RESEARCH ARTICLE

Heavy-tailed distribution of the number of papers within scientific journals

Robin Delabays[1] **and Melvyn Tyloo**[2]

[1]Center for Control, Dynamical Systems and Computation, UC Santa Barbara, Santa Barbara, CA 93106 USA
[2]Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545 USA

Keywords: cumulative advantage, heavy-tail, preferential attachment, publications, scholarly journals

ABSTRACT

Scholarly publications represent at least two benefits for the study of the scientific community as a social group. First, they attest to some form of relation between scientists (collaborations, mentoring, heritage, ...), useful to determine and analyze social subgroups. Second, most of them are recorded in large databases, easily accessible and including a lot of pertinent information, easing the quantitative and qualitative study of the scientific community. Understanding the underlying dynamics driving the creation of knowledge in general, and of scientific publication in particular, can contribute to maintaining a high level of research, by identifying good and bad practices in science. In this article, we aim to advance this understanding by a statistical analysis of publication within peer-reviewed journals. Namely, we show that the distribution of the number of papers published by an author in a given journal is heavy-tailed, but has a lighter tail than a power law. Interestingly, we demonstrate (both analytically and numerically) that such distributions match the result of a modified preferential attachment process, where, on top of a Barabási-Albert process, we take the finite career span of scientists into account.

1. INTRODUCTION

One of the core mechanism in the practice of science is the self-examination of a field of research. The validation of a scientific result is always collective, in the sense that it has been scrutinized, criticized, and (hopefully) validated by a sufficient number of peers. Furthermore, any scientific result is permanently subject to new evaluation and might be replaced by more accurate work. At the level of a community, scientists are then used to criticize the work of colleagues and to have their work criticized by them. It is then not surprising that some scientists started to study (and thus somehow critically assess) the scientific community itself (Price, 1963).

The quantitative study of the scientific community, sometimes referred to as *Science of Science* (Fortunato, Bergstrom et al., 2018; Narin, 1976; Price, 1976; van Raan, 2019), is a key step to unravel the underlying behaviors of its composing agents (authors, journals, institutions, etc.). Pioneered by the early works of Lotka (1926), the science of science gained a lot of momentum in the second half of the 20th century, with the creation of the first databases of scientific publications (Garfield, 1955; Merton, 1968; Price, 1965). More recently, the scientometric investigations have been significantly eased by the emergence of large online databases of scientific publications (Web of Science, PubMed, arXiv, ...) and the ever-increasing computation power

an open access journal

Check for updates

Citation: Delabays, R., & Tyloo, M. (2022). Heavy-tailed distribution of the number of papers within scientific journals. *Quantitative Science Studies*, *3*(3), 776–792. https://doi.org/10.1162/qss_a_00201

DOI: https://doi.org/10.1162/qss_a_00201

Peer Review: https://publons.com/publon/10.1162/qss_a_00201

Received: 18 February 2022
Accepted: 21 June 2022

Corresponding Author:
Robin Delabays
robindelabays@ucsb.edu

Handling Editor:
Ludo Waltman

The MIT Press

of modern computers. These improvements have allowed the analysis of scientometric indicators on a larger scale (Frandsen & Nicolaisen, 2017; Wang & Waltman, 2016) and with finer resolution in terms of publication units (considering single articles instead of whole journals (e.g., Waltman & van Eck, 2012) and time (Newman, 2001; Egghe & Rousseau, 2000). For a clear historical overview of scientometrics, we refer to van Raan (2019).

The science of science has the potential to help maintaining the quality of research, and is thus a good use of public funding. There are nowadays an increasing number of scientific papers (Bornmann & Mutz, 2015; Price, 1965), combined with the ubiquitous presence of *predatory journals* which publish the papers they receive, charging publication fees, but without performing the fundamental editorial work that guarantees the papers' quality (e.g., quality and pertinence check, referee process; Bohannon, 2013; Sorokowski, Kulczycki et al., 2017). In such a context, distinguishing bad practices from honest work in scientific publishing becomes more and more challenging. Understanding the underlying dynamics of scientific publication will be instrumental in this endeavor.

The fight against predatory publishing has benefited from the effort of many dedicated citizens, whose initiatives have shown their efficacy (Butler, 2013; Grudniewicz, Moher et al., 2019), as well as their limits (Beall, 2017). With regard to the proliferation of predatory journals, the task of identifying all of them unequivocally is overwhelming. In such a context, the ability to perform a preliminary data-based sanity check of a given journal would allow resources to be focused on the more problematic venues. However, such an approach requires an accurate understanding of the quantitative and qualitative characteristics of scientific journals, which is still scarce.

The quality of a scientist's work is commonly quantified by two different, but related, measures, namely, their number of papers and the number of citations thereof (summarized in the *h*-index [Hirsch, 2005; Siudem, Żogała-Siudem et al., 2020]). The vast majority of investigations about the scientific publication process are focused on the citation side. These analyses mostly aim to describe how the citation network impacts the number of citations a given paper is (and therefore its authors are) likely to receive. In particular, evidence suggests that citations follow a *cumulative advantage* or *preferential attachment* process, where the more citations a scientist has, the more likely they are to get new citations (Price, 1976). This process leads to a power law (PL) distribution of citations (Eom & Fortunato, 2011; Waltman, van Eck, & van Raan, 2012) or other heavy-tailed distributions (Thelwall, 2016). Indeed, preferential attachment has been proven to lead to heavy-tailed distributions (Krapivsky, Redner, & Leyvraz, 2000), with some refinements to account for the lifetime of a paper (Parolo, Pan et al., 2015).

As early as 1926, Lotka showed that, in the field of chemistry, the number of scientists having published N papers is proportional to N^{-2} (Lotka, 1926). In other words, he showed that the distribution of the number of papers published by scientists follows a PL. Later on, the same analysis was extended to other fields of science (e.g., Barrios, Borrego et al., 2008; Gupta & Karisiddappa, 1996; Huber & Wagner-Döbler, 2001a, 2001b; Newby, Greenberg, & Jones, 2003; Pal, 2015; Sutter & Kocher, 2001; Wagner-Döbler & Berg, 1999) and refined to more elaborate distributions, such as the *power law with cutoff* (*PLwC*) (Kretschmer & Rousseau, 2001; Saam & Reiter, 1999; Smolinsky, 2017) or the *stretched exponential* distribution (Laherrère & Sornette, 1998). Despite this early start, the number of papers published by a scientist has been less investigated than the number of citations that a paper or a scientist gets.

With the objective of refining these past analyses, in this article we focus on the distribution of the number of papers published by scientists within a given peer-reviewed journal. The distribution of the number of papers is both easily accessible (through any scientific publication

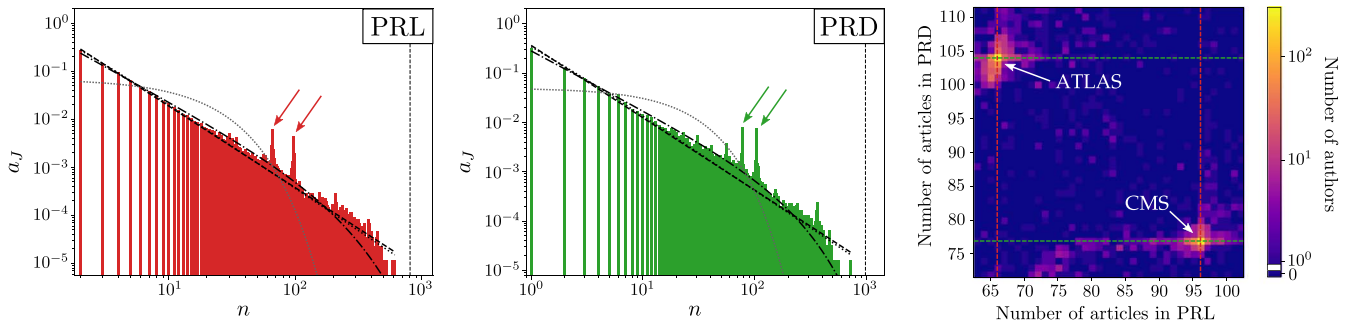

Figure 1. Left and center: Histograms of the number of papers n published in *Physical Review Letters* (PRL) and *Physical Review D* (PRD) among the authors who published in these journals. For each value of n, the height of the bar gives the proportion of authors who published n articles in the corresponding journal. Best distribution fits (see Section 2.1) are displayed for an exponential distribution (gray dotted), a power law (dashed black), a power law with cutoff (dash-dotted black), and a Yule-Simon distribution (dotted black). The arrows indicate significant peaks in the number of authors, corresponding to the ATLAS and CMS experiments at the CERN. Right: Two-dimensional, color-coded histogram of the number of authors with respect to the number of papers published in PRL (horizontal axis) and PRD (vertical axis).

database) and informative. Indeed, various characteristics of the publication dynamics within a journal can be extracted from the aforementioned distribution. We illustrate this claim in the striking examples of *Physical Review Letters* and *Physical Review D*, shown in Figure 1, where the analysis of the distribution emphasizes an underlying *preferential attachment* dynamics; the finiteness of scientific careers; and the presence of (very) large groups of scientists in the related fields of physics (see the caption of Figure 1 for a detailed discussion).

