YING CHEN[1], THORSTEN KOCH[2], NAZGUL ZAKIYEVA[3],
KAILIANG LIU[4], ZHITONG XU[5], CHUN-HOUH CHEN[6],
JUNJI NAKANO[7], KEISUKE HONDA[8]

Article's Scientific Prestige: Measuring the Impact of Individual Articles in the Web of Science

[1] 0000-0002-2577-7348
[2] 0000-0002-1967-0077
[3] 0000-0001-9106-9916
[4] 0000-0001-6324-1475
[5] 0000-0003-1737-2831
[6] 0000-0003-0899-7477
[7] 0000-0002-6829-2914
[8] 0000-0002-7674-2081

Article's Scientific Prestige: Measuring the Impact of Individual Articles in the Web of Science

Ying Chen[a,b], Thorsten Koch[c,d,*], Nazgul Zakiyeva[c,d],
Kailiang Liu[e], Zhitong Xu[f], Chun-houh Chen[g], Junji Nakano[h],
Keisuke Honda[i]

[a]*Department of Mathematics, National University of Singapore, Block S17, Level 4, 2 Science Drive 2, 117543 Singapore*
[b]*Risk Management Institute, National University of Singapore, 21 Heng Mui Keng Terrace, 04-03, 119613 Singapore*
[c]*Technische Universität Berlin, Chair of Software and Algorithms for Discrete Optimization, Straße des 17. Juni 135, 10623 Berlin, Germany*
[d]*Zuse Institute Berlin, Applied Algorithmic Intelligence Methods Department, Takustraße 7, 14195 Berlin, Germany*
[e]*National University of Singapore (Chongqing) Research Institute, Building 5B, Chongqing Internet Institute, No. 16 South Huashan Road, Chongqing Liang Jiang New Area, Chongqing, China*
[f]*Department of Statistics and Data Science, National University of Singapore, Block S16, Level 7, 6 Science Drive 2, 117546 Singapore*
[g]*The Institute of Statistical Science, Academia Sinica, 128 Academia Road, Section 2, Nankang, 11529 Taipei, Taiwan*
[h]*Department of Global Management, Chuo University, 742-1 Higashinakano, Hachioji, 192-0393 Tokyo, Japan*
[i]*Institute of Statistical Mathematics, 10-3 Midorichō, Tachikawa, 190-0014 Tokyo, Japan*

Abstract

We performed a citation analysis on the Web of Science publications consisting of more than 63 million articles and over a billion citations on 254 subjects from 1981 to 2020. We proposed the Article's Scientific Prestige (ASP) metric and compared this metric to number of citations (#Cit) and journal grade in measuring the scientific impact of individual articles in the large-scale hierarchical and multi-disciplined citation network. In contrast to #Cit, ASP, that is computed based on the eigenvector centrality, considers both direct and indirect citations, and provides steady-state evaluation cross different disciplines. We found that ASP and #Cit are not aligned for most articles, with a growing mismatch amongst the less cited articles. While both metrics are reliable for evaluating the prestige of articles such as Nobel Prize winning articles, ASP tends to provide more persuasive rankings than #Cit when the articles are not highly cited. The

---

[*]Corresponding author
Email address: `koch@zib.de` (Thorsten Koch)

journal grade, that is eventually determined by a few highly cited articles, is unable to properly reflect the scientific impact of individual articles. The number of references and coauthors are less relevant to scientific impact, but subjects do make a difference.

1. Introduction

Known as the most popular deliverable of scientific research, the peer reviewed article is considered a main carrier of new knowledge and information, presenting innovative findings, demonstrating unique contributions, and promoting openness and transparency in science. It is

5　apparent that individual articles have different scientific impacts. Given the ever growing number of publications in science, quantifying an article's scientific prestige has been an important topic to fairly evaluate its contribution to the scientific progress, see, for example, Krattenthaler (2021); Chu & Evans (2021); Chang et al. (2019); Zhao & Feng (2022); Li et al. (2019b); Nie et al. (2019); Tahamtan et al. (2016); Xiao et al. (2016).

10　Given that all parts of science compete on the available research funding resources, and that universities, and even countries try to evaluate and compare their respective scientific impact, we consider the problem of measuring the scientific prestige of individual articles in a setting where the size of the citation network is huge — that is, where the number of nodes (articles) and edges (citations/references) is at the million/billion level and all the disciplines in science

15　are considered. We propose the Article's Scientific Prestige (ASP) metric, based on the recent advances in eigenvector centrality (or Pagerank) and optimization to address this large-scale data analysis challenge with computational tractability. More importantly, we attempt to perform a comprehensive citation analysis of all the published articles in various disciplines and over time, and provide a scientific comparison of several citation metrics at the level of individual articles.

20　Our approach is motivated by a specific application: measuring the scientific impact of each individual article in the Web of Science (WoS) citation network. The top influential papers are easy to spot. They introduce new terms and names and initiate research in a new area. The least influential papers are also easy to identify, as those articles are never cited and thus have

negligible impact. Evaluating the remaining articles is however challenging and how this should
be done remains an open question. The number of citations (#Cit) and the journal grade have
long been used as metrics to show how much attention an article has received in the science
community. Note that the popular metrics are counting the number of times an entity (e.g., a
scholar, an institute, or a journal), rather an article, has been cited, see, e.g., the Science Citation
Index (SCI) by Garfield (1955), CiteScore by Garfield (1972), H-index by Hirsch (2005), and the
SCImago Journal Rank (SJR) by González-Pereira et al. (2010). It is hypothesised that the
more citations an entity obtains, the higher scientific impact it has. Statistically speaking, such
an aggregation lowers the randomness and misjudge chance for an entity with many publications
compared to an individual article. Simultaneously, it trivializes the individual impact of each
scientific work. Alternatively, it becomes common and recognized to judge an article's scientific
value by journal grade, i,e. which journal the article is published. Scholars partition journals in
classes like A*, A, B, C, and imply that at least on average the grade of the journal reflects the
quality of the articles it publishes.

Our main question of interest is to what extent the #Cit and journal grade are helpful to
assess the impact of an individual article given a citation network. Though popularly adopted in
all kinds of evaluations, one can easily find counter examples where either way fails. It has been
acknowledged that #Cit, though direct and convenient, is not comparable across disciplines
and over time given different publication frequency and citation duration. It has also been
argued that self-citation (i.e., author cites their own articles in another article) or community-
citation (different authors, yet with strong academic connections) can easily abuse the metric.
As for journal grade, a large portion of the articles have a much lower impact than the journal's
average given the extremely skewed distribution of citations. Even the top-tiered journals have a
substantial number of articles that are not cited at all, implying one should not judge an article
(solely) based on which journal publishes it.

We perform a large-scale analysis on the Web of Science (WoS) citation network, with more
than 63 million articles and over a billion citations on 254 subjects from 1981-2020[1]. To demon-

---

[1] After removing self-citations and articles without references or without a subject, we look at articles from
254 subjects. There are 255 subjects in WoS. However, articles in the subject "Planning and development" do

strate the spectrum of citations on various disciplines, we compute the ASP of all the articles together, with which we assess the scientific influence of individual articles in the network. To obtain an accurate quantitative measure of scientific prestige, we must solve two technical challenges. The first is the large scale of the citation network, which requires an efficient optimization

⁵⁵ approach with computational tractability. Second is the hyperparameter choice in the eigenvector centrality computation such as the damping factor, ensuring a stable performance and also fair comparisons among various disciplines over time.

We implement a parallel Jacobi iterations based algorithm on sparse data-structures to compute a steady-state solution for the ASP values, see Golub & Loan (2013) and Srivastava et al.

⁶⁰ (2019). Running on an 8-core Intel Core i7-9700K CPU at 3.60GHz, the algorithm takes less than 2 seconds per iteration and converges in less than 20 iterations. The efficient algorithm allows for a wide range search of hyperparameters, even for large-scale citation networks. Specifically, we determine the damping parameter to 0.5 and also adopt a citing window of 5 and 10 years for the optimal stability of scientific contributions over disciplines and time.

⁶⁵ We found that ASP and #Cit are not aligned for most articles, with a growing mismatch amongst the less cited articles, although the two metrics display similar ranks among the top 10% highly cited articles and are identical for the bottom 20-30% of articles (as those are never cited). The journal grade, that is eventually determined by a few highly cited articles, is unable to properly reflect the scientific impact of individual articles. When aggregating to the journal

⁷⁰ level, ASP is more consistent with the journal grade than #Cit. Moreover, we found that articles with the largest ASP and #Cit were in the subjects of Science, Biology, and Geography, and the smallest in Social Science, Arts, Law & Policy, and Education. The number of references and coauthors are less relevant to scientific impact, but subjects do make a difference.

We build our analysis on pioneering works. Many aspects of the current work, including data

⁷⁵ (the size, time interval, and diversity), algorithm (to estimate the eigenvector centrality metric), and the empirical investigations at article level are however novel with respect to the prior works. Massucci & Docampo (2019) considered the citation network dataset of 5 disciplines

not have references.

4

– Dentistry, Oral Surgery & Medicine; Business & Finance; Information Science & Library Science; Telecommunication; and Veterinary Sciences – from 2010 to 2014 provided by Clarivate Analytics, and analyzed citation patterns at a university level. Ma et al. (2008) studied 236,517 articles in Biochemistry and Molecular Biology from 2003 to 2005 based on the Institute for Scientific Information (ISI) database, see also Palacios-Huerta & Volij (2004). In terms of data size, Chu & Evans (2021) conducted also a large-scale citation analysis with WoS data from 1960 to 2014. The focus is to show that the gigantic increase of articles may impede the rise of new ideas instead of promoting the rate of scientific progress. Our paper is also related to other works on eigenvector centrality or Pagerank based metrics.

Bergstrom (2007) and González-Pereira et al. (2010) computed a citation metric for academic journal evaluations (i.e., at journal level) and the latter demonstrated the application on articles from 296 subjects but published in year 2007 only with the Scopus database. Waltman & Yan (2014) discussed the application of the PageRank algorithm in the citation network. Ding (2011) proposed weighted PageRank to investigate the popularity and prestige of academic scholars based on the Web of Science citation data in the Information Retrieval for the period of 1956 and 2008. To the best of our knowledge, we are the first study to conduct scientific impact measured at individual article level over a large-scale citation network (63,092,643 articles, 953,967,411 citations, 254 disciplines over the period of 40 years from 1981 to 2020). For such a large dataset, article-level algorithms have to be reasonably convergent and thus there are high requirements for computers with large memory to run parallel computations. This possibly explains why large-scale citation analysis across all scientific disciplines is important but rare in the literature.

Our paper contributes a multi-disciplined citation analysis via connectivity in an extensive citation network. For using eigenvector centrality, we measure the influence of individual articles and provide a scientific comparison and statistics summary of several citation metrics at the level of individual articles. The framework we have developed can be applied to a broad class of citation analysis problems whose goal is to quantify the impact of an entity in a high-dimensional setting. Meanwhile, we are limited to the references within our database. By incorporating articles from online platforms such as *arXiv.org* and *Social Science Research Network (SSRN)*, or data from *crossref.org*, we can update the citation analysis in the future.

The paper is organized as follows. Section 2 presents the Web of Science data. Section 3 details the method and the implementation algorithm. Section 4 implements the ASP to evaluate the scientific contribution of articles. Section 5 discusses the comparison of ASP with respect to #Cit and journal grade, as well as relation to coauthors and references. Section 6 draws a conclusion.

2. Web of Science Data

Our primary source is the citation data of the Web of Science (WoS). We obtained the digital data from Clarivate Analytics via the Institute of Statistical Mathematics, Japan. WoS is an internet search platform that provides comprehensive citation data for 254 academic disciplines, including Natural Science, Technology, Social Sciences, Humanities, Arts, and so on. The WoS citation data contains 63,092,643 unique articles published in 65,045 journals[2] with at least 953,967,411 citations over 40 years from 1981 to 2020. Each article contains a number of attributes, including information on the article (UID, Document type, DOI, Language, Title, Abstract, Discipline), author (Name and Affiliation), journal (Publisher Name, Journal Name, Year, Issue, Volume, Pages), and a list of references (citations received after publication and references cited in the article). See Appendix A.1 for a sample observation of an article titled "Basic local alignment search tool" by Stephen Frank Altschul, Gish Warren, and others, published in 1990 in the *Journal of Molecular Biology* which received 10,277 citations within the 10 years after publication.

The input of our main analysis is a directed hierarchical graph, where each node (vertex) represents an article and each arc (link/arrow) represents a reference/citation. In contrast to our expectations, the graph resulting from the data was not a direct acyclic graph (DAG) implied by the topological order. The reason for this is that articles in the same year might reference each other, or, due to the different delays in review and publication, an article may reference a future article, leading to directed cycles in the graph.

---

[2]The average number of articles per journal per year is 94. The maximum number was 31,273 articles published in PLOSone in 2013.

A unique feature is that there should be no backward citation in the citation network, that is, the article cites a reference that was published after its own publication. In other words, the citation network must be unidirectional, with no cycles. However, it is well known that cycles can appear in large-scale citation networks due to data errors. We make an effort to exclude the cycles (by excluding references to articles after the publication date of the referencing article) so that the citation network is only unidirectional at any time point. Before creating the citation tree, we performed the following pre-processing steps: 1) Restrict the analysis to articles published in the time frame of 1990 to 2010 only, but their references still traced back to 1981 and citations up to 2020 in the computation; 2) Ignore 33,107,872 articles out of year range; 3) Ignore further 27,569 articles without subject information and 469,718 articles without references in the range; 4) Ignore 93,640,704 reference links to publications outside of the range. The resulting citation network contained 29,984,771 articles with 270,872,067 references and 376,519,109 citations.

