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

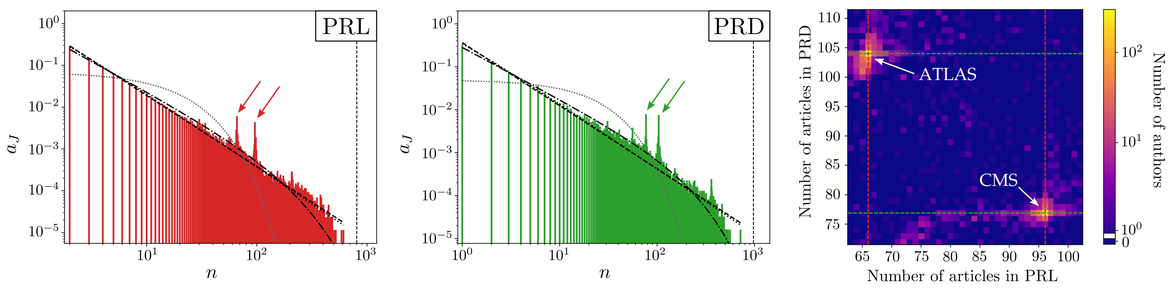
Robin Delabays[1] and Melvyn Tyloo[2]

[1] *School of Engineering, University of Applied Sciences of Western Switzerland Valais-Wallis, Sion, Switzerland.*
[2] *Theoretical Division, Los Alamos National Laboratory, USA.*

robin.delabays@hevs.ch

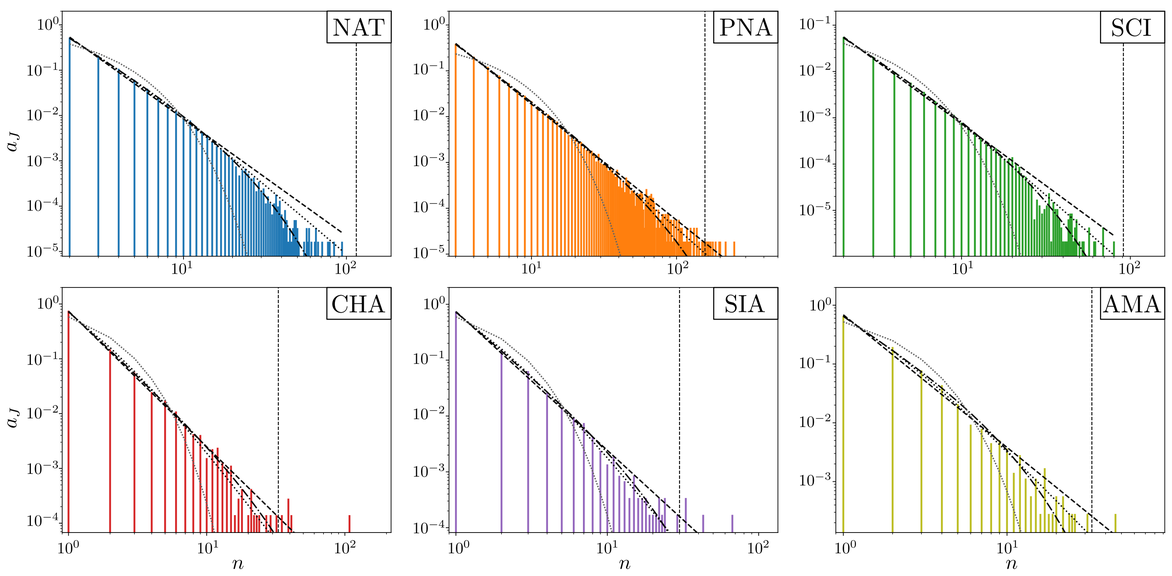
Introduction

Scholarly publications represent at least two benefits for the study of the scientific community as a social group. First, they attest to some form of relation between scientists, useful to determine and analyze social subgroups. Second, most of them are recorded in large databases, easily accessible and including a lot of pertinent information, easing the quantitative and qualitative study of the scientific community. In this work, we perform a statistical analysis of publication within peer-reviewed journals.
Namely, we show that the distribution of the number of papers published by an author in a given journal is heavy-tailed, but has a lighter tail than a power law. Interestingly, we demonstrate (both analytically and numerically) that such distributions match the result of a modified preferential attachment process, where, on top of a Barabási-Albert process, we take the finite career span of scientists into account.



Empirical and fitted distributions

Publication data are presented as histograms of the number of papers published by an author, in a given journal. The tail of the empirical distribution is heavier than an exponential distribution (best fit in dashed gray).



Despite a rather good visual fit, the *power law* (PL, dashed black), *power law with cutoff* (PLwC, dash-dotted black), and *Yule-Simon* (Y-S, dotted black) distributions are usually not a very good fit.