As interestingly pointed out by Sekara, Deville et al. (2018), publishing in a peer-reviewed journal (especially in high-impact ones) is more likely if one author of the manuscript has already published in the same journal. Such a process can be interpreted as *preferential attachment*, and an expected outcome of such an observation is a high representation of a few authors in a given journal (Krapivsky et al., 2000). Furthermore, a scientist whose field of research is well aligned with a journal topic is likely to publish a large proportion of their work in this journal, leading again to high representation of a few specialized authors in a given journal.

The heavy-tailedness of the distribution of the number of papers is striking in the histograms (see Figures 1 and 2). Indeed, the tail of the histogram is stronger than the best exponential fit to the data (gray dotted line). However, as we show below, the famous *PL* is not a good fit to the data either, and the actual distribution lies somewhere between an exponential and a PL. In addition to our analysis of the distribution, we propose an adaptation of the preferential attachment law that models the evolution of the number of papers of a set of authors within a journal.

2. EMPIRICAL AND FITTED DISTRIBUTIONS

We consider an arbitrary selection of 14 peer-reviewed journals (Table 1), whose data are available on the Web of Science data base (WoS, www.webofscience.com). The selected journals vary in age (from a few decades to more than a century) but are not too young, in order to have sufficiently many papers available, and all of them are still publishing nowadays. Whereas the choice of journals is arbitrary and limited, we tried to cover a diversity of disciplines of the natural sciences and various time spans. The limited sample of journals does not allow us to claim any universality in our results, but we argue that it demonstrates the pertinence of our approach in the quantitative analysis of the scientific publication process.

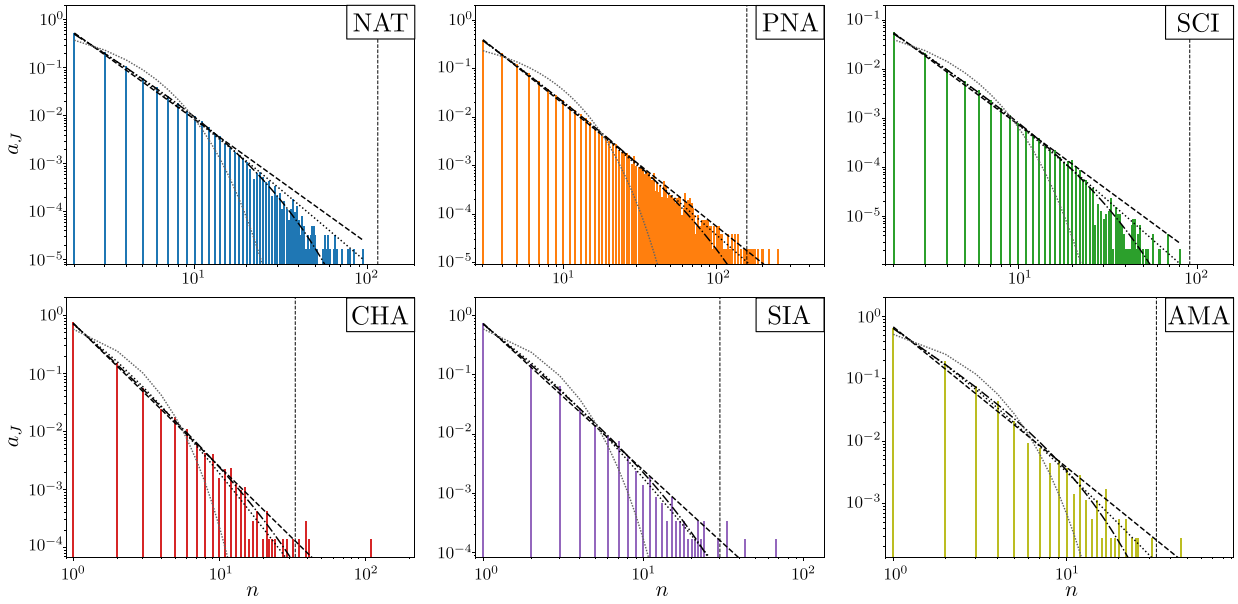

Figure 2. Histograms of the number of papers n published in the six journals indicated in the insets, among the authors who published in these journals (see Table 1 for legends). As in Figure 1, for each value of n, the height of the bar gives the proportion of authors who published n articles in the corresponding journal. The gray dotted line is an exponential fit of the data, emphasizing that the distribution is heavy-tailed. We also show the best fit (MLE), discussed in Section 2.1, for a power law distribution (dashed black), power law with cutoff (dash-dotted black), and Yule-Simon distribution (dotted black). The vertical dashed line indicates the theoretical maximum number of papers if the distribution was the fitted power law (see Section 4). The same plots for the other journals are available in Figure 1 and in Figure A.1.

Table 1. Labels, names, and number of authors in the journals considered. In parentheses is given the reduction year (discussed in Section 4) and the number of authors up to this year. One (resp. two) asterisk(s) indicate the journals where authors with one (resp. two) paper(s) are discarded.

Label	Journal name (reduction year)	# authors (reduction)
NAT	*Nature** (1950)	63,791 (3,374)
PNA	*Proceedings of the National Academy of Sciences of the USA*** (1950)	55,849 (2,495)
SCI	*Science** (1940)	48,928 (4,788)
LAN	*The Lancet** (1910)	33,416 (3,015)
NEM	*New England Journal of Medicine** (1950)	27,078 (3,842)
PLC	*Plant Cell* (2000)	20,649 (4,712)
ACS	*Journal of the American Chemical Society** (1930)	82,223 (5,301)
TAC	*IEEE Transactions on Automatic Control* (2000)	8,911 (3,603)
ENE	*Energy* (2005)	28,920 (4,491)
CHA	*Chaos*	7,409
SIA	*SIAM Journal on Applied Mathematics*	6,106
AMA	*Annals of Mathematics*	3,679
PRD	*Physical Review D*	64,922
PRL	*Physical Review Letters**	90,993

We denote by $\mathcal{J}$ = {NAT, PNA, ..., PRL} the set of journals considered (see Table 1 for the list of labels). Within each journal $J \in \mathcal{J}$, we index authors by an integer $i = 1, ..., A_J^{\text{tot}}$, A_J^{tot} being the number of authors who published in journal J. Then for each author $i = 1, ..., A_J^{\text{tot}}$, we count the number n_i^J of papers published by author i in journal J up to year 2017 in the whole WoS database (meaning from year 1900 or the year of the journal's creation, whichever is the later). This process yields the set of data $\mathcal{D}_J = \{n_i^J : i = 1, ..., A_J^{\text{tot}}\}$, which is a set of A_J^{tot} integer numbers. We restrict our investigation to papers labeled as "*Article*" in the WoS data base, to focus on peer-reviewed papers.

From the data set $\mathcal{D}_J$ we can compute the number and proportion of authors who published n papers

$$A_J(n) = \#\{i : n_i^J = n\}, \qquad a_J(n) = A_J(n)/A_J^{\text{tot}}, \tag{1}$$

and by definition, $\sum_n a_J(n) = 1$. The proportion a_J is represented on logarithmic scales in Figures 1, 2, and A.1, each panel corresponding to a different journal.

Remark. Note that we did not take into account the fact that the different papers are co-signed by multiple authors. Consequently, different papers have different "weights" in the data set. We are mostly interested in the number of papers from the point of view of the authors; it is then adequate to count, for each author, the number of papers they signed, independently of the number of coauthors. Refining the analysis and taking into account the number of coauthors on each paper would be the purpose of future work.

Note also that we do not take into account papers published anonymously, which represent a large number of papers in medicine journals in particular.