We restricted the analysis time period of articles to 1990-2010 to avoid boundary bias due to incomplete citations/references. The boundary effect is particularly severe in the earlier years such as 1981, where references are completely missing leading to broken edges, and more recently, where articles published in, for example, 2020 are cited less often than those published in, for example, 1990. Intuitively, a fixed window size standardizes the time frame of citation metrics and allows the comparison of scientific contributions fair between articles published a long time ago and those published recently. It also means one focuses on the relevant immediate scientific impact of an article over a certain time interval after its publication. We chose both 5 and 10 years as the citing windows. Although that it may not favour certain types of articles or disciplines such as pure theoretical articles or arts works which usually need a longer time to exhibit impact on science, it can reflect the impact of citation window on article's scientific prestige. Data seems to support the choice given that the average age for an article receiving its first citation is 2.3 years. Moreover, the choice of 5 years is occasionally consistent with the common evaluation period adopted by many academic entities.

Table 1 presents statistics of #Cit in 5 and 10 years after publications, References, and Coauthors of the WoS data from 1990 to 2010 at article level. In general, all features are right skewed distributed. The median of #Cit of 10-year citation window is 5 per article meaning that

7

Table 1. Statistical summary for #Cit (with 5- and 10-year citation windows), References, and Coauthors for articles between 1990 and 2010.

		Range	Median	Mean
#Cit	in 5 years	[0; 29,905]	3	10.6
	in 10 years	[0; 55,610]	5	20.4
References		[0; 4,049]	5	14.0
(Co-)authors		[0; 5,576]	2	3.7

50% of the articles receive 5 or fewer citations within 10 years after publication, while its mean is almost four times this with a value of 20.4. The average #Cit of 10-year citation window is twice as big as the average #Cit of 5-year citation window. The skewness is caused due to extremes. It is worth noting that 38.58% of the articles of between 1990 and 2010 are not cited at all within 10 years after publication. In contrast, the maximum citation counts reaches up to 55,610, which is twice as large as that within 5-year citation window. It shows the high sensitivity of #Cit to a range of citation windows. Analogously, the distributions of the number of references and coauthors are right skewed too but with less extreme values. About 50% of articles contain 5 references or less and the maximum references is 4,049. There is a symmetry between the References and #Cit. The median reference count of 5 per article is equal to the median #Cit of 10-year citation window. The number of coauthors remains low, with a median of 2, at least 50% of articles are written by less than 2 authors. Although, in Physics collaborations such as Atlas (Switzerland) and Compact Muon Solenoid, published articles have more than 1,000 coauthors. The article entitled "International prevalence, recognition, and treatment of cardiovascular risk factors in outpatients with atherothrombosis" by Deepak L. Bhatt et al., and REACH Registry Investigators (2006) has 5,576 authors and was published in the *Journal of the American Medical Association*.

To further understand the features of articles and their citations in different disciplines, we performed statistical analysis according to the academic subject an article belongs to. Given 254 subjects, the idea was to cluster closely related disciplines, such as Chemistry Analytical and Chemistry Applied, or Engineering Ocean and Engineering Marine, into a higher level scientific cluster. We identified 14 clusters based on the intensity of the cross-citations and our knowledge

of the scientific disciplines. Specifically, we adopted graphical clustering, where the distance between any two subjects is measured by the intensity of cross-citations. Suppose an article belongs to N_1 subjects $S_1, \ldots, S_{N_1}$ and has N_2 references. And suppose one of the references belongs to N_3 subjects: $S'_1, \ldots, S'_{N_1}$. The intensity between S_i and S_j contributed by this article and this reference is defined as $\frac{1}{N_1} \times \frac{1}{N_2} \times \frac{1}{N_3}$. The cross-citation intensity is calculated by summing over all the articles and their references in the 2 subjects. Note that all ASP calculations are still based on the article level. Only for visualization and discussion purposes, we present statistics at the cluster level, such as median or mean. These statistics are calculated based on articles in subjects that are classified in the same cluster.

Given the intensity matrix, we adopted the graphical clustering approach (Wu et al. 2010) to form clusters according to the proximity measures, where subjects with high cross-citation intensity are grouped together, and subjects with low cross-citation intensity are separated using the elliptical separation algorithm and the property of a converging sequence of iteratively formed correlation matrices, Chen (2002). Next, we manually fine tuned the clustering by merging subjects with similar topics. We obtained 14 scientific clusters and presented a comparison among disciplines at the cluster level. Appendix A.2 lists the subjects contained in each cluster as well as the corresponding number of articles, references and citations.

Figure 1 displays the cross-citation intensity matrix among the 254 subjects (panel a) and the cross-citation chord diagram among the 14 clusters (panel b). A visualization clustering package GAP[3] is used to arrange the 254 disciplines according to the proximity measures. In the heatmap, the colors correspond to different quartiles of cross-subject citation intensity, i.e., blue represents the range of intensity between the minimum and Q1 (25% quartile = 1.5), pink represents between Q1 and median (9.6), green represents the range between the median and Q3 (75% quartile = 74.8), and red represents the range between Q3 and the maximum high citation intensity. The graphical clustering method helps to group subjects with high citation intensity across disciplines, indicating that they are closely linked (citing each other) and therefore have a high chance of belonging to the same cluster. We then manually looked at the description

---

[3]The software is downloaded from: http://gap.stat.sinica.edu.tw/GAP/index.htm
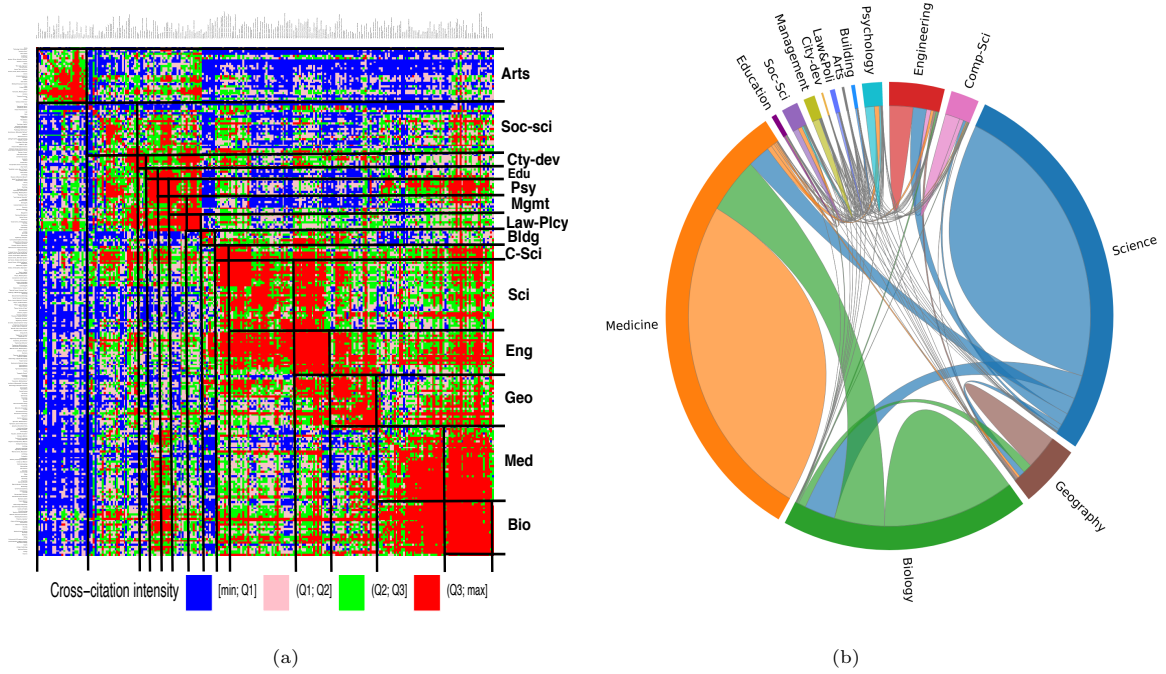
Figure 1. Panel (a) Heatmap of the intensity of cross subject citation. The quantiles show the intensity of the cross citations between subjects. Panel (b) Chord diagram for the cross-citations among the 14 clusters

of each subject and group subjects with high citation intensity (e.g., with red colors) in the same cluster. While graphical clustering provides a simple way to group topics, there is also the possibility that some subjects, while numerically "close" with high cross-subject citation intensity, are eventually belonging to different clusters. For example, 'Chemistry, Applied' was categorised to 'Engineering' given the numerical measure of cross-citations only, while it should belong to "Science". We manually moved these subjects into an appropriate cluster.

In the intensity matrix, some cells along the diagonal demonstrate a high intensity of cross citations (coloured in red), with value > 74.8, the upper quartile of intensity, whereas some cells representing e.g. Arts and Science off the diagonal have low a intensity of cross citations (coloured in blue), with value < 1.5, the lower quartile of intensity. This is consistent with the chord diagram in panel (b), which also gives a good impression of the proportion of the clusters regarding the number of articles. About half of the articles in WoS belong to either the Medicine or Biology cluster. There is high cross-citation among subjects in the common

cluster, where links are circled back to the same cluster, with the line thickness reflecting the strength of interdisciplinary cross citations. The chord diagram also shows that certain clusters such as Science and Biology do influence other clusters. Specifically, Science (blue area) is cited intensively by Biology, Medicine, Engineering, Computer Science and others, with blue linked projects to these areas. Biology (green area) is cited by Medicine and Geography. Medicine (orange area) is cited by Science and Biology. Other clusters, in contrast, have less cross-disciplinary citations.

3. Method and Measure of ASP

3.1. Eigenvector centrality

We propose ASP to evaluate the scientific prestige of an article, based on eigenvector centrality or Pagerank. The idea of using eigenvector centrality to analyze a citation network is not new (see, e.g., Ma et al. 2008). However, we are now able to compute Pagerank on the millions of node-graphs spanning all science disciplines in reasonable computational time. Articles together with the references build a mostly acyclic graph, where the direction of arcs is important. The challenge however exists in the boundary, where the leaves (newest publications) have no incoming arcs. To remedy this border effect, we performed our computations on the full graph from 1981 to 2020, but only looked at the results of articles from 1990 to 2010.

The ASP of an article i in a citation network of size N is defined as follows:

$$ASP_i = (1 - d) + d\sum_{j=1}^{N} ASP_j L_{ij}/m_j \tag{1}$$

where $L_{ij} = 1$ if an article j cites an article i and $L_{ij} = 0$ otherwise, $m_j = \sum_k L_{kj}$ is the total number of articles that j links to. In other words, the ratio L_{ij}/m_j denotes the fraction of references article j has cited. The damping factor d influences how much "prestige" of an article is passed on to the references. If d is larger, more is passed on to the referenced (older) articles. As d gets smaller, the benefit of being cited decreases. The minimum value of ASP is $1 - d$, which means that the article is not cited. We argue that an article that is never cited, or equivalently has the minimum value of ASP, has negligible scientific impact.

11

Present in matrix form, ASP is eventually an eigenvector of a Markov matrix. Let

$$ASP_{N\times 1} = \begin{bmatrix} ASP_1 \\ ASP_2 \\ \vdots \\ ASP_N \end{bmatrix}, \ \mathbf{L}_{N\times N} = \begin{bmatrix} L_{11} & L_{12} & \dots & L_{1N} \\ L_{21} & L_{22} & \dots & L_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ L_{N1} & L_{N2} & \dots & L_{NN} \end{bmatrix}, \ \mathbf{M}_{N\times N} = \begin{bmatrix} m_1 & 0 & \dots & 0 \\ 0 & m_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & m_N \end{bmatrix},$$

and $\mathbf{\Omega} = \frac{1-d}{N}\mathbf{E} + d\mathbf{L}\mathbf{M}^{-1}$ is a strongly connected Markov chain with a transition matrix $\mathbf{\Omega}^{\top}$, $\mathbf{E}$ is a $N \times N$ matrix of $1's$ and the damping factor $0 < d < 1$. From (1), we have

$$ASP = \mathbf{\Omega}ASP,$$

where ASP is the eigenvector of the matrix $\mathbf{\Omega}$ with an eigenvalue 1. The matrix $\mathbf{\Omega}$ follows the Markov Chain with

$$P(\text{go from j to i}) = \begin{cases} (1-d)/N + d/m_j, & \text{if j cites i} \\ (1-d)/N, & \text{if j does not cite i} \end{cases},$$

which means that the chain moves from state j to state i with probability $(1-d)/N+d/m_j$, if the paper j cites the paper i, and with probability $(1-d)/N$, otherwise. The transition probability is the mixture of either randomly starting from a new article with the probability $1/N$, or follow one of the references of the paper j with the probability $1/m_j$, respectively. If an article j has many references, the probability of going from j to a certain article i in the reference becomes low.

There has been rich literature on theoretical properties of PageRank, which is applicable for networks, usually with cycles. Brezinski & Redivo-Zaglia (2006) give a theoretical justification for acceleration methods proposed for accelerating the convergence of the power method. Pinto (2018) discuss various acceleration methods of computing PageRank problems, where the LumpingE method was proposed for a faster convergence in large-scale networks compared to the classical Power method. Ipsen & Wills (2006) analyse the sensitivity of PageRank to changes in the network, including addition and deletion of links in the web graph and present error

bounds for the iterates of the power method and for their residuals. Gu & Wang (2013) shows that convergence tends to slow down noticeably when the damping factor is very close to 1. The choice of damping factor is important in PargeRank calculation. Srivastava et al. (2017) conduct an experimental analysis on the damping factor value and observe that for a damping factor of 0.7, Pagerank method takes fewer numbers of iterations to converge than that of 0.85. Boldi et al. (2005) give mathematical analysis of Pagerank with respect to damping factor by providing a closed-form formula for Pagerank derivatives of any order and approximate with extension of Power method. Tang et al. (2021) present quantum PageRank by using the Runge-Kutta method and TensorFlow to conduct GPU parallel computing for the USA major airline network with up to 922 nodes which take no more than 100 seconds to converge. Fountoulakis & Yang (2022) discuss an open problem on running time complexity of accelerated l1-regularised PageRank with accelerated proximal gradient method. Hajarathaiah et al. (2022) propose the nearest neighbourhood trust PageRank based on the degree ratio, the similarity between nodes, the trust values of neighbours, and the nearest neighbours, and compare it with the maximum influence of the existing basic centrality measures. Yan & Ding (2011b) considered three methods to handle dangling nodes on citation networks using the PageRank, namely, retaining, deleting and clustering all the dangling nodes. Zhao et al. (2022) propose a scalable deep network for graph clustering via personalised Pagerank by utilising the combination of multi-layer perceptrons and linear propagation layer based on personalised Pagerank as the backbone network and employ a deep neural network module for auto-encoder to learn different dimensions embeddings.

