Journal of Informetrics 16 (2022) 101341

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Accidentality in journal citation patterns

Maciej J. Mrowinski[a,*], Marek Gagolewski[b,c], Grzegorz Siudem[a]

[a] Warsaw University of Technology, Faculty of Physics, ul. Koszykowa 75, Warsaw 00-662, Poland
[b] Deakin University, School of IT, Geelong, VIC 3220, Australia
[c] Systems Research Institute, Polish Academy of Sciences, ul. Newelska 6, Warsaw 01-447, Poland

ARTICLE INFO

Keywords:
Complex networks
DBLP
CORE
Rich get richer
Pareto principle

ABSTRACT

We study an agent-based model for generating citation distributions in complex networks of scientific papers, where a fraction of citations is allotted according to the preferential attachment rule (rich get richer) and the remainder is allocated accidentally (purely at random, uniformly). Previously, we derived and analysed such a process in the context of describing individual authors, but now we apply it to scientific journals in computer and information sciences. Based on the large DBLP dataset as well as the CORE (Computing Research and Education Association of Australasia) journal ranking, we find that the impact of journals is correlated with the degree of accidentality of their citation distribution. Citations to impactful journals tend to be more preferential, while citations to lower-ranked journals are distributed in a more accidental manner. Further, applied fields of research such as artificial intelligence seem to be driven by a stronger preferential component – and hence have a higher degree of inequality – than the more theoretical ones, e.g., mathematics and computation theory.

1. Introduction

Citation networks – graphs in which vertices represent publications and edges correspond to the citations between them – are governed by many latent variables (research fields with their specific publishing practices, journals policies, co-authorship patterns, etc.). These networks are rich and complex structures, usually too sophisticated to explain them in their entirety. Therefore, one by necessity needs to focus only on small subsets of citation networks or aggregate their characteristics and consider more coarse information granules.

There are numerous aggregation procedures which can be applied on the whole, raw graph representing a citation network. The most microscopic (i.e., most fine-grained) approaches concentrate on the properties of individual papers, e.g., the distribution of the number of citations (Brzezinski, 2015; Redner, 1998; Thelwall, 2016a; 2016b; 2016c). There are, however, also some higher description levels such as per-author and per-institution ones (Chatterjee, Ghosh, & Chakrabarti, 2016; Egghe, 2009; Néda, Varga, & Biró, 2017; Siudem, Żogała-Siudem, Cena, & Gagolewski, 2020) as well as the journal-level setting, which we explore in this very contribution.

Since the first work concerned with journal description (Gross & Gross, 1927), important problems about the citation patterns (Chatterjee et al., 2016), database completeness (Liu & Fang, 2020), and their time-dependence (Fang, 2020) remain open. Moreover, there are other topics not directly related to citations (e.g., questions of the language of published papers, see Kulczycki et al., 2020). However, since the seminal paper Garfield, 1972, which introduced journals' Impact Factor, most of the research focuses on this metric (Brzezinski, 2014; Liu & Fang, 2020; Mansilla, Köppen, Cocho, & Miramontes, 2007) or suggests its modifications (e.g., Fang, 2020;

* Corresponding author.
E-mail address: maciej.mrowinski@pw.edu.pl (M.J. Mrowinski).

https://doi.org/10.1016/j.joi.2022.101341
Received 22 February 2022; Received in revised form 20 September 2022; Accepted 13 October 2022

Kim, Portenoy, West, & Stovel, 2020; Zeng & Shi, 2021). Also, there are purely data-related works on Impact Factors (Fang, 2020; Liu & Fang, 2020), analytical studies (e.g., the search for the Impact Factor distributions Brzezinski, 2014; Huang, 2017), articles on properties of other metrics (Prathap, 2020) or proposals of models for Impact Factors (Mansilla et al., 2007) with theoretical discussions (Sarabia, Prieto, & Trueba, 2012). The Impact Factor is a measure aggregating journals' citations vectors (numbers of citations received by every paper published in the journal) into one number. Yet, research on the properties of unaggregated data is much less advanced. There were some attempts to apply the Discrete Generalised Beta Distribution to non-aggregated citation vectors of journals (Campanario, 2010), dynamical two-state Markovian (Delbianco, Fioriti, Hernandez-Chanto, & Tohmé, 2020), or universal (i.e., nonparametric) (Chatterjee et al., 2016) approaches, but more studies are necessary to capture the complexity of these networks.

The research we present in this paper lies somewhere in-between these two extremes – the dynamical approach which requires more parameters on the one side and the universal approach which attempts to be parameter-free on the other. We apply the model introduced in Siudem et al. (2020) to entire scientific journals. Its original purpose was to describe the individual authors' citation vectors. Most importantly, it is governed by only three simple parameters: the number of published papers N, the total number of received citations C, and the preferential-to-accidental attachment ratio ρ, which describes the ratio between the citations allocated based on the rich get richer principle (Perc, 2014) (success breads success Price, 1963, the Matthew effect Merton, 1968) and those allotted purely at random.