Finally, for some journals, the number of authors is too large to be downloaded from the WoS database. As a consequence, authors who have published only one or two papers in these journals have to be removed from the data (e.g., NAT, PNA, or SCI, indicated by asterisks in Table 1).

2.1. Distribution Fitting

Because of the apparent heavy-tailedness of the distribution, it is tempting to fit a PL. However, as pointed out by Clauset, Shalizi, and Newman (2009), such fitting should be done with care in order to avoid spurious conclusions (Broido & Clauset, 2019). We therefore fit three heavy-tailed distributions and assess the goodness-of-fit of our fitting following Clauset et al. (2009), which is encoded in a p-value. Numerical results are summarized in Table 2.

For each empirical distribution of the number of papers published by an author i in journal J, we fit an exponential distribution (gray dotted lines in Figures 1 and 2) to emphasize their heavy-tailed behavior. The three heavy-tailed distribution that we fit are

- A *PL distribution* (black dashed lines in the figures),

$$P_{\text{pl}}(n_i^J = n; \alpha) = C_\alpha n^{-\alpha}, \tag{2}$$

  with $\alpha > 1$ and $C_\alpha \in \mathbb{R}$ normalizing the distribution;
- A *PLwC* (black dash-dotted lines in the figures),

$$P_{\text{plc}}(n_i^J = n; \beta, \gamma) = C_{\beta,\gamma} n^{-\beta} e^{-\gamma n}, \tag{3}$$

  with $\beta > 1$, $\gamma > 0$, and normalizing constant $C_{\beta,\gamma} \in \mathbb{R}$; and

Table 2. Fitted parameters and *p*-value of the goodness-of-fit for power law (PL), power law with cutoff (PLwC), and Yule-Simon (Y-S) distributions. No set of data is well fitted by a PL distribution. However, the PLwC seems to be a good fit for three journals (SCI, PLC, CHA), and the Yule-Simon distribution seems to correctly fit the distribution of NEM and SIA. For the other journals, none of the distributions seem to fit the data appropriately.

	PL		PLwC			Y-S	
	α	p (%)	β	γ	p (%)	ρ	p (%)
NAT	2.58	0.0	2.11	0.07	0.0	3.10	0.0
PNA	2.53	0.0	2.30	0.02	0.0	2.83	0.0
SCI	2.68	0.0	**2.30**	**0.06**	**16.64**	3.28	0.02
LAN	2.47	0.0	2.09	0.05	0.18	2.90	0.0
NEM	2.76	0.0	2.36	0.07	0.2	**3.43**	**8.82**
PLC	2.30	0.0	**1.92**	**0.10**	**13.42**	3.01	0.92
ACS	2.11	0.0	1.95	0.01	0.0	2.32	0.0
TAC	2.08	0.0	1.84	0.04	0.0	2.51	0.02
ENE	2.36	0.0	2.12	0.06	0.12	3.15	0.0
CHA	2.47	0.0	**2.28**	**0.05**	**80.84**	3.43	0.0
SIA	2.49	0.0	2.20	0.08	2.24	**3.49**	**9.06**
AMA	2.26	0.0	1.72	0.14	0.18	2.95	0.0
PRD	1.49	0.0	1.24	0.005	0.02	1.55	0.0
PRL	1.73	0.0	1.52	0.005	0.12	1.80	0.0

- A *Yule-Simon distribution* (black dotted lines in the figures),

$$P_{\mathrm{ys}}\left(n_i^J = n; \rho\right) = C_\rho(\rho - 1)\mathrm{B}(n, \rho), \tag{4}$$

with $\rho > 0$, $C_\rho \in \mathbb{R}$ is the normalizing constant, and where $\mathrm{B}(x, y)$ is the *Euler beta function*.

We perform the distribution fitting by optimizing the parameters α, β, γ, and ρ with a Maximum Likelihood Estimator (Clauset et al., 2009). The curves of the fitted distributions are plotted in Figures 1, 2, and A.1, and the fitted parameters are given in Table 2. Other distributions (such as log-normal, Lévy, Weibull) were tested and discarded because they were far from matching the data.

2.2. Goodness of Fit

To evaluate the goodness of our fits, we again follow Clauset et al. (2009), to which we refer for an in-depth discussion of heavy-tailed distribution fitting. The whole goodness-of-fit estimation is summarized in Figure 3.

Let us denote by θ_J the parameters of the distribution $P(X; \theta)$ (e.g., $\theta_J = \alpha$ for the PL distribution), fitted to the data set $\mathcal{D}_J$. We generate 5,000 sets of synthetic data $\widetilde{\mathcal{D}}_i$, $i = 1, \ldots, 5{,}000$, each of them composed of $A_J^{\mathrm{tot}} = |\mathcal{D}_J|$ integer numbers, drawn randomly from the probability

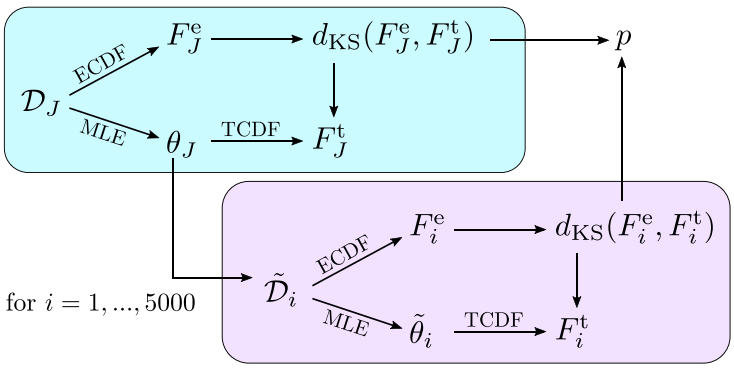

Figure 3. Scheme of the goodness-of-fit computation. For a given journal J, the data set $\mathcal{D}_J$ is fitted with a distribution whose parameters are θ_J, and we compute the Kolmogorov-Smirnov (KS) distance between its empirical and theoretical cumulative distribution functions (TCDFs). Then, based on the parameters θ_J, we generate 5,000 synthetic data sets $\widetilde{\mathcal{D}}_i$ for $i = 1, ..., 5{,}000$, on which we repeat the same process. Finally, the p-value is the proportion of synthetic data sets whose empirical and TCDF are closer to each other (in the KS sense) than for the original data set $\mathcal{D}_J$.

distribution $P_J = P(X; \theta_J)$. For each of these synthetic data sets $\widetilde{\mathcal{D}}_i$, we perform again an MLE to fit the same distribution $P(X; \theta)$, yielding parameters $\tilde{\theta}_i$ and the distribution $P_i = P(X; \tilde{\theta}_i)$.

The goodness-of-fit then relies on how well F^{e}, the empirical cumulative distribution function (ECDF) for a given set of data, matches F^{t}, the theoretical cumulative distribution function (TCDF) of its fitted distribution. We define

$$F_i^{\mathrm{e}}(k) = \frac{\#\left\{n \in \widetilde{\mathcal{D}}_i : n \leq k\right\}}{\#\widetilde{\mathcal{D}}_i}, \qquad F_i^{\mathrm{t}}(k) = P(n \leq k; \theta_i), \tag{5}$$

and F_J^{e} and F_J^{t} are defined similarly with the data set $\mathcal{D}_J$.

The p-value of the goodness-of-fit is then given by

$$p = \frac{\#\left\{i : d_{\mathrm{KS}}\left(F_i^{\mathrm{e}}, F_i^{\mathrm{t}}\right) > d_{\mathrm{KS}}\left(F_J^{\mathrm{e}}, F_J^{\mathrm{t}}\right)\right\}}{5000}, \tag{6}$$

where the *Kolmogorov-Smirnov distance* between two cumulative distribution functions F_1 and F_2 is defined as the maximum difference between them:

$$d_{\mathrm{KS}}(F_1, F_2) = \max_k |F_1(k) - F_2(k)|. \tag{7}$$

Namely, p is the proportion of synthetic data sets that are further from the theoretical distribution (in the Kolmogorov-Smirnov sense) than the analyzed data set. The fit is rejected if $p < 5\%$, and considered as *good* otherwise (see Clauset et al. (2009) for more details).

This goodness-of-fit estimation is performed for each journal $J \in \mathcal{J}$ and each distribution listed above (PL, PLwC, and Yule-Simon). The results are presented in Table 2 and the resulting distributions together with the data are shown in Figures 1, 2, and A.1.

