1]$$

3. Constant and Decreasing Returns to scale: [14] In the phase of constant returns, an increase in one input may yield an increase in corresponding output in the same proportion. The 3D plots obtained are concave.
   Whereas, In decreasing returns to scale the deployment of an additional input will result into increase in output but at a diminishing rate or lower ratio.

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

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

C: Class of continuous and first order differential functions,

Proof: Using the definition of concave functions;

$$f(\alpha(x+\theta)+(1-\alpha)x) \ge \alpha f(x+\theta)+(1-\alpha)f(x)$$
$$\Rightarrow f(x+\alpha\theta)-f(x) \ge \alpha(f(x+\theta)-f(x))$$
$$\Rightarrow f(x)+\frac{f(x+\alpha\theta)-f(x)}{\alpha} \ge f(x+\theta)$$
$$\Rightarrow f(x)+\nabla f(x)\theta \ge f(x+\theta) \quad as \quad \alpha \to 0$$

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]

Proof: f is concave, for some $x \in U$ and some $\theta \neq 0$, consider the Taylor expansion;

$$f(x+\alpha\theta) = f(x) + \nabla f(x)(\alpha\theta) + \frac{(\alpha\theta)^2}{2} D^2 f(x+t\theta) \;\; for\, some\; 0 < t < \theta$$

By lemma;

$$\frac{(\alpha\theta)^2}{2} D^2 f(x+t\theta) \leq 0$$

Consider an arbitrary $\alpha \to 0$, and $t \to 0$

$$\theta^2 D^2 f(x) \leq 0 \Rightarrow D^2 f(x) \leq 0 \Rightarrow H\; is\; negative\; semi-definite.$$

6.4 Implications of Theorem 1:

Cobb-Douglas is concave for conditions on elasticity, thus for such values of elasticity, the Hessian Matrix of the function is negative semi-definite and therefore concave and attains a global maxima.
Now consider, the Cobb-Douglas function;$f(x_1,x_2) = kx_1^\alpha x_2^\beta withk, \alpha, \beta > 0 fortheregionx_1 > 0andx_2 > 0$

$$H = \begin{bmatrix} \alpha(\alpha-1)kx_1^{\alpha-2}x_2^\beta & \alpha\beta kx_1^{\alpha-1}x_2^{\beta-1} \\ \alpha\beta kx_1^{\alpha-1}x_2^{\beta-1} & \beta(\beta-1)kx_1^\alpha x_2^{\beta-2} \end{bmatrix}$$

First order principal minors [13] of H are:

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

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 $M_1 \leq 0,\; M_1' \leq 0\; and\; M_2 \geq 0$
For decreasing and constant returns to scale: $\alpha+\beta \leq 1$, 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

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

Significance of concavity:
The extrema of the function, $f(x,y)$ used to model ”internationality” is useful in finding a global maximal value of the ”internationality” indicator. The modeling paradigm is based on the fact that, there exists a maximum internationality score and the score/values in the neighborhood could be classified as the levels of internationality. It is, in this context,

we explore if the maxima given by the concave function, i.e.Cobb-Douglas is the global maxima.

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

Proof: x^* is a critical point. Therefore; $Df(x^*) = 0$ [D: first order partial derivative]
Using a well known result about concave functions;
$f : R^2 \rightarrow R$ is concave iff $f(x_2) - f(x_1) \leq Df(x_1)(x_2 - x_1)\ \ \forall x_1, x_2 \in U$;
Therefore;
$f(x_2) - f(x_1) \leq \frac{\partial f(x_1)}{\partial (x_1)}(x_2' - x_1') + \ldots \frac{\partial f(x_2)}{\partial (x_2)}(x_2^2 - x_1^2)$
Since,
$Df(x^*) \equiv 0$ using the inequality
$f(x_2) - f(x^*) \leq Df(x^*)(x_2 - x^*) \Rightarrow f(x_2) \leq f(x^*)\ \ \forall x_2 \in U$

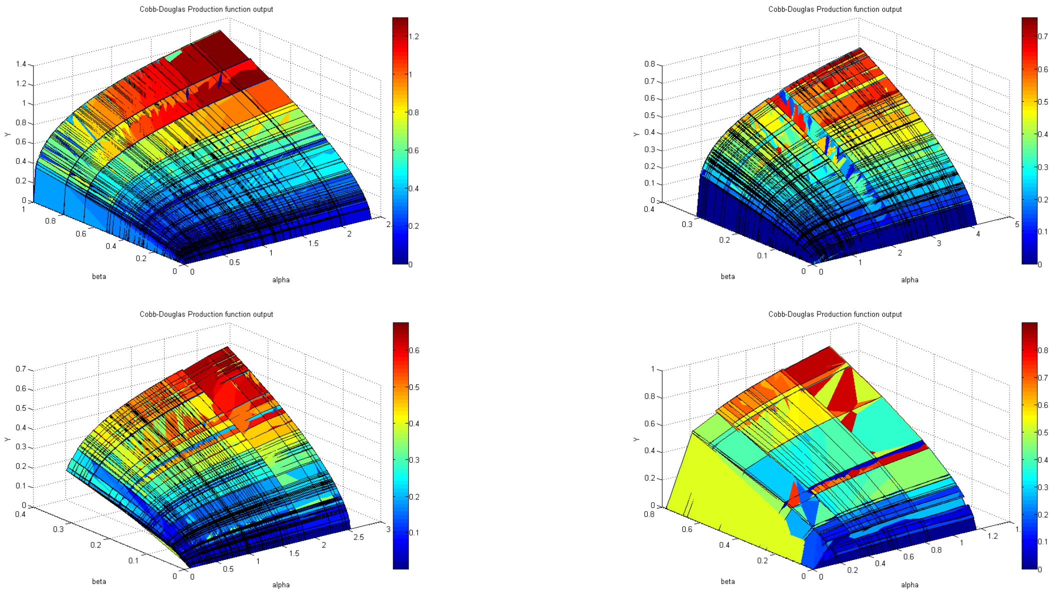


Fig. 10: y, internationality values from Cobb-Douglas Production function at various instances [27]

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 $f : U \subset R^n \rightarrow R$ & the global maxima holds.
2. U doesn't necessarily be open, the global maxima is guaranteed to be on the closed set as well, since the search for global maxima is allowed on the boundary.
3. Any Cobb-Douglas 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 corroborates **Lemma II** proved above.