		PL		PLwC			Y-S	
		α	p [%]	β	γ	p [%]	ρ	p [%]
NAT	Nature	2.58	0.0	2.11	0.07	0.0	3.10	0.0
PNA	PNAS	2.53	0.0	2.30	0.02	0.0	2.83	0.0
SCI	Science	2.68	0.0	**2.30**	**0.06**	**16.64**	3.28	0.02
LAN	The Lancet	2.47	0.0	2.09	0.05	0.18	2.90	0.0
NEM	New England Journal of Medicine	2.76	0.0	2.36	0.07	0.2	**3.43**	**8.82**
PLC	Plant Cell	2.30	0.0	**1.92**	**0.10**	**13.42**	3.01	0.92
ACS	Journal of the American Chemical Society	2.11	0.0	1.95	0.01	0.0	2.32	0.0
TAC	IEEE Transactions on Automatic Control	2.08	0.0	1.84	0.04	0.0	2.51	0.02
ENE	Energy	2.36	0.0	2.12	0.06	0.12	3.15	0.0
CHA	Chaos	2.47	0.0	**2.28**	**0.05**	**80.84**	3.43	0.0
SIA	SIAM Journal on Applied Math	2.49	0.0	2.20	0.08	2.24	**3.49**	**9.06**
AMA	Annals of Mathematics	2.26	0.0	1.72	0.14	0.18	2.95	0.0
PRD	Physical Review D	1.49	0.0	1.24	0.005	0.02	1.55	0.0
PRL	Physical Review Letters	1.73	0.0	1.52	0.005	0.12	1.80	0.0

Fitting is done following the recommendation of Ref. [2].

Heavy-tailed distributions

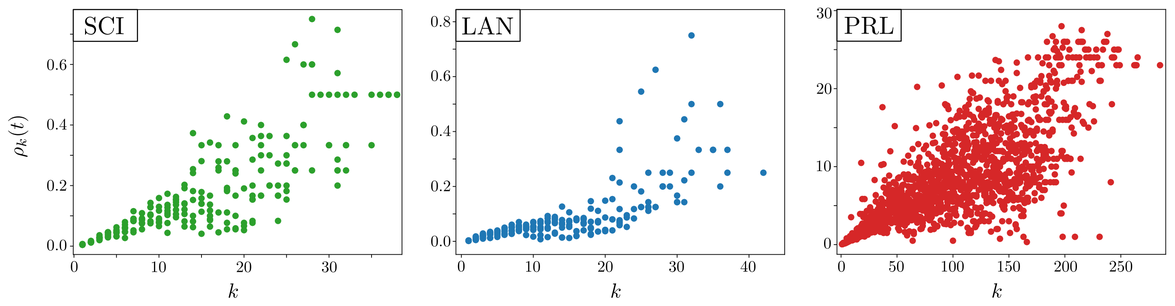
The three heavy-tailed distribution that we fit are:

- A *power law distribution* (black dashed lines in the figures),
$$P_{\mathrm{pl}}(n_i^J = n; \alpha) = C_\alpha n^{-\alpha}\,,$$
with $\alpha > 1$ and $C_\alpha \in \mathbb{R}$ normalizing the distribution;
- A *power law with cutoff* (black dash-dotted lines in the figures),
$$P_{\mathrm{plc}}(n_i^J = n; \beta, \gamma) = C_{\beta,\gamma} n^{-\beta} e^{-\gamma n}\,,$$
with $\beta > 1$, $\gamma > 0$, and normalizing constant $C_{\beta,\gamma} \in \mathbb{R}$;
- A *Yule-Simon distribution* (black dotted lines in the figures),
$$P_{\mathrm{ys}}(n_i^J = n; \rho) = C_\rho(\rho - 1)\mathrm{B}(n, \rho)\,,$$
with $\rho > 0$, $C_\rho \in \mathbb{R}$ is the normalizing constant, and where $\mathrm{B}(x, y)$ is the *Euler beta function*.

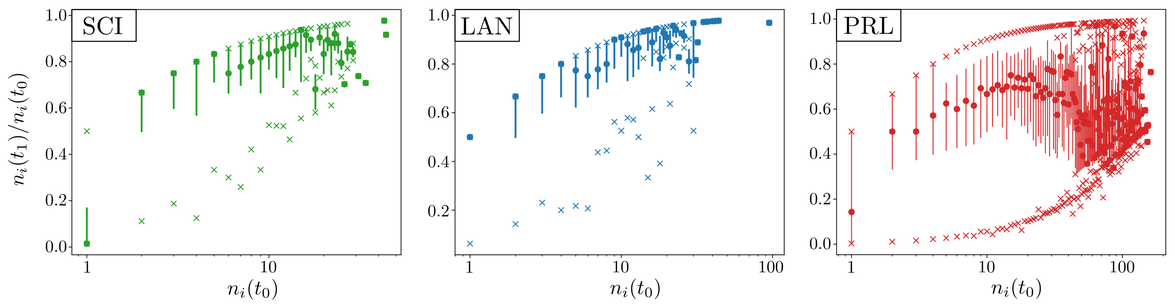
Preferential attachment or cumulative advantage

We argue that the heavy-tailedness of the distributions follows from a *preferential attachment* or *cumulative advantage* process. This claim relies on two observations.