As can be seen in Figures 1, 2, and A.1, the PL distribution is a poor fit for all data, its p-value being zero for all journals. Indeed, for most of the journals, the tail of the data set is lighter than the tail of its PL fit (black dashed lines). For three journals (SCI, PLC, CHA), the p-value of the PLwC is larger than 5% and it seems to be a rather good fit, and for two others (NEM and SIA), the Yule-Simon distribution cannot be excluded.

3. GENERAL DYNAMICS

We argue that the heavy-tailedness observed in the previous section is likely to be a consequence of a *preferential attachment* or *cumulative advantage* process. Many social processes are ruled by so-called preferential attachment (Jeong, Néda, & Barabási, 2003), also called *cumulative advantage*. Scientific coauthorship (Barabási, Jeong et al., 2002), citations (Eom & Fortunato, 2011; Price, 1976), and performance of scientific institutions (van Raan, 2007) are apparently no exception to the rule. For instance, according to Eom and Fortunato (2011), the probability that a paper will get a new citation at time t is proportional to the number of citations this paper already has at time t.

Such processes naturally lead to PLs in the relations between characteristics of the systems of interest. For instance, Katz (1999) showed that the number of citations a scientific community gets is a PL of the number of publications in this community, with positive exponent (≈ 1.27). More recently, Bettencourt, Lobo et al. (2010) illustrate that the *Gross Metropolitan Product* of a city is a PL of its population, with positive exponent (≈ 1.126). In a similar spirit, Barabási and Albert (1999) showed that the empirical probability that a web page is targeted by k other pages follows a PL with negative exponent (≈ -2.1).

It is reasonable to expect that the evolution of the number of papers published by an author in a given journal is described by a similar preferential attachment process. We support the hypothesis of a preferential attachment or cumulative advantage process by two distinct but similar analysis of publication data.

Remark. Notice that even though we refer to the two analyses below as *preferential attachment* and *cumulative advantage*, respectively, these two denominations fundamentally refer to the same general process (Perc, 2014). The main reason for us to use these two denominations is to distinguish the two analyses. Furthermore, the line of reasoning underlying each of our analysis is inspired by the definition of the corresponding notion ("preferential attachment" or "cumulative advantage").

3.1. Preferential Attachment

Heuristically, our first argument is that if an author published a lot of papers in a journal, it means (a) that they write a lot of papers and (b) that their research topic is well aligned with the scope of the journal (for specialized journals), or that the scientific impact of this author's research matches the standards of the journal (for interdisciplinary journals). Assumptions (a) and (b) together imply that this author is likely to publish again in this journal. We refer to this process as *preferential attachment*.

The above heuristic can be made more rigorous. For a given journal and for k, $t \in \mathbb{Z}_{\geq 0}$, we define

- $\mathcal{S}(k, t)$: the set of all authors who have published k papers on December 31 of year $t-1$;
- $A_k(t) = \#\mathcal{S}(k, t)$: the number of authors in the set $\mathcal{S}(k, t)$;
- $N_k(t)$: the number of papers published during year t by all the authors in the set $\mathcal{S}(k, t)$; and
- $\rho_k(t) = N_k(t)/A_k(t) \in \mathbb{R}$: the average number of papers published during year t, by the authors in the set $\mathcal{S}(k, t)$.

In Figure 4, we plot the values of $\rho_k(t)$ with respect to the number of papers k for years $t \in \{1999, \ldots, 2008\}$ for SCI, LAN, and PRL (each point corresponds to one year t and one number of papers k). For each of the three journals, these values have a linear correlation coefficient

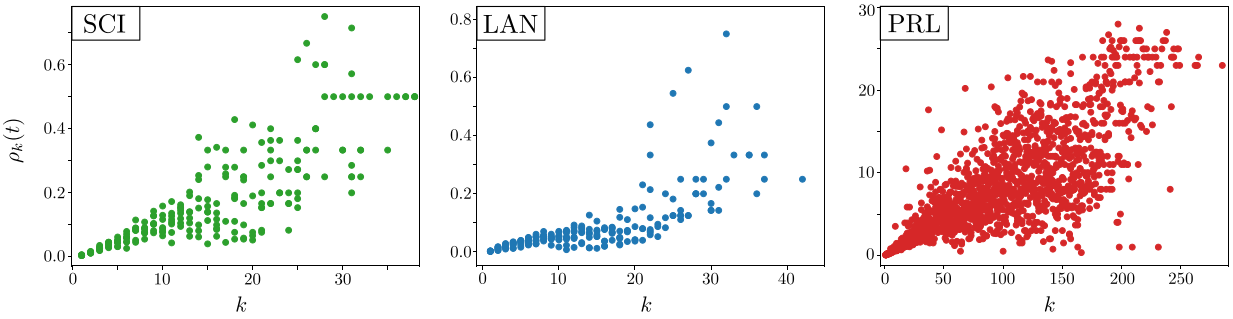

Figure 4. Average number of papers published within year $t \in \{1999, ..., 2008\}$, for authors in the set $\mathcal{S}(k, t)$, as a function of k, for SCI, LAN, and PRL. Each point corresponds to one of the years in {1999, ..., 2008} (hence multiple points for the same value of k). The Pearson correlation coefficients of the point clouds are respectively $r_{SCI} \approx 0.714$, $r_{LAN} \approx 0.707$, and $r_{PRL} \approx 0.763$, all larger than 0.7, suggesting a relation close to linear. For SCI (resp. LAN and PRL), 14 points (resp. 12 points and two points) are left out of the frame, for sake of readability.

larger than 0.7, supporting a fairly good linear dependence,

$$\rho_k(t) \sim k. \tag{8}$$

Note that, for each year considered, we do not take into account authors who did not publish, because the majority of those are not active anymore.

The empirical probability that a new paper is signed by an author with k papers is then close to being proportional to k. Krapivsky et al. (2000) rigorously proved that, if the relation in Eq. 8 was exactly proportional, then after a long enough time, the distribution of the number of papers over the set of authors would be a PL with exponent $\alpha \leq -2$. The fact that the relation 8 is not exactly proportional but close to it probably explains that the observed distributions have tails that are heavy, but lighter than the PL, as suggested in Figures 1 and 2.

3.2. Cumulative Advantage

The concept of *cumulative advantage*, which is directly related to preferential attachment, has been derived from the seminal work of Merton (1968, 1988) and Price (1976), and the follow-up by Katz (1999). Cumulative advantage emphasizes that an initial advantage leads to a disproportionate advantage in the future. For instance, it has been shown that, if author i has twice as many publications as author j, then they are likely to get more than twice as many citations (Katz, 1999).

In the context of interest for this article, cumulative advantage translates as follows. Assume that author i and author j have respectively $n_i(t_0)$ and $n_j(t_0)$ papers in a journal at time t_0, with a ratio $\eta_{ij}(t_0) = n_i(t_0)/n_j(t_0) > 1$. Then cumulative advantage means that, at a later time $t_1 > t_0$, the ratio $\eta_{ij}(t_1) \geq \eta_{ij}(t_0)$, implying that author i gains a disproportional advantage over time. Mathematically speaking, cumulative advantage implies the following equivalences:

$$n_i(t_0) \geq n_j(t_0) \Leftrightarrow \frac{n_i(t_0)}{n_j(t_0)} \leq \frac{n_i(t_1)}{n_j(t_1)} \Leftrightarrow \frac{n_i(t_1)}{n_i(t_0)} \geq \frac{n_j(t_1)}{n_j(t_0)} \Leftrightarrow \xi_i(t_0, t_1) \geq \xi_j(t_0, t_1), \tag{9}$$

where we defined $\xi_i(t, s) = n_i(s)/n_i(t)$, and where equalities hold if the relation in Eq. 8 is exact.

To support the presence of a cumulative advantage in the publication within the journals SCI, LAN, and PRL, we computed $\xi_i(1999, 2008)$ for each author who published between 1999 and 2008. The statistics of ξ_i are shown in Figure 5 as a function of the initial number of papers $n_i(1999)$. Even though the data are not perfectly conclusive, we clearly observe an increasing trend of ξ_i as a function of n_i, suggesting that the relation of Eq. 9 may be satisfied.