A 3-D graph of Cobb Douglas function with output measured along the vertical axis is shown in Fig 10. The graph is a part of an AVI file whose frame are created in MATLAB to demonstrates the quasiconcave nature of Cobb Douglas model and to show how y reaches its maximum value at certain input values of x_1, x_2 α and β. The Matlab code can be viewed on GitHub, Appendix I [32]. The input parameters (x_1, x_2) are SNIP and other-citations/total-citations and the elasticity coefficients (α, β) are taken along X, Y coordinates. Lower values of y is indicated in blue region which increases and at certain values of elasticity coefficients, reaches to its maximum as marked in red. This is a sample representation and can't include more than two input parameters.

7 SNIP and Non-Local Influence Quotient, a new metric definition

The authors have taken a four-pronged approach to thoroughly validate the use of SNIP and NLIQ in the Cobb-Douglas model. First, we shall show the merits of Source-Normalized Impact per Paper (SNIP) over Impact Factor (IF) [28]. Then, the algorithm used to compute SNIP is described and verified with a sample data set. The third sub-section looks at citation patterns with increased granularity; namely inter-journal and intra-journal collaboration which will help show that SNIP is a good indicator of collaboration at the journal level. Lastly, we will look at Non-Local Influence Quotient (NLIQ) described in section 5.3.4 and justify its usage in calculating the internationality of a journal and show why SNIP alone should not be used to determine the relative ranks of journals in academia.

7.1 Comparison of SNIP and IF

Now, Source-Normalized Impact per Paper (SNIP) measures a source's contextual citation impact. It takes into account characteristics of the source's subject field, especially the frequency at which authors cite other papers in their reference lists, the speed at which citation impact grows, and the extent to which the database used in the evaluation covers the fields documents. SNIP is the ratio of a source's average citation count per paper, and the citation potential of its subject field. It aims to allow direct comparison of sources in different subject fields due to the subject-field normalization that takes place in calculating it.

The impact factor (IF) of an academic journal is a measure reflecting the average number of citations to recent articles published in that journal. In any given year, the impact factor of a journal is the average number of citations received per paper published in that journal during the two or five preceding years.

SNIP offers several advantages:

1. **Openly Available and Greater Coverage** How IF is calculated and the source database for citations is known only to Thomson Reuters (ISI Web of Science) which means journals not present in their database are not assigned an IF value. Also, not all journals indexed by them are provided an IF. This disallows researchers from comparing journals which are not indexed.
   Scopus, on the other hand, provides journal metrics values to all peer-reviewed journals indexed in their database which is comparably larger. Furthermore, SNIP can be calculated from any Open Access journal using the white paper describing the calculation of SNIP. This allows one to compare journals, however, the types of citations taken into account from Open Access journals must be kept in mind to give as fair a comparison as possible.
2. **Subject Field Normalization** Life Sciences have a much higher IF as compared to Mathematical journals due to the differences in citation behavior between the two fields.The quality of a journal cannot be derived from its Impact Factor. Due to the fact that SNIP inherently normalizes for differences in citation practices across subject fields, comparison of the prestige of two journals belonging to difference subject fields is possible.
3. **Citation Window** SNIP has an ideal citation window, in the authors opinion. A three-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.
4. **More Difficult to Game the System** A journals impact factor is derived from citations of all types of content - including non-peer reviewed material such as editorials. On the other hand, SNIP is derived only from citations of peer-reviewed content and directed to peer-reviewed content, which makes it much more difficult to game the system as the content goes through some form of scrutiny vis-a-vis editorials.
   Further, given the dramatic increase in predatory journals in recent years who merely charge a fee for publishing an authors paper albeit with deceitful tactics; and their the proportional increase in their IF values - it is clear that IF is not a suitable metric for measuring the 'prestige' of a journal.

7.2 Algorithm to Compute SNIP

As shown in the original SNIP indicator designed by Henk F. Moed [29] the SNIP indicator is defined as the ratio of a journal's raw impact per paper (RIP) and a journal's database citation potential in its subject field (DCP), that is the RIP value of a journal equals the average number of times the publications of that journal were cited in the three years the year of analysis. For example, if 200 publications were present in a journal from 2009 to 2011 and if these publications were cited 400 times in 2012, the RIP value of the journal for 2012 would be 400 / 200 = 2. In calculating RIP, both citing and cited publications are included only if they have the Scopus document type article, conference paper or review - i.e peer reviewed material. RIP is similar to journal impact factor (IF), although RIP uses three instead of two years of cited publications and only includes citations to the previously mentioned document types. RIP does not account for differences in citation practices among different journals.
The DCP value of a journal is equal to the average number of references in the publications belonging to the journal's subject field, where the average is calculated as the arithmetic mean. By finding the ratio of a journal's RIP to it's DCP, we can compare journals belonging to two different fields in a more fair manner. Algorithm 8 shows how to calculate the same.
Although there are certain differences between the original SNIP indicator [29] and the revised SNIP indicator, the authors decided to forgo the latter. This is because an empirical analysis was done between the two and Ludo et al. stated that "from an empirical point of view the differences between the original SNIP indicator and the revised one are relatively small" [12].

Algorithm 8 $calc\_SNIP(cites[][], Jpub[], Jsize)$: Algorithm to calculate SNIP

```
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
$journal \leftarrow Jname$
$year \leftarrow$ read $year\_to\_be\_computed\_for$
$citation\_count \leftarrow 0$
for all paper in Jpub do
    $citation\_count \leftarrow citation\_count$ + count of papers published in year - 1, year - 2, year - 3
    $num\_papers \leftarrow num\_papers + 1$
end for
$RIP \leftarrow citation\_count / num\_papers$
$DCP \leftarrow$ Average number of 1-3 year old cited references contained in papers in the dataset citing the target journal
$median \leftarrow$ median DCP of all journals
$RDCP \leftarrow DCP / median$
$SNIP \leftarrow RIP / RDCP$
return $SNIP$
```

A random set of journals in Computer Science and Mathematics were selected and Algorithm 8 was used to calculate SNIP for these journals. The SNIP values thus obtained for the year 2010 with citation data taken from the Aminer Citation Network Data Set [26] were compared with their actual values for the same journals and same year provided by Journal Metrics [3, 4]. The values thus obtained were not on par in terms of sheer magnitude with the corresponding values provided by Journal Metrics due to two main reasons. Firstly, our database is a fraction of the one used by Journal Metrics in terms of size, and further, some citations in our database may not have been included by Journal Metrics in calculating SNIP - and vice-versa.

