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

Robin Delabays[a,b] and Melvyn Tyloo[a]

[a] *School of Engineering, University of Applied Sciences of Western Switzerland (HES-SO), CH-1951 Sion, Switzerland;*

[b] *Center for Control, Dynamical Systems, and Computation, University of California at Santa Barbara (UCSB), Santa Barbara, 93106-9560 California, USA*

(Dated: November 11, 2020)

The community of scientists is characterized by their need to publish in peer-reviewed journals, in an attempt to avoid the "perish" side of the famous maxim. Accordingly, almost all researchers authored some scientific articles. Scholarly publications represent at least two benefits for the study of the scientific community as a social group. First, they attest of some form of relation between scientists (collaborations, mentoring, heritage,...), useful to determine and analyze social subgroups. Second, most of them are recorded in large data bases, 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, in addition to its interest from the social science point of view, can contribute to maintaining a high level of research, by identifying good and bad practices in science. In this manuscript, we attempt to advance this understanding by a statistical analysis of publications within peer-reviewed journals. Namely, we show that the distribution of the number of articles published by an author in a given journal is heavy-tailed, but has lighter tail than a power law. Moreover, we observe some anomalies in the data that pinpoint underlying dynamics of the scholarly publication process.

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 always subject to new evaluation and might eventually be replaced by a 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 surprizing that some scientists started to study (and thus somehow criticize) the scientific community itself [1]. The study of the the scientific community, sometimes refered to as *Science of Science* [2, 3], is a key step to unravel the underlying behaviors of its members and draw some lessons about it. In the last decades, such an investigation has been significantly eased by the emergence of large data bases of scientific publications (Web of Sience, PubMed, arXiv,...). It allowed for instance to build the time-evolving collaboration network of scientists [4].

Such an approach and associated tools has the potential to help maintaining the quality of research, and thus a good use of public funding. Indeed, in the current context of increasing number of scientific publications [5, 6] in parallel to the ubiquitous presence of *predatory journals* [7, 8], 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 task.

A scientist's work is commonly evaluated by two different, but related, quantities, namely, their number of publications and the number of citations thereof. These quantities are summarized in the criticized, but widely spread, h-index [9, 10]. Naturally, a vast majority of investigations about the scientific publication process is focussed on the citation side. It mostly aims at describing how the citation network impacts the number of citations a given paper is (and therefore its authors are) likely to recieve. In particular, evidence suggests that citations follow a *rich-get-richer* or *preferential attachment* process, where the more citations a scientist has, the more likely they are to get new citations [11], leading to a power law distribution of citations [12, 13] or other heavy-tailed distributions [14]. Indeed, preferential attachement has been proven to lead to heavy-tailed distributions [15], with some refinements to account for the life-time of a publication [16].

Compared to the number of citations that an article or a scientist gets, the number of articles published by a scientist has been much less investigated, even though publishing papers is a *sine qua none* to get cited. In this manuscript, we focus on the distribution of articles published by a scientist within a given peer-reviewed journal. As interestingly pointed out by Sekara et al. [17], publishing in a peer-reviewed journal (especially in high-impact ones) is more likely if one author of the manuscript already published in the same journal. Such a process can be viewed as some sort of preferential attachment, and an expected outcome of such an observation is a high representation of a few authors in a given journal [15]. 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 a high representation of a few specialized authors in a given journal.

We support these expectations, showing that in a selection of fourteen journals (listed in Table I) the distribution of the number of articles published by an author

within a journal has a heavy tail. It appears however that this distribution have a tail weaker than a power law. We argue that this distribution can be explained by a preferential attachement process, which is backed up by evidence. On top of that, in some of the selected journals, we observe some interesting anomalies for which we give an explanation.

RESULTS

For each journal in Table I, we consider the list of all authors who published in it and the number of articles published by each of them up to 2017. From this, one can plot the empirical distribution of the number of articles published by an author in a given journal (Fig. 1). On these data, we fit three heavy-tailed distributions, namely a *power law* (Eq. 2), a *power law with cutoff* (Eq. 3), and a *Yule-Simon distribution* (Eq. 4), using a Maximum Likelihood Estimator (see Methods). We then assess the goodness-of-fit of our fitting following [18], which is encoded in a p-value (see Methods). The results of each fit and goodness-of-fit tests are presented in Table II, and the resulting distributions together with the data are shown in Figs. 1, 3, and 5. Clearly, the power law distribution is a poor fit for all data, its p-value being zero for all journals. This can be seen in Figs. 1, 3, and 5, where for most of the journals, the tail of the data set is lighter than the tail of its power law fit (black dashed lines). For three journals (namely SCI, PLC, CHA), the p-value of the power law with cutoff 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.