3.2. Algorithm

To solve the above equation system, we implement parallel Jacobi iterations on sparse data-structures, resulting in a steady-state solution for the ASP computation, see Golub & Loan (2013) and Srivastava et al. (2019). The procedure is formulated in Algorithm 1. It begins by assigning an identical amount of prestige to each article. Next, this weight is redistributed in an iterative process whereby the articles transfer their attained weight to each other through the citations. The process ends when the difference between articles' prestige values in consecutive iterations does not surpass a pre-established threshold. Although in our study, ASP is computed in citation networks that exclude cycles, the proposed computation/algorithm is general and it

can be used to measure the impact of individual nodes in networks with cycles. After setting up the data structures, the algorithm typically takes less than 2 seconds per iteration and converges in less than 20 iterations when running on an 8 core Intel Core i7-9700K CPU at 3.60GHz.

Algorithm 1 ASP computation

Input: $d = 0.5$, N, $\mathbf{\Omega} \in \mathbb{R}^{N \times N}$, $\mathbf{E} \in \mathbb{R}^{N \times N}$, $\mathbf{M} \in \mathbb{R}^{N}$, $ASP^{(0)} \in \mathbb{R}^{N}$ initialized equal to 1, $\epsilon = 0.01$, $k = 0$
Output: ASP
1: **procedure** Jacobi-Iteration:
2:     **while** $\max |\epsilon| \geq 0.01$
3:     $ASP^{(k+1)} \leftarrow (1 - d) + d\mathbf{A}\mathbf{M}^{-1} \times ASP^{(k)}$
4:     $\epsilon = ASP^{(k+1)} - ASP^{(k)}$
5: **end procedure**

3.3. Damping factor and citing window

There are two hyperparameters to choose in our algorithm. The damping factor d decides how much of the incoming weight to a node is passed along to the referenced nodes. If a too-high damping factor is chosen, the oldest articles would receive most of the ASP, since they have no outgoing references within the data. When the damping factor is 1, the Markov chain matrix $\mathbf{\Omega}$ becomes reducible, meaning article j cannot reach article i in a finite number of steps. For example, the articles published in later years cannot be reached (cited) by earlier published articles in the network. Therefore, if $d = 1$, ASP converges to zero. If a too-low value is chosen, all the weights would stay with the article, and very little would be conferred to the references.

Analogous to the damping factor, the choice of citation window may have an effect on an article's prestige in the citation network. It is known that PageRank favors the older pages than the newer ones. To conduct a fair evaluation, we conduct analysis based on two fixed citation windows of 5- and 10 years. As mentioned above, there is a practical problem that the very new articles have not yet received reasonable citations due to time constraints. To compensate for this "newborn" problem, we did not include the latest 10 years in our analysis, i.e. we only considered the articles from 1990 to 2010, and compared the 5-year and 10-year windows for articles published in the same period.

14

The obvious question is what is a good damping factor, together with which citing window? By assuming that no subject should be better than another in terms of scientific contribution, we chose the hyperparameters that lead to the minimum variations among the 254 subjects. Specifically, we computed the average value of ASP in each subject. The difference is measured between the subject average ASP and the average ASP among all articles. We conduct the above computations for each year to avoid time impact. It shows that the choice of $d = 0.5$ led to the minimum deviation among the scientific disciplines. This choice is also consistent with the fact that an article usually traces up to two consecutive articles (Chen et al. 2007). In the following, we conduct the citation analysis based on that choice in our study. Appendix A.3 details the choice.

4. ASP

We summarize the statistics and distributional properties of ASP in this section. Table 2 summarises the statistics for the 14 clusters of ASP and #Cit in the 5-year and 10-year citation windows, respectively. In both citation windows, these metrics are right-skewed in distribution and have extreme values. For ASP, Biology, Science, Medicine, and Psychology lead with the largest mean values in both citation windows. In contrast, Arts have the lowest ASPs, with 50% of articles cited less than 2-3 times. In principle, the ASP statistics remained at the same level, see Science, Medicine, Biology, etc., while the mean, median, and max #Cit doubled when the citation window changed from 5 to 10 years. This suggests that ASPs are less sensitive to the choice of citation time window. Notably, although Psychology has a larger max #Cit than Computer Science, its max ASP is much lower compared to the max ASP of Computer Science in both citation windows.

For dynamic comparison, we prepare plots to show some statistics of various clusters/subjects over time. Specifically, Figure 2 present the dynamic evolution of the average ASP and #Cit of the 14 clusters from 1990 to 2010 for a citation window of 5 years and 10 years[4]. Again, #Cit has larger values of e.g. averages, medians, max, for 10 years. There is a much smaller

---

[4]Figure A.2 visualise the medians of ASP and #Cit. Again, the ASP statistics remained at the same level, while the value for #Cit doubled when the citation window changed from 5 to 10 years.

difference in ASP between the two citation windows, with similar dynamic patterns among most cases. We noticed that the statistics of ASPs with a citation window of 10 years have an obvious increase for Computer Science (1998 to 2010), Management (2001 to 2003) and Building (2004), respectively, compared to those with a citation window of 5 years. The variations are triggered by different reasons, e.g. Computer Science is more by the expected life expectancies of an article (i.e. the potential impact of an article may last over longer years), while Management and Building are very likely due to the impact of certain events (e.g. Management articles on 2000-2002 dot com bubble may be cited even after some years given it is a special event).

There are also larger deviations in magnitude among the clusters. The difference between #Cit and ASP can be further illustrated using a single article as an example. The article "The Pascal Visual Object Classes (VOC) Challenge" by Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn and Andrew Zisserman (2010), which was published in *International Journal of Computer Vision* has the top 3rd ASP of 607.14. The article belongs to the Computer Science cluster. Although it received comparably less #Cit of 3,424, which is much less than the highest #Cit, 55,610, its scientific prestige is higher due to indirect citations. Specifically, this article was cited by other articles with high impact, which eventually enhanced its influence in the citation network. In other words, while a direct count of citations for the article is not the highest, there is an impact via indirect citations too. And the impact of the indirect citation is only considered in the computation of ASP.

	5-year citation window						10-year citation window					
	#Cit			ASP			#Cit			ASP		
Cluster	Range	Median	Mean	Range	Median	Mean	Range	Median	Mean	Range	Median	Mean
Medicine	[0; 6,834]	6	11.7	[0.5; 332.4]	0.6	0.86	[0; 12,904]	10	22.3	[0.5; 355.7]	0.6	0.89
Science	[0; 29,905]	5	11.0	[0.5; 2,751.8]	0.6	0.95	[0; 55,610]	8	20.6	[0.5; 2,403.2]	0.7	1.01
Biology	[0; 10,822]	7	13.8	[0.5; 658.4]	0.7	0.96	[0; 21,471]	13	27.1	[0.5; 697.6]	0.7	1.00
Engineering	[0; 1,544]	3	6.3	[0.5; 208.8]	0.5	0.79	[0; 4,285]	5	12.5	[0.5; 380.8]	0.5	0.88
Social Science	[0; 526]	3	5.7	[0.5; 63.7]	0.5	0.60	[0; 2,606]	4	11.6	[0.5; 132.2]	0.5	0.63
Geography	[0; 1,942]	5	7.9	[0.5; 128.9]	0.6	0.88	[0; 4,054]	9	17.3	[0.5; 228.7]	0.7	0.95
Arts	[0; 144]	1	2.4	[0.5; 17.4]	0.5	0.53	[0; 472]	2	3.9	[0.5; 51.4]	0.5	0.54
Computer Science	[0; 1,649]	3	5.8	[0.5; 275.0]	0.5	0.85	[0; 7,460]	4	11.8	[0.5; 607.1]	0.5	0.98
Psychology	[0; 1,780]	5	8.7	[0.5; 137.6]	0.5	0.86	[0; 8,848]	9	20.6	[0.5; 198.2]	0.6	0.93
Management	[0; 866]	3	6.5	[0.5; 101.0]	0.5	0.85	[0; 2,785]	7	16.6	[0.5; 136.9]	0.5	0.98
Law and Policy	[0; 316]	3	5.0	[0.5; 39.9]	0.5	0.66	[0; 1,246]	4	9.8	[0.5; 59.3]	0.5	0.71
Building	[0; 254]	3	4.8	[0.5; 33.1]	0.5	0.69	[0; 1,509]	5	10.9	[0.5; 75.5]	0.5	0.79
Education	[0; 240]	3	4.8	[0.5; 35.8]	0.5	0.74	[0; 1,622]	5	10.5	[0.5; 67.6]	0.5	0.83
City Development	[0; 283]	3	4.9	[0.5; 28.2]	0.5	0.73	[0; 1,208]	5	11.5	[0.5; 33.3]	0.5	0.83
Total	[0; 29,905]	1	10.6	[0.5; 2,751.8]	0.5	0.85	[0; 55,610]	1	20.4	[0.5; 2,403.2]	0.5	0.91

Table 2. Distribution of #Cit and ASP at cluster level with 5-year and 10-year citation window between 1990 and 2010
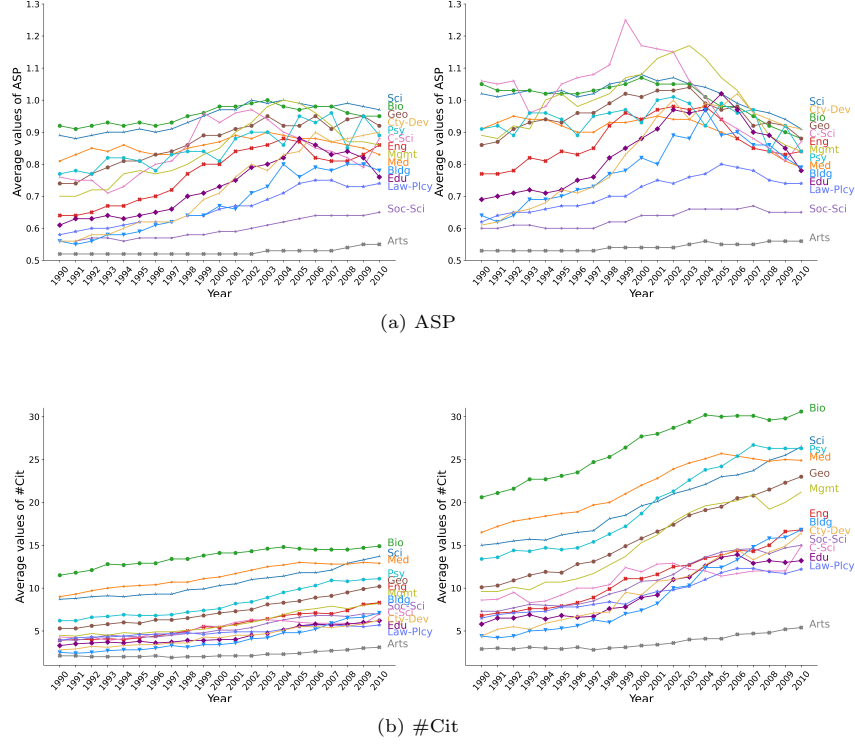
(a) ASP

(b) #Cit

Figure 2. Average values of ASP and #Cit for all the 254 subjects (Left: with citation window of 5 years; Right: with citation window of 10 years).

Moreover, Figures 3 present the median ASP and #Cit for each of the 254 subjects for
citation windows of 5 and 10 years. It shows that #Cit is sensitive to the choice of the citation
window, while the citation window has little effect on the median ASP.

Analogous to the Pareto principle in economics, right skewed distribution implies that, in
the citation network, very few articles have the most scientific influence or citations, while the
rest are rarely cited or not cited at all. We use the Pareto distribution to approximate the tail
behaviour of ASP, whose probability density function is defined as:

$$p(x) = \alpha x_{\min}^{\alpha}/x^{1+\alpha}, \ x \geq x_{\min} > 0.$$

The shape parameter α, also known as the tail index, describes the heaviness of the tail. The
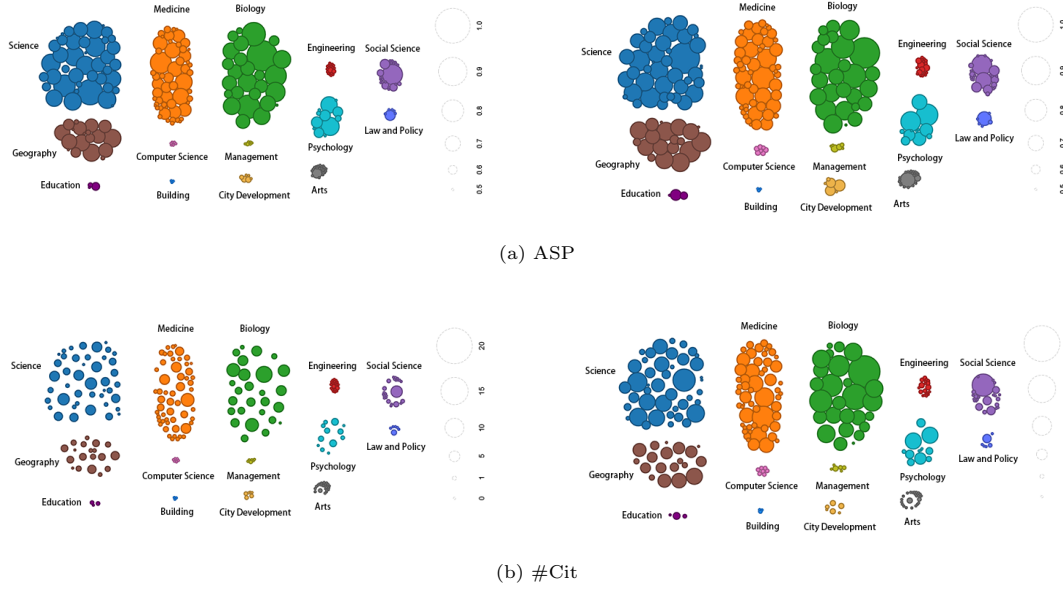
17

(a) ASP

(b) #Cit

Figure 3. Medians of ASP and #Cit for all the 254 subjects (Left: with citation window of 5 years; Right: with citation window of 10 years).

larger the tail index, the smaller the proportion of extreme values and vice versa for thinner tails. We estimate the tail index with a threshold of $x_{\min}$ to be the top 25% percentile of ASP and #Cit, respectively. Figure 4 display the tail index of ASP and #Cit, where the value of Arts, City Development, and Building has a big change and the values of the rest clusters have a slight decrease. The dynamic pattern of the tail index of ASP and #Cit basically unchanged.