However, on further analysis using regression, we were able to show that there does indeed exist a strong correlation between the two values - SNIP calculated by us and the actual SNIP values provided by Journal Metrics. Figure 11 shows a linear regression line, fitting calculated SNIP and actual SNIP. We obtain an R-squared value of 0.7363 meaning 73.63% of the variance in actual SNIP is accounted for by the variance in calculated SNIP.

To further validate our algorithm, we calculated SNIP for 189 journals which were common with both the Aminer data set and Journal Metrics' data set (from Elsevier). In this case, simple linear regression would not suffice; support vector regression (SVR) was used instead [23]. Consider a set of linearly separable points, then the support vectors are those points which are difficult to classify and have a direct influence on the optimal location of the

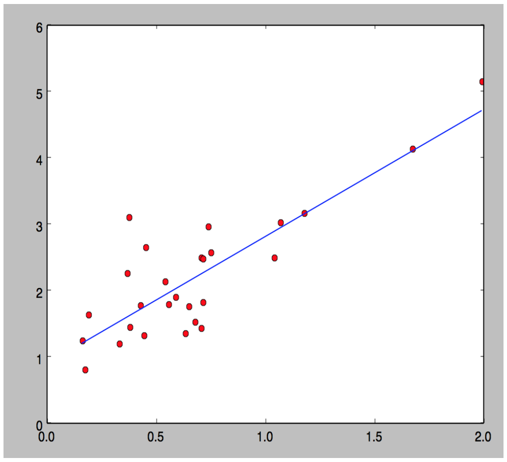
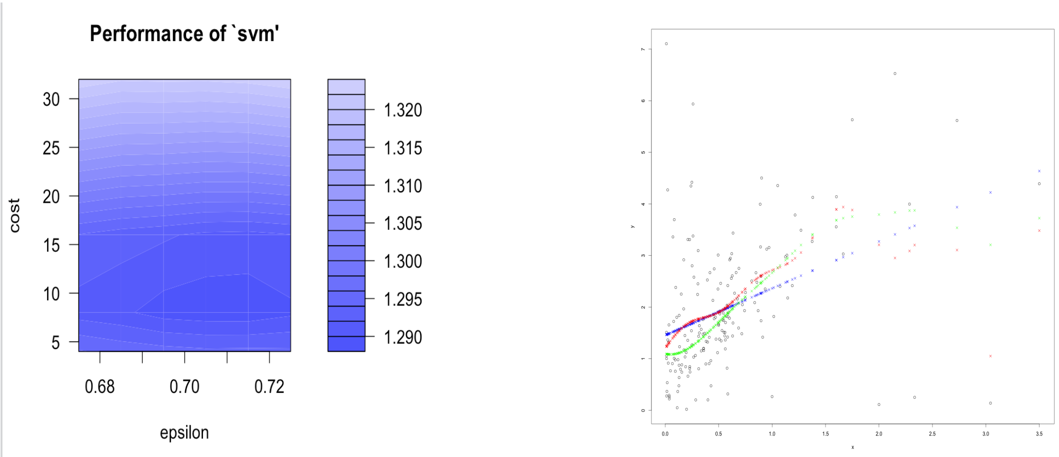


Fig. 11: Actual SNIP v/s Computed SNIP

decision boundary. SVR is designed to find an optimal hyperplane which divides the two sets of linearly separable points such that an ε-margin from either of these support vectors is obtained.

Initially, an RMSE of 1.162693 was obtained for linear regression and 1.163683 for SVR, without any tuning. Further tuning was performed by changing the values of ε and cost. The range of ε was narrowed from (0,2) to (0.68,0.72) with the cost parameter narrowed from 2^9 to 2^2. As a result, an RMSE of 1.116192 was obtained. Figure 12(a) shows how performance varies with ε and cost, with darker blue areas indicating optimal performance.



(a) Performance of SVR

(b) The tuned SVR model

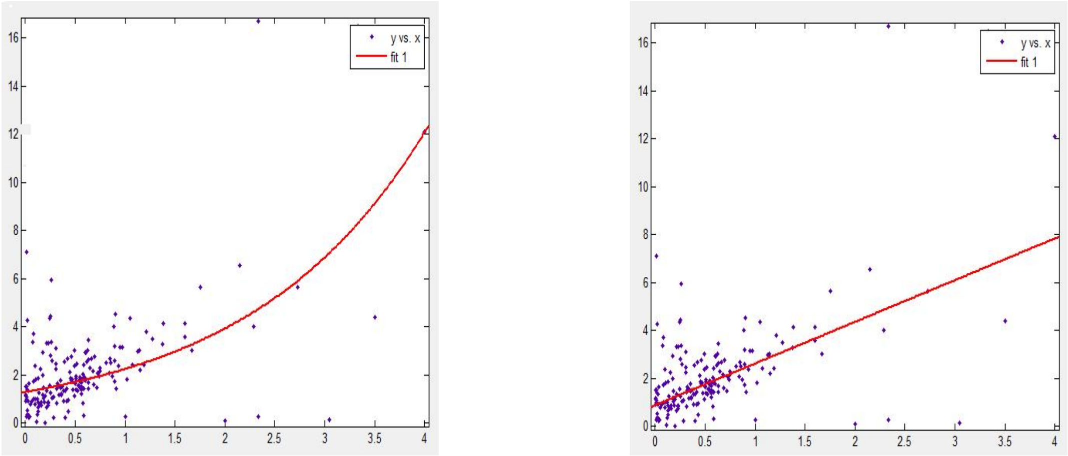
Fig. 12: Results of Support Vector Regression

The graph in figure 12(b) shows the linear regression model in blue, untuned SVR model in red and the tuned SVR model in green which gave us the best fit.
To corroborate the results from SVR, exponential and polynomial regression was also performed on the same data set. The exponential model shown in Fig.13(a) returned an R^2 value of 0.9868. The relationship between the Y (aSNIP) and X (cSNIP) along with the coefficients (with 95% confidence bounds) is best represented by the following equation:

$$Y = 1.273 * e^{0.5626*X} \tag{7}$$

The polynomial regression model shown in Fig.13(b) also resulted in an R^2 value of 0.9866. The equation obtained from polynomial regression is:

$$Y = 1.741 * X + 0.8641 \tag{8}$$



(a) Exponential Regression

(b) Polynomial Regression

Fig. 13: Curve fit of Exponential and Polynomial Regression

We can confidently conclude - with the help of the derived RMSE and R-squared values, graphs and equations - that support vector regression, polynomial regression and exponential regression are all suitable for predicting actual SNIP values from their calculated values, with a very high degree of certainty.