General explanation

We propose the following explanation for this heavy-tailedness. Many social processes are ruled by the so called *preferential attachment* [19]. Scientific collaborations [20] and citations [12] are apparently no exception to the rule. Namely, the probability that an author will create a new scientific collaboration at time t is proportional to the number of scientific collaboration they have. It is reasonable to assume that the evolution of the number of articles published by an author in a given journal is described by a similar preferential attachment process. In other words, it would mean that the probability that a new article published in a given journal is signed by an author is proportional to the number of articles published by this author in the very same journal.

Heuristically, our argument is that if an author published a lot of article in a journal, it means (i) that they write a lot of papers, and (ii) that their research topic is well-aligned with the topics covered by 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 (i) and (ii) together imply that this author is likely to publish again in this journal.

This intuition can be made more rigorous. For three journals (SCI, LAN, and PRL), we refined the data to account for the time evolution of the number of articles published by each author. It turns out that, on average, the number of articles published during a year, by an author having already published k articles, is close to be proportional to k (details are found in the Methods section). According to [15], if it was exactly proportional k, the final distribution would be a power law. The fact that the relation is not exactly proportional, but close to be, probably explains the lighter-than-power-law tails observed in Figs. 1, 3, and 5.

Observations

Aside of these general considerations, we note three interesting observations in the data. First, for journals with a large number of authors and published articles, the tail of the histogram drops dramatically. Second, some authors apprear to be stronger than the power law. And third, some very large groups of authors can be identified even in long term aggregated data.

Decay in long-life journals. We observe in Figs. 1, 3, and 5 that for old journals where a lot of articles 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, Fig. 3). We explain this by the fact that the number of pulications of a given author depends on two parameters, namely 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 articles 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.

Key players. The general distribution of the number papers per author is quite clear in our analysis, it seems to be somewhere between an exponential distribution and a power law. The power law having the heaviest tail of the four distributions considered (exponential, power law, power law with cutoff, and Yule-Simon), we use it to estimate an upper bound on the number of articles published by an author for each journal, shown as the vertical dashed lines in Figs. 1, 3, and 5 (more details in the Methods section). In some journals (see e.g., PNA, CHA, SIA, and AMA in Fig. 1, and NEM and ACS in the Appendix, Fig. 5), it appears that, some authors, which we refer to as *key players*, publish significantly more articles in a journal than what the power law 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. In all cases presented here, there is a unique person appearing in the

Label	Journal name (red. year)	# authors (red.)
NAT	Nature* (1950)	63'791 (3'374)
PNA	Proc. Natl. Acad. Sci. 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	J. of the American Chemical Society* (1930)	82'223 (5'301)
TAC	IEEE Trans. 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

Table I. Labels, names, and number of authors in the journals considered. In parenthesis is given the reduction year (when applicable) and the number of authors up to this year. One (resp. two) asterisc(s) indicate the journals where authors with one (resp. two) publication(s) are discarded (see the Methods section for details).