It is worth noting that Peterson, Pressé, & Dill, 2010 introduced a similar model to Siudem et al. (2020) but its authors considered the recursive search rather than the preferential mechanism and solved the model in the distribution- rather than ranks-domain (see, e.g., Bertoli-Barsotti & Lando, 2019). Also, authors of Peterson et al. (2010) tested their results only for one journal (*Physical Review D*). In this paper, we test the model using the entire DBLP database (Tang et al., 2008) of computer and information science papers.

This way, we are able to gain some interesting insights into both the dimensionality of the citation space of journals and measures of their scientific impact. From Siudem et al. (2020), as we have already mentioned, we know that three numbers (N, C, ρ) can describe the citation vectors at the author level reasonably well. The search to quantify the degree of luck vs. reason has a long history in bibliometric research, as it dates back to the classical work Price, 1976; similar models were investigated later in Ionescu & Chopard (2013); Siudem, Nowak, & Gagolewski (2022); Żogała-Siudem, Siudem, Cena, & Gagolewski (2016), amongst others. Recently, the question of the role of randomness in scientific success (Janosov, Battiston, & Sinatra (2020)) became an active thread in bibliometric research (Heesen, 2017; Pluchino, Biondo, & Rapisarda, 2018; Pluchino et al., 2019) and the so-called Science of Science (Sinatra, Wang, Deville, Song, & Barabási, 2016). With this work, we would like to expand this list to include a study of the structure of randomness in citation patterns at the journal level as well as the features of the journals' citation vectors. In particular, we are concerned with the question whether the most highly ranked journals receive citations based on the said rich get richer principle, which is known to lead to high inequalities in the citation distributions. Also, we shall take a look at the differences between specific research subfields, e.g., computation theory, informetrics, and artificial intelligence.

2. Model

The model introduced in Siudem et al. (2020) describes an iterative procedure where citations are distributed amongst scientific papers published by individual scientists. In every time step, one paper is added to the system, and the articles already published receive new citations according to a mixture of the preferential attachment rule and sheer chance. The procedure stops once all papers have been output and every citation has been allocated.

Due to the assumptions of the model, each author is characterised by three parameters: the total number of published papers N that received at least one citation, the total number of citations C to be distributed, and the fraction of citations distributed according to the preferential attachment rule $\rho \in (0, 1)$. Hence, $(1-\rho)100\%$ of citations are distributed accidentally (uniformly).

The model recreates the citation vector of an author $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, where X_k is the number of citations of the k-th most cited paper and $\sum_{k=1}^{N} X_k = C$.

To avoid ambiguity, we will mark the predictions generated by the model with the hat symbol, i.e.,

$$\hat{X}_k(N, C, \rho) = \frac{1-\rho}{\rho}\frac{C}{N}\left(\prod_{i=k}^{N}\frac{i}{i-\rho} - 1\right) = \frac{1-\rho}{\rho}\frac{C}{N}\frac{(k)_{N-k+1} - (k-\rho)_{N-k+1}}{(k-\rho)_{N-k+1}}, \tag{1}$$

where the Pochhammer symbol $(k)_n$, see Olver, 2021, Sec. 5.2, is defined by means of the Euler gamma function as:

$$(k)_m = \frac{\Gamma(k+m)}{\Gamma(k)} = k(k+1)\cdots(k+m-1). \tag{2}$$

While the original formulation of the model was restricted ρ to (0,1), in Gagolewski, Żogała-Siudem, Siudem, & Cena (2022) we noted that it is actually possible to consider $\rho \in (-\infty, 1)$, with $\rho = -\infty$ yielding the same number of citations per paper. The value of ρ (including negative ones) can be estimated from a sample using the following equation (see the aforementioned article for details)

$$\rho_R = \frac{N - 4er(\mathbf{X}) + 3}{N - 2er(\mathbf{X}) + 1}, \tag{3}$$

where the expected rank is given by

$$er(\mathbf{X}) = \sum_{i=1}^{N} i\frac{X_i}{C}. \tag{4}$$

We shall use this estimator to obtain all the results presented below.

It is worth noting that the interpretation of ρ presented in this section – as a parameter quantifying the intensity of the preferential attachment – must be reexamined if one allows for negative values of ρ. With $\rho < 0$, a fraction of citations is still distributed uniformly amongst all papers published up to a certain point. Such a process alone would lead to a distribution of citations skewed towards older papers. This, however, is counterbalanced by the fact that with $\rho < 0$, a fraction of citations is also preferentially *removed* from articles. It follows that as ρ tends to $-\infty$, the distribution of citations is becoming more and more uniform.

This description – or rather, the definition of the model itself – shows that while negative values of ρ are valid, in such cases ρ loses its interpretation as a natural measure of rich-get-richer-ness. Still, this is only the case for few journals in our database.