7.3 Journal Collaboration

7.3.1 Inter-Journal Collaboration

The same journals used in section 7.2 Fig. 11 are considered in this section as well. The citation data was taken from Aminer Citation Network Data Set [26] and correlated with journals having SNIP values taken from Journal Metrics [3, 4]. The citation network was then constructed as follows - the nodes are taken as journals, the size of each node is relative to its SNIP value taken from Scopus. The edge between two journals is a citation between a paper in one journal to a paper in another journal, where the color gradient of the edge is relative to the number of citations in total using Algorithm 9, as shown below in Fig. 14.
Figure 15 shows the citation network when we separate the edges on the basis of the table is given in Fig. 16, namely, those originating from and ending at a journal having a SNIP value above the median, and the same for below the median, as well as those going from a journal having a SNIP value above the median to below and vice versa.

When we split the citations up into four groups (see Fig. 15, 16) - namely, out of 19,359 inter-journal citations, a very large majority - 57.962% - are between journals having SNIP values above the median value versus only 5.103% between journals below the median. The citation network between journals of low SNIP value is quite sparse while that between journals of high SNIP value is dense - this indicates there is far more collaboration among journals of higher prestige or ranking, and little to no collaboration between those journals having a lower SNIP value, despite the fact that half of the journals taken into account were those below the median. This justifies our use of SNIP as a metric for collaboration - higher the SNIP, more the collaboration.

One must be careful to note, these numbers could imply that not only do authors tend to favor their papers being published in prestigious journals, but they also cite those papers present in journals of similar level, or papers of

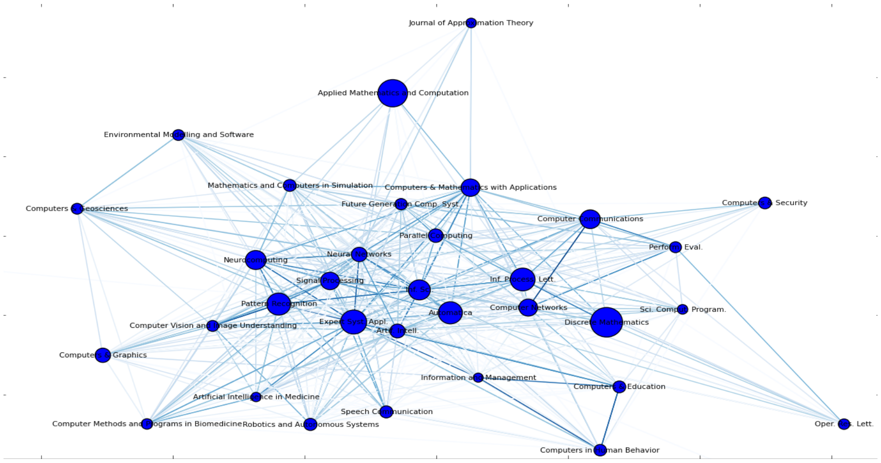
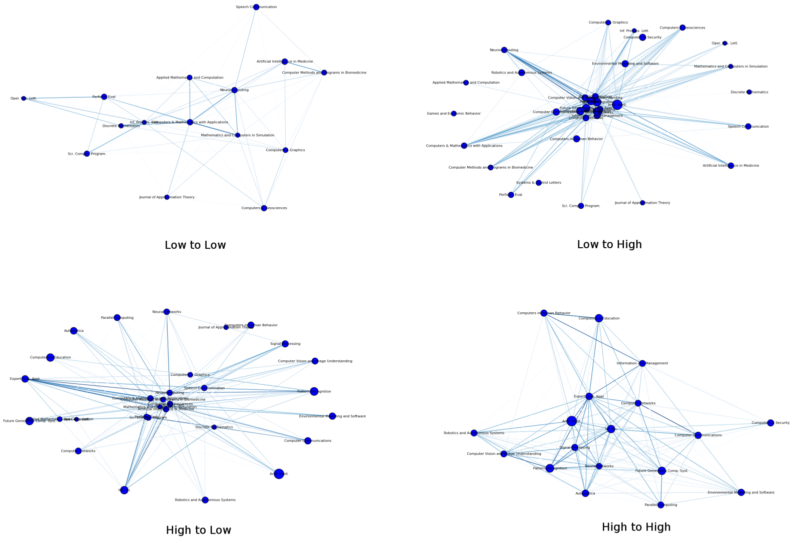

Journal of Appr mation Theory
Applied Mathem nd Computation
Environmental M ling and Software
Mathematics and C uters in Simulation
Computers & Mathe s with Applications
Computers eosciences
Future Generat Comp. Syst.
Computer C munications
Compute Security
Parallel puting
Perfor Eval.
Neural works
Neur uting
Signal ssing
Inf. P Lett
Patter gnition
Computer Vision an age Understanding
Exper Syst Appl.
Art ell
Comput tworks
Sci. Comp Program.
Discre matics
Computer Graphics
Information a Management
Compute Education
Artificial Intelli ce in Medicine
Speech C unication
Computer Methods and rograms in Biomedicine
Robotics and Au omous Systems
Oper. Lett.
Computers in man Behavior

Fig. 14: Citation network between all journals

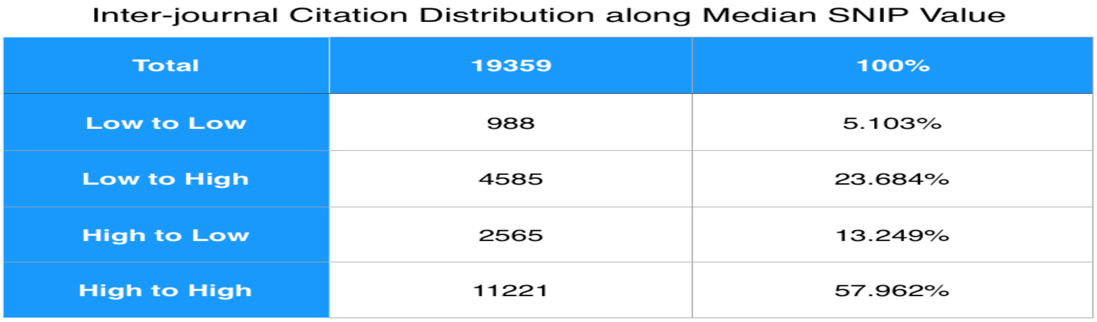

Low to Low

Low to High

High to Low

High to High

Fig. 15: Citation network between journals separated along median SNIP value


Inter-journal Citation Distribution along Median SNIP Value

Total	19359	100%
Low to Low	988	5.103%
Low to High	4585	23.684%
High to Low	2565	13.249%
High to High	11221	57.962%