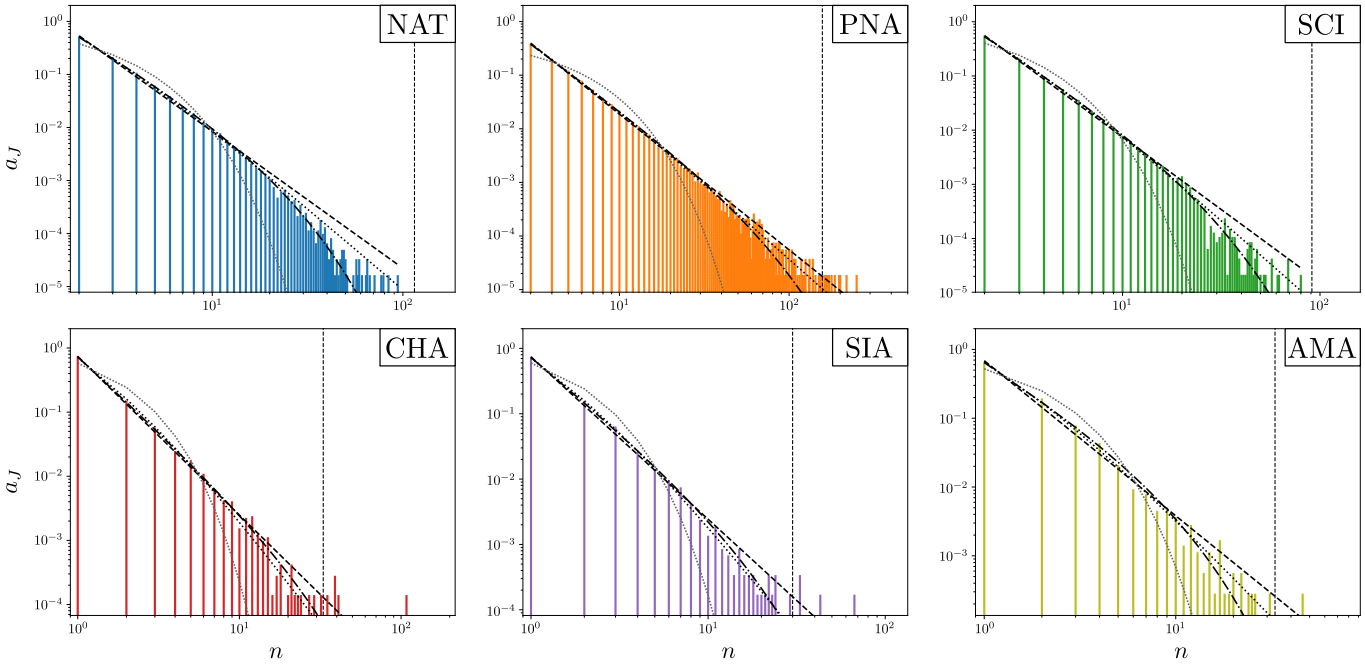

Figure 1. Histograms of the proportion of authors a_J with respect to the number of articles published, for the six journals, indicated in the insets. The grey dotted line is an exponential fit of the data, emphasizing that the distribution is heavy-tailed. We also show the best fit (MLE)f for a power 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 publications if the distribution was the fitted power law (see Eq. 8). The same plots for the other journals are available in Fig. 3 and in the Appendix, Fig. 5.

authors' list of a very large number of papers.

In order to make the data more comparable, we restrict our investigation to the early years between 1900 (earliest possible in WoS) and the year in parenthesis in the second column of Table I for our first nine journals in the table. This yields a number of authors comparable to the following three journals in the table (reduced number of authors is given in parenthesis in the third column of Table I). The resulting distributions are depicted in Fig. 2 and in the Appendix, Fig. 6, and the fitted parameters are detailed in Table III. It appears from Figs. 2 and 6 that for such reduced number of authors, the overshoot

	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

Table II. Fitted parameters and $\boldsymbol{p}$-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 power law distribution. However, the power law with cutoff 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	2.32	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

Table III. Fitted parameters and $\boldsymbol{p}$-value of the goodness-of-fit for power law (PL) and power law with cutoff (PLwC), and Yule-Simon (Y-S) distributions. We see that the only data that are well-approximated by the power law are for NAT when reduced to the first 3374 entries of WoS. The power law with cutoff, 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 power law with cutoff. 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.

of some authors is more systematic, suggesting that in the early years of scientific journals, there is usually a few very prolific authors publishing in it at a rather high rate.

Considering the resluts of the fitting, in Table III, 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.

Peaks in PRL and PRD. In Fig. 3, we observe two peaks in the empirical distributions of PRL (around 66 and 96) and PRD (around 77 and 104). Crossing the lists of authors for each number of articles between 63 and 102 for PRL (resp. 72 and 111 for PRD), we get the right panel of Fig. 3. The fact that the authors composing a peak in PRL are also the ones composing one of the peaks in PRD suggests that these authors are all part of a large group publishing together.

A quick search, indicates that the peaks correspond to the research groups of the experiments ATLAS and CMS at the CERN. These two experiments are so big and gather so many authors that they can be seen, even in the data used in our analysis, aggregated throughout the whole history of PRL (since 1958) and PRD (since 1970).

DISCUSSION

Our analysis reveals a series of interesting, even though not surprizing, dynamics ruling the process of publication within scientific journals. The main observation is the heavy-tailed shape of the distribution of publications, which we explain by a preferential attachment process. We showed that the preferential attachment dynamics is heuristically meaningful, in the sense that if an author publishes a lot of papers and if their profile aligns with the journal's profile, they are likely to publish in this journal and at the same time they are likely to have already published in the same journal. Moreover, we also backed up the preferential attachment process by data-based evidence, where we show that the proportion of articles published in a journal by the authors with already k articles (in this journal) is approximately proportional to k. An exact proportionality would lead, according to Ref. [15], to a power law distribtion. Of course, in the long run, scientists cannot published an unbounded number of articles, due to finiteness of their careers. This translates, in our analysis, as a drop in the tail of the distribution for older journals, which then do not follow a power law. Apparently, a power law with cutoff or a Yule-Simon distribution are better suited to describe the data.