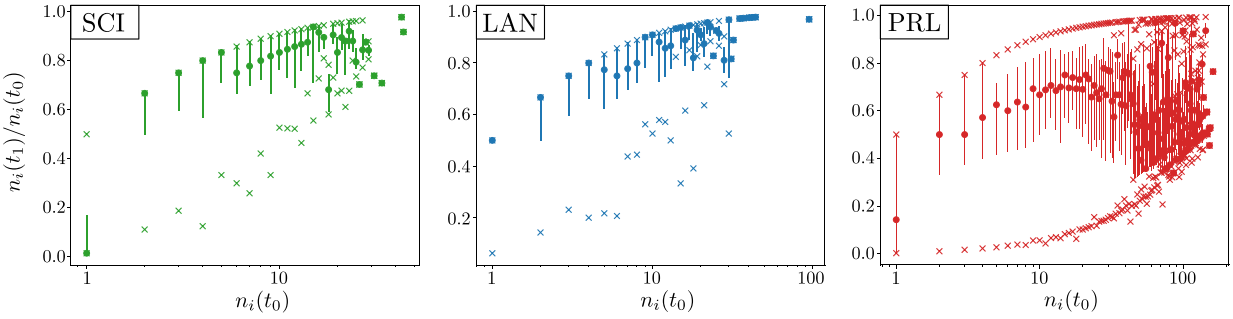

Figure 5. Statistics of the ratio ξ_i between the number of papers in 1999 and in 2008 as a function of the number n_i of papers in 1999, in the three journals SCI (left), LAN (center), and PRL (right). For each value of n_i(1999), there are multiple authors with this number of papers in 1999. Among these authors, the dots show the median value of ξ_i, the bar covers the second and third quartiles, and the crosses are the maximal and minimal values. Despite no exact increase of the values, there is an increasing trend of ξ_i with respect to n_i, supporting the presence of a cumulative advantage process.

This observation supports (at least partly) a cumulative advantage process, and henceforth the presence of a PL.

The increasing trends in Figure 5 even suggest a superlinear cumulative advantage (Krapivsky & Krioukov, 2008; Zhou, Wang et al., 2007). Indeed, as mentioned above, if the relation Eq. 8 was exact, $\xi_i(t_0, t_1)$ would be constant with respect to $n_i(t_0)$. In such a case, the heavy-tailed distribution observed in Figures 1, 2, and A.1 would be the transient state of the distribution discussed by Krapivsky and Krioukov (2008). A more in-depth analysis of the possibility of a superlinear cumulative advantage could be done, following the calibration approach proposed by Zadorozhnyi and Yudin (2015), but goes beyond the purpose of this article and will be treated in future work.

4. KEY PLAYERS

The general distribution of the number of papers per author is quite clear in our analysis: It seems to be somewhere between an exponential distribution and a PL. The PL having the heaviest tail of the three distributions considered (PL, PLwC, and Yule-Simon), we use it to estimate an upper bound on the number of papers published by an author for each journal. Assuming that the data are well described by the PL distribution in Eq. 2, one can compute the number of authors with n papers in journal J, $A_n \approx A_J^{\text{tot}} C_\alpha n^{-\alpha}$. Setting this number to $A_n = 1$, the maximum number of papers is given by $n_{\max} \approx (A_J^{\text{tot}} C_\alpha)^{\frac{1}{\alpha}}$, determining a theoretical upper bound on the number of papers published by an author for each journal, shown as the vertical dashed lines in Figures 1, 2, and A.1.

In some journals (see e.g., PNA, CHA, SIA, and AMA in Figure 2, and NEM and ACS in Figure A.1), it appears that, some authors, which we refer to as *key players*, publish significantly more papers in a journal than the PL would predict. Note that we checked that these key players are not artifacts due to multiple authors having the same name, which would count as the same person.

To make the data of different journals more comparable, we restricted our investigation to the early years between 1900 (earliest possible in WoS) and the year in parentheses in the second column of Table 1 for our first nine journals in the table. This yields a number of authors comparable to the three following journals in Table 1 (CHA, SIA, and AMA). The reduced number of authors is given in parentheses in the third column of Table 1. The resulting

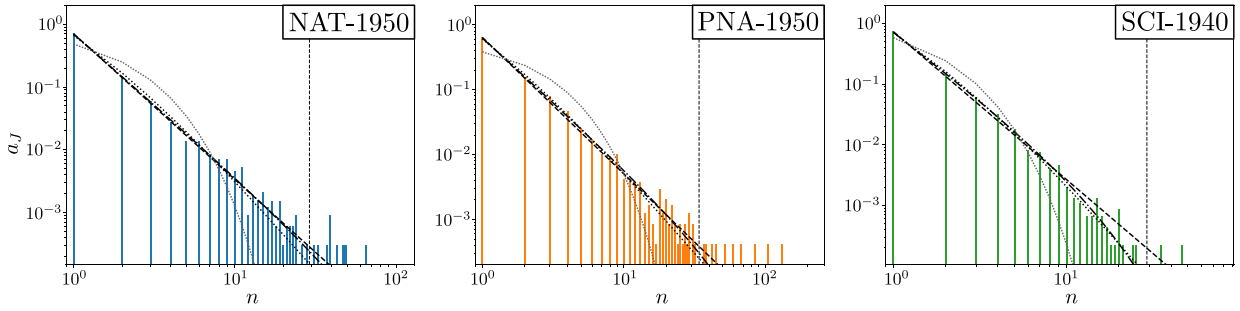

Figure 6. Histograms of the number of papers n published in the six journals indicated in the insets, among the authors who published in these journals (see Table 1 for legends). Data are restricted to the years between 1900 (earliest possible in WoS) and the years indicated in the insets. The number of authors covered is given in parentheses in the third column of Table 1. As in Figures 1 and 2, for each value of n, the height of the bar gives the proportion of authors who published n articles in the corresponding journal. We show the best fit for a power law distribution (dashed black), power law with cutoff (dash-dotted black), and Yule-Simon distribution (dotted black). The vertical dashed line indicates the theoretical maximal number of published papers if the distribution was the fitted power law (see Section 4). We observe an almost systematic exceeding of the number of papers published by some authors. The same plot for other journals is available in Figure A.2.

distributions are depicted in Figure 6 and in Figure A.2, and the fitted parameters are detailed in Table 3. It appears from Figures 6 and A.2 that for such reduced number of authors, the overshoot of some authors is more systematic, suggesting that in the early years of scientific journals, there are usually a few very prolific authors publishing in it at a rather high rate.

Considering the results of the fitting, in Table 3, we observe better agreements than for the full data sets. This probably indicates that the sample size is not large enough to accurately fit heavy-tailed distributions, which obviously need large samples. The fact that NAT and PNA are well fitted by two distributions also indicates that the reduced data sets are not large enough to be conclusive.

Table 3. Fitted parameters and p-value of the goodness-of-fit for power law (PL), power law with cutoff (PLwC), and Yule-Simon (Y-S) distributions, for the nine journals with reduced time span. We see that the only data that are well-approximated by the PL are for NAT when reduced to the first 3,374 entries of WoS. The PLwC, however, seems to be a good fit for the reduced data of six journals (NAT, PNA, SCI, LAN, TAC, and ENE). ENE is particularly well-fitted by the PLwC. Finally, the Yule-Simon distribution seems to correctly fit the distribution of PAN, PLC, and ACS. For the other journals, none of the distributions seem to fit the data appropriately. Remark that the reduced data of NAT and PNA are correctly fitted for two distributions, indicating that the amount of data is probably not sufficient for a good fit.

	PL		PLwC			Y-S	
	α	p (%)	β	γ	p (%)	ρ	p (%)
NAT	**2.32**	**29.4**	**2.23**	**0.016**	**6.0**	2.98	0.0
PNA	2.10	0.1	**1.96**	**0.02**	**15.0**	**2.55**	**6.3**
SCI	2.44	0.0	**2.13**	**0.09**	**72.0**	3.37	4.7
LAN	2.25	0.0	**1.81**	**0.11**	**30.2**	2.91	2.5
NEM	2.27	0.9	2.06	0.04	4.4	2.91	0.0
PLC	2.59	0.0	2.12	0.16	0.3	**3.82**	**54.7**
ACS	2.06	0.0	1.89	0.02	0.1	**2.46**	**64.0**
TAC	2.32	0.0	**2.06**	**0.06**	**23.7**	3.04	0.1
ENE	2.69	0.8	**2.50**	**0.06**	**94.5**	4.06	0.0

5. MODELING

We observe in Figures 1, 2, and A.1 that for old journals where a lot of papers are published, the tail of the histogram has a rather fast decay after a heavy-tailed regime (this is particularly striking in PRL and PRD, Figure 1). We explain this observation by the fact that the number of publications of a given author depends on two parameters: their publication rate and the length of their career. Both these quantities are bounded in practice, and even if it is possible to publish a very large number of papers in a given journal, there is a practical limit to this number. We hypothesize that the decay in the histograms of long-living journals comes from the finiteness of publication rates and career lengths.