Fig. 16: Inter-Journal Citation Distribution

Algorithm 9 *InterJ_Collaboration* : Algorithm to show Inter-Journal Collaboration graph

```
Input: $databse\ of\ citations$
Output: $graph\ and\ adjacency\ matrix\ of\ inter-journal\ collaboration$
$data \leftarrow read\ aminer\_cites$
$journals\_low, journals\_high \leftarrow$ SplitByMedian(data)
$G \leftarrow DiGraph$
for all publication in data do
    $papers \leftarrow data[publication]$
    for all paper in papers do
        $cites \leftarrow data[publication][paper]$
        for all cite in cites do
            $src \leftarrow publication$
            $dest \leftarrow cite['publication']$
            if src = dest then                                   ▷ self cite within publication
                continue
            end if
            if src in journals_low and dest in journals_low then
                $type \leftarrow 1$
            end if
            if src in journals_low and dest in journals_high then
                $type \leftarrow 2$
            end if
            if src in journals_high and dest in journals_low then
                $type \leftarrow 3$
            end if
            if src in journals_high and dest in journals_high then
                $type \leftarrow 4$
            end if
            for all type $\leftarrow$ (1, 2, 3, 4) do
                if edge(src, dest) in G  then
                    $G[src][dest]['weight'] \leftarrow G[src][dest]['weight'] + 1$
                else
                    $G[src][dest]['weight'] \leftarrow 1$
                end if
            end for
        end for
    end for
end for
$G1 \leftarrow G$
for all edge_weight in G1 $\leftarrow$ edges do
    $edge\_weight \leftarrow log10(edge\_weight)$                ▷ Normalize weight
end for
for all type $\leftarrow$ (1, 2, 3, 4) do
    PlotNodes($G1$)
    PlotEdges($G1$)
    PlotAdjacencyMatrix($G$)
end for
```

the same journal itself. In turn, a cycle is created - authors who publish in prestigious journals are cited more often than those who publish in less prestigious one - thereby increasing the apparent prestige of that journal due to the increased citation count. Whether these citations are genuine or simply reciprocal in nature is not known.

This factor also fuels the growth of predatory journals with nary an oversight in terms of authentic peer-review - less prestigious journals exist with minimal collaboration simply because there was a low bar for a paper to be accepted.

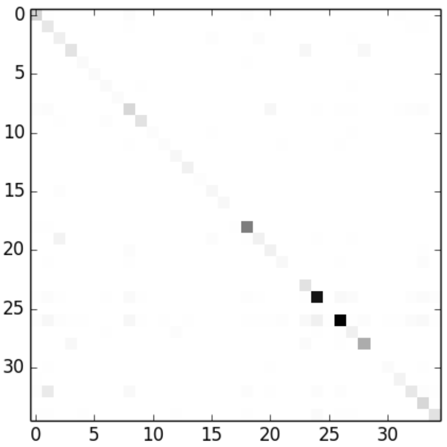


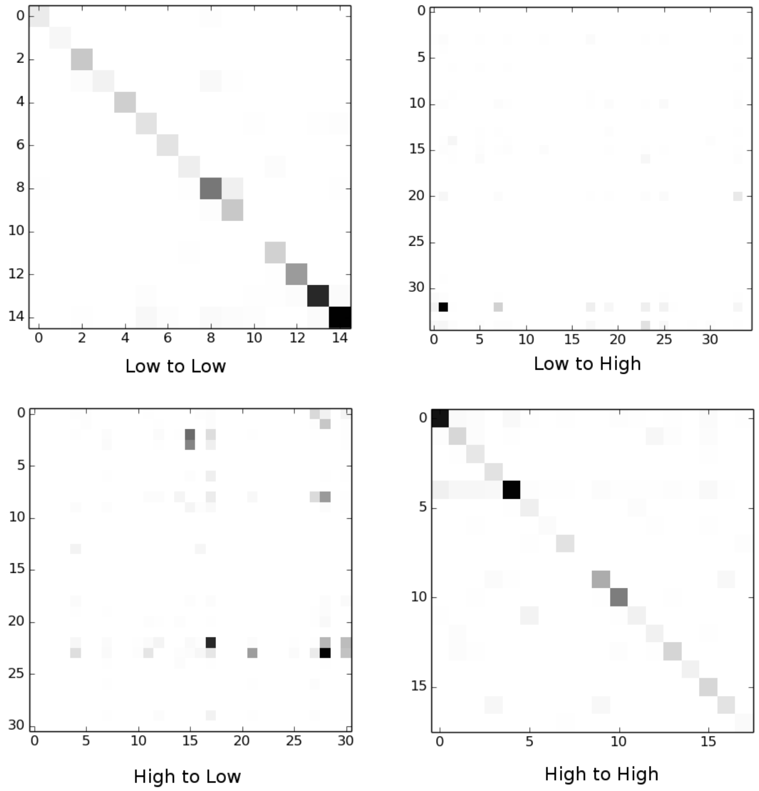
Fig. 17: Adjacency matrix for citations between all journals

7.3.2 Intra-Journal Collaboration

The adjacency matrix between the journals shown in Fig. 14 and 15 are given in Fig. 17 and 18, with the color gradient being relative to the total number of citations between the journals. It is observed that there exist darker points along the principal diagonal indicating more collaboration within a journal as opposed to inter-journal collaboration.

Papers published in one journal cite papers from the same journal much more often than those from different journals, regardless of the journal's SNIP value. This, too, leads to a cycle wherein an individual journal's prestige is increased by virtue of increased citations from within. It should be noted that journals of higher SNIP value have a lower NLIQ value as shown in Fig. 19, compared to journals of lower SNIP value - meaning citations are mostly restricted to the same journal they originate from. This in no way implies that there is a correlation between the two (as shown in section 7.4); merely revealing that journals most people would consider to be highly ranked (i.e by having higher SNIP values) exhibit only a low level of non-local influence. Evidently, information about the internationality of these journals is incomplete - whether the authors are from the same institution or the same country or merely citing their previous works or those of colleagues due to reciprocity, as previously mentioned - is not known.