On top of this general dynamics, our analysis displays some interesting anomalies that point towards specific underlying dynamics. First, the data show that in the early decades of existence of a journal, a small number

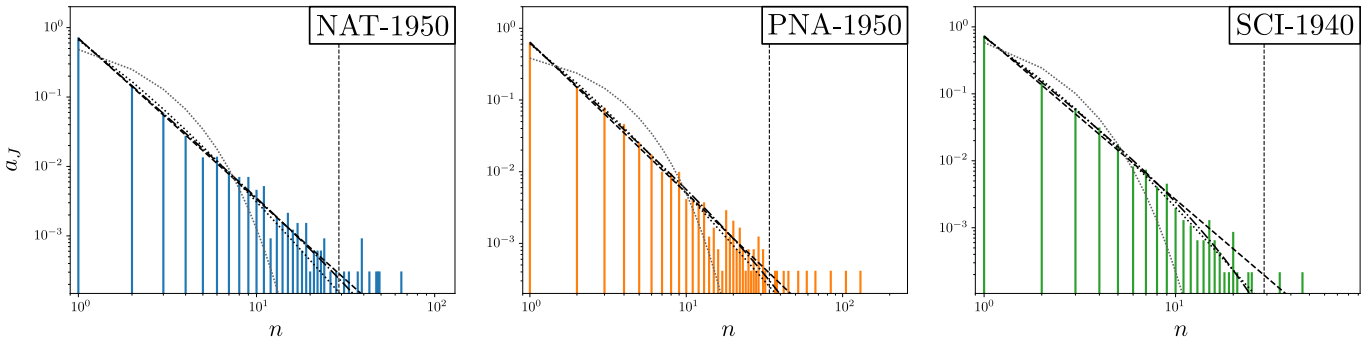

Figure 2. Histograms of the number of authors a_J with respect to the number of articles published, for the first three journals of Table I, with data restricted to the years between 1900 (earliest possible in WoS) and the years indicated in the insets. The number of authors covered is given in parenthesis in the third column of Table I. 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 Eq. 8). We observe an almost systematic exceeding of the number of articles published by some authors. The same plot for other journals is available in the Appendix, Fig. 6.

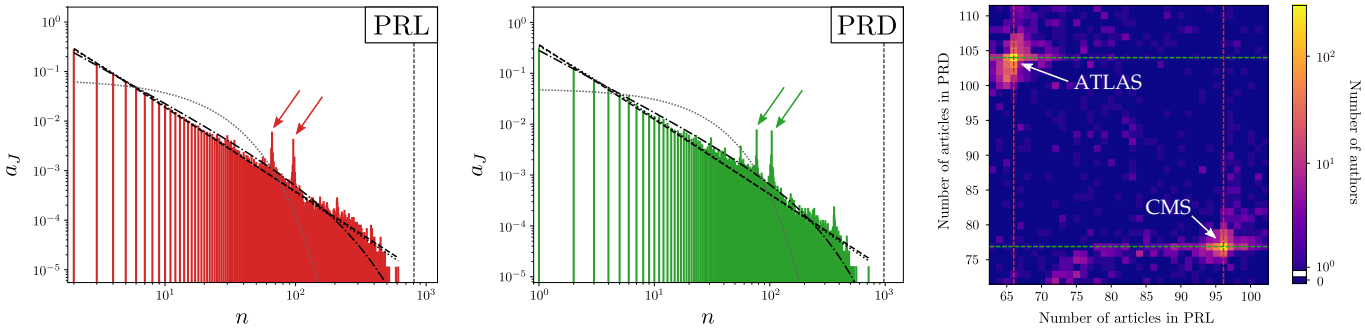

Figure 3. Analysis of PRL and PRD. Left and center: Same figures as in Fig. 1 for PRD and PRL respectively. The arrows indicate the increased number of authors corresponding to the ATLAS and CMS experiments at the CERN. Right: Two-dimensional, color-coded histogram of the number of authors with respect to the number of articles published in PRL (horizontal axis) and PRD (vertical axis). The peak centered at (96,77) is the CMS experiment and the one at (66,104) is the ATLAS experiment, both at the CERN.

of authors are extremely influencial. This translates as some authors having much more publications than what a power law distribution would predict, given that the power law already has an heavier tail than our data. Such authors, which we refer to as key players, are likely to be some very influencial scientists in the topic(s) covered by the journal.

