The Risks of Scientific Gerontocracy

Antoine Houssard, Floriana Gargiulo, Gabriele Di Bona, Tommaso Venturini, Paola Tubaro

The Risks of Scientific Gerontocracy

Antoine Houssard[1†], Floriana Gargiulo[2*†], Gabriele Di Bona[2,3], Tommaso Venturini[4], Paola Tubaro[5]

[1]CIS, CNRS, 59 rue Pouchet, Paris, 75017, France.
[2*]GEMASS, CNRS, 59 rue Pouchet, Paris, 75017, France.
[3]Sony Computer Science Laboratories Rome, Joint Initiative CREF-Sony, Centro Ricerche Enrico Fermi, Via Panisperna 89/A, Rome, I-00184, Italy .
[4]Medialab, Université de Geneve, 24 rue du Général-Dufour, Geneve, 1211, Switzerland.
[5]CREST, ENSAE, Institut Polytechnique de Paris, Palaiseau, 91120, France.

*Corresponding author(s). E-mail(s): floriana.gargiulo@cnrs.fr;
Contributing authors: antoine.houssard@cnrs.fr; gabriele.dibona.work@gmail.com; tommaso.venturini@cnrs.fr; paola.tubaro@cnrs.fr;
†These authors contributed equally to this work.

Abstract

While much has been written about the problem of information overload in news and social media, little attention has been paid to its consequence in science. Scientific literature, however, has witnessed decades of exponential growth, to the point that the publications of the last twenty years now constitute 60% of all academic literature. This information overload is not without consequence. Our analysis reveals that, unlike other cultural products, scientific publications face unique challenges: the decreasing proportion of papers capturing large shares of researchers' attention and the slow turnover of influential papers lead to a disproportionate prominence of established works, resulting in stagnation and aging of scientific canons. To determine whether scientific hypergrowth is responsible for such "gerontocratization of science", we propose a generative model of paper citations based on random discovery and cumulative advantage, with a varying number of new papers each year. Our findings show that, as exponential growth intensifies, gerontocratization appears and becomes increasingly pronounced. Recognizing and understanding this mechanism is hence essential for

developing targeted strategies to counteract this trend and promote a balanced and healthy renewal of scientific canons.

Keywords: Science of science, Information overload, Attention cycles

1 Introduction

In 2024, scientific literature continues to grow at an unprecedented rate. More than 60% of all scientific publications have been produced after the year 2000 [1]. This exponential growth of scientific production is marked by an increase in the average individual output, the number of authors per paper, and the total number of people pursuing academic careers [2].

Already in the early 1600s Barnaby Rich [3] observed that:

> *"One of the diseases of this age is the multiplicity of books; they doth so overcharge the world that it is not able to digest the abundance of idle matter that is every day hatched and brought forth into the world".*

This "disease" was addressed through the establishment of scientific societies and journals, but ironically these institutions have then become major factors in the overgroth of academic publications. In the 1960s, Derek De Solla Price highlighted the risk that *science hypergrowth* could overwhelm researchers, making it difficult to keep pace with new developments and to attract attention to their new work [3–5]. Nonetheless, following Malthusian ideas [6], he also predicted that the initial expansion would eventually give way to a saturation or stationary phase. However, the expansion of research and higher education investments, the development of technologies that facilitate the writing and dissemination of papers [7], the rise of a market for academic publishers (as well as of "mega-" and "predatory journals" [8]), and the consolidation of a "publish or perish" ethos [9] proved him wrong.

Although the phenomenon of hypergrowth is well-recognized, its concrete impacts on the structure and dynamics of scientific ecosystems remain underexplored. While extensively discussed by classic authors in the sociology of science [10, 11], this issue has not been fully examined in its current and concrete implications on the structure and dynamics of the scientific publication system.

The hypergrowth dynamics of the scientific ecosystem resembles, to some extent, to the acceleration of information flows that has been observed in other cultural sectors with the advent of digital media [12, 13] and post-modern societies [14]. In media studies, the impact of information overload has been widely discussed in relation to the affordances of social media and the effects of their recommendation algorithms. It has been linked to different phenomena including political polarization, fake news diffusion, and the formation of echo chambers and filter bubbles.