To support our hypothesis, we propose a model to generate data sets that mimic the distributions observed above. As discussed, this model is built on two main dynamics. Fundamentally, it is a *preferential attachment* process, where the likelihood that a researcher is in the author's list of a new paper is proportional to the number of papers this researcher already has in this journal. But in addition, it is refined with a *limited career span*, requiring that after some time, the likelihood that a researcher publishes a new paper decreases to reach zero after they retire.

The model is based on five parameters:

- $N_y \in \mathbb{Z}_{\geq 0}$: The number of years (i.e., number of iterations) over which the model is run;
- $N_p \in \mathbb{Z}_{\geq 0}$: The number of papers that are published every year in the synthetic journal;
- $\rho_0 \in [0, 1]$: The proportion of papers that are authored by new researchers who have not yet published in the synthetic journal; and
- $T_{\min}, T_{\max} \in \mathbb{Z}_{\geq 0}$: The likelihood that an author publishes a new paper decreases linearly after their $T_{\min}$th year of activity, until reaching zero at their $T_{\max}$th year of activity. We illustrate this likelihood in Figure 7.

The model is arbitrarily initialized with some number of authors each with a few papers in the synthetic journal, gathered in the data set $\mathcal{D}(0) = \{n_1(0), n_2(0), \ldots, n_{A(0)}(0)\}$. Then for each year $t \in \{1, \ldots, N_y\}$ where the model is run, N_p papers are attributed randomly either to new authors (i.e., who have not yet published) with probability ρ_0, or to an existing author with probability $1 - \rho_0$. If it is attributed to an existing author, the probability that it is attributed to author i is:

- proportional to $n_i(t)$, the number of papers published by i at year t; and
- linearly decreasing for $T_i(t) \in [T_{\min}, T_{\max}]$, where $T_i(t)$ is the "*academic age*" of i, which is the number of iterations between t and the first publication year of i.

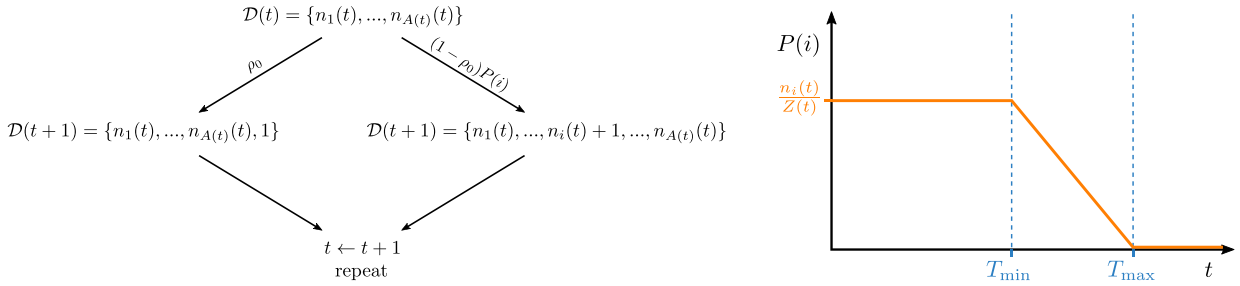

Figure 7. Left: Scheme of the iterative process generating the synthetic distribution of number of publication per author in a journal. Right: Illustration of the probability that a new paper is attributed to author i, knowing that they have already published in the past.

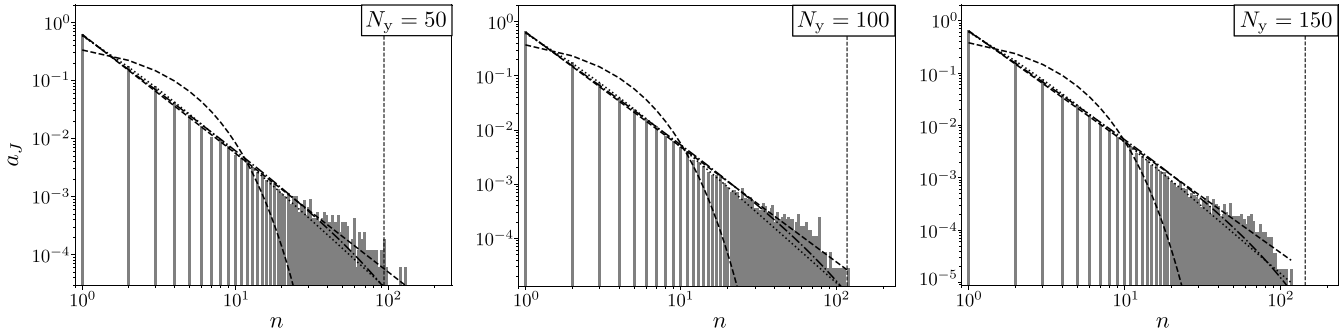

Figure 8. Histograms of the outcome of our synthetic data generator for different value of the journal life spa N_y. Fixed parameters are N_p = 1,000, $\rho_0 = 0.5$, $T_{min} = 20$, $T_{max} = 60$. There is a clear similarity between the shapes of these synthetic distributions and those of the actual data.

Table 4. Fitted parameters and p-value of the goodness-of-fit for power law (PL) and power law with cutoff (PLwC), and Yule-Simon (Y-S) distributions on the synthetic histograms of Figure 8. None of the goodness-of-fit tests are conclusive, but the values of the fitted parameters are very similar to what is observed in actual data.

	PL		PLwC			Y-S	
	α	p (%)	β	γ	p (%)	ρ	p (%)
$N_y = 50$	2.05	0.0	1.94	0.013	0.0	2.44	0.2
$N_y = 100$	2.12	0.0	2.03	0.01	0.0	2.58	0.0
$N_y = 150$	2.12	0.0	2.01	0.02	0.0	2.58	0.06

Mathematically, knowing that the new paper is attributed to an existing author, the probability that it is attributed to author i at year t is given by

$$P(i) = \frac{1}{Z(t)} n_i(t) \min\left\{1, \frac{T_{max} - T_i(t)}{T_{max} - T_{min}}\right\}, \tag{10}$$

where $Z(y)$ is the appropriate normalizing factor. The actual implementation of this model is available online (Delabays, 2022).

Histograms of the outcome of this model are illustrated in Figure 8 and the fitted parameters are in Table 4. We observe a clear similarity between the histograms for synthetic and real data. Namely, for short lifetime (N_y = 50), some authors beat the PL and exceed the number of papers that would be expected, as is observed in Figure 2 for CHA, SIA, and AMA. For longer lifetime (N_y = 150) the tail of the distribution decays and loses its heaviness, similar to PRL and PRD in Figure 1.

These observations advocate in favor of the hypothesis that the two main ingredients in the description of the evolution of the authorship within journals are both *preferential attachment* and *finiteness of careers*.

6. DISCUSSION

The main observation of our article is the heavy-tailed shape of the distribution of papers, which we explain by a preferential attachment or cumulative advantage process. Heavy-tailedness in distributions related to scientific publications, especially in citation or

collaboration networks, has widely been documented (Eom & Fortunato, 2011; Price, 1976). We showed that heavy-tailedness is preserved when restricting the analysis to a single journal.

Interestingly, our analysis suggests that the distribution does not follow a PL, but has a slightly lighter tail. Whereas we have not been able to unequivocally identify a canonical distribution, we demonstrated that a PLwC or a Yule-Simon distribution seem to be better fits to the data than the PL.

We argue that the observed heavy-tailedness of the distribution follows from a preferential attachment process through three pieces of evidence. First, we showed that the probability that an author gets a new paper in a given journal at time t is approximately proportional to the number of papers they already have in the very same journal. According to Krapivsky et al. (2000), exact proportionality would lead to a PL. Therefore, it is likely that an approximate proportionality leads to a heavy-tailed distribution.

Second, we emphasized an approximate cumulative advantage process, which also leads to PL behaviors. Whereas both what we refer to as preferential attachment and cumulative advantage are closely related, they display two underlying mechanisms explaining the heavy-tailedness of the distributions.

Finally, we provided a mathematical model for generating synthetic data of number of papers in a given journal, where preferential attachment plays a crucial role. The similarity between the obtained distribution and the observed distributions also supports the claim of the heavy tails being driven by preferential attachment.