Second, we realized that some huge scientific project can impact the distribution of publications even on large scale aggregated data. In our samples, this is seen for the journals Physical Review Letters (PRL) and Physical Review D (PRD), which publish the outcomes of the large experiments ATLAS and CMS at the CERN, gathering thousands of scientists. Our approach was then able to pinpoint further dynamics taking place in nowadays science.

As seen in Table II, the fitting of the data by a power law with cutoff or a Yule-Simon distribution is not perfect. More advanced fitting techniques might be able to identify a common distribution for all journals, provided that one exists. From a social science point of view, 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 manuscript. This is work for a future research.

MATERIALS AND METHODS

Data sets

We consider an arbitrary selection of 14 peer-reviewed journals (see Table I), whose data are available on the Web of Science data base (WoS). 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 publications available, and all of them are still publishing nowadays. We denote by $\mathcal{J} \coloneqq \{\mathrm{NAT}, \mathrm{PNA}, ..., \mathrm{PRL}\}$ the set of journals considered (see Table I for the list of labels).

Within each journal $J \in \mathcal{J}$, we index authors by an integer and for each author $i = 1, ..., N_J$, we count the number n_i^J of articles published by i in J up to year 2017, which gives the set of data $\mathcal{A}_J = \{n_i^J\}$. We restrict our investigation to publications labelled as *"Article"* in the WoS data base, to focus on peer-reviewed articles and to discard editorial material for instance. For some journals, the number of authors was too large to be downloaded from the WoS data base. As a consequence, the authors having published only one or two articles in these journals had to be removed from the data (e.g., NAT, PNA, or SCI, indicated by asteriscs in Table I). Note also that we do not take into account articles published anonymously, which represent a large number of articles in medicine journals in particular.

From the data set $\mathcal{A}_J$ we can compute the proportion of authors who published $n \in \mathbb{N}$ articles

$$a_J(n) \coloneqq \frac{|\{i \colon n_i^J = n\}|}{N_J} \,. \tag{1}$$

These values are represented in logarithmic scales in Figs. 1, 3, and 5, each panel corresponding to a different journal.

Distribution fitting

For each empirical distribution in Figs. 1, 3, and 5, we fit an exponential distribution (grey dotted lines) to emphasize their heavy-tailed behavior. With this observation, it is tempting to fit a *power law distribution* (black dashed lines),

$$\mathbb{P}_{\rm pl}(a_J = n) = C_1 \cdot n^{-\alpha} \,, \tag{2}$$

with $\alpha > 1$ and $C_1 \in \mathbb{R}$ normalizing the distribution. However, as pointed out in [18], fitting a heavy-tailed distribution is not trivial and should be done carefully, the risk being to derive spurious conclusions [21]. Following recommendations in Ref. [18], we also try to fit other heavy-tailed distributions, such as the *power law with cutoff* (black dash-dotted lines),

$$\mathbb{P}_{\rm plc}(a_J = n) = C_2 \cdot n^{-\beta} e^{-\gamma n} \,, \tag{3}$$

with $\beta > 1$, $\gamma > 0$, and normalizing constant $C_2 \in \mathbb{R}$, and the *Yule-Simon distribution* (black dotted lines),

$$\mathbb{P}_{\rm ys}(a_J = n) = C_3 \cdot (\rho - 1)\mathrm{B}(n, \rho) \,, \tag{4}$$

with $\rho > 0$, $C_3 \in \mathbb{R}$ 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 [18]. Other distribtions (such as log-normal, Lévy, Weibull) were tested and discarded because they were far from matching the data.

Goodness-of-fit

To evaluate the goodness of our fitting, we again follow the recommendations of [18]. We generate 5000 sets of synthetic data $\tilde{\mathcal{A}}_i$, $i = 1, ..., 5000$, with the same number of elements $|\tilde{\mathcal{A}}_i| = N_J$ and following the distribution whose goodness-of-fit is to be tested. For each of these data sets, we define its associated empirical cumulative distribution function (CDF)

$$S_i(k) \coloneqq \frac{|\{x \in \tilde{\mathcal{A}}_i \colon x \leq k\}|}{|\tilde{\mathcal{A}}_i|} \,, \tag{5}$$

and denote by S_J the empirical CDF of $\mathcal{A}_J$. We denote by P_i the CDF of the best fitted distribution associated to $\tilde{\mathcal{A}}_i$ (P_J for $\mathcal{A}_J$). The p-value of the goodness-of-fit is then given by

$$p \coloneqq \frac{|\{i \colon d_{\rm KS}(S_i, P_i) > d_{\rm KS}(S_J, P_J)\}|}{5000} \,, \tag{6}$$

where the *Kolmogorov-Smirnov distance* between two CDFs Q_1 and Q_2 is defined as the maximum difference between them, i.e.,

$$d_{\rm KS}(Q_1, Q_2) \coloneqq \max_k |Q_1(k) - Q_2(k)| \,. \tag{7}$$

Namely, p is the proportion of synthetic data sets that are further from the theoretical distribtion (in the Kolmogorov-Smirnov sense) than the data set investigated. The fit is rejected if $p < 5\%$, and considered as *good* otherwise [see [18] for more details].

