The K-player: some authors beat the power law

A. B.[1,2], M. N.[2,3], and X. Y.[3,1]
[1] *Concordia Research Station, Antarctica.*
[2] *International Space Station (ISS), Low Orbit.*
[3] *Professor Khromov, Research Vessel, Arctic Ocean.*
(Dated: September 11, 2019)

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

INTRODUCTION

Read Bryson74 and Corral19;
Redo all figures (main 12 journals, pref-attach, PRL-PRD-cross, appendix other journals);
Review all the text, especially intro.

As pointed out by Sekara et al. [1], publishing in a peer-reviewed journal is more likely if one author of the manuscript already published in the same journal. An outcome that could be expected of such an observation is a high representation of a few authors in a given journal.

In line with this, a scientist whose research topic is well-aligned with a journal topic is likely to publish a large proportion of their work in this journal. Leading to a high representation of this author.

In this manuscript, we support these expectations, showing that in a selection of twelve journals, the distribution of the number of authors with respect to the number of articles published within a journal is close to a power-law, and in particular has a heavy tail. Furthermore, we observe that whereas in general this distribution has a slightly thinner tail than a power-law, in some journals, there exists a few authors whose number of publications is significantly larger than what the power-law would predict. We refer to those authors as *K-players* to emphasize their preponderant role in the journal.

We relate this power-law-like distribution to a mechanism that can be compared to *preferential attachement* in the context of network evolution. It has been shown that such a mechanism of network construction leads to a degree distribution that follows a power-law [2].

METHODOLOGY

We consider an arbitrary selection of 12 peer-reviewed journals (see Table I), whose data are available on the Web of Science database (WoS) [3]. Each journal considered is sufficiently old (at least 20 years???) and is still publishing articles nowadays. We denote by $\mathcal{J} := \{\text{BMJ}, \text{CHA}, ..., \text{TAC}\}$ the set of journals considered.

	Label	Journal name
1	NAT	Nature*
2	PNA	Proc. of the Natural Academy of Sciences**
3	SCI	Science*
4	LAN	The Lancet*
5	NEM	New England Journal of Medicine*
6	PLC	Plant Cell
7	PCA	Journal of Physical Chemistry A
8	TAC	IEEE Transactions on Automatic Control
9	ENE	Energy
10	CHA	Chaos
11	SIA	SIAM Journal on Applied Mathematics
12	AMA	Annals of Mathematics
13	PRD	Physical Review D
14	PRL	Physical Review Letters*

Table I. Labels and names of the journals considered. Journals where the authors with one (resp. two) articles published were discarded are indicated by * (resp. **).

Within each journal JOU $\in \mathcal{J}$ and for each author i who published in JOU, we count the number n_i^{JOU} of articles published by i, which gives the set of data $N_{\text{JOU}} = \{n_i^{\text{JOU}}\}$. We restricted our investigation to publications labelled as "Article" in the WoS database, to focus on peer-reviewed articles and to discard editorial material. For some journals, the number of authors was too large to be downloaded from the Web of Science database. As a consequence, some authors having published only one or two articles in these journals had to be removed from the data (e.g., PRL). When this happens, it is indicated by asteriscs in Table I. Note also that we did not take into account articles published anonymously, which represent a large number of articles in medicine journals in particular. Do the same for the time-splitted