5. ASP and the alternative metrics

5.1. ASP and #Cit

#Cit is possibly the most commonly used evaluation metric for individual articles. Though direct and convenient, citation count has been criticized for several shortcomings. As mentioned, the citation metric is not comparable across disciplines where the frequency of citation differs. For example, physics articles are published at a much higher frequency and are more likely to have higher citation counts than mathematics articles. Also, self-citation or community-citation can easily abuse citation counts. Even excluding self-citations, it is sometimes still unclear
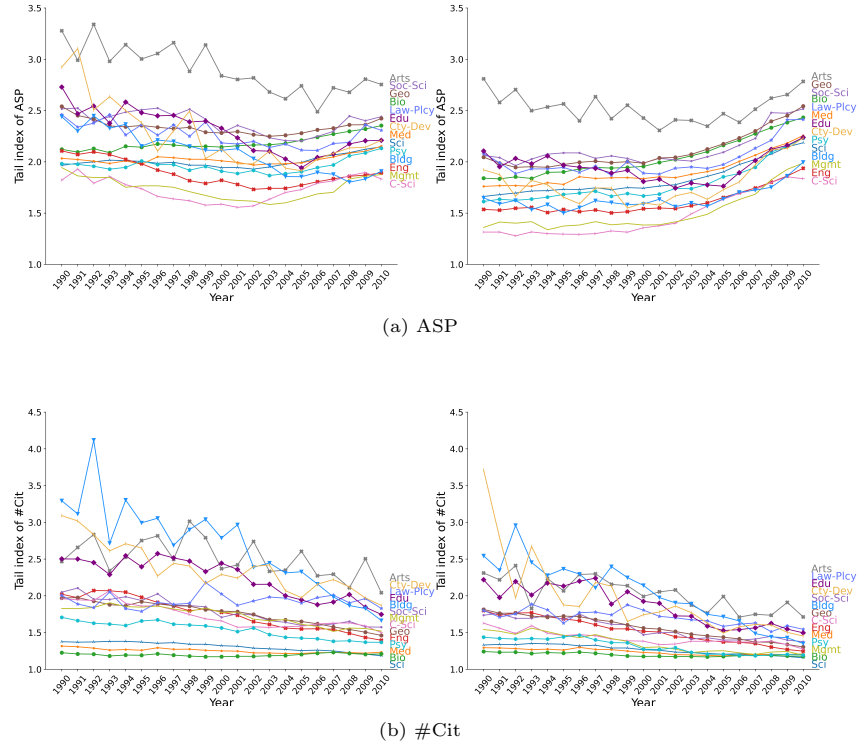
18

(a) ASP

(b) #Cit

Figure 4. Tail index of ASP and #Cit (Left: with citation window of 5 years; Right: with citation window of 10 years).

whether an article is more important by judging with higher citation counts alone. First, it depends on the type of article. A survey article usually includes a lot references on average and will possibly be cited more often. In some sense, a survey article may dilute the importance (if reflected by citation counts) of some articles, as the latter would be cited as a whole as in the survey article. Second, it also depends on the wide recognition of the work. Newton's gravitational law would be directly used without citing the original work. A recent example is Roger Penrose's work "Gravitational Collapse and Space-Time Singularities" published in 1965. This article later won Penrose the "2020 Nobel Prize in Physics", but received in total only 153 #Cit after publication up to December 2020 according to MathSci. More importantly, it lacks sufficient empirical evidence on how reliable the citation metric is at the article level.

19

Given that ASP is computed based on citation counts but also considers the sequential impact of the article via indirect citations, i.e., influence of other articles that cite, questions arise: 1) what is the relationship between the two citation metrics and 2) which metric is more reliable and under which situations. We computed the Spearman Rank correlations between ASP and #Cit.

₃₈₅ In the computation, we remove articles without citations, which correspond to about 38.58% articles as the values of #Cit are always zero and ASP always 0.5, leading to meaningless perfect correlation. Appendix A.4 presents the statistics of the articles without citation. We found that although the articles with the top 10% #Cit, i.e., the articles with high citations, have similar ranks with ASP, the remaining 90% articles differ significantly. It means it would be relatively

₃₉₀ safe to evaluate the scientific prestige of articles with either ASP or #Cit, but only for the top 10% articles.

Figure 5 panel (a) presents the scatterplot of the ranks according to #Cit vs ASP. First glance shows strong positive correlation. One would expect that there is little difference of the two metrics for evaluating an article's prestige. The scatter plot shows positive parabola shape,

₃₉₅ which means even though low ranked #Cit papers have high ranks in ASP, it is not true for the opposite case. Given the long tails of both metrics and the sensitivity of the correlation coefficients to outliers, we divided articles into deciles according to their sorted ranks according to #Cit in each cluster. The first group contains the top 10% articles with the highest #Cit in each cluster, and the last group (#10) has the last 10% of articles with the lowest #Cit for each

₄₀₀ cluster. The boxplot of the Spearman rank correlation coefficients between ASP and #Cit is displayed in panel (b) for each of the 10 groups. Except the top 10% articles have a high average correlation at 0.61, the remaining 90% articles have, on average, correlations between 0 and 0.27. The correlation, in general, drops further when the decile increases. The correlations reach to the lowest values in the group with the lowest citation, accompanied with large variations. It is

₄₀₅ consistent to Yan & Ding (2011a) where the PageRank values of the authors in lower citation level have lower correlation with their number of citations. Figure A.5 presents the scatterplot and the Pearson correlation coefficients between ASP and #Cit for each of the 10 groups, where correlation is high among the top 10% articles, and low for the rest.
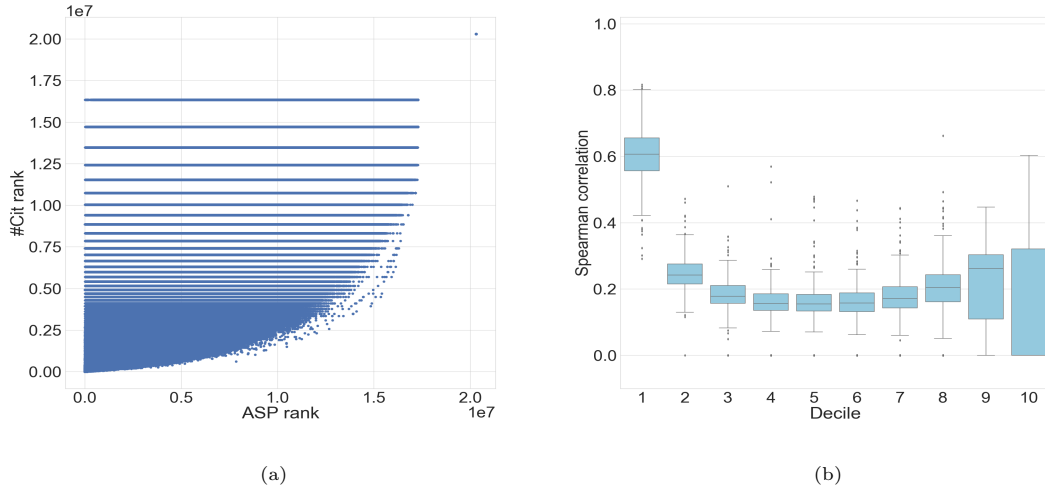
20

Figure 5. Panel (a). Scatterplot of the ranks according to #Cit vs the ASP. Panel (b). Spearman correlation of the ranks according to ASP and #Cit in 10 groups over years between 1990 and 2010. Each group includes the correlation values with respect to subjects and deciles over years. Each decile is obtained by dividing articles in each subject into 10 equal groups according to their sorted ranks according to #Cit.

5.1.1. Coincidence among the "top" articles

To verify the relationship between the two metrics, we select some individual articles to perform a detailed investigation. Table 3 lists the top 20 articles in each metric and their corresponding ranks. There are 10 articles that appear in both top 20 rankings, including the article with the maximum #Cit of 55,610 (ASP of 2,403.19 ranked #1). Meanwhile, there are articles, within the top 10%, exhibit differences in ranks. The article "The Pascal Visual Object Classes (VOC) Challenge" with ASP=607.14 and #Cit =3,424 is ranked as 3rd according to its ASP but 321st for its #Cit. In contrast, the article "A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu" with a relatively lower ASP=152.48 but larger #Cit =12,310 is ranked 201st according to its ASP and 14th according to its #Cit. The comparison reconfirms the strong correlation of the two metrics for highly influential and highly cited articles within the top group.

As another example of the high correlation between the top articles, we consider the publication records of Nobel laureates in Physics, Chemistry, and Physiology or Medicine, in a total

21

Table 3. The sets of article with either highest 20 ASP or highest 20 #Cit in WoS dataset, where n is the rank ASP and k is the rank #Cit.

n	ASP	k	#Cit	Title	Year	Source	Cluster
1	2403.19	1	55,610	A short history of SHELX	2008	Acta Crystallogr. Sect. A	Science
2	697.57	3	18,909	Gapped BLAST and PSI-BLAST: a new generation of...	1997	Nucleic Acids Res.	Biology
3	607.14	321	3,424	The Pascal Visual Object Classes (VOC) Challenge	2010	Int. J. Comput. Vis.	Computer Science
4	605.37	6	14,389	Processing of X-ray diffraction data collected ...	1997	Methods Enzymol.	Biology
5	593.17	12	12,632	Electric field effect in atomically thin carbon...	2004	Science	Science
6	578.00	2	21,471	MEGA4: Molecular evolutionary genetics analysis...	2007	Mol. Biol. Evol.	Biology
7	576.64	28	9,379	Single-crystal structure validation with the pr...	2003	J. Appl. Crystallogr.	Science
8	535.80	84	6,382	PHASE ANNEALING IN SHELX-90 - DIRECT METHODS FO...	1990	Acta Crystallogr. Sect. A	Science
9	530.02	285	3,614	Histograms of oriented gradients for human dete...	2005	PROC CVPR IEEE	Computer Science
10	513.65	4	17,517	The rise of graphene	2007	Nat. Mater.	Science
11	486.78	15	12,265	CLUSTAL-W - IMPROVING THE SENSITIVITY OF PROGRE...	1994	Nucleic Acids Res.	Biology
12	456.69	192	4,262	Distinctive image features from scale-invariant...	2004	Int. J. Comput. Vis.	Computer Science
13	456.38	30	9,246	MEGA3: Integrated software for molecular evolut...	2004	Brief. Bioinform.	Biology
14	455.45	114	5,392	NavOptim coding: Supporting website navigation ...	2004	IEEE/WIC/ACM INTERNATIONAL CONFERENCE ON WEB IN...	Computer Science
15	423.55	21	10,277	BASIC LOCAL ALIGNMENT SEARCH TOOL	1990	J. Mol. Biol.	Biology
16	419.78	57	7,460	Compressed sensing	2006	IEEE Trans. Inf. Theory	Computer Science
17	411.09	996	2,113	The capacity of wireless networks	2000	IEEE Trans. Inf. Theory	Computer Science
18	389.33	456	2,965	Space-time codes for high data rate wireless co...	1998	IEEE Trans. Inf. Theory	Computer Science
19	382.72	345	3,342	Collective dynamics of 'small-world' networks	1998	Nature	Science
20	380.75	289	3,592	A simple transmit diversity technique for wirel...	1998	IEEE J. Sel. Areas Commun.	Engineering
33	317.13	5	17,049	Fast and accurate short read alignment with Bur...	2009	Bioinformatics	Biology
163	171.68	7	14,247	Systematic and integrative analysis of large ge...	2009	Nat. Protoc.	Biology
166	169.45	8	13,568	QIIME allows analysis of high-throughput commun...	2010	Nat. Methods	Biology
62	246.10	9	12,904	Preferred Reporting Items for Systematic Review...	2009	PLos Med.	Medicine
61	251.02	10	12,901	Clustal W and clustal X version 2.0	2007	Bioinformatics	Biology
74	235.22	11	12,636	The electronic properties of graphene	2009	Rev. Mod. Phys.	Science
63	244.83	13	12,414	Analysis of relative gene expression data using...	2001	Methods	Biology
201	152.48	14	12,310	A consistent and accurate ab initio parametriza...	2010	J. Chem. Phys.	Science
135	185.19	16	11,491	PHENIX: a comprehensive Python-based system for...	2010	Acta Crystallogr. Sect. D-Struct. Biol.	Biology
167	169.41	17	10,690	Features and development of Coot	2010	Acta Crystallogr. Sect. D-Biol. Crystallogr.	Biology
50	273.23	18	10,569	Global cancer statistics, 2002	2005	CA-Cancer J. Clin.	Medicine
40	298.46	19	10,439	Coot: model-building tools for molecular graphics	2004	Acta Crystallogr. Sect. D-Struct. Biol.	Biology
35	308.71	20	10,302	Crystallography & NMR system: A new softwar...	1998	Acta Crystallogr. Sect. D-Biol. Crystallogr.	Biology

of 24 articles, from 1990 to 2010. The list is retrieved from the Harvard Nobel prize papers database by Li et al. (2019a). We found, among the Nobel prize winning papers, 21 articles

₄₂₅ are ranked in top 1% in both #Cit and ASP and 3 are higher ranked according to ASP, in the 88th, 76th, and 54th percentiles, and lower ranked according to #Cit, in the 55th, 67th, and 48th percentiles, supporting the argument that both metrics are reliable for evaluating the prestige of articles, although ASP provides more accurate rankings than #Cit when the articles are not highly cited, see Table 4.