Even though there seems to be a pattern in the data analyzed in this article, standard distributions (e.g., PLwC, Yule-Simon) do not perfectly fit the data. More advanced fitting techniques could identify a common distribution for all journals, provided that one exists. A more refined explanation of the approximate preferential attachment taking place in scientific publishing could unravel with more certainty the source of the distributions observed in this article. Even though the preferential attachment has been emphasized in the past, the underlying reasons for this bias are intricate. Disentangling the impact of scientific factors (quality and novelty of the research) and more social ones (rank and reputation of the authors) in the publication process will be a key step towards a fair and square evaluation of scientists and their work.

AUTHOR CONTRIBUTIONS

Robin Delabays: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing—original draft, Writing—review & editing. Melvyn Tyloo: Conceptualization, Methodology, Writing—review & editing.

FUNDING INFORMATION

Both authors were partly supported by the Swiss National Science Foundation under grant number 200020_182050. RD was supported by the Swiss National Science Foundation under grant number P400P2_194359.

COMPETING INTERESTS

The authors have no competing interests.

DATA AVAILABILITY

The data were extracted from www.webofscience.com and cannot be shared openly. The code for synthetic data generation is available online (Delabays, 2022).

REFERENCES

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*, 509–512. https://doi.org/10.1126/science.286.5439.509, PubMed: 10521342

Barabási, A.-L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A*, *311*, 590–614. https://doi.org/10.1016/S0378-4371(02)00736-7

Barrios, M., Borrego, A., Vilaginés, A., Ollé, C., & Somoza, M. (2008). A bibliometric study of psychological research on tourism. *Scientometrics*, *77*, 453–467. https://doi.org/10.1007/s11192-007-1952-0

Beall, J. (2017). What I learned from predatory publishers. *Biochemia Medica*, *27*, 273–278. https://doi.org/10.11613/BM.2017.029, PubMed: 28694718

Bettencourt, L. M. A., Lobo, J., Strumsky, D., & West, G. B. (2010). Urban scaling and its deviations: Revealing the structure of wealth, innovation and crime across cities. *PLOS ONE*, *5*, e13541. https://doi.org/10.1371/journal.pone.0013541, PubMed: 21085659

Bohannon, J. (2013). Who's afraid of peer review? *Science*, *342*, 60–65. https://doi.org/10.1126/science.2013.342.6154.342_60, PubMed: 24092725

Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, *66*, 2215–2222. https://doi.org/10.1002/asi.23329

Broido, A. D., & Clauset, A. (2019). Scale-free networks are rare. *Nature Communications*, *10*, 1–10. https://doi.org/10.1038/s41467-019-08746-5, PubMed: 30833554

Butler, D. (2013). Investigating journals: The dark side of publishing. *Nature*, *495*, 433–435. https://doi.org/10.1038/495433a, PubMed: 23538810

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, *51*, 661–703. https://doi.org/10.1137/070710111

Delabays, R. (2022). ADGenerator: Authors Distribution Generator (v1.0). *Zenodo*. https://zenodo.org/record/6030303

Egghe, L., & Rousseau, R. (2000). The influence of publication delays on the observed aging distribution of scientific literature. *Journal of the American Society for Information Science and Technology*, *51*, 158–165. https://doi.org/10.1002/(SICI)1097-4571(2000)51:2<158::AID-ASI7>3.0.CO;2-X

Eom, Y.-H., & Fortunato, S. (2011). Characterizing and modeling citation dynamics. *PLOS ONE*, *6*, e24926. https://doi.org/10.1371/journal.pone.0024926, PubMed: 21966387

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., ... Barabási, A.-L. (2018). Science of science. *Science*, *359*, eaao0185. https://doi.org/10.1126/science.aao0185, PubMed: 29496846

Frandsen, T. F., & Nicolaisen, J. (2017). Citation behavior: A large-scale test of the persuasion by name-dropping hypothesis. *Journal of the Association for Information Science and Technology*, *68*, 1278–1284. https://doi.org/10.1002/asi.23746

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, *122*, 108–111. https://doi.org/10.1126/science.122.3159.108, PubMed: 14385826

Grudniewicz, A., Moher, D., Cobey, K. D., Bryson, G. L., Cukier, S., ... Lalu, M. M. (2019). Predatory journals: No definition, no defence. *Nature*, *576*, 210–212. https://doi.org/10.1038/d41586-019-03759-y, PubMed: 31827288

Gupta, B. M., & Karisiddappa, C. R. (1996). Author productivity patterns in theoretical population genetics (1900–1980). *Scientometrics*, *36*, 19–41. https://doi.org/10.1007/BF02126643

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the USA*, *102*, 16569–16572. https://doi.org/10.1073/pnas.0507655102, PubMed: 16275915

Huber, J. C., & Wagner-Döbler, R. (2001a). Scientific production: A statistical analysis of authors in mathematical logic. *Scientometrics*, *50*, 323–337. https://doi.org/10.1023/A:1010581925357

Huber, J. C., & Wagner-Döbler, R. (2001b). Scientific production: A statistical analysis of authors in physics, 1800–1900. *Scientometrics*, *50*, 437–453. https://doi.org/10.1023/A:1010558714879

Jeong, H., Néda, Z., & Barabási, A.-L. (2003). Measuring preferential attachment in evolving networks. *Europhysics Letters*, *61*, 567–572. https://doi.org/10.1209/epl/i2003-00166-9

Katz, J. S. (1999). The self-similar science system. *Research Policy*, *28*, 501–517. https://doi.org/10.1016/S0048-7333(99)00010-4

Krapivsky, P., & Krioukov, D. (2008). Scale-free networks as preasymptotic regimes of superlinear preferential attachment. *Physical Review E*, *78*, 026114. https://doi.org/10.1103/PhysRevE.78.026114, PubMed: 18850904

Krapivsky, P. L., Redner, S., & Leyvraz, F. (2000). Connectivity of growing random networks. *Physical Review Letters*, *85*, 4629–4632. https://doi.org/10.1103/PhysRevLett.85.4629, PubMed: 11082613

Kretschmer, H., & Rousseau, R. (2001). Author inflation leads to a breakdown of Lotka's law. *Journal of the American Society for Information Science and Technology*, *52*, 610–614. https://doi.org/10.1002/asi.1118

Laherrère, J., & Sornette, D. (1998). Stretched exponential distributions in nature and economy: "Fat tails" with characteristic scales. *European Physical Journal B*, *2*, 525–539. https://doi.org/10.1007/s100510050276

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of Washington Academy of Sciences*, *16*, 317–323.

Merton, R. K. (1968). The Matthew effect in science: The reward and communication systems of science are considered. *Science*, *159*, 56–63. https://doi.org/10.1126/science.159.3810.56, PubMed: 5634379

Merton, R. K. (1988). The Matthew effect in science, II: Cumulative advantage and the symbolism of intellectual property. *Isis*, *79*, 606–623. https://doi.org/10.1086/354848

Narin, F. (1976). *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity*. Washington, DC: National Science Foundation.

Newby, G. B., Greenberg, J., & Jones, P. (2003). Open source software development and Lotka's law: Bibliometric patterns in programming. *Journal of the American Society for Information Science and Technology*, *54*, 169–178. https://doi.org/10.1002/asi.10177

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the USA*, *98*, 404–409. https://doi.org/10.1073/pnas.98.2.404, PubMed: 11149952

Pal, J. K. (2015). Scientometric dimensions of cryptographic research. *Scientometrics*, *105*, 179–202. https://doi.org/10.1007/s11192-015-1661-z

Parolo, P., Pan, R. K., Ghosh, R., Huberman, B. A., Kaski, K., & Fortunato, S. (2015). Attention decay in science. *Journal of Informetrics*, *9*, 734–745. https://doi.org/10.1016/j.joi.2015.07.006

Perc, M. (2014). The Matthew effect in empirical data. *Journal of the Royal Society Interface*, *11*, 20140378. https://doi.org/10.1098/rsif.2014.0378, PubMed: 24990288

Price, D. de Solla. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science and Technology*, *27*, 292–306. https://doi.org/10.1002/asi.4630270505

Price, D. J. de Solla. (1963). *Little science, big science*. Columbia University Press. https://doi.org/10.7312/pric91844

Price, D. J. de Solla. (1965). Networks of scientific papers. *Science*, *149*, 510–515. https://doi.org/10.1126/science.149.3683.510, PubMed: 14325149

