$9^{th}$ International Conference on Computational Social Science IC$^2$S$^2$
July 17-20, 2023, Copenhagen, Denmark

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

*Keywords: Scientometrics, scientific publication, power law, heavy tails, modeling.*

## Extended Abstract

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 [1, 2]). A vast majority of investigations about the scientific publication process is focused on the citation side. These analysis mostly aim at describing 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 [3]. This process leads to a power law distribution of citations [4, 5] or other heavy-tailed distributions [6]. Indeed, preferential attachment has been proven to lead to heavy-tailed distributions [7], with some refinements to account for the life-time of a paper [8].

As early as 1926, Lotka showed that, in the field of chemistry, the number of scientists having published $N$ papers is proportional to $N^{-2}$ [9]. In other words, he showed that the distribution of the number of papers published by scientists follows a power law. Later on, the same analysis has been extended to other fields of science [e.g., [10, 11, 12, 13, 14, 15] and refined to more elaborate distributions, such as the *power law with cutoff* [16, 17, 18] or the *stretched exponential* distribution [19]. 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 analysis, in this article, we focus here 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 data base) 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 Fig. 1, where the analysis of the distribution emphasizes: (i) an underlying *preferential attachment* dynamics; (ii) the finiteness of the scientific careers; and (iii) the presence of (very) large groups of scientists in the related fields of physics (see caption of Fig. 1 for a detailed discussion).

As interestingly pointed out by [20], 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 interpreted as *preferential attachment*, and an expected outcome of such an observation is a high representation of a few authors in a given journal [7]. 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.

The heavy-tailedness of the distribution of the number of papers is striking in the histograms (see Fig. 1). 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 *power law* is not a good fit to the

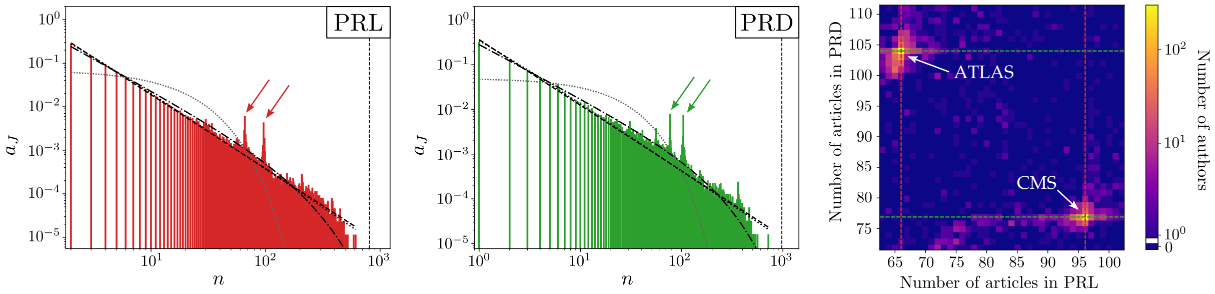

Figure 1: Left and center: histograms of the number of papers $n$ published in *Phys. Rev. Lett.* (PRL) and *Phys. Rev. D* (PRD) among the authors who published in these journals. For each value of $n$, the height of the bar gives the proportion of authors who published $n$ articles in the corresponding journal. Best distribution fits are displayed for an exponential distribution (gray dotted), a power law (dashed black), an 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). One realizes the impact of such large scale experiment on the publication landscape.

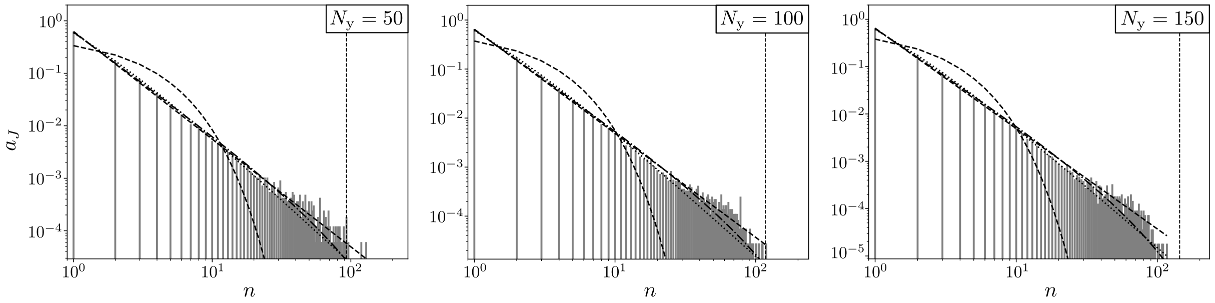

Figure 2: Histograms of the outcome of our synthetic data generator, for different value of the journal life span $N_\mathrm{y}$ years. There is a clear similarity between the shapes of these synthetic distributions and those of the actual data.

data neither, and the actual distribution lies somewhere between an exponential and a power law.