The explosion of user-generated content circulating online, combined with a growing reliance on metrics such as clicks and views [15–17], has been connected to the acceleration of news cycles [12, 13] and to a superficial mode of consumption characterized by a quick turnover of "opinion" leaders [15, 18]. The attention regime typical

of online platforms has been in fact described as dominated by "junk news bubbles", i.e., sudden bursts of interest around pieces of news or matters of discussion that capture large shares of collective attention, but only for a very short time. The object of these attention bubbles tend to be sensational yet ephemeral objects that are entertaining but of poor informational quality (hence the name "junk news"). Castaldo & al. [19] have shown that attention bubbles can be generated by the influence of recommendation algorithms and their preference for trendy contents. It has also been shown that this phenomenon is amplified when the information flow is larger.

While the popularity of online content is measured through views and comments, the popularity of a scientific publication is generally measured by the number of citations it attracts. While this metric is not exempt from criticism [20], citations allow gauging *the intensity* with which arguments, ideas, articles, people, research programs, disciplines, etc. resonate in the scientific community [21].

Research on the evolution of citations has produced diverging results. While some studies suggested a lengthening in citation life cycle [7, 22, 23], others showed constant or accelerating attention cycles [12]; some highlighted an increase in citation inequality [24], others a decrease [25]; some observed the persistence of key concepts [26] even after they stop being directly referenced [27], and others the decreasing disruptiveness of the new publications [28].

None of these studies, however, has specifically investigated the link between literature hypergrowth and attention disorders in the scientific system. To zoom in on this question, we analyze the citation patterns in different disciplines, using different datasets extracted from OpenAlex. We observe a pattern that is distinctively different from that of news and digital media: while we observe an increasing concentration of academic attention on a minority of publications, we also observe that such publications become more and more persistent in the popularity ranking. This low turnover rate of the most popular references suggests a phenomenon of "standardization of scientific canons" and of "aging of the scientific elites".

Hypergrowth, however, is not the only possible culprit for this gerontocratization. Other factors, such as the platformization of science and the action of recommendation algorithms in academic platforms, the advent of predatory journals (and the different trust regime that they creates) and individual behaviors in the research community, may also contribute to the sclerotization of the cumulative advantage process.

To disentangle these concurrent mechanisms, we employ simulations designed to test whether scientific hypergrowth is directly responsible for gerontocratization. Our model, introduced and analyzed in section 3, shows that, excluding all other factors, hypergrowth can lead to stagnation within the scientific ecosystem.

2 Gerontocratization of science

2.1 Rising average impact despite science hypergrowth

Following the results of De Solla Price [3], we first quantitatively analyze the relative growth of the number of papers in time for various disciplines (see section 5 for details

on the data). In Fig. 1A, we plot the normalized number of papers $n_i(t)$ published every year in each discipline i, that is

$$n_i(t) = \frac{N_i(t)}{\sum_t N_i(t)}, \tag{1}$$

where $N_i(t)$ is the number of papers published in discipline i in year t.

As already observed in other papers [1, 4, 5], the normalized number of publications increases in time with an exponential growth regime, i.e., we can fit such growth with the expression

$$n(t) = B + Ae^{\alpha}t, \tag{2}$$

where the exponents and coefficients depend on the specific field of study. The different disciplines we examine exhibit the same structural behavior, with exponents varying between 0.005 for Molecular Physics to 0.14 for Biochemical Engineering.

Averaging across all disciplines, we find that 91% of papers receive their first citation in the first two years of publication, with limited variations between each discipline. Therefore, to measure the attraction of citations of each paper, we focus on the first two years after publication, in order to fairly compare more and less recent papers. In particular, we count the average number of citations, $C_i^{2Y}(t)$, that the papers in discipline i published in year t received in the first two years after publication ($[t, t+2]$). In Fig. 1B, we can observe that, as the year of publication moves forward, papers receive more and more citations within the first two years after publication. Similarly, in Fig. 1C we can see that the portion $f_0^{2Y}(t)$ of "invisible literature", namely the fraction of papers published in year t that do not receive any citation in the first two years of life, decreases linearly in time for all the disciplines.

We could conclude from these first observations that, although the number of papers published each year is exponentially growing, research outcomes have gained over time a larger capacity to circulate and to be recognized, through citations, in the academic ecosystem soon after being published.