3. Dataset

We used the AMiner Citation Network Dataset (DBLP-Citation-network V12) (Tang et al., 2008), which provides data on ca. 5 million computer science papers and 45 million citation relationships. The metadata of most articles available in the dataset contains the title and type of the outlet in which an article was published (e.g., “Book”).

We filtered out all papers that were not published in journals or conference proceedings, and for which the year of publication, venue name, or DOI were not available. This reduced the dataset to about 3.5 million articles. Next, using the *crossrefapi* library[1] which provides a Python API to Crossref, we matched the DOIs of the articles to the ISSNs of the outlets. This was possible in 60% of cases. Based on the ISSNs, we created a list of about 6000 publication venues (each corresponding to a set of ISSNs acquired during the previous step), and assigned articles to these venues accordingly. As for the remaining 40% of the papers for which the ISSNs could not be found using the aforementioned library, we managed to match about 86% of them to the outlets using the venue names provided by DBLP. The remaining papers were discarded.

It is worth noting that we used citation counts available directly in the metadata of articles, instead of counting citations in the citations network that can be recreated using the dataset. The former approach is more reliable, as the dataset does not contain complete information about citation relationships.

Furthermore, we matched the journals (once again, using the ISSNs) to scores available in the CORE (Computing Research and Education Association of Australasia) Journal Ranking Portal.[2] Journals listed in the CORE ranking are assigned one of four ranks – A* (the highest rank), A, B, or C (the lowest rank). Detailed descriptions of these ranks are available on the CORE website.[3] CORE also assigns a primary field of research (FoR[4]) code to each journal, which allows for the outlet’s subject classification. We found 687 journals from our dataset to be listed in the CORE ranking.

Similarly, using ISSNs, we found some of the journals (3987) in the 2021 SCImago Journal & Country Rank database and assigned the two following indicators[5] to each of these journals:

SJR (SCImago Journal Rank)– expresses the average number of weighted citations received in a selected year by the documents published in a given journal in the three preceding years, i.e., weighted citations received in year X to documents published in the journal in years $X-1$, $X-2$, and $X-3$.

Cites per Document (2 years)– average citations per document in a 2 year period. It is computed based on the number of citations received by a journal in the current year and the documents published in the two previous years, i.e., ratio of the citations received in year X to the number of documents published in years $X-1$ and $X-2$.

We also used the information available in SCImago to separate journals into open access (607 in total) and non-open access (3380) venues.

4. Results

In Siudem et al., 2020, the model given by Eq. (1) was used to estimate the ρ parameter of individual authors. However, much like a single author, an entire scientific journal can be characterised by a citation vector created by grouping all the articles published therein. Such journal citation vectors are a starting point for the calculation of many indicators of scientific impact (e.g., how the Impact Factor changes over time).

It should be noted, though, that these indicators usually aggregate the data further and operate on macroscopic parameters, such as the total number citations or the number papers. However, Gagolewski (2013); Siudem et al. (2020), amongst others, argued that no single aggregated measure can accurately describe the scientific success (or productivity/impact) of a scientist, which is a consequence of the multidimensionality of the citation space. We believe that the same argument holds for scientific journals. In order to test this generalisation, for each journal we computed its N (i.e., number of papers published therein) and C (total citations to papers in that outlet). Then, we estimated ρ for each journal using Eq. (3).

It is worth emphasising that citation vectors for journals are much longer than those of the individual authors – and of course more data means that the results are expected to be much more reliable.

[1] https://github.com/fabiobatalha/crossrefapi.
[2] http://portal.core.edu.au/jnl-ranks/.
[3] https://www.core.edu.au/conference-portal/journal-rankings-history.
[4] https://www.abs.gov.au/statistics/classifications/australian-and-new-zealand-standard-research-classification-anzsrc/latest-release.
[5] https://www.scimagojr.com/help.php.