Maximum number of articles

Based on Eq. 2, one can compute x_n, the number of authors with n publications in J if the distribution followed a power law. Setting this number to $x_n = 1$, the maximal number of articles is given by

$$x_n \approx N_J C_1 n^{-\alpha} \qquad \Longrightarrow \qquad n_{\max} \approx (N_J C_1)^{\frac{1}{\alpha}} \,. \tag{8}$$

This determines a theoretical upper bound on the number of articles published by an author for each journal, shown as the vertical dashed lines in Figs. 1, 3, and 5.

Number of articles published every year

For three journals (SCI, LAN, and PRL) we compare the number of authors having published k articles at the begining of year t with the number of articles published by these authors during year t. We define:

- $N_k(t)$: the number of authors who have published k articles on December 31st of year $t-1$;
- $m_k(t)$: the number of articles published during year t by all the authors with k articles on December 31st of year $t-1$.

In Fig. 4, we plot the values of $m_k(t)/N_k(t)$ with respect to k for years $t \in \{1999, ..., 2008\}$ for SCI, LAN, and PRL. 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 (retired or dead). For each of the three journals, these values have a linear correlation coefficient larger than 0.7, supporting a fairly good linear dependence,

$$m_k(t) \sim k \cdot N_k(t)\,. \tag{9}$$

The probability that a new paper is signed by an author with k publications is then close to be proportional to k. According to [15], if it was exactly proportional, after a long enough time, the distribution of N_k would follow a power law. The fact that the relation 9 is not exact and that our samples are limited to a finite time horizon, explains that we do not obtain exactly a power law. However, the good correlation between $m_k(t)/N_k(t)$ and k tells us that the distribution should not be too far away from a power law, in agreement with our observation of Table II.

DATA AVAILABILITY

The data are available from WoS. The study used no special computer code.

ACKNOWLEDGMENTS

RD and MT were supported by the Swiss National Science Foundation under grant number 2000020_182050. RD was supported by the Swiss National Science Foundation under grant number P400P2_194359.

APPENDIX

We show here the figures not displayed in the Results section.

[1] D. J. de Solla Price, *Little Science, Big Science* (Columbia University Press, 1963).
[2] A. Clauset, D. B. Larremore, and R. Sinatra, Science **355**, 477 (2017).
[3] S. Fortunato, C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, A. M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, A. Vespignani, L. Waltman, D. Wang, and A.-L. Barabási, Science **359**, eaao0185 (2018).
[4] M. E. J. Newman, Proc. Natl. Acad. Sci. USA **98**, 404 (2001).
[5] D. J. de Solla Price, Science **149**, 510 (1965).
[6] L. Bornmann and R. Mutz, J. Assoc. Inf. Sci. Tech. **66**, 2215 (2015).
[7] J. Bohannon, Science **342**, 60 (2013).
[8] P. Sorokowski, E. Kulczycki, A. Sorokowska, and K. Pisanski, Nature **543**, 481 (2017).
[9] J. E. Hirsch, Proc. Natl. Acad. Sci. USA **102**, 16569 (2005).
[10] G. Siudem, B. Żogaĺa Siudem, A. Cena, and M. Gagolewski, Proc. Natl. Acad. Sci. USA **117**, 13896 (2020).
[11] D. de Solla Price, J. Am. Soc. Inf. Sci. **27**, 292 (1976).
[12] Y.-H. Eom and S. Fortunato, PLoS ONE **6**, e24926 (2011).
[13] L. Waltman, N. J. van Eck, and A. F. J. van Raan, J. Am. Soc. Inf. Sci. Tech. **63**, 72 (2012).
[14] M. Thelwall, J. Infometr. **10**, 336 (2016).
[15] P. L. Krapivsky, S. Redner, and F. Leyvraz, Phys. Rev. Lett. **85**, 4629 (2000).
[16] P. Parolo, R. K. Pan, R. Ghosh, B. A. Huberman, K. Kaski, and S. Fortunato, J. Infometr. **9**, 734 (2015).
[17] V. Sekara, P. Deville, S. E. Ahnert, A.-L. Barabási, R. Sinatra, and S. Lehmann, Proc. Natl. Acad. Sci. USA **115**, 12603 (2018).
[18] A. Clauset, C. R. Shalizi, and M. E. J. Newman, SIAM Review **51**, 661 (2009).
[19] H. Jeong, Z. Néda, and A. L. Barabási, Europhys. Lett. **61**, 567 (2003).
[20] A. L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek, Physica A **311**, 590 (2002).
[21] A. D. Broido and A. Clauset, Nature Comm. **10**, 1 (2019).