First, we see a good correlation (> 0.7) between $\rho_k(t)$, the number of papers published within a year t by a given author, and k, the number of papers published by this author up to year t.



Second, we observe that between year $t_0 = 1999$ and year $t_1 = 2008$, the proportion of papers gained is larger for authors with more papers in year t_0. Therefore, authors with more papers gain a disproportionate amount of papers over time.

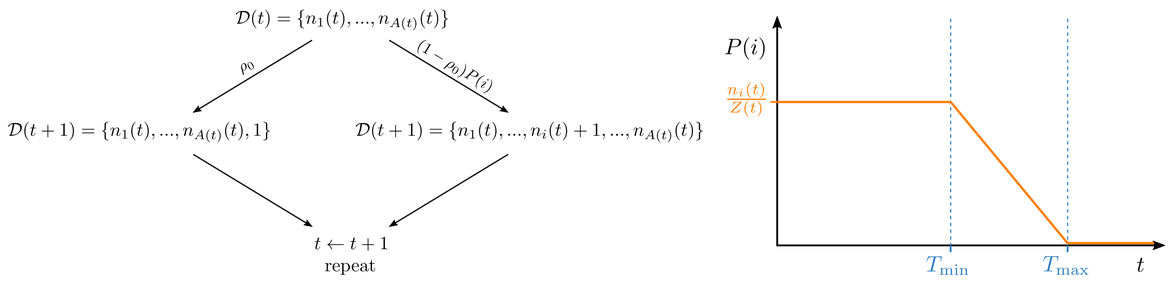


It is known [3] that if the process was a perfect preferential attachment/cumulative advantage, the induced distribution would be a perfect power law. The process not being a perfect preferential attachment, there is no surprise that the distribution deviates from a power law.

Synthetic data generator

In order to explain the apparent cutoff that we see in some of the empirical distributions, we propose a synthetic publication generator [4] relying on two main ingredients:

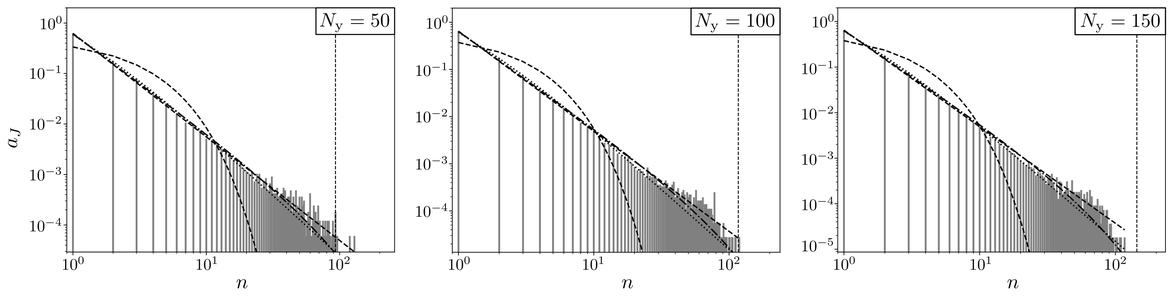
- A preferential attachment process;
- A finite lifespan for authors.



In summary, the algorithm attributes new papers to authors iteratively and randomly. At each step, the probability that the new paper is attributed to an author is:

- proportional to the number of papers this author has;
- linearly decreasing with the ”academic age” of the author.

The resulting distributions are quite similar to the observed empirical distributions, suggesting that the finite lifespan of authors is an important ingredient.



References

[1] R. Delabays and M. Tyloo, *Heavy-tailed distribution of the number of papers within scientific journals*, Quantitative Science Studies **3**, 776 (2022). https://doi.org/10.1162/qss_a_00201

[2] A. Clauset, C. R. Shalizi, and M. E. J. Newman, *Power-law distributions in empirical data*, SIAM Review **51**, 661 (2009). https://doi.org/10.1137/070710111

[3] P. L. Krapivsky, S. Redner, and F. Leyvraz, *Connectivity of growing random networks*, Physical Review Letters **85**, 4629 (2000). https://doi.org/10.1103/PhysRevLett.85.4629

[4] R. Delabays, *ADGenerator: Authors Distribution Generator (v1.0)*, Zenodo. https://zenodo.org/record/6030303