2.2 Science is not getting "junk"

Considering the exponential growth of the scientific literature, here we extend our previous analysis to find out if the increased attention received right after being published could be a sign of the presence of "junk science bubbles", namely ephemeral scientific contents characterized by the capacity of attracting a large share of attention within a short time period, with a fast turnover, similarly to what observed for other cultural products (videos, songs, etc.) [12, 19].

We define the attention on a paper x from papers of discipline i at year t as the number of citations $k_x^i(t)$ received by the paper in papers of discipline i published in year t. To remove the bias due to the exponential expansion of the scientific ecosystem, for each paper, we measure the attention share $\xi_x^i(t)$ for each year t as follows:

$$\xi_x^i(t) = k_x^i(t) \Big/ \sum_{y \in i} k_y^i(t). \tag{3}$$

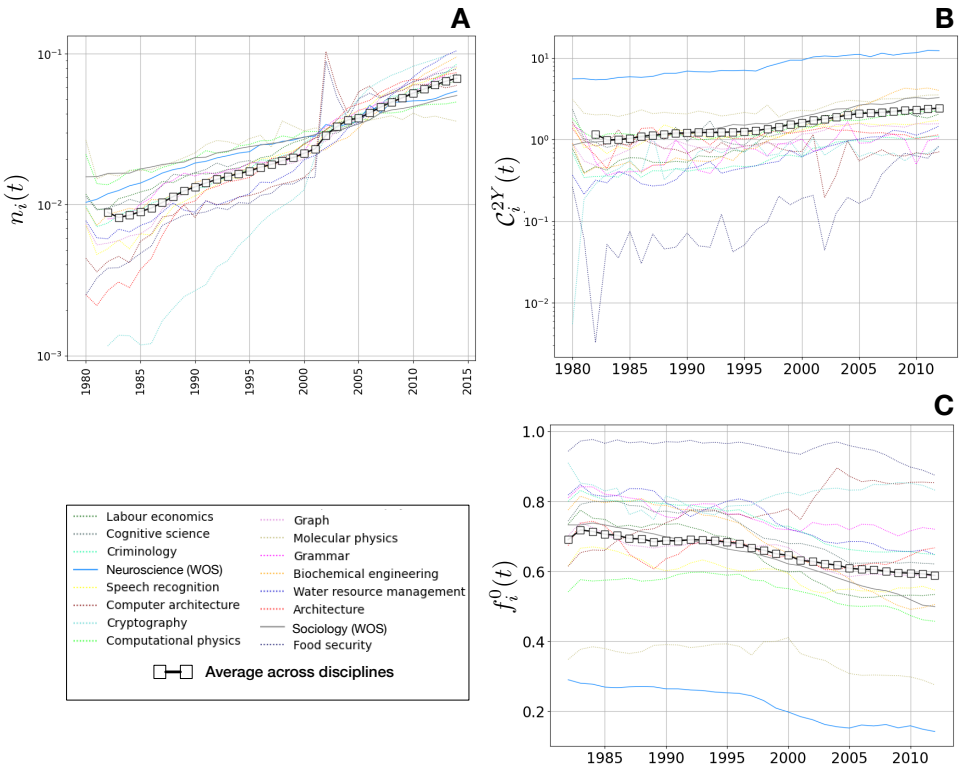

Fig. 1 **(A)** Relative number of publications by year for all the disciplines. **(B)** Average number of citations received by papers appeared in year t, after 2 years from their publication. **(C)** Fraction of papers, appeared in year t, not having any citation after 2 years.
Each discipline is characterized by a different color, while the square points represent the average measures among all datasets.

We begin by examining how attention is distributed among competing contents within each disciplinary arena. Specifically, we measure the evolution of the Gini index of the citation share $\xi_x^i(t)$ for each discipline on a year-by-year basis. Ranging from 0 (perfect equality) to 1 (extreme inequality), the Gini index is a widely used synthetic indicator in sociology and economics that measures the level of inequality for a specific variable within a given population [29]. In the context of our study, it allows us to quantify the inequality in the distribution of citation shares among publications within each discipline. In Fig. 2A we show that the Gini index ($Gini(\xi)$) increases for all the disciplines, indicating a more and more unequal distribution of the citation shares, similarly to what previously observed in different case studies [30]. This suggests that, much like other forms of cultural contents, scientific literature exhibits an increasingly unequal structure, with fewer papers able to monopolize the attention space.