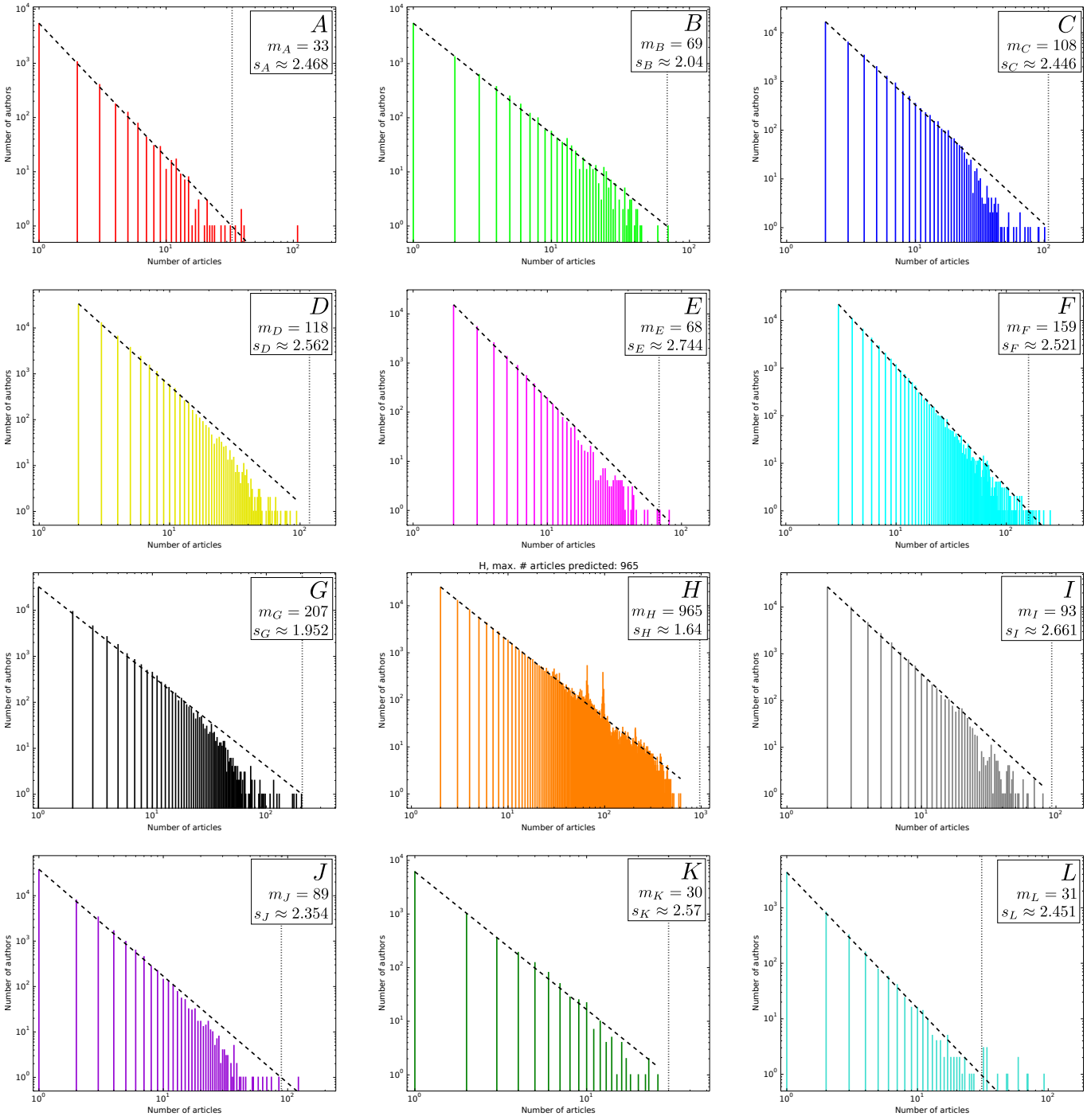

Figure 1. Blah...

data !!! From these data, within each journal JOU $\in \mathcal{J}$, we can compute for each number of articles published n, the number of authors who published n articles

$$a_{\mathrm{JOU}}(n) \coloneqq \frac{\#\{i\colon n_i^{\mathrm{JOU}} = n\}}{N}\,, \tag{1}$$

The distributions of this value are represented in logarithmic scales in Fig. 1, each panel corresonding to a different journal.

To the eye, these distributions have a heavy tail and it is tempting to fit a power law to it,

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

However, as pointed out by Broido and Clauset [4], caution is needed when fitting a distribution to heavy tailed data. Thus we also tried to fit other heavy tailed distri-