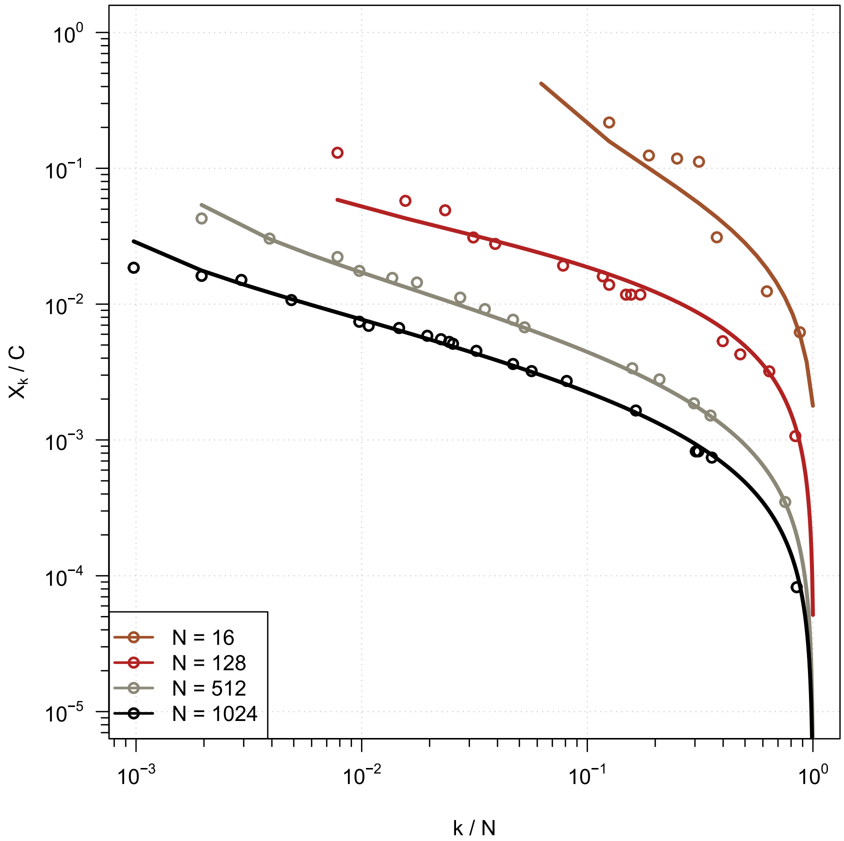

Fig. 1. Predictions generated by the model $\hat{X}_k$ (solid lines) vs. real data X_k (points; the number of citations X of the kth most cited paper) for four example journals with various numbers of published papers (each set of points corresponds to an individual journal). Given enough data samples N, the model fits data very well.

Fig. 1 shows examples of predictions of the model given by Eq. (1) for four example journals/conference proceedings with various numbers of published papers (IEEE Nuclear Science Symposium with $N = 16$, International Journal of Interactive Multimedia and Artificial Intelligence with $N = 128$, Distributed Computing in Sensor Systems with $N = 512$, International Journal of Foundations of Computer Science with $N = 1024$). The citation vectors obtained from the model are in good agreement with the data, especially in the case of journals with higher numbers of papers, which agreed with the findings presented in Siudem et al. (2020) in the context of describing the individuals.

Next, let us inspect how the ρ parameter is distributed across our data sample. Fig. 2(a) presents the violin plots (box and whisker charts together with kernel density estimators) for all journals. Also, the information available in the SCImago Journal & Country Rank database allowed us to divide the journals into two subsets – open (average $\rho = 0.29$) and non-open access (average $\rho = 0.34$) ones; see Fig. 2(b). We used the Mann–Whitney U Test to compare these subsets: the hypothesis that the distributions of ρ in both subsets are the same must be rejected ($U = 101657$, p-value $\simeq 0.00$).

Furthermore, let us study how the ρ parameter relates to some measures of journal impact. While the Impact Factor (IF) seems like the most obvious choice, it would be difficult to calculate it for all journals using the incomplete citation relationships available in our dataset (the database does not contain the entire citation network and the number of citations calculated by counting edges in the network can differ significantly from the actual one). Instead, in order to quantify the impact of all the journals, we decided to employ a simpler measure: the average number of citations per article (C/N).

Fig. 3 shows a positive correlation between the average number of citations per article and ρ (the Spearman rank correlation coefficient $r_S = 0.56$). Note that the red points correspond to the average value of C/N calculated by assigning each journal with $\rho \in [0, 1]$ to one of 10 bins of equal width defined on the interval $[0, 1]$ and computing the geometric mean of C/N in each bin (red points are placed in the middle of each bin). The red lines simply connect the pairs of adjacent points.

Overall, it seems that the preferential attachment rule plays a more important role for more impactful venues (impactful in terms of the selected metric). On the other hand, the degree of randomness (random distribution of citations without any bias) in the citation distribution is more prevalent for less impactful venues.

Fig. 4 (a) and (b) show that SJR and Cites per Document (2 years) indicators exhibit a small positive correlation with ρ (Spearman's rank correlation coefficients $r_S = 0.18$ and $r_S = 0.21$, respectively; the positions of the red points, which correspond to the average