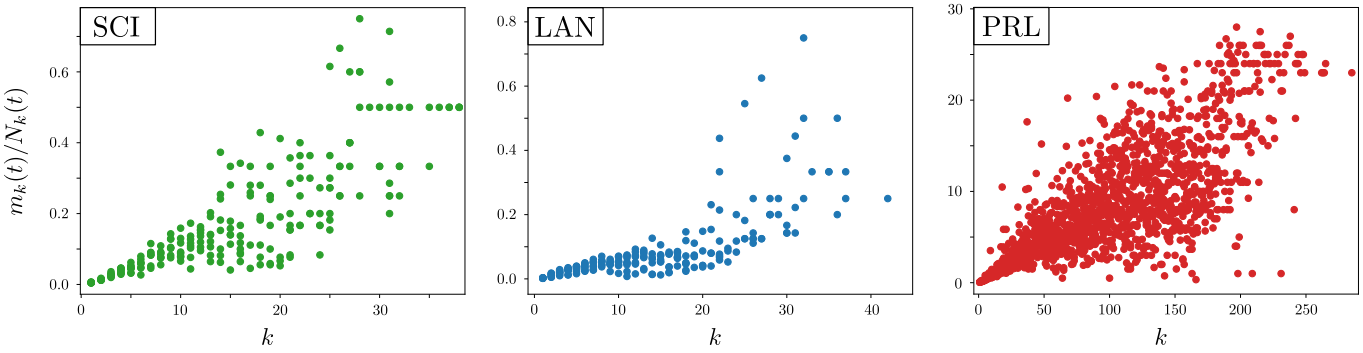

Figure 4. Average number of publication within year t for authors with k publication at the begining of year t, with respect to k, for years $t \in \{1999, ..., 2008\}$ and for the three journals SCI, LAN, and PRL. The Pearson correlation coefficients are respectively $r_{\mathrm{SCI}} \approx 0.714$, $r_{\mathrm{LAN}} \approx 0.707$, and $r_{\mathrm{PRL}} \approx 0.763$, all larger than 0.7, suggesting a relation close to linear. For SCI (resp. LAN and PRL), 14 points (resp. 12 points and 2 points) are left out of the frame, for sake of readability.