		PL		PLwC			Y-S	
		α	p	β	γ	p	ρ	p
1	NAT	2.58	0.00	0.00	0.32	0.00	3.10	0.00
2	PNA	2.53	0.00	0.00	0.17	0.00	2.83	0.00
3	SCI	2.68	0.00	0.00	0.34	0.00	3.28	0.00
4	LAN	2.49	0.00	0.00	0.29	0.00	2.90	0.00
5	NEM	2.76	0.00	0.00	0.36	0.00	3.43	0.094
6	PLC	2.30	0.00	1.92	0.10	0.659	3.01	0.007
7	PCA	2.04	0.00	1.78	0.04	0.927	2.45	0.00
8	TAC	2.08	0.00	1.84	0.04	0.078	2.51	0.0004
9	ENE	2.36	0.00	2.12	0.06	0.199	3.15	0.00
10	CHA	2.47	0.00	2.28	0.05	0.979	3.43	0.00
11	SIA	2.49	0.00	2.20	0.08	0.429	3.49	0.082
12	AMA	2.26	0.00	1.72	0.14	0.206	2.95	0.00
13	PRD	$***$	$***$	1.27	0.005	0.00	$***$	$***$
14	PRL	1.73	0.00	0.91	0.02	0.00	1.80	0.00

Table II. Re-enter new values of PLwC. Fitted parameters and p-value of the goodness-of-fit for power law and power law with cutoff distributions.

butions. We tried the *power law with cutoff* [4],

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

and the Yule-Simon distribution,

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

where $\mathrm{B}(x,y)$ is the beta function. We performed the distribution fitting by optimizing the parameters α, β, γ, and ρ with a Maximum Likelihood Estimator (MLE) [5].

To evaluate the goodness of our fitting, we proceeded as follows. We generated 2500 sets of synthetic data D_i, $i = 1, ..., 2500$, following the distribution obtained. Comparing our data $D_0 = \{a_{\rm JOU}(n)\}$ to the sets of synthetic data, we can compute p, the proportion of synthetic data that are further than D_0 (in the Kolmogorov-Smirnov sense REF) from the theoretical distribution. The fit is considered as *good* if $p > 0.05$.

RESULTS

The results of each fit and goodness-of-fit tests are presented in Table II. Clearly, the power law distribution is a poor fit for all data, its p-value being (almost) zero for all journals. This can be seen in Fig. 1 as for most of the journals, the tail of the data set is lighter that the tail of its power law fit (black dashed line). However, for some journals (namely BLI, BLA, BLO), the p-value of the power law with cutoff is high and it seems to be a rather good fit (black dotted line in Fig. 1).

We tried to fit other heavy tailed distributions (such as BLI, BLA, BLO), but none of them performed better than the power law with cutoff under our goodness-of-fit test.

Note that exponential distribution is taken into account in the *power law with cutoff*. But in any case where the data were better fitted by an exponential, the KS test discarded it.

General explanation. The better explanation we found to this heavy tail behavior is the following. It has been shown/postulated [REF], that the number of coauthorship of an author is ruled by *preferential attachement*. 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 means 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 given 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. Assumptions (i) and (ii) together imply that this author is likely to published again in this journal.

To make this more rigorous, for three journals (BLI, BLA, and BLO) we compared the number of authors having published k articles at year t with the number of articles published by these authors between years t and $t+1$. Defining $m_k(t,s)$ as the number of articles published between years t and s by the authors with k articles at time t, we plot in Fig. 2 the values of $m_k(t,t+1)/N_k(t)$ with respect to k for years $t \in \{1999, ..., 2008\}$. Note that, for each year considered, we did 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 0.5, supporting a fairly good linear dependence,

$$m_k(t,t+1) \approx k \cdot N_k(t)\,. \tag{5}$$

Restrict the data to a time span of max. 30 years to remove authors who are not publishing anymore.

The probability that a new paper is signed by an author with k publications is then close to be proportional to k. According to [2], 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 (5) is not exact and that our samples are limited to a finite time horizon, explain that we do not obtain exactly a power law. However, the good correlation between $m_k(t,t+1)/N_k(t)$ and k tells us that the distribution should not be too far away from a power law, in agreement with our observation of a power law with cutoff.