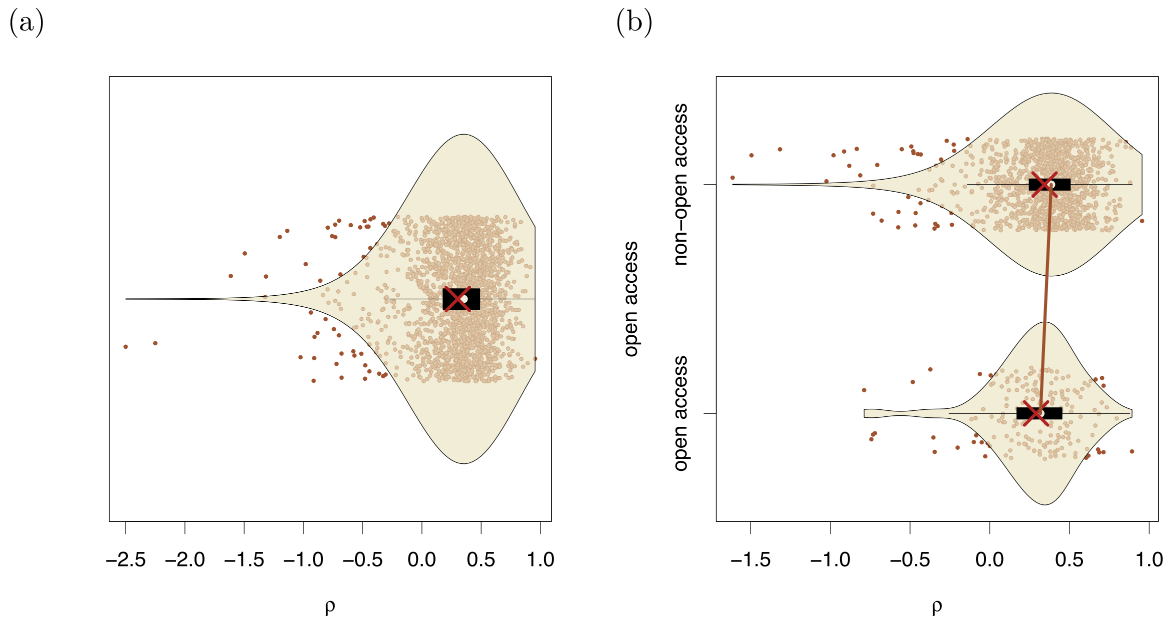

Fig. 2. Distributions (violin plots) of the ρ parameter for a) all journals and for b) open and non-open access journals.

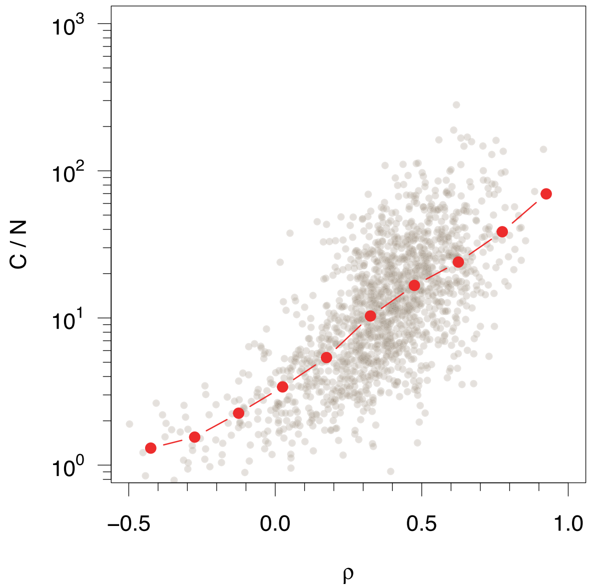

Fig. 3. The number of citations per article C/N as a function of the ρ parameter for journals that published more than 100 papers. The more cited outlets tend to have, on average, more preferentially-attributed citations.

values, were calculated in the same way as in Fig. 3). This strengthens our previous statement that citations to more cited journals are preferential and citations to less impactful venues are distributed more accidentally.

Fig. 5(a) shows the distribution of ρ for journals assigned to each rank separately, while examples of journals with the highest/lowest values of ρ for each CORE rank can be found in Table 1. These results support the above hypothesis. Journals with higher ranks are characterised by higher values of ρ, which indicates that the distribution of citations (i.e., accidental or preferential) is indeed correlated with the impact of a journal. This is supported by the Mann–Whitney U Test. Assuming a significance level of 5%, the null hypothesis that the underlying ρ distributions are the same cannot be rejected for the A* vs. A ($U = 1935$, p-value $= 0.24$) and A vs. B ($U = 7612$, p-value $= 0.13$) pairs. However, for all the remaining pairs, it must be rejected in favour of the alternative.

The CORE database allowed us to perform a similar analysis of the distribution of ρ for journals that belong to different fields of research (Fig. 5(b)). More applied fields seem to have higher ρs, however there is some inherent variability even within each FoR; see Table 2 for a sample of journals in library and information studies (including the current outlet).