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. The two main ingredients of our evolution model are (i) a preferential attachment process and (ii) a limited career length for authors. These two aspects of the model have a contradicting influence on the number of papers of each author and therefore a significant impact of the shape of the final distribution. Indeed, the preferential attachment drives the distribution towards a power law, that would be reached if the authors had infinite career-span. Counterbalancing the preferential attachment, the limited career length of each author prevents the tail of the distribution to be tooo heavy, leading to the typical cutoff seen in Fig. 1, which we reproduce with synthetic data in Fig. 2.

# References

[1] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci. USA*, 102(46):16569–16572, 2005.

[2] G. Siudem, B. Żogała Siudem, A. Cena, and M. Gagolewski. Three dimensions of scientific impact. *Proc. Natl. Acad. Sci. USA*, 117(25):13896–13900, 2020.

[3] D. de Solla Price. A general theory of bibliometric and other cumulative advantage processes. *J. Am. Soc. Inf. Sci.*, 27:292–306, 1976.

[4] Y.-H. Eom and S. Fortunato. Characterizing and modeling citation dynamics. *PLoS ONE*, 6(9):e24926, 2011.

[5] L. Waltman, N. J. van Eck, and A. F. J. van Raan. Universality of citation distributions revisited. *J. Am. Soc. Inf. Sci. Tech.*, 63(1):72–77, 2012.

[6] M. Thelwall. The discretised lognormal and hooked power law distributions for complete citation data: Best options for modelling and regression. *J. Infometr.*, 10:336–346, 2016.

[7] P. L. Krapivsky, S. Redner, and F. Leyvraz. Connectivity of growing random networks. *Phys. Rev. Lett.*, 85(21):4629–4632, 2000.

[8] P. Parolo, R. K. Pan, R. Ghosh, B. A. Huberman, K. Kaski, and S. Fortunato. Attention decay in science. *J. Infometr.*, 9:734–745, 2015.

[9] A. J. Lotka. The frequency distribution of scientific productivity. *J. Washington Acad. Sci.*, 16(12):317–232, 1926.

[10] B. M. Gupta and C. R. Karisiddappa. Author productivity patterns in theoretical population genetics (1900–1980). *Scientometrics*, 36(1):19–41, 1996.

[11] R. Wagner-Döbler and J. Berg. Physics 1800–1900: A quantitative outline. *Scientometrics*, 46(2):213–285, 1999.

[12] J. C. Huber and R. Wagner-Döbler. Scientific production: A statistical analysis of authors in physics, 1800-1900. *Scientometrics*, 50(3):437–453, 2001.

[13] J. C. Huber and R. Wagner-Döbler. Scientific production: A statistical analysis of authors in mathematical logic. *Scientometrics*, 50(2):323–337, 2001.

[14] M. Sutter and M. G. Kocher. Power laws of research output. Evidence for journals of economics. *Scientometrics*, 51(2):405–414, 2001.

[15] M. Barrios, A. Borrego, A. Vilaginés, C. Ollé, and M. Somoza. A bibliometric study of psychological research on tourism. *Scientometrics*, 77(3):453–467, 2008.

[16] N. J. Saam and L. Reiter. Lotka's law reconsidered: The evolution of publication and citation distributions in scientific fields. *Scientometrics*, 44(2):135–155, 1999.

[17] H. Kretschmer and R. Rousseau. Author inflation leads to a breakdown of Lotka's law. *J. Amer. Soc. Inf. Sci. Tech.*, 52(8):610–614, 2001.

[18] L. Smolinsky. Discrete power law with exponential cutoff and Lotka's law. *J. Assoc. Inf. Sci. Tech.*, 68(7):1792–1795, 2017.

[19] J. Laherrère and D. Sornette. Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales. *Eur. Phys. J. B*, 2(4):525–539, 1998.

[20] V. Sekara, P. Deville, S. E. Ahnert, A.-L. Barabási, R. Sinatra, and S. Lehmann. The chaperone effect in scientific publishing. *Proc. Natl. Acad. Sci. USA*, 115(50):12603–12607, 2018.