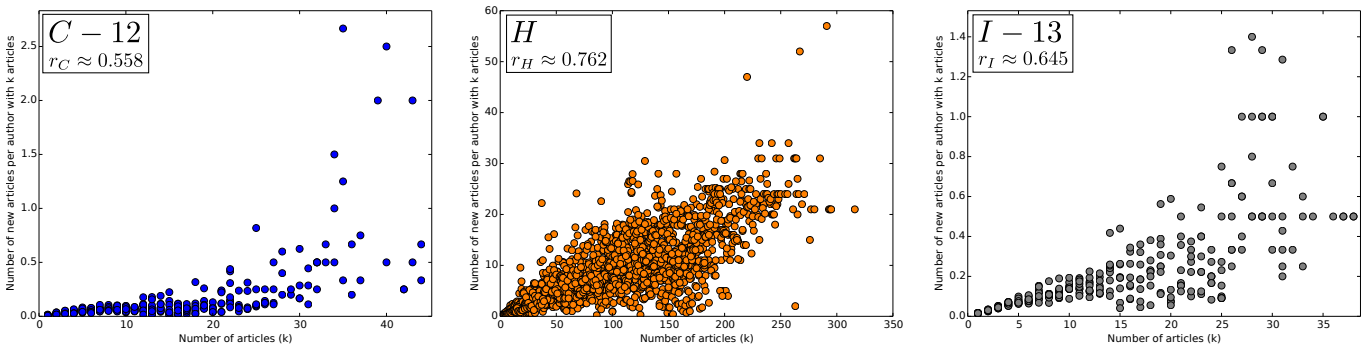

Figure 2. Blih...

List of discarded distribution

Exonential: light tail.

Poisson: light tail.

Strechted exponential: light tail.

Log-normal: shape does not match, in log-log scale, it is exactly a parabola, for some of our data, it could fit the tail only (moreover, there is no closed form for the normalizing factor in the discrete version, so very difficult to implement MLE).

Parabolic fractal: the tail is a straight line in log-log scale, so it will not do better than the power-law.

Lévy: the tail is a straight line in log-log scale, so not better than power-law.

Weibull: in log-log scale, it is very flat at the beggining a drops very fast, i.e., the shape does not match at all.

OBSERVATIONS

Aside of these general considerations, we note two interesting observations in the data. Namely, some authors are stronger that the power law, and some very large experiments can be seen even in aggregated data.

Exceptions

The general distribution of the number of authors with respect to the number of paper per author is quite clear in our analysis. However, in some journals, we observe anomalies (see journals BLI,BLA, and BLO in Fig. 1). It appears that sometimes, some authors publish significantly more articles in a journal than what the power law with cutoff would predict, and somtimes even more articles that what a power law would predict. These authors, who we refer to as *K-players*, are supposedly some very influencial scientists in the journal considered, and they literally *beat the power law*.

Remark. *We emphasize that we checked that these K-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 authors' list of a very large number of papers.*

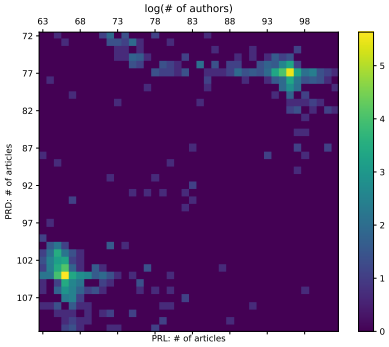

Figure 3. Beuh.

Peaks in PRL and PRD

We observe two peaks in the distribution of PRL (around 70 and 96) and PRD (around 77 and 104). Crossing the lists of authors for each number of articles between 63 and 102 (resp. 72 and 111) for PRL (resp. PRD), we get Fig. 3. The fact that the authors composing a peak in PRL are the same composing one of the peaks in PRD suggests that these authors are all part of a large group publishing together. After a quick search, we realize that the peaks correspond to the research groups of the experiments ATLAS and CMS at the *Centre Européen de Recherche Nucléaire* (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).

CONCLUSION

[1] V. Sekara, P. Deville, S. E. Ahnert, A.-L. Barabási, R. Sinatra, and S. Lehmann, PNAS **115**, 12603 (2018).

[2] P. L. Krapivsky, S. Redner, and F. Leyvraz, Phys. Rev. Lett. **85**, 4629 (2000).

[3] http://apps.webofknowledge.com .

[4] A. D. Broido and A. Clauset, Nature Comm. **10**, 1 (2019).

[5] A. Clauset, C. R. Shalizi, and M. E. J. Newman, SIAM Review **51**, 661 (2009).