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: February 25, 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

As pointed out by Sekara et al. [1], publishing in a peer-reviewed journal is more likely if one author of the paper 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 journals, labelled by capital letters, $\mathcal{J} = \{A, B, ..., L\}$, available on the Web of Science database (WoS) [3]. Each element of $\mathcal{J}$ corresponds to a peer-reviewed journal with a significant number of publications within the last decades. We do not explicitly give the journals' names for privacy reasons.

Within each journal $j \in \mathcal{J}$, we count the number n_i^j of articles published by author i, which gives the set of data $\{n_i^j\}$. We restricted our investigation to publications labelled as "Article" in the WoS database. 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. As this might influence our results, we indicate when this was the case in the captions. Note also that we did not take into account articles published anonymously. Do the same for the time-splitted data !!!

From these data, we can the compute

$$a_j(n) := \#\{i \colon n_i^j = n\}\,, \tag{1}$$

which are represented in logarithmic scales in Fig. 1, for each journal. We then fit a power law (black dashed lines in Fig. 1) to the data of each journal $j \in \mathcal{J}$. The exponent s_j of the power law

$$z_j(n) = C_j n^{-s_j} \tag{2}$$

is obtained by a maximum likelihood estimator, following [4], and C_j is the constant normalizing the distribution. Finally, we compute the theoretical maximum number of articles, m_j, which is the value satisfying $N_j z_j(m_j) \approx 1$, where N_j is the total number of articles published in journal j.

Following the warnings of Broido and Clauset [5], about the over-representation of power-laws data analysis, we have carefully considered other possible distributions, namely BLAH, BLAH, AND BLAH. The best estimate of the distributions observed were the power-law with exponents given in Fig. 1.

RESULTS

To the eye, the histograms for all journals considered follow a power law, also called a Zipf's law for discrete-valued variables. Performing a Kolmogorov-Smirnov test

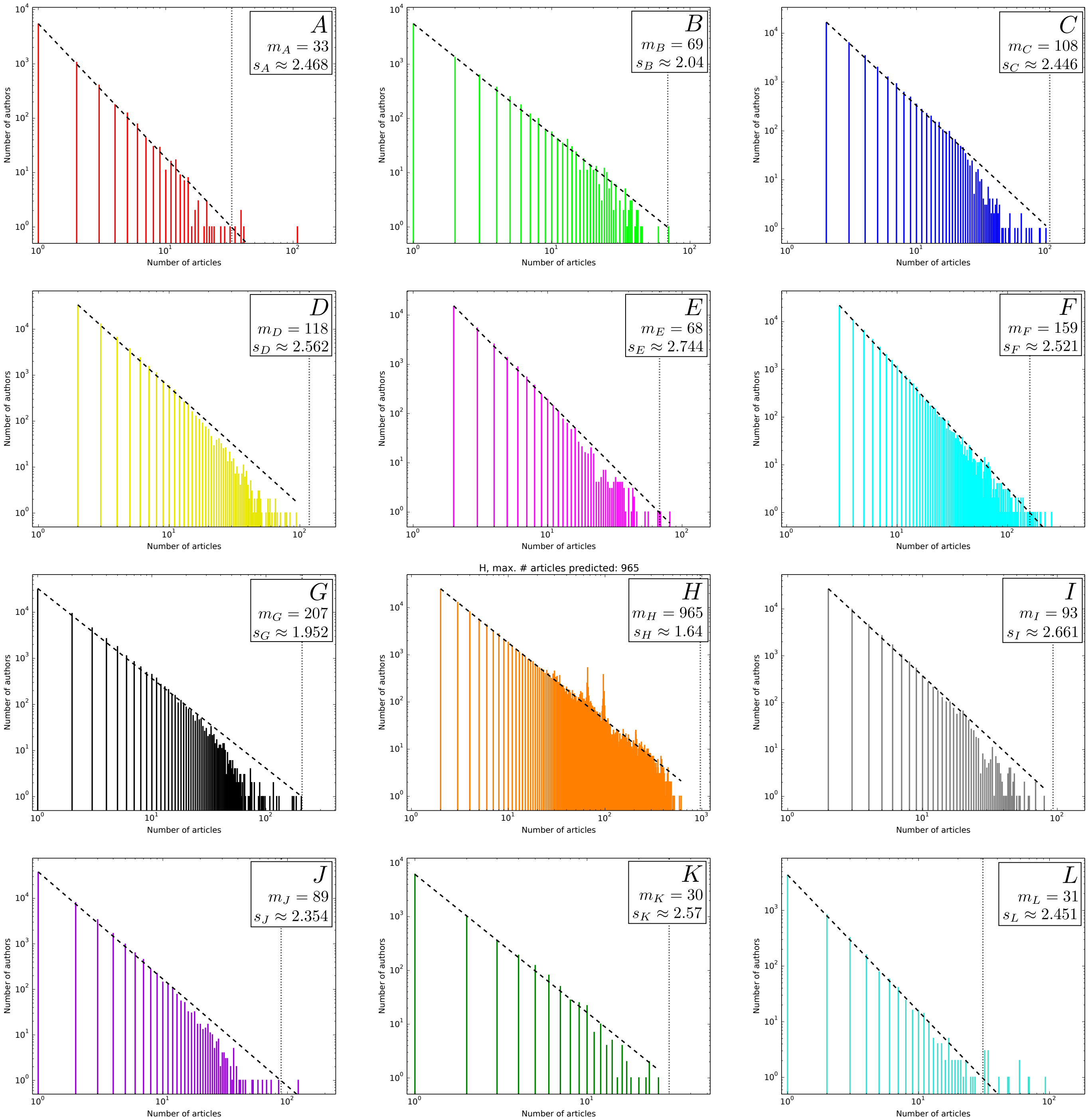

Figure 1. Blah...