₄₃₀ Moreover, it seems that ASP alleviates citation inflation towards certain types of articles. We use Computer Science as an example. Table 5 lists the articles that appeared in the top 20 rankings of ASP and #Cit in the cluster. Articles in the top ASP ranking come from more concentrated topics: machine learning, computer vision, and so on. In contrast, articles on the #Cit are led by applied papers published in interdisciplinary fields, including Biomedical Informatics,

₄₃₅ Statistics and so on, which usually receive many more citations compared to pure Computer Science articles. This may imply that ASP instead of #Cit makes for fairer comparisons when evaluating articles from different research orientations.

Table 4. The 24 Nobel Prize winning papers, ranked by ASP with percentile (%ile) in each metric.

ASP %ile	ASP	#Cit %ile	#Cit	Title	Year	Source	Cluster
99	593.18	99	12,632	Electric field effect in atomically thin carbon films	2004	Science	Science
99	145.96	99	8,007	Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors	2006	Cell	Biology
99	122.19	99	7,213	Induction of pluripotent stem cells from adult human fibroblasts by defined factors	2007	Cell	Biology
99	109.95	99	3,274	Observational evidence from supernovae for an accelerating universe and a cosmological constant	1998	Astron. J.	Science
99	104.54	99	3,260	Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans	1998	Nature	Science
99	83.70	99	2,444	Quantum phase transition from a superfluid to a Mott insulator in a gas of ultracold atoms	2002	Nature	Science
99	83.40	99	2,403	Evidence for oscillation of atmospheric neutrinos	1998	Phys. Rev. Lett.	Science
99	44.56	99	1,614	The complete atomic structure of the large ribosomal subunit at 2.4 angstrom resolution	2000	Science	Science
99	39.47	99	828	The dorsoventral regulatory gene cassette spatzle/Toll/cactus controls the potent antifungal response in Drosophila adults	1996	Cell	Biology
99	36.08	99	1,382	Direct evidence for neutrino flavor transformation from neutral-current interactions in the Sudbury Neutrino Observatory	2002	Phys. Rev. Lett.	Science
99	31.04	99	654	Observing the progressive decoherence of the "meter" in a quantum measurement	1996	Phys. Rev. Lett.	Science
99	28.24	99	801	Discovery of a supernova explosion at half the age of the Universe	1998	Nature	Science
99	26.54	99	930	The structural basis of ribosome activity in peptide bond synthesis	2000	Science	Science
99	24.45	99	474	Direct link between microwave and optical frequencies with a 300 THz femtosecond laser comb	2000	Phys. Rev. Lett.	Science
99	22.29	99	525	Generation of nonclassical motional states of a trapped atom	1996	Phys. Rev. Lett.	Science
99	20.57	99	776	Microstructure of a spatial map in the entorhinal cortex	2005	Nature	Science
99	17.83	99	316	Single photons on demand from a single molecule at room temperature	2000	Nature	Science
99	15.93	99	459	Collapse and revival of the matter wave field of a Bose-Einstein condensate	2002	Nature	Science
99	14.77	99	617	Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics	2000	Nature	Science
99	14.41	99	488	Structure of functionally activated small ribosomal subunit at 3.3 angstrom resolution	2000	Cell	Biology
99	11.22	99	389	Visualizing secretion and synaptic transmission with pH-sensitive green fluorescent proteins	1998	Nature	Science
99	7.12	99	452	Induction of pluripotent stem cells from fibroblast cultures	2007	Nat. Protoc.	Biology
88	1.62	55	7	Discovery of a supernova explosion at half the age of the Universe (vol 391, pg 51, 1998)	1998	Nature	Science
76	1.07	67	12	Generation of nonclassical motional states of a trapped atom (vol 76, pg 1796, 1996)	1996	Phys. Rev. Lett.	Science
54	0.72	48	5	Direct evidence for neutrino flavor transformation from neutral-current interactions in SNO	2002	AIP CONF PROC	Science

Table 5. The sets of articles with either highest 20 ASP or highest 20 #Cit in Computer Science cluster, ranked by ASP, where n is the rank ASP and k is the rank #Cit.

n	ASP	k	#Cit	Title	Year	Source	Cluster
1	607.14	15	3,424	The Pascal Visual Object Classes (VOC) Challenge	2010	Int. J. Comput. Vis.	Computer Science
2	530.02	11	3,614	Histograms of oriented gradients for human detection	2005	PROC CVPR IEEE	Computer Science
3	456.69	9	4,262	Distinctive image features from scale-invariant keypoints	2004	Int. J. Comput. Vis.	Computer Science
4	455.45	3	5,392	NavOptim coding: Supporting website navigation optimisation using effort minimisation	2004	IEEE/WIC/ACM INTERNATIONAL CONFERENCE ON WEB INTELLIGENCE (WI 20	Computer Science
5	419.78	1	7,460	Compressed sensing	2006	IEEE Trans. Inf. Theory	Computer Science
6	411.09	38	2,113	The capacity of wireless networks	2000	IEEE Trans. Inf. Theory	Computer Science
7	389.33	22	2,965	Space-time codes for high data rate wireless communication: Performance criterion and code construction	1998	IEEE Trans. Inf. Theory	Computer Science
8	364.07	20	3,208	Wireless sensor networks: a survey	2002	Comput. Netw.	Computer Science
9	329.59	4	5,177	Cooperative diversity in wireless networks: Efficient protocols and outage behavior	2004	IEEE Trans. Inf. Theory	Computer Science
10	308.01	6	4894	Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information	2006	IEEE Trans. Inf. Theory	Computer Science
11	285.09	61	1,584	A fast learning algorithm for deep belief nets	2006	Neural Comput.	Computer Science
12	282.65	297	678	SUPPORT-VECTOR NETWORKS	1995	Mach. Learn.	Computer Science
13	276.91	189	878	Wireless integrated network sensors	2000	Commun. ACM	Computer Science
14	258.09	77	1,437	Ad-hoc On-Demand Distance Vector Routing	1999	WMCSA '99, SECOND IEEE WORKSHOP ON MOBILE COMPUTING SYSTEMS AND	Computer Science
15	257.79	27	2,556	OpenFlow: Enabling innovation in campus networks	2008	ACM SIGCOMM Comp. Commun. Rev.	Computer Science
16	256.10	7	4,763	Speeded-Up Robust Features (SURF)	2008	Comput. Vis. Image Underst.	Computer Science
17	244.39	45	1,952	A tutorial on Support Vector Machines for pattern recognition	1998	Data Min. Knowl. Discov.	Computer Science
18	241.72	8	4,760	The Internet of Things: A survey	2010	Comput. Netw.	Computer Science
19	239.61	19	3,246	A fast and elitist multiobjective genetic algorithm: NSGA-II	2002	IEEE Trans. Evol. Comput.	Computer Science
20	239.03	26	2,742	Space-time block codes from orthogonal designs	1999	IEEE Trans. Inf. Theory	Computer Science
39	170.46	2	6,079	Research electronic data capture (REDCap)-A metadata-driven methodology and workflow process for providing translational research informatics support	2009	J. Biomed. Inform.	Computer Science
36	174.07	5	5,015	Robust Face Recognition via Sparse Representation	2009	IEEE Trans. Pattern Anal. Mach. Intell.	Computer Science
250	66.51	10	3,641	Conducting Meta-Analyses in R with the metafor Package	2010	J. Stat. Softw.	Computer Science
74	127.63	12	3,608	A Survey on Transfer Learning	2010	IEEE Trans. Knowl. Data Eng.	Computer Science
82	121.17	13	3,577	Regularization Paths for Generalized Linear Models via Coordinate Descent	2010	J. Stat. Softw.	Computer Science
22	228.31	14	3,482	Image quality assessment: From error visibility to structural similarity	2004	IEEE Trans. Image Process.	Computer Science
29	194.15	15	3,424	A View of Cloud Computing	2010	Commun. ACM	Computer Science
72	129.02	17	3,401	A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems	2009	SIAM J. Imaging Sci.	Computer Science
47	156.20	18	3,248	TreeView: An application to display phylogenetic trees on personal computers	1996	Comput. Appl. Biosci.	Computer Science

5.1.2. Examples of other groups

The story is different for other groups where the two metrics are not compatible. Heuristically, an article can be considered influential if it is cited by many articles, or if it, though not directly cited by many, is cited by other influential article(s) with many citations. As mentioned, #Cit counts only direct citations, while ASP is able to reflect the prestige including these indirect citations. An Engineering article "Blind decorrelation and deconvolution algorithm for multiple-input multiple-output system: I. Theorem derivation" was ranked as high in the 99th percentile (top 1%) with ASP value (21.75) but relatively low in the 23rd percentile by #Cit (with 1 count). The Computer Science article "A computer algebra system based on order-sorted algebra" was ranked in the 99th percentile by the ASP (19.94) but in the 30th percentile by #Cit (with 2

counts). The Social Science article "Vision and the autonomous symbol in the works of Lorrain, Jean-stage sets and obstacles" was ranked in the 99th percentile by the ASP (17.34), but in the 23rd percentile by #Cit (with 1 count). Although the above may be argued as special cases, there are more examples of articles with small #Cit but high ASP. Specifically, 43.83% of articles have high ASP values and a small number of citation counts. In contrast, there are 33.62% articles with low ASP but high #Cit.

5.2. ASP and Journal Grade

Historically, journals are used to publishing (printing and distributing) scientific results but less so in this millennium. One can find important publications in *arXiv.org* even though they are not published in known (refereed) journals. Journals are also used to cluster/sort/filter articles regarding particular topics. If one looks into "Applied Statistics," articles in the journals, one expects articles to be about applied statistics. Over the years, there might have been a shift in the topics, so the journal name might not exactly match its contents anymore. Important journals (by the usual metrics) like *Science* or *Nature* are not sorted by topic at all. Since the mid-20th century, under the hypothesis that the more citations an entity (e.g., a scholar, an institute, or a journal) receives, the higher scientific impact it has, citation-based metrics have been developed and became popular, particularly for evaluating a journal's scientific prestige at the journal level, where the total citation counts of all articles published in the journal are considered over a certain time period. See, for example, Science Citation Index (SCI), CiteScore, Impact Factor (IF), Hirsch's bibliometrics index (H-Index), and SCImago Journal Rank (SJR). Given the publicly available journal-level citation metrics, it becomes common and recognized to judge an article's scientific value based on which journal publishes it.

Admittedly, an article published in a highly regarded journal is more likely to be read and cited, increasing its chances of becoming "important" and influential in scientific society. However, it is often misleading to evaluate an article's scientific prestige indirectly based on a journal's rank. The quality of an article is unlikely to change with a journal's impact. Instead, a journal's value will be improved if it publishes important articles. Meanwhile, the distribution of #Cit is right-skewed with a long tail slowly fragmenting towards the extremely large citations, implying that the majority of papers published in the top tiered journal are overvalued when judging with

24

journal prestige. Figure 6 presents the empirical density of #Cit in *Nature* (ISSN: 0028-0836, 1476-4687) from 1981 to 2020. The journal issued by Nature Research is a prestigious journal in multidisciplinary science and has been well recognized by the journal-level citation metrics, with an IF of 42.778 in 2019, H-index of 1226, and SJR of 15.993 in 2020. According to journal grade information, its max, min, median of #Cit are 3157, 1, and 4, respectively and ASP are 60.68, 0.5, and 0.5, respectively. Given that *Nature* is considered as one of the most prestigious journals, all the papers should have higher chances of being cited. Nevertheless, among the 121,107 articles published in the journal obtained from the WoS data between 1981 and 2020, 36.56% were never cited, yet all would be considered top publications if journal grade is used as an evaluation metric.
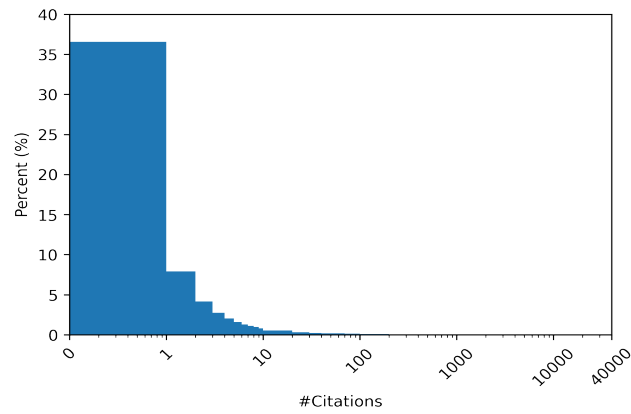

Figure 6. Histogram for #Cit for articles published between 1981 and 2020 in *Nature* (ISSN: 0028-0836, 1476-4687)

Simultaneously, an important article may be undervalued if it is not published in a prestigious journal. Some essential works have been known, to introduce innovations that are too advanced, were rejected by conservative referees and published in less prestigious journals. A famous example is the article "The market of lemons," written by George Akerlof in 1966 which won the 2001 Nobel Memorial Prize in Economic Sciences. The article was rejected by three renowned journals and was finally accepted and published upon the fourth submission in 1978 in *The Quarterly Journal of Economics.*

We considered the 65,045 journals in the WoS data and categorized these journals according to

