The K-player: some authors beat the power law

R. Delabays[a,b,1] **and M. Tyloo**[a,c]

[a]School of Engineering, University of Applied Sciences of Western Switzerland HES-SO CH-1951 Sion, Switzerland.; [b]Automatic Control Laboratory, Swiss Federal Institute of Technology (ETH) Zürich, Switzerland.; [c]Institute of Physics, EPF Lausanne, CH-1015 Lausanne, Switzerland.

January 15, 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.

Keyword 1 | Keyword 2 | Keyword 3 | ...

1. Introduction

Since the emergence of digitalized databases of scientific publications, large-scale statistics about the peer-reviewed publication system can be computed easily. The availability of such data opened the way to new investigations of scientists on their own activity. One of the first objects that comes to mind when talking about scienific publications is the graph of scientific collaborations, where each vertex represents a scientist, and two scientists are connected if they cosigned a article. This graph has been largely studied, and various graph measures have been used on it to determine important scientists or the emergence of a new field of research [REF]. A significant amount of the literature is concerned about the growth of this graph, which is constantly growing in terms of vertices (new authors publish every year) and of edges (new collaborations are developped every year). According to most of the literature, the graph of scientific collaborations is scale-free [REF], due to the *preferential attachement* process that is supposed to rule its growth.

In this brief note, we are interested in the distribution, within a journal, of the number of authors with respect to the number of articles published. To be more precise, for a journal j, we are interested in the distribution of $a_j(n)$, the number of authors having published n articles in j, as a function of n. To the best of our knowledge, such statistics have not yet been investigated.

Based on the data of the Web of Science database (1), we show that for all journals investigated, the distribution of a_j is well approximated by a power law. We can reasonably consider that preferential attachement leads to this distribution. Even if the ubiquity of power law distributions in real-world systems has been revised recently(2), in the world of scientific publications, many quantities happen to follow such a distribution [REF]. Distribution are even usually scale-free, meaning that the exponenent of the power law is in the interval [bli,bla] according to the literature [REF]. For the journals studied here, the distribution of a_j does not contradict this rule.

We observe however a surprizing anomaly in the data. For some journals, the distribution of a_j is well approximated by a power law until it takes value 1, but then a few authors *beat the power law*, namely, they have published significatly more articles than what would be expected following the power law.

2. Methodology

We consider an arbitrary selection of 12 journals, labelled by capital letters, $\mathcal{J} = \{A, B, C, D, E, F, G, H, I, J, K, L\}$, available on the Web of Science database (1).

Remark. *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 $\{n_i^j\}$. From these data, we can the compute

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

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

$$z_j(n) = C_j n^{-s_j} \qquad [2]$$

is obtained by a maximum likelihood estimator, following (3), 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.

Remark. *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*

Significance Statement

Authors must submit a 120-word maximum statement about the significance of their research paper written at a level understandable to an undergraduate educated scientist outside their field of speciality. The primary goal of the Significance Statement is to explain the relevance of the work in broad context to a broad readership. The Significance Statement appears in the paper itself and is required for all research papers.

Please provide details of author contributions here.

Please declare any conflict of interest here.

[1]A.O.(Author One) and A.T. (Author Two) contributed equally to this work (remove if not applicable).