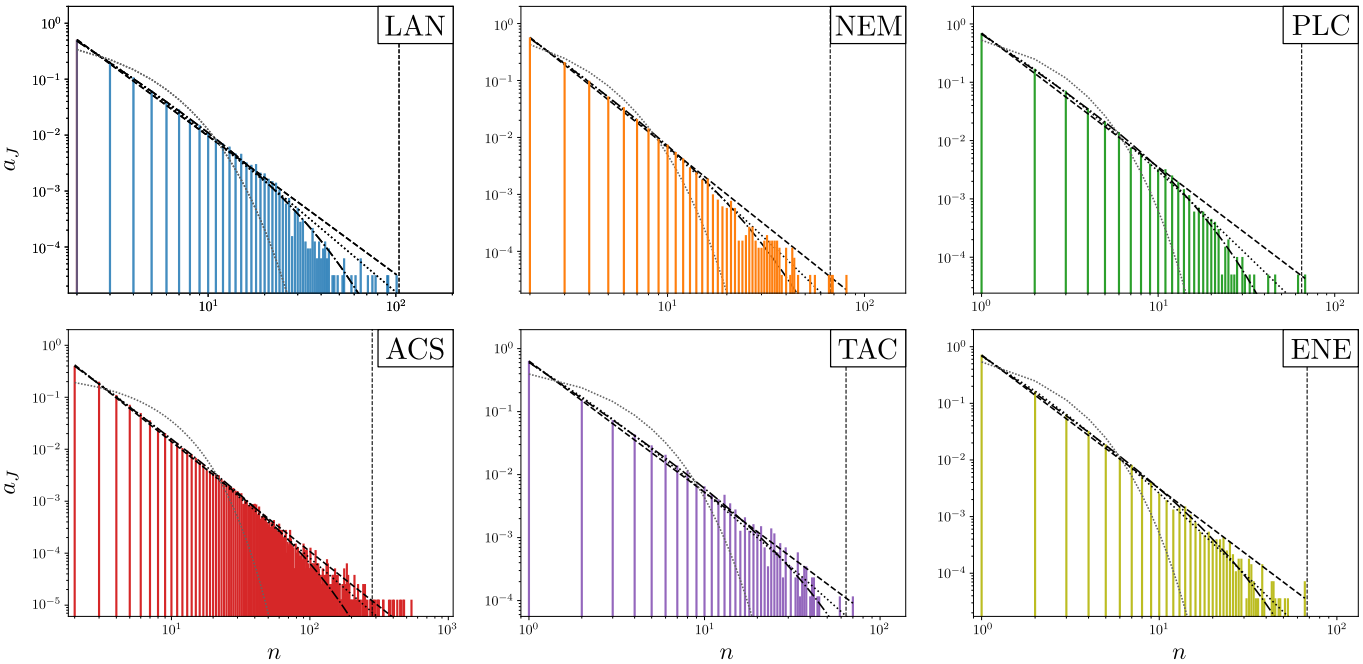

Figure 5. Histograms of the proportion of authors a_J with respect to the number of articles published, for the six journals indicated in the insets. The grey dotted line is exponential fit of the data, emphasizing that the distribution is heavy-tailed. We show the best fit for a power law distribution (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.

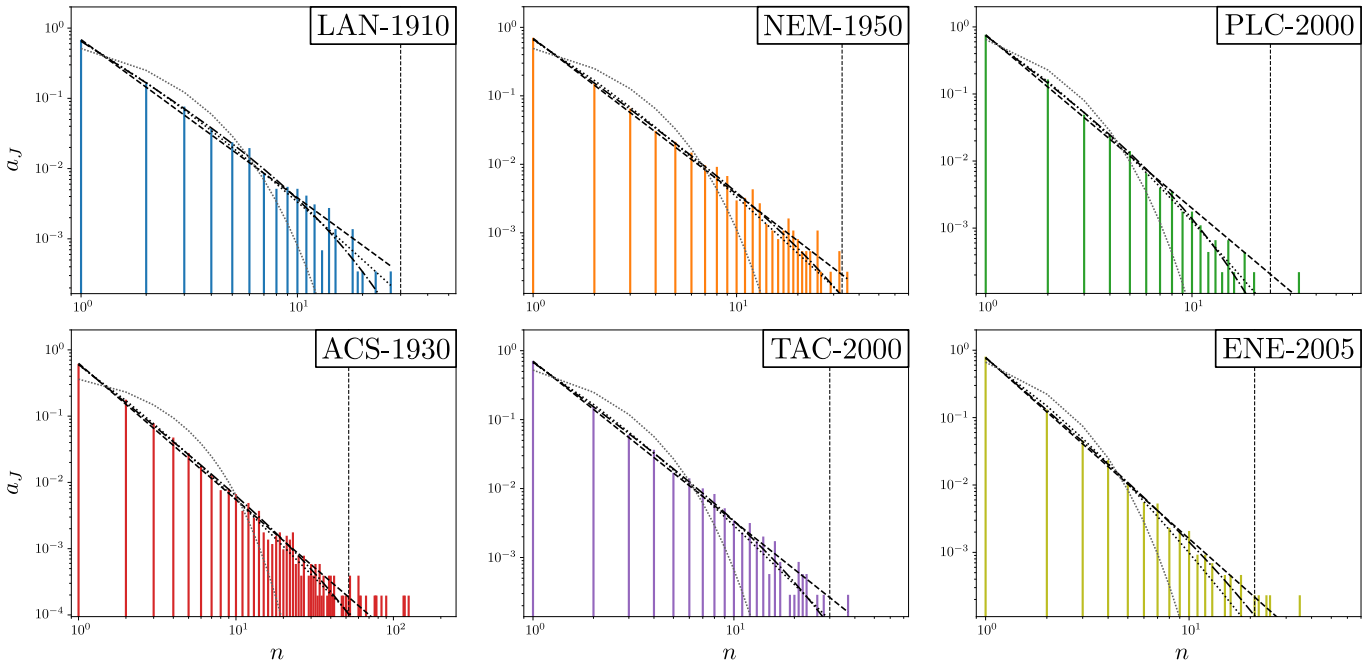

Figure 6. Histograms of the number of authors a_J with respect to the number of articles published, for the six journals indicated in the insets, with data restricted to the years between 1900 (earliest possible in WoS) and the years indicated. The number of authors covered is given in parenthesis in the third column of Table 1 in the Main Text. 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 articles published by some authors.