These are all factors that can be heavily gamed to enhance the prestige and rank of an author as well as the journal their papers are published in. Hence, we proposed NLIQ in section 5, which favors inter-journal collaboration as opposed to intra-journal collaboration thereby accounting for non-local diffusion of influence and fortifying our definition of internationality.



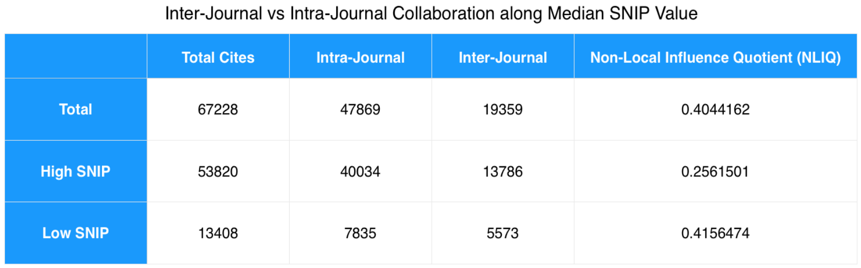
Low to Low

Low to High

High to Low

High to High

Fig. 18: Adjacency matrix for citations between journals separated along median value



Inter-Journal vs Intra-Journal Collaboration along Median SNIP Value

	Total Cites	Intra-Journal	Inter-Journal	Non-Local Influence Quotient (NLIQ)
Total	67228	47869	19359	0.4044162
High SNIP	53820	40034	13786	0.2561501
Low SNIP	13408	7835	5573	0.4156474

Fig. 19: Non-Local Influence Quotient Statistics

7.4 Non-Local Influence Quotient, NLIQ

Reiterating the definition from section 5.3.4, NLIQ is the number of citations made by articles published in a journal X to articles published in different journals divided by the total number of citations made by all papers in that journal X. Clearly, higher the number of external citations made by articles in a journal, higher the NLIQ of that journal.

In Fig. 20, we see a plot of SNIP on X-axis versus NLIQ on the Y-axis. Even though it appears at first that journals with low SNIP values tend to have higher NLIQ, once we look into the goodness of fit and correlation statistics, we see that there is an insignificant relationship between the two. A linear regression line is fit and the R-squared value obtained is 0.1681 and the cross-correlation coefficient as -0.41 at lag 0 and near 0 at lag -1 and +1. A cross correlation value not close to -1 or +1 indicates that there is little to no correlation between SNIP and NLIQ. Similar R-squared values were obtained for higher degree polynomial regression models.

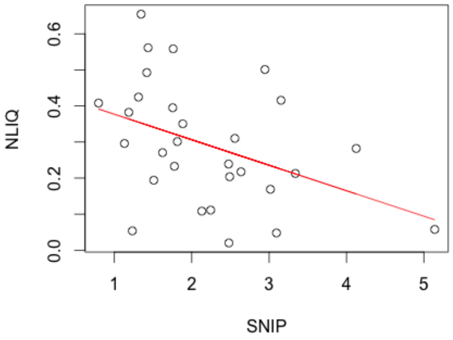


Fig. 20: SNIP vs NLIQ

Further, even though we don’t possess the entire database of citations, we have proven that the Aminer data set has sufficient and even coverage which allowed us to calculate SNIP to a high degree of accuracy - and thus calculating NLIQ with a complete database will not vary by much, either.

Clearly, SNIP is a good indicator of impersonal influence and therefore is used as one of the parameters for computing JIMI, but does not distinguish between the type of collaboration; intra-journal or inter-journal. NLIQ on the other hand is able to differentiate between the two types. It is vital to differentiate between the two; take for example a set of authors who constantly publish papers in one particular journal. They are only collaborating with people in the same area. On the other hand, if papers in different journals and subject areas are cited, cross-collaboration is established. This form of collaboration is indicated by NLIQ, which will be used alongwith SNIP in the Cobb-Douglas model for computing the internationality of a journal.

8 Discussion, Conclusion and Future Work

Internationality has thus been defined as the degree to which a journal transcends local communities and boundaries, with respect to the quality of publication and influence. The methods illustrated in the paper ensures disposition of any kind of local influence that unreasonably boosts scholarly impact of journals. The paper meticulously defines internationality of peer-reviewed journals as a measure of influence that spreads across boundaries and attempts to capture different and hitherto unperceived aspects of a journal for computing internationality.The current work utilizes parameters like ”International Collaboation Ratio” which incorporates participation of authors from different demographic regions. Authors humbly submit that ”true” internationality of a scholarly publication is necessarily contextual and must be devoid of local or community influence. Source-Normalized Impact per Paper (SNIP) is another parameter taken into account which normalizes the citation pattern within a subject field allowing the comparison of journals belonging to two different domains. This is one of the reasons why authors preferred SNIP over Thomson Reuters Impact Factor (details in section 7.1).

Authors wish to articulate that SNIP alone is not sufficient to compute influence and hence the necessity of defining a metric, NLIQ arises, which disregards the local influence diffusion within a journal. Authors believe that any artificial enhancement of a journals influence through coercive citation can be effectively subsided by NLIQ. The paper also uses a parameter - other-citation/total citation which reflects a journals integrity owing to the fact that no legitimate journal will promote authors and allow them to indiscreetly cite their own work. Parameters are chosen to ensure that every induced boost to a journals influence through coercive citations, extensive self-citations or through any other mechanism is negated by the authors model of internationality computation wherein other-citations are promoted.

Exclusive algorithms are written to scrape metric information from the web. Algorithms to sweep journal, author and article level information needed scrutinization of web pages. However, thanks to the simplicity of Python, it made most of the scraping task hassle-free incurring lesser overhead. Acquisition of journal names, origin, article names, author names and their affiliations were the first few steps in tailoring the parameters of internationality index. These attributes don't contribute to the model directly but are pivotal to creating the indigenous database for subsequent computation. Data acquisition and build are therefore significant contributions of the paper and should not be overlooked.

The acquired data was validated using different regression techniques. It was observed that values obtained from SNIP algorithm were not very close to the original ones. This is because the native acquisition algorithms could not have scraped through all databases due to the prohibitory firewalls built in several of those. Predictive analytic techniques are used to overcome such barriers so that data recorded in our database is reasonably accurate, endowed with appreciable " goodness of fit " statistic. The derived RMSE and R-squared values from support vector regression (SVR), linear regression, polynomial and exponential regression were found to be satisfactory. This concludes that any of these regression methods can be used to predict original SNIP (for detailed analysis, refer section 7.2).