Table 6. Journal grade: ASP and #Cit

		SJR				H-index			
		ASP		#Cit		ASP		#Cit	
		range	mean	range	mean	range	mean	range	mean
min	Q1/H1	[0.50; 0.57]	(0.51)	[0; 2]	(0.02)	[0.50; 0.63]	(0.53)	[0; 5]	(0.32)
	Q2/H2	[0.50; 0.51]	(0.50)	[0; 1]	(0.01)	[0.50; 0.53]	(0.50)	[0; 1]	(0.03)
	Q3/H3	[0.50; 0.50]	(0.50)	[0; 0]	(0.01)	[0.50; 0.51]	(0.50)	[0; 1]	(0.02)
	Q4/H4	[0.50; 0.50]	(0.50)	[0; 0]	(0.04)	[0.50; 0.50]	(0.50)	[0; 0]	(0.01)
mean	Q1/H1	[0.50; 21.05]	(1.18)	[0.12; 561.80]	(20.66)	[0.50; 19.42]	(1.12)	[0.10; 510.23]	(19.16)
	Q2/H2	[0.50; 8.53]	(0.84)	[0.07; 197.91]	(9.68)	[0.50; 3.88]	(0.72)	[0.07; 66.06]	(5.14)
	Q3/H3	[0.50; 5.53]	(0.71)	[0.01; 103.88]	(5.32)	[0.50; 2.04]	(0.58)	[0.01; 15.29]	(1.07)
	Q4/H4	[0.50; 3.15]	(0.60)	[0.01; 50.80]	(2.39)	[0.50; 1.41]	(0.53)	[0.01; 6.02]	(0.23)
median	Q1/H1	[0.50; 9.70]	(1.03)	[0; 241]	(15.30)	[0.50; 9.60]	(0.98)	[0; 240]	(14.15)
	Q2/H2	[0.50; 4.93]	(0.81)	[0; 109]	(8.32)	[0.50; 2.84]	(0.69)	[0; 47]	(4.12)
	Q3/H3	[0.50; 2.99]	(0.68)	[0; 56]	(4.39)	[0.50; 1.76]	(0.57)	[0; 11]	(0.66)
	Q4/H4	[0.50; 1.93]	(0.58)	[0; 20]	(1.48)	[0.50; 1.25]	(0.52)	[0; 4]	(0.13)
max	Q1/H1	[2.03; 2,403.19]	(10.88)	[118; 55,610]	(394.50)	[2.77; 2,403.19]	(10.88)	[177; 55,610]	(394.50)
	Q2/H2	[2.77; 257.79]	(6.02)	[177; 7,768]	(177.00)	[2.04; 176.19]	(2.70)	[118; 7,768]	(118.00)
	Q3/H3	[0.56; 225.47]	(2.70)	[3; 2,797]	(33.01)	[0.55; 12.37]	(1.04)	[1; 188]	(14.47)
	Q4/H4	[0.50; 73.08]	(1.41)	[1; 2,572]	(27.60)	[0.79; 8.11]	(0.79)	[0; 53]	(2.50)
Remark		number of journals Q1: 5,117, Q2: 3,242, Q3: 1,791, Q4: 664				number of journals H1: 6,621, H2: 3,123, H3: 792, H4: 313			

the SJR (Scientific Journal Ranking) in 2020 and the H-index[5] respectively. When merging the data according to ISSN, 35,952 journals had complete information, among which 10,814 journals have both citation and ASP records. We then conducted a statistical analysis to investigate the relation between ASP and journal grade. Table 6 presents the range of mean, median, min, and max of ASP and #Cit of journals according to the journal grade. It shows that, in terms of average value, ASP and #Cit, when aggregated to journal level, are consistent to the SJR journal grade. However, the minimum average of #Cit increases from Q1 to Q4 level journals, meaning that #Cit is not aligned to journal grade, where Q4 journals supply the minimum average of 0.04. The inconsistency between #Cit and journal grade disappears when the H-index is used. This is no surprise as the H-index essentially reflects the same information as the total number of citations according to probability theory, see Krattenthaler (2021).

Nevertheless, we argue that journal grade is not the right metric to evaluate an article's scientific prestige. Figure 7 demonstrates the distribution of ASP in 4 journals in the cluster of Medicine with different SCImago Journal Ranks (SJR) by González-Pereira et al. (2010). The histograms show the ASP distribution in Medicine journals, namely *Molecular Therapy* with $SJR_{2020} = 3.871$, *Pharmacoepidemiology and Drug Safety* with $SJR_{2020} = 1.023$, *Veterinary*

---

[5]We followed SJR ranking to separate journals into 4 groups: Q1 to Q4. We separated the H-index according to quartile, leading to 4 groups labelled as H1 to H4.

26

Record with $SJR_{2020} = 0.261$, and *Deutsche Medizinische Wochenschrift* with $SJR_{2020} = 0.151$. It shows that the ASP follows logarithmic law distribution regardless of the journal grade, where 60.74%, 72.59%, 61.43% and 50.60%, respectively, are not cited. This means that no matter where an article is published, there is a chance of no scientific influence. For the articles published in grade I journals, this is more questionable given that the high journal grade enhances the visibility of articles. In short, an article should not be judged solely based on the grade of the journal it is published in.
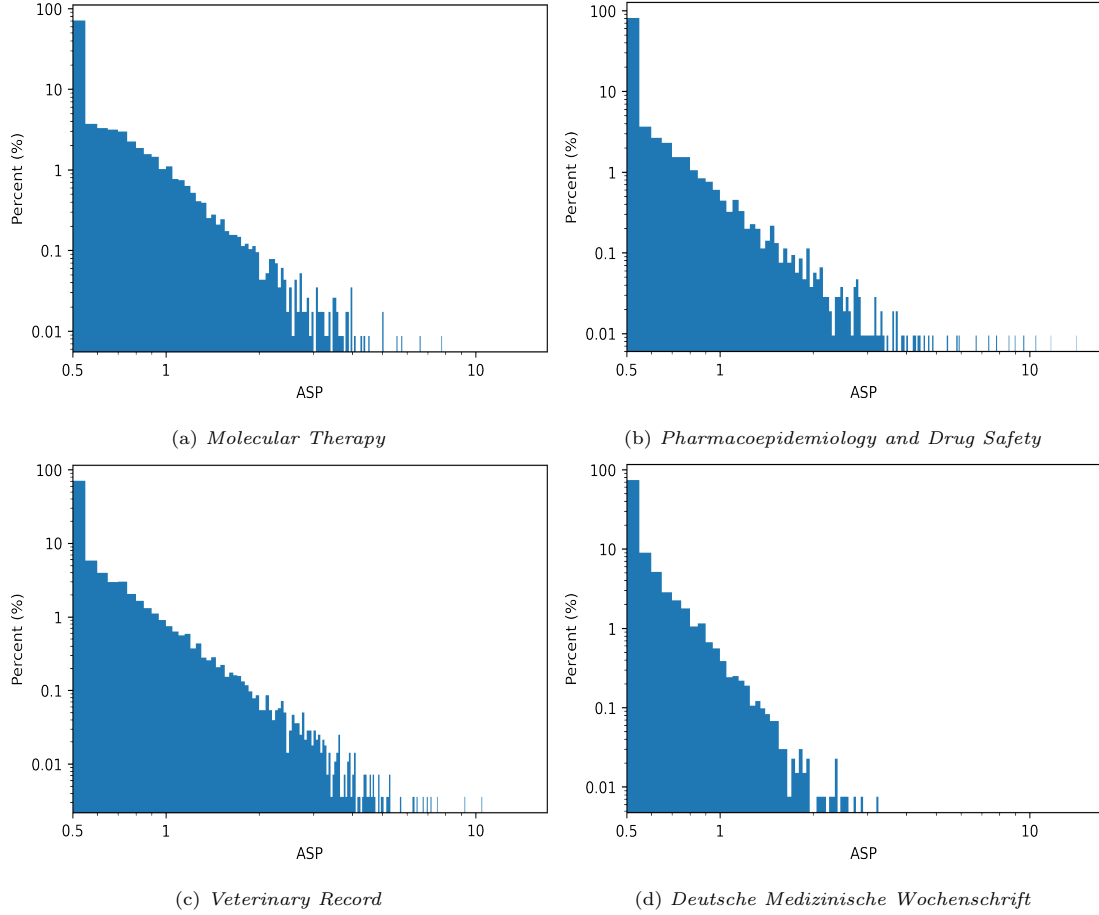

(a) *Molecular Therapy*

(b) *Pharmacoepidemiology and Drug Safety*

(c) *Veterinary Record*

(d) *Deutsche Medizinische Wochenschrift*

Figure 7. ASP distribution for Medicine articles published in four journals in different quartile of SJR_{2020} ranking. Articles are published between 1990 and 2010

27

The WoS data shows that 50% articles are coauthored by no more than two authors and have no more than 5 references. Figure 8 presents the median of references and coauthors per article for the articles published from 1990-2010. In terms of references, Biology leads with 50% of articles referring to 6-9 previous articles, followed by Medicine with 5-7 references. In general, there is a mild increase of references per article in almost all the clusters, yet at different rates. Science exhibits a dramatic increase from four references per article in 1990 to seven in 2010. Geography and Psychology display significant increases too, though at a slower speed. Arts, Education, and Building have the smallest number of references. Computer Science appears at the bottom, possibly because conference proceedings rather than articles are more recognized in this cluster.

Regarding the number of coauthors, Medicine on average involves a bigger team, and the median increases over time and has reached five coauthors in recent years. On the other hand, Social Science, Arts, and Law & Policy have smaller size, where at least 50% of articles are sole-authored. Science, Computer Science, Engineering, and Geography have had increasingly more coauthors over the recent years. The median number of authors in Computer Science increased from one in 1990 to four in 2000 on median.

To what extent is references or number of coauthors helpful to improve ASP? Figure 9 presents the scatter plot of the ASP versus the number of coauthors (panel a) and references (panel b). We find there is no evidence that more references or more coauthors improve ASP.

6. Conclusion

We analyzed a large-scale WoS citation network with millions of articles from 254 subjects published between 1981 and 2020. We proposed the ASP index to evaluate the scientific importance of individual articles in the network using the eigenvector centrality metric. We found that there is a high correlation between the ASP and the #Cit among the top 10% of articles but a significantly minor dependence for the rest articles. There is little evidence of influence of the number of references and coauthors on the article's scientific quality. Furthermore, ASP minimizes the difference in scientific quality distribution among the disciplines. In consistent to
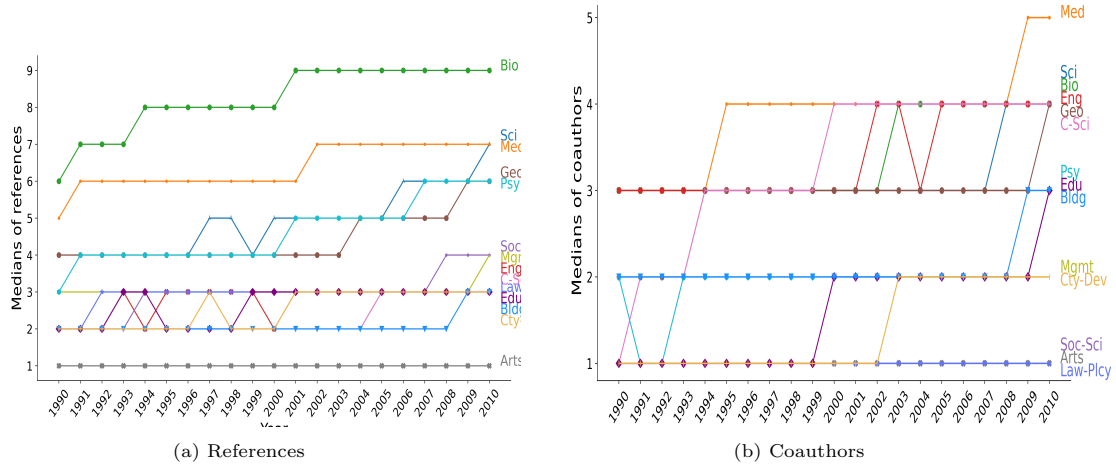
28

(a) References

(b) Coauthors

Figure 8. Medians of References and Coauthors per article for 14 clusters between 1990 and 2010.

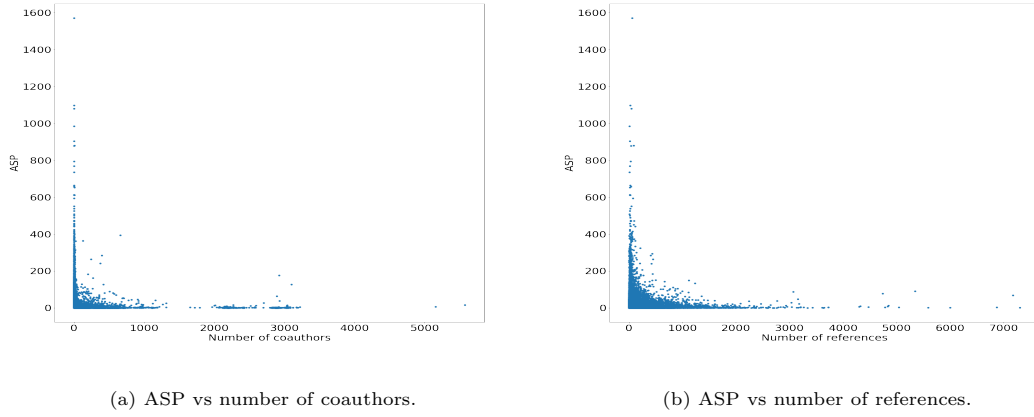(a) ASP vs number of coauthors.

(b) ASP vs number of references.

Figure 9. ASP vs number of coauthors and references

the fact that the quality distribution of journal articles is dramatically right-skewed, the articles' scientific prestige should not be judged based on the journal grades, which is supported by our analysis. With a parallel algorithm on sparse data-structures, we can obtain the ASPs for nearly 30 million articles in a few seconds, demonstrating that it is computationally feasible to evaluate all articles individually. Without question, there is still room for improvement in evaluating the prestige and impact of scientific articles. Our analysis showcases that there is no computational

29

hurdle from including further aspects. For example, the increasing use of unique and well defined IDs like OrcID will allow in the future a reliable evaluation of author/co-authorship relations over multiple articles and citations. Meanwhile, the quality of the data is of crucial importance. We noticed that it seems very likely that a considerable number of references is missing from WoS, though it corresponds to small percentage in the large-scale citation network. Publications channels continue to expand, the importance of Proceedings and Open Access repositories e.g. *arXiv.org*, or self publishing via Social Media like ResearchGate is constantly increasing. Maybe it is time to end judging a publication by where it is published but to compute individually how much "prestige" it manages to attract. As an additional benefit this would make the introduction of new publication outlets much easier.

7. Acknowledgement

Appendix A

A.1   Sample of citation data

Figure A.1 illustrates the raw information of an article entitled "Basic local alignment search tool" by Altschul Stephen F. and Gish Warren and others, published in 1990 in the *Journal of Molecular Biology*.