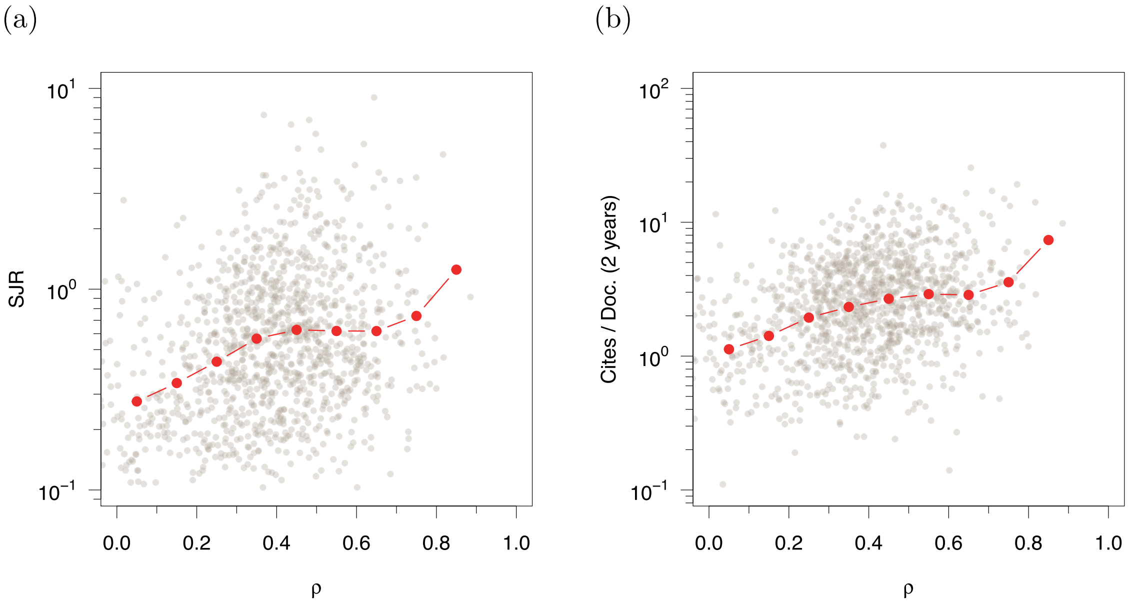

Fig. 4. SJR (a) and Cites per Document (2 years) (b) as functions of ρ (with negative values removed for clarity) for journals found in the SCImago Journal & Country Rank database that published more than 100 papers.

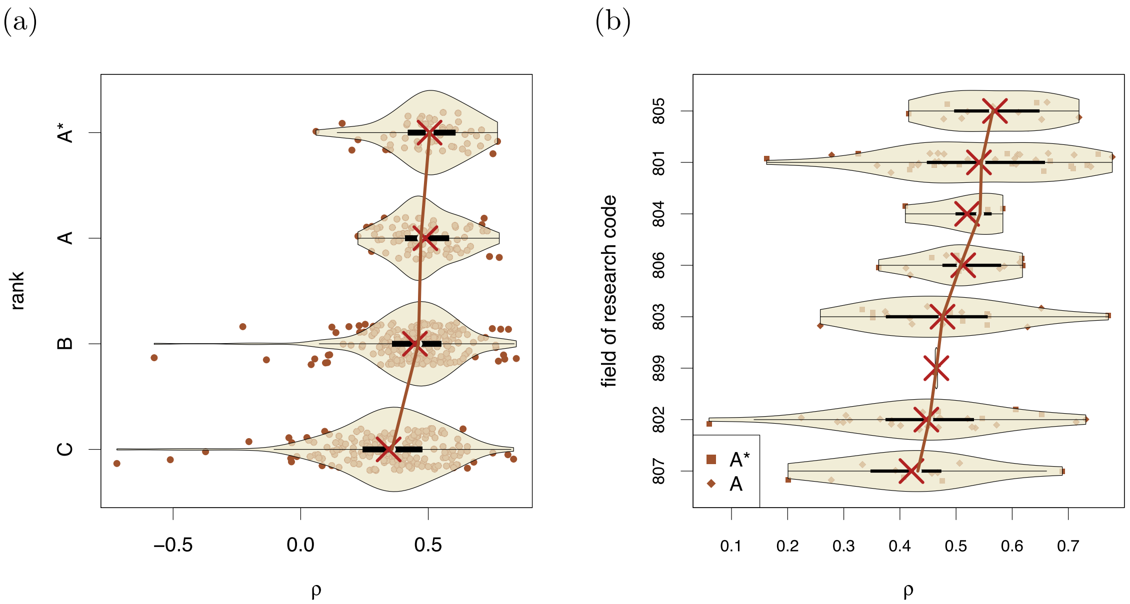

Fig. 5. Distribution of ρ according to (a) journal rank and (b) field of research (FoR, A* and A-ranked journals only), where the codes are: 801 – Artificial Intelligence and Image Processing, 802 – Computation Theory and Mathematics, 803 – Computer Software, 804 – Data Format, 805 – Distributed Computing, 806 – Information Systems, 807 – Library and Information Studies, 899 – Other Information and Computing Sciences.

5. Discussion and conclusion

Our research indicates that the model presented in Siudem et al. (2020) can be successfully applied to scientific journals. The model accurately recreates journal citation vectors (which are the basis for many bibliometric indicators). It also allows for the estimation of the degree to which the citations to the articles published in a given journal are preferential vs. accidental.

This degree of accidentality (randomness), provides us with an interesting insight into the nature of the measures of scientific impact, the dimensionality of the space of citations, and the dynamics of the process of generating citation distribution. Our results show that citations to more “impactful” journals (i.e., journals that attract more citations per article) are less accidental than citations to less cited journals. As such, the parameter ρ seems to be, on average, correlated with the impact of a journal.