Cobb-Douglas Production Function is used for the first time to model internationality of journals. Appropriateness and adequacy of the model is evaluated in section 6. It is shown that a function is strictly concave if the Hessian Matrix of the second order partial derivatives is negative semi-definite (Theorem 1). The property holds true for the production function implying that the function is strictly concave in nature given that certain conditions on elasticity are met. The importance of concavity lies in validating that the maximum value obtained by the production function is actually a global maxima (Theorem 2 proves this) and the search for such maxima via the model is complete once the maxima is found,by simulation and otherwise (Fig. 10). This maxima is then used as an indicator of highest internationality score and the subsequent neighborhood values may define lower international levels for the same journal. This process is iterated for all journals sweeping through the database.

Painstaking care has been exercised in creating a knowledge base of citation pattern followed by authors when they publish their work. To investigate the pattern, inter-journal collaboration network was created from Aminer and Journal Metric datasets. Dense citation network between journals of high SNIP values validated the fact that authors are not only tempted to publish their work in prestigious journals but are also inclined to cite papers of journals having a higher SNIP value. By doing so, receiving citation in large numbers is assured. The intra-journal collaboration network, on the other hand, is a reflection of authors tendency to cite the papers published in the same journal, suggesting signs of community behavior practiced within journals. In an attempt to disregard such publishing practice, authors, while computing internationality have considered parameters that precisely and unambiguously define, measure and render new meaning to the internationality of journals. Thus, Non-Local influence Quotient, **NLIQ** is a major contribution for computing internationality and could potentially be a metric to be used by peers, the authors believe!

Commensurate with the current work, author's research contributions may be summarized as follows.

- Quantification of 'internationality' of peer-reviewed journals as a measure of influence, introducing a novel treatment by defining new parameters and acquiring new data.
- Definition of Non-Local Influence Quotient (NLIQ): It is determined by computing the ratio of journal's "non-local" citations to its "local" citations. The parameter signifies the spread of a journal's influence outside its boundaries. It is not documented, but common knowledge that external and internal factors are at play to ramp up impact factors of journals, in the form of suggestions to cite articles from the same journal. This explains the importance of Non-Local Influence Quotient **NLIQ** as it could enunciate the **bias-corrected** impact of journals by boasting of greater number of non-local citations. Therefore the diffusion of a journal's internationality is not manipulated by local factors if it possesses greater NLIQ i.e closer to 1. NLIQ is thus, a reasonably trustworthy indicator of internationality and a significant outcome of the manuscript.
- Definition of Other-Citations Quotient: If a journal's self-citations/total citations ratio is high, then papers in a given journal more frequently cite other articles in the same journal, than articles in different journals. That is, a high level of intra-journal collaboration is exhibited as opposed to inter-journal collaboration. Hence, journals with

a high self-citations/total citations value cannot be rewarded a high internationality score. In fact, such journals must be penalized. Self-citations/total citations needs to be low in order to appreciate influence diffusion. This prompted us to define ”Other-Citations Quotient” as $1-$ (self-citations/total citations); if self-citations equal total citations for a journal, then a journal's internationality score shall be rendered **ZERO** since such a trend reflects closed-community behavior and not true internationality, as defined by the authors.

- Definition of ”internationality” as a metric that shows evidence of non-local diffusion. To effectively reduce the effect of localization, SNIP is considered as a parameter for influence calculation.
- Novel Algorithms: Developing algorithms to compute International Collaboration, Other-Citations Quotient, NLIQ and SNIP as part of research carried out by the authors. The algorithms scrape and compute the required parameters to be fed into Cobb-Douglas model as a part of internationality computation.
- Predictive Analytics: Extensive validation process is carried out on the values obtained from scraping algorithms particularly for SNIP. Regression analysis and support vector regression is performed to confirm these values and the results are found to meet the expected level. Elaborate simulation and testing support the validity of our results.
- Normalization: The input parameters fed to the Cobb Douglas function for computing internationality score are normalized. For example, NLIQ takes into consideration the ratio of citations (external to total) and not raw numbers. This practice allows for a fair comparison between different subject fields such as Computer Science, the Social Sciences and Mathematics where collaboration and citation trends differ remarkably.

As mentioned earlier, exclusive algorithms are written to extract journal names, country information and author's affiliation to meet the requirement for computation of International Collaboration, Other-Citations Quotient and Non-Local Influence Quotient (NLIQ). These algorithms are written to create and develop a platform for **ScientoBASE**, a repository, which will consist of international journals by subject category with ranks and scores of internationality and necessary metric information by using various web-scraping and parsing techniques. Enabled with real-time search, the software will be an end-to-end product comparable with Scopus and ISI's web of Science but positioned in a distinct space and cater to the needs of the underprivileged researchers in developing countries.

The broader aim of our research is to define a yardstick of scientific contribution and international diffusion; especially in niche areas such as Astroinformatics, Computational Neuroscience, Industrial Mathematics and Data Science from India, as well as other countries across the globe. The outcome of our research will pave way for data and model validation and construction of a data visualization and web interface tool (ScientoBASE Toolkit), an open source web interface, that will compute the scores and provide visualizations of all essential parameters of internationality. This tool can be used as a web-kit to measure/analyze the growth of Indian as well as global Scientometry in state of the art and emerging areas in Science and Technology.

In future, it is possible to increase the number of parameters and differentiate between number of survey and original research article citations of a journal, number of article downloads per country and average cites per country along with the ones already included in the model. The model and data acquisition methods can be extended further to visualize growth of a subject based on region (cartogram), author (geospatial influence), topic and journal (spatial diffusion temporal invariant model). Further, this may be extrapolated to include normalization of scientific contributions and diffusion of scientometric indices in niche areas.

Authors do realize that there exists a plethora of metrics for ranking and scoring mechanisms. A practical approach would be to propose one, supported by the two powerful models, Multiple Linear Regression (used in JIS) for general influence and Cobb Douglas Model (in JIMI) for international influence. Authors intend to compute a single score, **RAGIS -Reputation and Global Influence Score**, y_{ragis} , as a convex combination of JIS and JIMI. JIS ([32], Appendix II in the repository contains details of JIS) computes influence score for journals that are indexed in SJR, Scopus and Web of Science. Computation of JIMI brings many other journals under it's fold. **RAGIS** would facilitate clustering of journals as demonstrated by supplementary data provided in the authors repository (see note below). This is set as a future goal. The authors endeavor to pursue this line of reasoning, hoping for proliferation to a comprehensive set of journals and to cater to a much larger audience.