Saam, N. J., & Reiter, L. (1999). Lotka's law reconsidered: The evolution of publication and citation distributions in scientific fields. *Scientometrics*, *44*, 135–155. https://doi.org/10.1007/BF02457376

Sekara, V., Deville, P., Ahnert, S. E., Barabási, A.-L., Sinatra, R., & Lehmann, S. (2018). The chaperone effect in scientific publishing. *Proceedings of the National Academy of Sciences of the USA*, *115*, 12603–12607. https://doi.org/10.1073/pnas.1800471115, PubMed: 30530676

Siudem, G., Żogała-Siudem, B., Cena, A., & Gagolewski, M. (2020). Three dimensions of scientific impact. *Proceedings of the National Academy of Sciences of the USA*, *117*, 13896–13900. https://doi.org/10.1073/pnas.2001064117, PubMed: 32513724

Smolinsky, L. (2017). Discrete power law with exponential cutoff and Lotka's law. *Journal of the Association for Information Science and Technology*, *68*, 1792–1795. https://doi.org/10.1002/asi.23763

Sorokowski, P., Kulczycki, E., Sorokowska, A., & Pisanski, K. (2017). Predatory journals recruit fake editor. *Nature*, *543*, 481–483. https://doi.org/10.1038/543481a, PubMed: 28332542

Sutter, M., & Kocher, M. G. (2001). Power laws of research output. Evidence for journals of economics. *Scientometrics*, *51*, 405–414. https://doi.org/10.1023/A:1012757802706

Thelwall, M. (2016). The discretised lognormal and hooked power law distributions for complete citation data: Best options for modelling and regression. *Journal of Informetrics*, *10*, 336–346. https://doi.org/10.1016/j.joi.2015.12.007

van Raan, A. F. J. (2007). Bibliometric statistical properties of the 100 largest European research universities: Prevalent scaling rules in the science system. *Journal of the American Society for Information Science and Technology*, *59*, 461–475. https://doi.org/10.1002/asi.20761

van Raan, A. F. J. (2019). Measuring science: Basic principles and application of advanced bibliometrics. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer handbook of science and technology indicators* (pp. 237–280). Cham: Springer. https://doi.org/10.1007/978-3-030-02511-3_10

Wagner-Döbler, R., & Berg, J. (1999). Physics 1800–1900: A quantitative outline. *Scientometrics*, *46*, 213–285. https://doi.org/10.1007/BF02464778

Waltman, L., & van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science: A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*, 2378–2392. https://doi.org/10.1002/asi.22748

Waltman, L., van Eck, N. J., & van Raan, A. F. J. (2012). Universality of citation distributions revisited. *Journal of the American Society for Information Science and Technology*, *63*, 72–77. https://doi.org/10.1002/asi.21671

Wang, Q., & Waltman, L. (2016). Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus. *Journal of Informetrics*, *10*, 347–364. https://doi.org/10.1016/j.joi.2016.02.003

Zadorozhnyi, V. N., & Yudin, E. B. (2015). Growing network: Models following nonlinear preferential attachment rule. *Physica A*, *428*, 111–132. https://doi.org/10.1016/j.physa.2015.01.052

Zhou, T., Wang, B.-H., Jin, Y.-D., He, D.-R., Zhang, P.-P., ... Liu, J.-G. (2007). Modelling collaboration networks based on nonlinear preferential attachment. *International Journal of Modern Physics C*, *18*, 297–314. https://doi.org/10.1142/S0129183107010437

APPENDIX

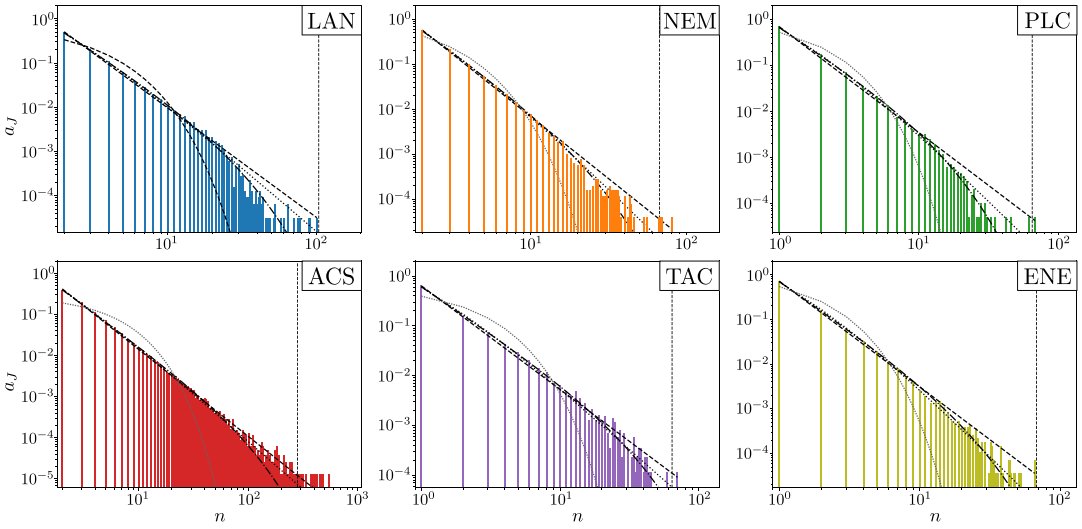

Figure A.1. Histograms of the number of papers n published in the six journals indicated in the insets, among the authors who published in these journals (see Table 1 for legends). As in Figures 1 and 2, for each value of n, the height of the bar gives the proportion of authors who published n articles in the corresponding journal. The gray dotted line is the exponential fit of the data, emphasizing that the distribution is heavy-tailed. We show the best fit for a power law distribution (dashed black), power law with cutoff (dash-dotted black), and Yule-Simon distribution (dotted black). The vertical dashed line indicates the theoretical maximal number of published papers if the distribution was the fitted power law.

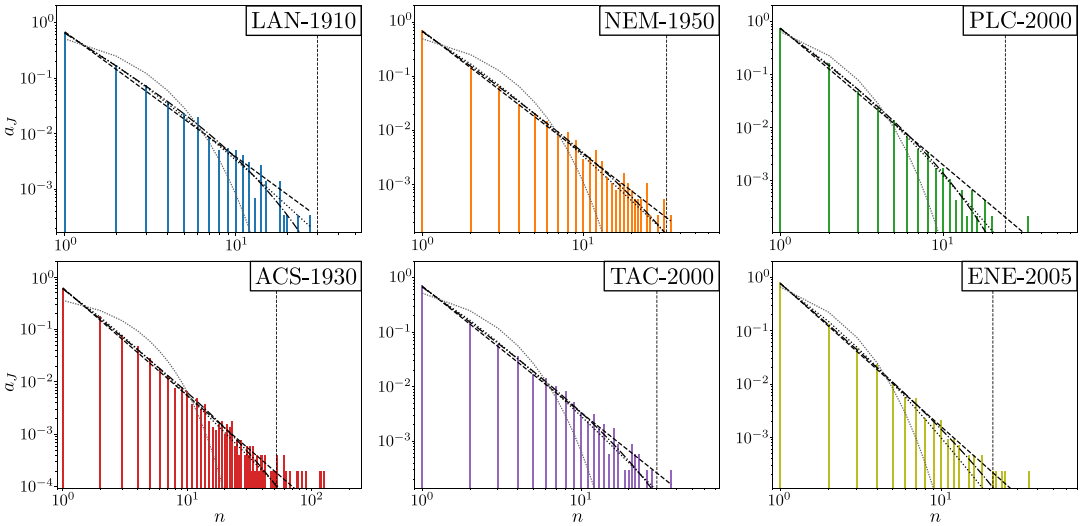

Figure A.2. Histograms of the number of papers n published in the six journals indicated in the insets, among the authors who published in these journals (see Table 1 for legends). Data are restricted to the years between 1900 (earliest possible in WoS) and the years indicated in the insets. The number of authors covered is given in parentheses in the third column of Table 1. As in Figures 1 and 2, for each value of n, the height of the bar gives the proportion of authors who published n articles in the corresponding journal. We show the best fit for a power law distribution (dashed black), power law with cutoff (dash-dotted black), and Yule-Simon distribution (dotted black). The vertical dashed line indicates the theoretical maximal number of published papers if the distribution was the fitted power law. We observe an almost systematic exceeding of the number of papers published by some authors.