on the data set, we observe that this fit does not perform very good. However, trying to fit other standard (and less standard) distributions, we were unable to find a better fit, mostly because the distribution tails of other distributions are too thin. Thus even if the data does not follow precisely a power-law, they are close to one, with a thinner tail. These discrepancies might be due to the size of the data sets, which are rather small compared to what would be needed to accurately give a power-law [REF].

General explanation. Our explanation to this fact is that the evolution of the number of articles in a journal is described by a *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. Consequences (i) and (ii) together imply that this author is likely to published again in this journal.

To make this more rigorous, compared for two journals the number of authors having published k articles at time t with the number of articles published by these authors between times t and $t+1$. Defining $m_k(t,s)$ as the number of articles published between t and s by the authors with k articles at time t, we plot in Fig. Figure the values of $m_k(t,t+1)/N_k(t)$ with respect to k for years $t \in \{1999, ..., 2008\}$. These values have a linear correlation coefficient of BLAH, supporting a fairly good linear dependence,

$$m_k(t,t+1) \approx k \cdot N_k(t)\,. \tag{3}$$

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 proportional to k. The dynamics of the number of authors with k articles is then described by Eq. (1) in [2], with exponent $\gamma \approx 1$. According to [2], after a long enough time, the distribution of N_k follows a power law, which agrees with our observations.

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 Figs...). It appears that sometimes, some authors publish more articles in a journal than what our Zipf's law distribution would predict. These authors, who we refer to as *K-players*, are supposedly some very influencial scientists in the journal considered here, 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.*

PEAKS IN PHYS. REV. LETT. AND PHYS. REV. D

We observe two peaks in the distribution of Phys. Rev. Lett. (at 70 and 96) and Phys. Rev. D (at 77 and 104). Crossing the lists of authors for each number of articles between 63 and 102 (resp. 72 and 111) for Phys. Rev. Lett. (resp. Phys. Rev. D), we get Fig. 3. The fact that the authors composing a peak in Phys. Rev. Lett. are the same composing one of the peaks in Phys. Rev. D indicates that these authors are all part of a large group publishing together. After a quick search, we realize that the peak 1 corresponds to the research group Blah... in CERN and peak 2 blah blah.

CONCLUSION

The aim of this manuscript is to point out some puzzling observations that could lead to further investigations. Our investigation are limited to a rather small number of journals and would need a large scale study to confirm the general validity of this power law behavior. However, in our opinion, the regularity of our observations regardless of the size and age of the journal, as well as with respect of the time interval considered are sufficient evidence to formulate a conjecture.

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

[5] A. D. Broido and A. Clauset, arXiv preprint: 1801.03400 (2018).

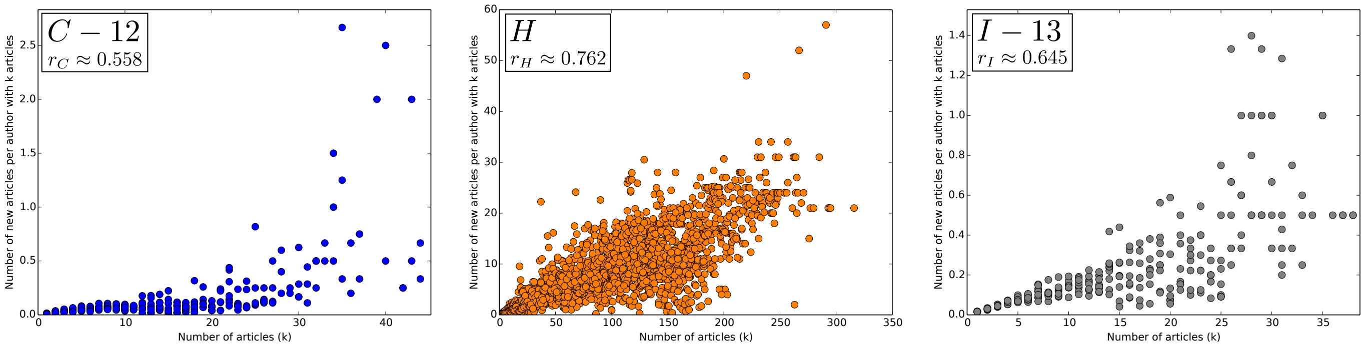

Figure 2. Blih...

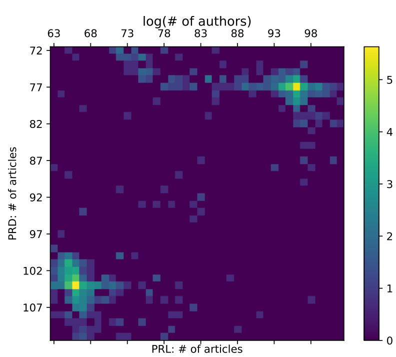

Figure 3. Beuh.