30

```
<REC r_id_disclaimer="ResearcherID data provided by Clarivate Analytics">
<UID>WOS:A1990ED16700008</UID>
<static_data>
<summary>
<EWUID>
<WUID coll_id="WOS"/>
<edition value="WOS.SCI"/>
</EWUID>
<pub_info sortdate="1990-10-05" pubyear="1990" has_abstract="N" coverdate="OCT 5 1990" pubmonth="OCT 5" vol="215" issue="3" pubtype="Journal">
<page begin="403" end="410" page_count="8">403-410</page>
</pub_info>
<titles count="6">
<title type="source">JOURNAL OF MOLECULAR BIOLOGY</title>
<title type="source_abbrev">J MOL BIOL</title>
<title type="abbrev_iso">J. Mol. Biol.</title>
<title type="abbrev_11">J MOL BIOL</title>
<title type="abbrev_29">J MOL BIOL</title>
<title type="item">BASIC LOCAL ALIGNMENT SEARCH TOOL</title>
</titles>
<names count="5">
<name seq_no="1" role="author" reprint="Y">
<display_name>ALTSCHUL, SF</display_name>
<full_name>ALTSCHUL, SF</full_name>
<wos_standard>ALTSCHUL, SF</wos_standard>
<first_name>SF</first_name>
<last_name>ALTSCHUL</last_name>
</name>
<name seq_no="2" role="author">
<display_name>GISH, W</display_name>
<full_name>GISH, W</full_name>
<wos_standard>GISH, W</wos_standard>
<first_name>W</first_name>
<last_name>GISH</last_name>
</name>
<name seq_no="3" role="author">
<display_name>MILLER, W</display_name>
<full_name>MILLER, W</full_name>
<wos_standard>MILLER, W</wos_standard>
<first_name>W</first_name>
<last_name>MILLER</last_name>
</name>
<name seq_no="4" role="author">
<display_name>MYERS, EW</display_name>
<full_name>MYERS, EW</full_name>
<wos_standard>MYERS, EW</wos_standard>
<first_name>EW</first_name>
<last_name>MYERS</last_name>
</name>
<name seq_no="5" role="author">
<display_name>LIPMAN, DJ</display_name>
<full_name>LIPMAN, DJ</full_name>
<wos_standard>LIPMAN, DJ</wos_standard>
<first_name>DJ</first_name>
<last_name>LIPMAN</last_name>
</name>
</names>
<doctypes count="1">
<doctype>Article</doctype>
</doctypes>
```

Figure A.1. Sample of the Web of Science dataset for the most cited article entitled "Basic local alignment search tool" published in 1990 in *Journal of Molecular Biology* with the UID "WOS:A1990ED16700008" and 5 coauthors.

A.2 The 14 clusters and 254 subjects

We group the disciplines into 14 scientific clusters where the cluster information is summarised in Table A.1. Medicine and Science form the two biggest groups, by 56 and 43 disciplines

31

respectively. The two clusters also have the largest number of articles with more than 7.2 millions for each. This is almost double of the 3rd largest cluster, Biology. In terms of citations, the 3rd largest cluster Biology stands out with a median of 12, while Medicine, the top one, have 10 citations on average within 10 years after publication, which is 2 citations less than Biology. Among the 14 clusters, Arts has 2 median citation, which means 50% of articles in the cluster have at most 2 citations or never cited. The Arts cluster also has the smallest number of references (2 per article).

In the analysis, each article is assigned to exactly one cluster according to the label of disciplines. While 79.26% of articles has one cluster, including articles with sole subject and articles with multiple subjects belonging to the same cluster, the rest belongs to multiple clusters. Among them, 7.24% articles are labelled to the cluster with the most common disciplines, and 13.49% articles with equal amount of subjects belonging to two and more clusters are labelled according to the first discipline in the WoS citation dataset.

Figure A.2 visualise the medians of ASP and #Cit. The ASP statistics remained at the same level, while the value for #Cit doubled when the citation window changed from 5 to 10 years.

A.3 Hyperparameters choice

We conduct sensitivity analysis given different combinations of damping factor $d \in (0.1, 0.9)$ and citing window size $\in [1, 10]$. To measure the stability, we compare the scaled average value of ASP over years. Specifically, we compute the average value of ASP in each subject. We display the sum over the difference between the subject ASP and the average ASP among all articles. To avoid the time impact, we conduct the computations for each year. As illustration, Figure A.3 shows that the choice of $d = 0.5$ and citing window of 5 years led to the minimum deviation among the scientific disciplines. By assuming that no subject is better than another in terms of scientific contribution, we chose the hyperparameters that lead to the minimum variations among the 254 subjects over years. Due to space limit, we omit other results on e.g. different citation windows, which is available upon request.

Table A.1. Classification of the 254 disciplines into 14 clusters (1990 and 2010).

Cluster	Disciplines	Sub.	Articles	Refs	#Cit
Medicine	Allergy; Anatomy & Morphology; Andrology; Anesthesiology; Audiology & Speech-Language Pathology; Cardiac & Cardiovascular Systems; Clinical Neurology; Critical Care Medicine; Dentistry, Oral Surgery & Medicine; Dermatology; Emergency Medicine; Endocrinology & Metabolism; Ethics; Gastroenterology & Hepatology; Genetics & Heredity; Geriatrics & Gerontology; Health Care Sciences & Services; Health Policy & Services; Hematology; Immunology; Infectious Diseases; Integrative & Complementary Medicine; Medical Ethics; Medical Informatics; Medical Laboratory Technology; Medicine, General & Internal; Medicine, Research & Experimental; Microscopy; Neuroimaging; Neurosciences; Nursing; Obstetrics & Gynecology; Oncology; Ophthalmology; Orthopedics; Otorhinolaryngology; Pathology; Pediatrics; Peripheral Vascular Disease; Pharmacology & Pharmacy; Physiology; Primary Health Care; Psychiatry; Public, Environmental & Occupational Health; Radiology, Nuclear Medicine & Medical Imaging; Rehabilitation; Respiratory System; Rheumatology; Sport Sciences; Surgery; Toxicology; Transplantation; Tropical Medicine; Urology & Nephrology; Veterinary Sciences; Virology	56	9,119,386	12	10
Science	Acoustics; Astronomy & Astrophysics; Chemistry, Analytical; Chemistry, Applied; Chemistry, Inorganic & Nuclear; Chemistry, Medicinal; Chemistry, Multidisciplinary; Chemistry, Organic; Chemistry, Physical; Crystallography; Electrochemistry; Engineering, Chemical; Imaging Science & Photographic Technology; Materials Science, Biomaterials; Materials Science, Ceramics; Materials Science, Characterization & Testing; Materials Science, Coatings & Films; Materials Science, Composites; Materials Science, Multidisciplinary; Materials Science, Paper & Wood; Materials Science, Textiles; Mathematics; Mathematics, Applied; Mathematics, Interdisciplinary Applications; Mechanics; Multidisciplinary Sciences; Nanoscience & Nanotechnology; Nuclear Science & Technology; Optics; Physics, Applied; Physics, Atomic, Molecular & Chemical; Physics, Condensed Matter; Physics, Fluids & Plasmas; Physics, Mathematical; Physics, Multidisciplinary; Physics, Nuclear; Physics, Particles & Fields; Polymer Science; Spectroscopy; Statistics & Probability; Thermodynamics; Quantum Science & Technology; Green & Sustainable Science & Technology	43	7,207,423	9	8
Biology	Biochemical Research Methods; Biochemistry & Molecular Biology; Biodiversity Conservation; Biology; Biophysics; Biotechnology & Applied Microbiology; Cell Biology; Cell & Tissue Engineering; Developmental Biology; Ecology; Entomology; Evolutionary Biology; Food Science & Technology; Horticulture; Limnology; Marine & Freshwater Biology; Mathematical & Computational Biology; Microbiology; Mycology; Nutrition & Dietetics; Oceanography; Ornithology; Parasitology; Plant Sciences; Reproductive Biology; Soil Science; Zoology	27	3,483,426	16	13
Engineer-ing	Automation & Control Systems; Energy & Fuels; Engineering, Aerospace; Engineering, Biomedical; Engineering, Electrical & Electronic; Engineering, Environmental; Engineering, Industrial; Engineering, Manufacturing; Engineering, Marine; Engineering, Mechanical; Engineering, Multidisciplinary; Engineering, Ocean; Engineering, Petroleum; Ergonomics; Instruments & Instrumentation; Metallurgy & Metallurgical Engineering; Remote Sensing; Robotics; Telecommunications	19	2,444,455	5	5
Social Science	Anthropology; Area Studies; Behavioral Sciences; Communication; Criminology & Penology; Demography; Ethnic Studies; Family Studies; Gerontology; History; History Of Social Sciences; Hospitality, Leisure, Sport & Tourism; Humanities, Multidisciplinary; Information Science & Library Science; Philosophy; Religion; Social Issues; Social Sciences, Biomedical; Social Sciences, Interdisciplinary; Social Sciences, Mathematical Methods; Social Work; Sociology; Substance Abuse; Women's Studies	24	1,689,969	5	4
Geography	Agricultural Engineering; Agriculture, Dairy & Animal Science; Agriculture, Multidisciplinary; Agronomy; Engineering, Geological; Environmental Sciences; Environmental Studies; Fisheries; Forestry; Geochemistry & Geophysics; Geography; Geography, Physical; Geology; Geosciences, Multidisciplinary; Meteorology & Atmospheric Sciences; Mineralogy; Mining & Mineral Processing; Paleontology; Water Resources	19	1,451,025	9	9
Arts	Archaeology; Art; Asian Studies; Classics; Cultural Studies; Dance; Film, Radio, Television; Folklore; Language & Linguistics; Linguistics; Literary Reviews; Literary Theory & Criticism; Literature; Literature, African, Australian, Canadian; Literature, American; Literature, British Isles; Literature, German, Dutch, Scandinavian; Literature, Romance; Literature, Slavic; Logic; Medieval & Renaissance Studies; Music; Poetry; Theater	24	1,322,894	2	2
Computer Science	Computer Science, Artificial Intelligence; Computer Science, Cybernetics; Computer Science, Hardware & Architecture; Computer Science, Information Systems; Computer Science, Interdisciplinary Applications; Computer Science, Software Engineering; Computer Science, Theory & Methods	7	1,253,026	4	4
Psychology	History & Philosophy Of Science; Psychology; Psychology, Applied; Psychology, Biological; Psychology, Clinical; Psychology, Developmental; Psychology, Educational; Psychology, Experimental; Psychology, Mathematical; Psychology, Multidisciplinary; Psychology, Psychoanalysis; Psychology, Social	12	578,906	10	9
Manage-ment	Business; Business, Finance; Economics; Management; Operations Research & Management Science; Public Administration	6	523,545	6	7
Law and Policy	Agricultural Economics & Policy; Industrial Relations & Labor; International Relations; Law; Medicine, Legal; Political Science	6	360,579	6	4
Building	Architecture; Construction & Building Technology; Engineering, Civil	3	257,992	3	5
Education	Education & Educational Research; Education, Scientific Disciplines; Education, Special	3	224,341	5	5
City Devel-opment	Planning & Development; Transportation; Transportation Science & Technology; Urban Studies; regional & urban planning; development studies	6	67,804	4	5

A.4 Articles without any citations

Figure A.4 presents the series of non-cited articles in the 14 clusters over time. Recall that 38.58% articles have never been cited, the distribution differs among clusters. Geography,
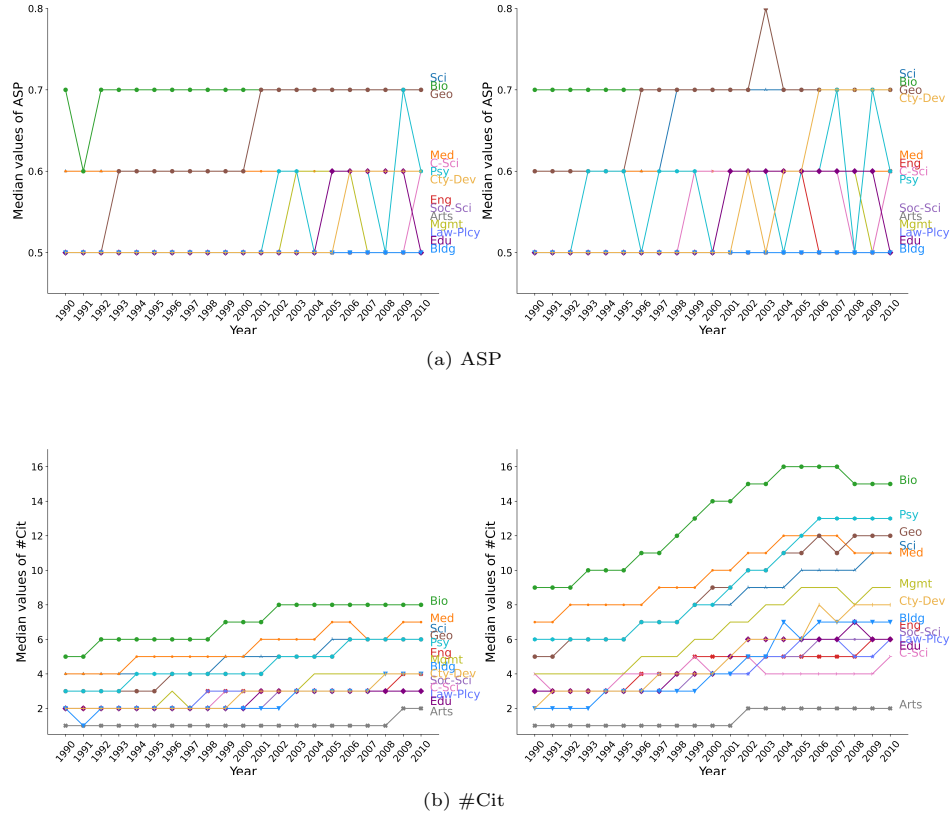
(a) ASP

(b) #Cit

Figure A.2. Median values of ASP and #Cit for all the 254 subjects (Left: with citation window of 5 years; Right: with citation window of 10 years).