[1]To whom correspondence should be addressed. E-mail: author.twoemail.com

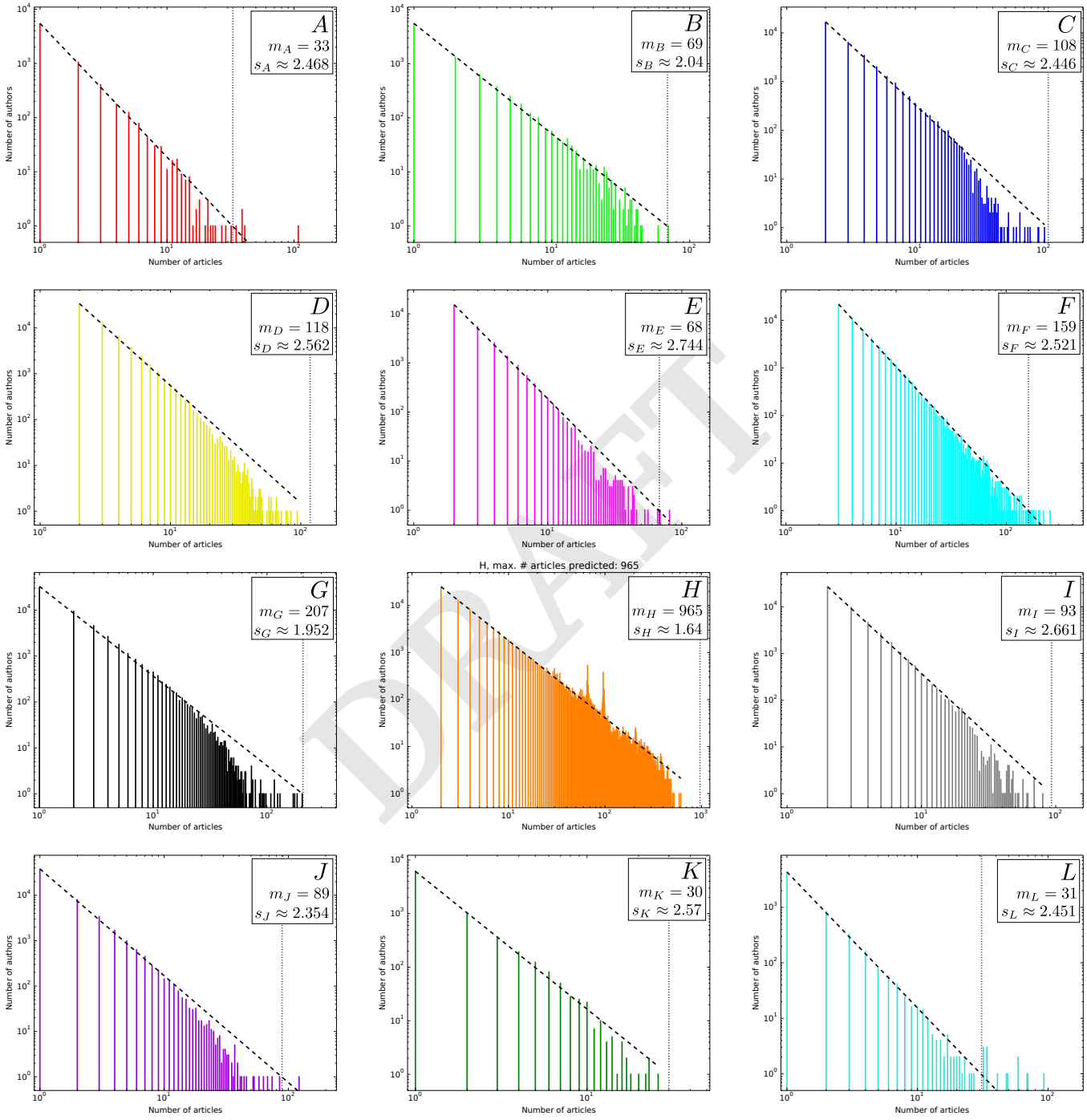

Fig. 1. Blah...

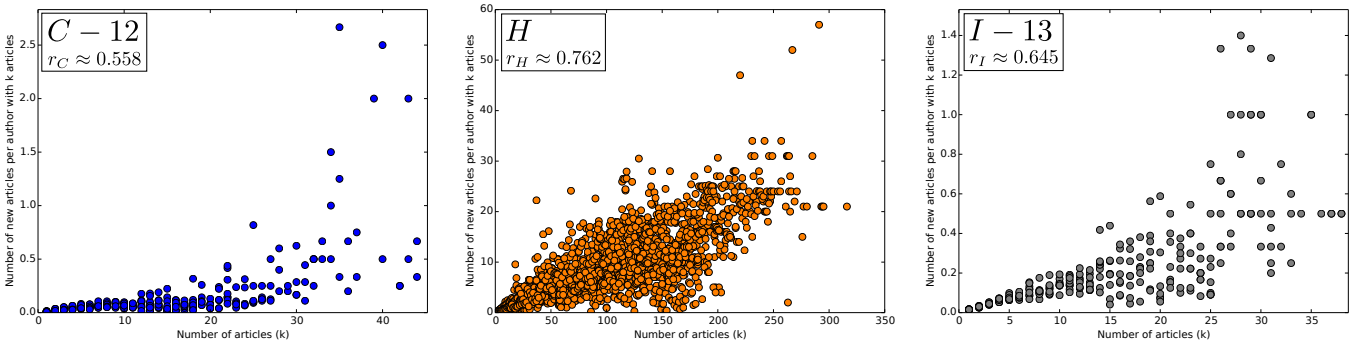

Fig. 2. Blih...

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 !!!

3. Results

To the eye, the histograms for all journals considered follow a power law, also called a Zipf's law for discrete-valued variables. The error of the fit is plotted in Fig. Figure, confirming that Zipf's law correctly describes the distribution of N_k^j.

The Zipf's law approximates the data well, slightly overestimating it for large values of n.

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,

$$m_k(t,t+1) \approx k \cdot N_k(t)\,. \qquad [3]$$

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 (4), with exponent $\gamma \approx 1$. According to (4), 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.*

4. Conclusion

The aim of this letter 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. http://apps.webofknowledge.com (2018).
2. Broido AD, Clauset A (year?) Scale-free networks are rare. *arXiv preprint: 1801.03400.*
3. Clauset A, Shalizi CR, Newman MEJ (2009) Power-Law Distributions in Empirical Data. *SIAM Review* 51(4):661–703.
4. Krapivsky PL, Redner S, Leyvraz F (2000) Connectivity of growing random networks. *Phys. Rev. Lett.* 85(21):4629–4632.