Table 1
A and A* journals with the lowest and the highest ρ; the FoR codes are explained in the caption of Fig. 5.

ρ	Rank	FoR	Journal
0.06	A*	802	Computers and Structures
0.16	A*	801	Cognition
0.20	A*	807	Annual Review of Information Science and Technology
0.22	A	802	Logical Methods in Computer Science
0.26	A	803	IEEE Transactions on Reliability
0.28	A	807	International Journal of Digital Curation
0.28	A	801	Artificial Intelligence in Medicine
0.29	A	802	Journal of Experimental Algorithmics
0.30	A	802	Journal of Graph Theory
0.30	A	802	Journal of Applied Logic
0.71	A*	801	IEEE Transactions on Evolutionary Computation
0.71	A	801	Evolutionary Computation
0.71	A	802	Theory of Computing
0.72	A	805	Computer Networks
0.73	A	802	Statistics and Computing
0.73	A*	801	Computational Linguistics
0.74	A	801	Data Mining and Knowledge Discovery
0.75	A*	801	International Journal of Computer Vision
0.77	A*	803	ACM Computing Surveys
0.78	A	801	Machine Learning

Table 2
Sample journals in FoR 807 – Library and Information Studies. For some journals the CORE rank was not available.

ρ	Rank	FoR	Journal
−0.23	B	807	Research Evaluation
0.28	A	807	International Journal of Digital Curation
0.20	A*	807	Annual Review of Information Science and Technology
0.39	A	807	Journal of Informetrics
0.41	–	807	Scientometrics
0.46	A	807	Information Sciences
0.41	A*	807	Journal of the Association for Information Science and Technology
0.61	–	807	Journal of Documentation
0.47	A*	807	Archival Science
0.61	B	807	First Monday
0.79	B	807	Education for Information

There might be an intuitive explanation for this phenomenon: when authors prepare lists of bibliographic references, they are of course more likely to include, in the first place, the most highly cited papers (which of course does not say anything about their true quality) in the leading journals. These references are chosen carefully and with intent, which is exactly what we call the preferential attachment. Still, the authors can also include some less cited but still relevant or even not relevant at all references. This in turn may be thought of as an accidental distribution of citations.

While it would certainly require further studies, our research suggests that parameter ρ could potentially be used as a measure of scientific impact (popularity), however, never on its own. The results presented in this paper indicate that the citation record space of scientific publications, much like the citation record space of individual authors, is multi-dimensional. As a result, no single bibliometric measure (e.g., the Impact Factor) can capture its complexity in its entirety. The appropriate choice of bibliometric measures, as well as the description of the relationship between ρ and scientific impact (underlying dynamical processes governing the evolution of citation networks), are open for further research.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by the Australian Research Council Discovery Project ARC DP210100227 (MG) and by the POB Research Centre Cybersecurity and Data Science of Warsaw University of Technology within the Excellence Initiative Program – Research University (ID-UB) (MM and GS).

We thank the reviewers for constructive remarks that led to the improvement of the manuscript as well as one of the anonymous referees of our previous paper (Gagolewski et al., 2022) for suggesting the ρ_R estimator.

References

Bertoli-Barsotti, L., & Lando, T. (2019). How mean rank and mean size may determine the generalised Lorenz curve: With application to citation analysis. *Journal of Informetrics, 13*(1), 387–396.

Brzezinski, M. (2014). Empirical modeling of the impact factor distribution. *Journal of Informetrics, 8*(2), 362–368.

Brzezinski, M. (2015). Power laws in citation distributions: Evidence from Scopus. *Scientometrics, 103*(1), 213–228.

Campanario, J. M. (2010). Distribution of ranks of articles and citations in journals. *Journal of the American Society for Information Science and Technology, 61*(2), 419–423.

Chatterjee, A., Ghosh, A., & Chakrabarti, B. K. (2016). Universality of citation distributions for academic institutions and journals. *PloS One, 11*(1), e0146762.

Delbianco, F., Fioriti, A., Hernandez-Chanto, A., & Tohmé, F. (2020). A Markov-switching approach to the study of citations in academic journals. *Journal of Informetrics, 14*(4), 101081.

Olver, F. W. J, Olde Daalhui, A. B, Lozier, D. W, Schneider, B. I., Boisvert, R. F, & Clark, C. W. et al. (2021). NIST digital library of mathematical functions. http://dlmf.nist.gov/.

Egghe, L. (2009). Lotkaian informetrics and applications to social networks. *Bulletin of the Belgian Mathematical Society - Simon Stevin, 16*(4), 689–703.

Fang, H. (2020). Investigating the journal impact along the columns and rows of the publication-citation matrix. *Scientometrics, 125*(3), 2265–2282.

Gagolewski, M. (2013). Scientific impact assessment cannot be fair. *Journal of Informetrics, 7*(4), 792–802. 10.1016/j.joi.2013.07.001.

Gagolewski, M., Żogała-Siudem, B., Siudem, G., & Cena, A. (2022). Ockham's index of citation impact. *Scientometrics, 127*(5), 2829–2845.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science, 178*(4060), 471–479.