Note: Additional file on GitHub [32] contains Matlab source code that generates an audio/video interface file. The file demonstrates frames of 3D plot of Cobb Douglas Production function. The file contains sample snapshots of the proposed toolkit, as well as other source code used in the course of this manuscript.

References

1. 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.
2. Leonard Heilig and Stefan Vo, *A Scientometric Analysis of Cloud Computing Literature IEEE Transactions on Cloud Computing*, Vol. 2, No. 3, July-September 2014
3. Website : http://www.journalmetrics.com/values.php *As accessed on 6/3/2016.*
4. Website : http://www.journalindicators.com/ *As accessed on 6/3/2016.*
5. Harzing, A.W. (2007) Publish or Perish, available from *http://www.harzing.com/pop.htm As accessed on 6/3/2016.*
6. Neelam Jangid, Snehanshu Saha, Anand Narasimhamurthy, Archana Mathur *Computing the Prestige of a journal: A Revised Multiple Linear Regression Approach (2015)*; WCI- ACM Digital library(accepted), Aug 10-13, 2015.
7. Neelam Jangid, Snehanshu Saha, Siddhant Gupta, Mukunda Rao J., *Ranking of Journals in Science and Technology Domain: A Novel And Computationally Lightweight Approach*; IERI Procedia, Elsevier, Vol 10(2014), pp 5762; doi:10.1016/j.ieri.2014.09.091
8. Website: http://www.scimagojr.com/journalrank.php *As accessed on 6/3/2016.*
9. Seyyed Mehdi Hosseini Jenab, Ammar Nejati, *Evaluation Of The Scientific Production Of Countries By A Resource Scaled Two-Dimensional Approach*; Journal of Scientometric Research, SepDec 2014, Vol 3, Issue 3.
10. Anup Kumar Das, Sanjaya Mishra, *Genesis Of Altmetrics Or Article-Level Metrics For Measuring Efficacy Of Scholarly Communications: Current Perspectives*, Journal of Scientometric Research, MayAug 2014, Vol 3, Issue 2
11. Gunther K. H. Zupanc, *Impact beyond the impact factor*, J Comp Physiology A (2014) 200:113116 Springer
12. 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
13. Gaby Haddow; Paul Genoni, *Citation Analysis And Peer Ranking Of Australian Social Science Journals*, Scientometrics, 85 (2) (2010) 471487.
14. Gualberto Buela-Casal, Pandelis Perkakis, Michael Taylor and Purificacion Checha, *Measuring Internationality: Reflections And Perspectives On Academic Journals*, Scientometrics, 67 (1) (2006) 45-65.
15. Chia-Lin Changa, Michael McAleer, Les Oxley, *Coercive journal self citations, impact factor, Journal Influence and Article Influence*, Mathematics and Computers in Simulation 93 (2013) 190197
16. Abrizah; A.N. Zainab, K. Kiran, R.G. Raj *LIS journals scientific impact and subject categorization: a comparison between Web of Science and Scopus*, Scientometrics (2013) 94:721740 DOI 10.1007/s11192-012-0813-7
17. Snehanshu Saha, Neelam Jangid, Archana Mathur, Anand M N *DSRS: Estimation and Forecasting of Journal Influence in the Science and Technology Domain via a Lightweight Quantitative Approach*, COLLNET Journal of Scientometrics and Information Management
18. http://www.crummy.com/software/BeautifulSoup/ *As accessed on 6/3/2016.*
19. http://scholarlyoa.com/publishers/ *As accessed on 6/3/2016.*
20. https://en.wikipedia.org/wiki/Jeffrey_Beall *As accessed on 6/3/2016.*
21. Beall J. *Predatory publishers are corrupting open access*, Nature. 2012;489:179
22. http://wokinfo.com/essays/impact-factor/
23. A Guide to Support Vector Regression http://web.mit.edu/6.034/wwwbob/svm-notes-long-08.pdf *As accessed on 7/3/2016.*
24. Walt Crawford, (July 2014), *”Journals, ’Journals’ and Wannabes: Investigating The List”*, Cites & Insights, 14:7, ISSN 1534-0937
25. Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su, *ArnetMiner: Extraction and Mining of Academic Social Networks*, Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD2008) pp.990-998
26. https://aminer.org/billboard/citation *As accessed on 21/1/2016.*
27. https://www.youtube.com/watch?v=8IuQzq8fEYU *As accessed on 5/3/2016.*
28. http://www.journalmetrics.com/faq.php *As accessed on 7/3/2016.*
29. Henk F. Moed *Measuring contextual citation impact of scientific journals,* Journal of Informetrics, Volume 4, Issue 3, July 2010
30. Snehanshu Saha, Avantika Dwivedi, Nandita Dwivedi, Gouri Ginde, Archana Mathur *JIMI,Journal Internationality Modelling Index-An Analytical Investigation,* Proceedings of the Fourth National Conference of Institute of Scientometrics, SIoT, August 2015
31. Gouri Ginde, Snehanshu Saha, Chitra Balasubramaniam, Harsha R.S, Archana Mathur, BS Dayasagar, Anand M N *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, August 2015
32. GitHub Repository for MATLAB code, information on JIMI - https://github.com/SciBase-Project/internationality-journals/blob/master/JIMI-JIS/ScientoBASE_appendix.pdf *As accessed on 7/3/2016.*
33. Chiang Kao, *The Authorship and Internationality of Industrial Engineering Journals*, Scientometrics, 80 (3) (2009) 123-136.
34. Liping Yu, Yuqing Chen, Yuntao Pan, Yishan Wu; *Research on the evaluation of academic journals based on structural equation modeling*, Journal of Informetrics 3(4):304-311, October 2009
35. Cobb, C.W.; Douglas, P.H.(1928) A Theory of Production, American Economic Review,18 (Supplement): 139165
36. Bao Hong,Tan (2008) Cobb-Douglas Production Function [Online Database] http://docentes.fe.unl.pt/ jamador/Macro/cobb-douglas.pdf, *As accessed on 09/03/2016.*
37. Snehanshu Saha, Jyotirmoy Sarkar, Avantika Dwivedi, Nandita Dwivedi, Anand M. Narasimhamurthy and Ranjan Roy, *A novel revenue optimization model to address the operation and maintenance cost of a data center*, Journal of Cloud Computing, Advances, Systems and Applications, 2016:1; DOI: 10.1186/s13677-015-0050-8