Science and Biology have a relatively low ratio of non-citations, i.e. 40% around 1990-1994 and continuously drops to less than 20% in 2010. Medicine keeps a stable ratio around 40%. Geography shows impressive improvement, with the ratio decreasing from 38% in 1990 to 18% in 2010. Another cluster City Development reduces the ratio even from 77% in 1990 to 25% in 2010. Arts and Social Science have the highest non-cited ratio, where most, e.g. more than 82% and 66% articles, are never cited within 10 years.
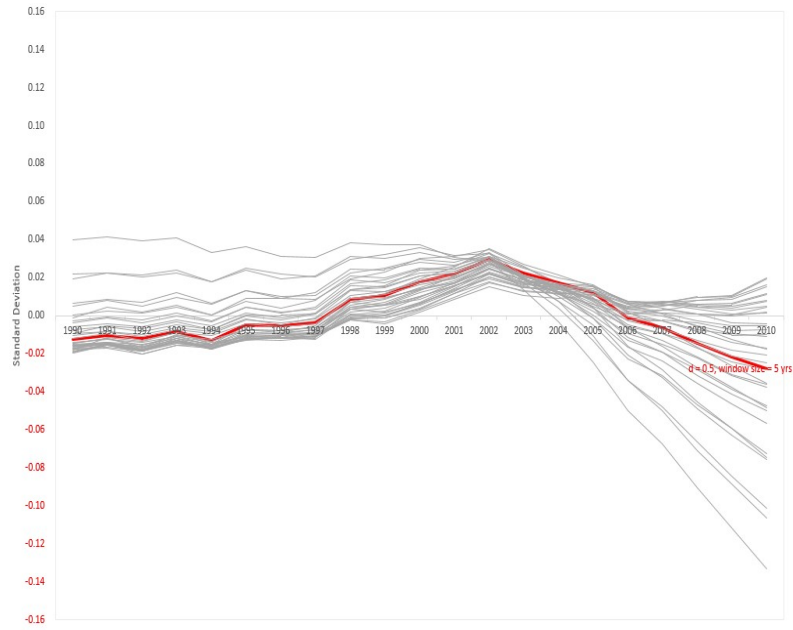
34

Figure A.3. Time evolution of subject scientific impact variations given various combination of damping factor and citing window between 1990 and 2010.
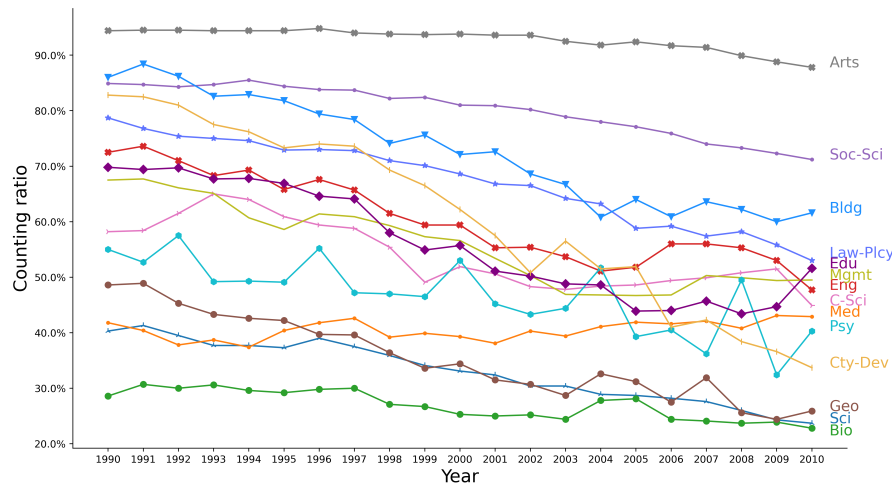

Figure A.4. Counting ratio of non-cited articles to total articles per cluster over years between 1990 and 2010.

35

Figure A.5 panel (a) presents the scatterplot of #Cit vs ASP. First glance shows strong positive correlation. One would expect that there is little difference of the two metrics for evaluating an article's prestige. Given the long tails of both metrics and the sensitivity of the correlation coefficients to outliers, we divided articles into deciles according to their sorted #Cit in each cluster, after removing the non-cited articles. The first group contains the top 10% articles with the highest #Cit in each cluster, and the last group (#10) has the last 10% of articles with the lowest #Cit for each cluster. The boxplot of the Pearson correlation coefficients between ASP and #Cit is displayed in panel (b) for each of the 10 groups. Except the top 10% articles have a high correlation at 0.76, the remaining 90% articles have, on average, correlations below 0.21. The correlation drops further to 0 for the least cited articles. There is generally a similar pattern for different clusters.

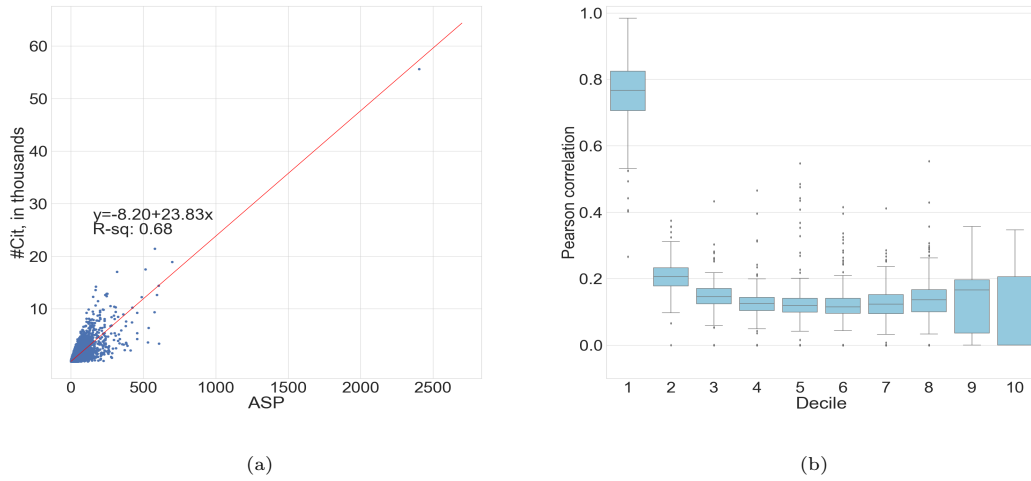(a)   (b)

Figure A.5. Panel (a). Scatterplot of #Cit vs ASP and the fitted line. Panel (b). Pearson correlation of ASP and #Cit in 10 groups and over years between 1990 and 2010. Each group includes the correlation values with respect to subjects and deciles over years. Each decile is obtained by dividing articles in each subject into 10 equal groups according to their sorted #Cit

References

Bergstrom, C. (2007). Eigenfactor: Measuring the value and prestige of scholarly journals. *College & research libraries news*, *68*, 314–316. Doi: https://doi.org/10.5860/crln.68.5.7804.

Boldi, P., Santini, M., & Vigna, S. (2005). Pagerank as a function of the damping factor. In *Proceedings of the 14th international conference on World Wide Web* (pp. 557–566).

Brezinski, C., & Redivo-Zaglia, M. (2006). The pagerank vector: properties, computation, approximation, and acceleration. *SIAM Journal on Matrix Analysis and Applications*, *28*, 551–575. Doi: https://doi.org/10.1137/050626612.

Chang, L. L.-H., Phoa, F. K. H., & Nakano, J. (2019). A new metric for the analysis of the scientific article citation network. *IEEE Access*, *7*, 132027–132032. Doi: https://doi.org/10.1109/ACCESS.2019.2937220.

Chen, C.-H. (2002). Generalized association plots: Information visualization via iteratively generated correlation matrices. *Statistica Sinica*, (pp. 7–29). Software: http://gap.stat.sinica.edu.tw/.

Chen, P., Xie, H., Maslov, S., & Redner, S. (2007). Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics*, *1*, 8–15. Doi: https://doi.org/10.1016/j.joi.2006.06.001.

Chu, J. S., & Evans, J. A. (2021). Slowed canonical progress in large fields of science. *Proceedings of the National Academy of Sciences*, *118*. Doi: https://doi.org/10.1073/pnas.2021636118.

Ding, Y. (2011). Applying weighted pagerank to author citation networks. *Journal of the American Society for Information Science and Technology*, *62*, 236–245.

Fountoulakis, K., & Yang, S. (2022). Open problem: Running time complexity of accelerated l_1-regularized pagerank. In *Conference on Learning Theory* (pp. 5630–5632). PMLR.

37

Garfield, E. (1955). Citation indexes for science. *Science*, *122*, 108–111. Doi: https://doi.org/10.1126/science.122.3159.108.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation: Journals can be ranked by frequency and impact of citations for science policy studies. *Science*, *178*, 471–479. Doi: https://doi.org/10.1126/science.178.4060.471.

Golub, G. H., & Loan, C. F. V. (2013). *Matrix Computations*. (4th ed.). John Hopkins University Press.

González-Pereira, B., Guerrero-Bote, V. P., & Moya-Anegón, F. (2010). A new approach to the metric of journals' scientific prestige: The SJR indicator. *Journal of informetrics*, *4*, 379–391. Doi: https://doi.org/10.1016/j.joi.2010.03.002.

Gu, C., & Wang, L. (2013). On the multi-splitting iteration method for computing pagerank. *Journal of Applied Mathematics and Computing*, *42*, 479–490. Doi: https://doi.org/10.1007/s12190-013-0645-5.

Hajarathaiah, K., Enduri, M. K., Anamalamudi, S., Subba Reddy, T., & Tokala, S. (2022). Computing influential nodes using the nearest neighborhood trust value and pagerank in complex networks. *Entropy*, *24*, 704. Doi: https://doi.org/10.3390/e24050704.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences*, *102*, 16569–16572. Doi: https://doi.org/10.1073/pnas.0507655102.

Ipsen, I. C., & Wills, R. S. (2006). Mathematical properties and analysis of google's pagerank. *Bol. Soc. Esp. Mat. Apl*, *34*, 191–196.

Krattenthaler, C. (2021). Was der h-index wirklich aussagt. *Mitteilungen der Deutschen Mathematiker-Vereinigung*, *29*, 124–128. Doi: https://doi.org/10.1515/dmvm-2021-0050.

Li, J., Yin, Y., Fortunato, S., & Wang, D. (2019a). A dataset of publication records for nobel laureates. *Scientific data*, *6*, 1–10. Doi: https://doi.org/10.1038/s41597-019-0033-6.

38

Li, M., Xu, J., Ge, B., Liu, J., Jiang, J., & Zhao, Q. (2019b). A deep learning methodology for citation count prediction with large-scale biblio-features. In *2019 IEEE International*

₆₇₅   *Conference on Systems, Man and Cybernetics (SMC)* (pp. 1172–1176). IEEE. Doi: https://doi.org/10.1109/SMC.2019.8913961.

Ma, N., Guan, J., & Zhao, Y. (2008). Bringing PageRank to the citation analysis. *Information Processing & Management*, *44*, 800–810. Doi: https://doi.org/10.1109/SMC.2019.8913961.

₆₈₀   Massucci, F. A., & Docampo, D. (2019). Measuring the academic reputation through citation networks via PageRank. *Journal of Informetrics*, *13*, 185–201. Doi: https://doi.org/10.1016/j.joi.2018.12.001.

Nie, Y., Zhu, Y., Lin, Q., Zhang, S., Shi, P., & Niu, Z. (2019). Academic rising star prediction via scholar's evaluation model and machine learning techniques. *Scientometrics*, *120*, 461–476.

₆₈₅   Doi: https://doi.org/10.1007/s11192-019-03131-x.

Palacios-Huerta, I., & Volij, O. (2004). The measurement of intellectual influence. *Econometrica*, *72*, 963–977. Doi: https://doi.org/10.1111/j.1468-0262.2004.00519.x.

Pinto, I. C. d. S. B. R. M. (2018). *PageRank: how to accelerate its computation*. Ph.D. thesis Universidade do Porto. https://hdl.handle.net/10216/118321.

₆₉₀   Srivastava, A. K., Garg, R., & Mishra, P. (2017). Discussion on damping factor value in pagerank computation. *International Journal of Intelligent Systems and Applications*, *9*, 19.

Srivastava, A. K., Srivastava, M., Garg, R., & Mishra, P. K. (2019). An Aitken-extrapolated Gauss-Seidel method for computing PageRank. *Journal of Statistics and Management Systems*, *22*, 199–222. Doi: https://doi.org/10.1080/09720510.2019.1580901.

₆₉₅   Tahamtan, I., Afshar, A. S., & Ahamdzadeh, K. (2016). Factors affecting number of citations: a comprehensive review of the literature. *Scientometrics*, *107*, 1195–1225. Doi: https://doi.org/10.1007/s11192-016-1889-2.

39

Tang, H., Shi, R., He, T.-S., Zhu, Y.-Y., Wang, T.-Y., Lee, M., & Jin, X.-M. (2021). Tensorflow solver for quantum pagerank in large-scale networks. *Science Bulletin*, *66*, 120–126. Doi: https://doi.org/10.1016/j.scib.2020.09.009.

Waltman, L., & Yan, E. (2014). Pagerank-related methods for analyzing citation networks. In *In Y. Ding, R. Rousseau, D. Wolfram (eds.) Measuring scholarly impact* (pp. 83–100). Springer International Publishing Switzerland.

Wu, H.-M., Tien, Y.-J., & Chen, C.-h. (2010). Gap: A graphical environment for matrix visualization and cluster analysis. *Computational Statistics & Data Analysis*, *54*, 767–778. Doi: https://doi.org/10.1016/j.csda.2008.09.029.

Xiao, S., Yan, J., Li, C., Jin, B., Wang, X., Yang, X., Chu, S. M., & Zha, H. (2016). On modeling and predicting individual paper citation count over time. In *IJCAI* (pp. 2676–2682).

Yan, E., & Ding, Y. (2011a). Discovering author impact: A pagerank perspective. *Information processing & Management*, *47*, 125–134.

Yan, E., & Ding, Y. (2011b). The effects of dangling nodes on citation networks. In *Proceedings of the 13th international conference on scientometrics and informetrics* (pp. 4–8).

Zhao, Q., & Feng, X. (2022). Utilizing citation network structure to predict paper citation counts: A deep learning approach. *Journal of Informetrics*, *16*, 101235. Doi: https://doi.org/10.1016/j.joi.2021.101235.

Zhao, Y., Li, X., Zhu, Y., Li, J., Wang, S., & Jiang, B. (2022). A scalable deep network for graph clustering via personalized pagerank. *Applied Sciences*, *12*, 5502. Doi: https://doi.org/10.3390/app12115502.