Gross, P. L., & Gross, E. M. (1927). College libraries and chemical education. *Science, 66*(1713), 385–389.

Heesen, R. (2017). Academic superstars: Competent or lucky? *Synthese, 194*(11), 4499–4518.

Huang, D.-W. (2017). Impact factor distribution revisited. *Physica A: Statistical Mechanics and its Applications, 482*, 173–180.

Ionescu, G., & Chopard, B. (2013). An agent-based model for the bibliometric H-index. *European Physical Journal B, 86*, 426.

Janosov, M., Battiston, F., & Sinatra, R. (2020). Success and luck in creative careers. *EPJ Data Science, 9*(1), 9.

Kim, L., Portenoy, J. H., West, J. D., & Stovel, K. W. (2020). Scientific journals still matter in the era of academic search engines and preprint archives. *Journal of the Association for Information Science and Technology, 71*(10), 1218–1226.

Kulczycki, E., Guns, R., Pölönen, J., Engels, T. C., Rozkosz, E. A., Zuccala, A. A., et al., (2020). Multilingual publishing in the social sciences and humanities: A seven-country European study. *Journal of the Association for Information Science and Technology, 71*(11), 1371–1385.

Liu, X. Z., & Fang, H. (2020). A comparison among citation-based journal indicators and their relative changes with time. *Journal of Informetrics, 14*(1), 101007.

Mansilla, R., Köppen, E., Cocho, G., & Miramontes, P. (2007). On the behavior of journal impact factor rank-order distribution. *Journal of informetrics, 1*(2), 155–160.

Merton, R. K. (1968). The Matthew effect in science. *Science, 159*(3810), 56–63. 10.1126/science.159.3810.56.

Néda, Z., Varga, L., & Biró, T. S. (2017). Science and facebook: The same popularity law!. *PloS One, 12*.

Perc, M. (2014). The Matthew effect in empirical data. *Journal of The Royal Society Interface, 11*(98), 20140378.

Peterson, G. J., Pressé, S., & Dill, K. A. (2010). Nonuniversal power law scaling in the probability distribution of scientific citations. *Proceedings of the National Academy of Sciences, 107*(37), 16023–16027.

Pluchino, A., Biondo, A. E., & Rapisarda, A. (2018). Talent versus luck: The role of randomness in success and failure. *Advances in Complex Systems, 21*(03n04), 1850014.

Pluchino, A., Burgio, G., Rapisarda, A., Biondo, A. E., Pulvirenti, A., Ferro, A., et al., (2019). Exploring the role of interdisciplinarity in physics: Success, talent and luck. *PLoS One, 14*(6), e0218793.

Prathap, G. (2020). Letter to the editor: Journal indicators from a dimensionality perspective. *Scientometrics, 122*(2), 1259–1265.

Price, D. J. (1963). *Little science, big science*. New York: Columbia Univ. Press.

Price, D. J. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science, 27*(5), 292–306.

Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B - Condensed Matter and Complex Systems, 4*(2), 131–134.

Sarabia, J. M., Prieto, F., & Trueba, C. (2012). Modeling the probabilistic distribution of the impact factor. *Journal of Informetrics, 6*(1), 66–79.

Sinatra, R., Wang, D., Deville, P., Song, C., & Barabási, A.-L. (2016). Quantifying the evolution of individual scientific impact. *Science, 354*(6312).

Siudem, G., Nowak, P., & Gagolewski, M. (2022). Power laws, the price model, and the Pareto type-2 distribution. *Physica A: Statistical Mechanics and its Applications, 606*, 128059. 10.1016/j.physa.2022.128059.

Siudem, G., Żogała-Siudem, B., Cena, A., & Gagolewski, M. (2020). Three dimensions of scientific impact. *Proceedings of the National Academy of Sciences, 117*(25), 13896–13900.

Thelwall, M. (2016a). Are the discretised lognormal and hooked power law distributions plausible for citation data? *Journal of Informetrics, 10*(2), 454–470.

Thelwall, M. (2016b). Are there too many uncited articles? Zero inflated variants of the discretised lognormal and hooked power law distributions. *Journal of Informetrics, 10*(2), 622–633.

Thelwall, M. (2016c). The discretised lognormal and hooked power law distributions for complete citation data: Best options for modelling and regression. *Journal of Informetrics, 10*(2), 336–346.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). Arnetminer: Extraction and mining of academic social networks. In *Proc. 14th ACM SIGKDD intl. conf. knowledge discovery and data mining* (pp. 990–998).

Zeng, Z., & Shi, L. (2021). A two-dimensional journal classification method based on output and input factors: Perspectives from citation and authorship related indicators. *Scientometrics, 126*(5), 3929–3964.

Żogała-Siudem, B., Siudem, G., Cena, A., & Gagolewski, M. (2016). Agent-based model for the h-index—Exact solution. *European Physical Journal B*, 21.