To test the junkisation hypothesis in science, we also need to analyze the attention cycles, namely the shape of the attention share patterns related to each paper. To compare papers appearing in different years, we consider the share curves limited to

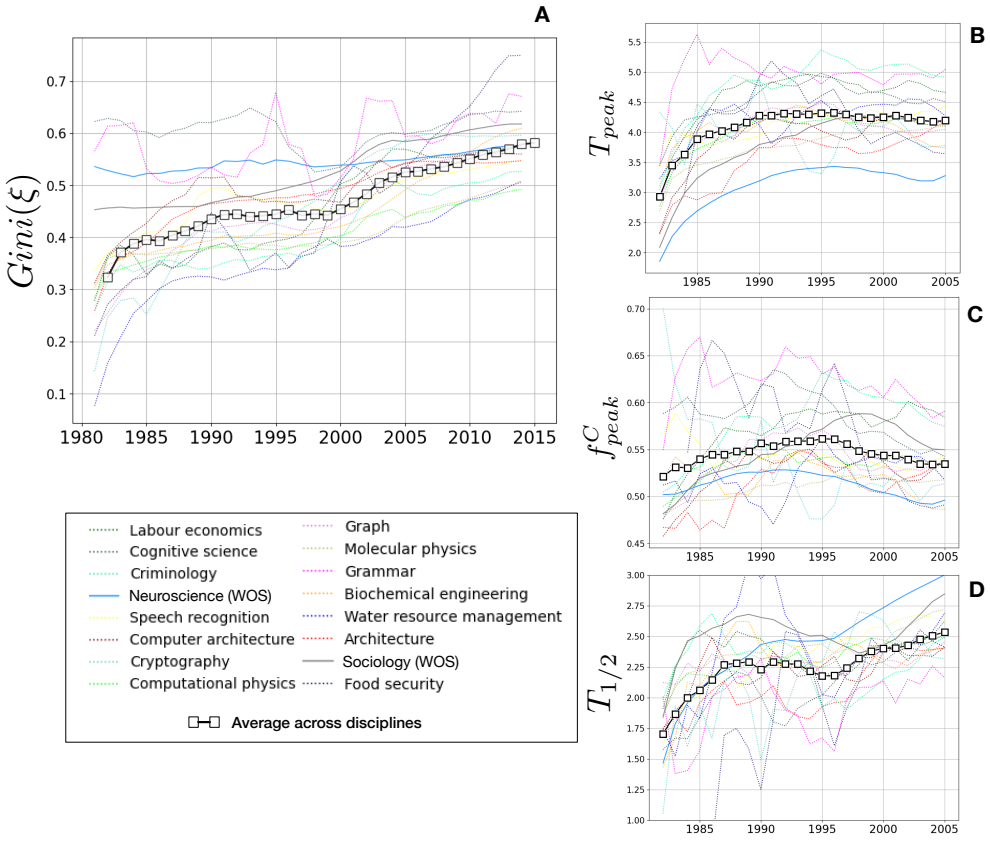

Fig. 2 **(A)** Gini index as a function of time. Each discipline is characterized by a different color, while the square points represent the average measures among all datasets. **(B)** Evolution of the time T_{peak} from the publication year to reach the peak with maximum share, averaged across papers published each year in each discipline. **(C)** Evolution of the cumulative fraction of share at the peak f^C_{peak}. **(D)** Evolution of the half-life time $T_{1/2}$ from the peak.

the first 10 years following publication. We normalize each share curve by dividing each point by the area under the curve, obtaining the attention share fraction in each year $\tilde{\xi}^i_x(t)$:

$$\tilde{\xi}^i_x(t) = \xi^i_x(t) \Big/ \sum_{t'<10} \xi^i_x(t'). \tag{4}$$

We characterize these patterns through three different measures [31]: (1) the time T_{peak} from the publication year to reach the peak with maximum share, averaged across papers published each year in each discipline; (2) the cumulative fraction of share at the peak, defined as $f^C_{peak} = \sum_{t \le T_{peak}} \tilde{\xi}^i_x(t)$; and (3) the half-life time $T_{1/2}$ from the peak, namely the time after the peak needed to reach the last point where the curve exhibits a value equal to the half of the maximum.

To reduce statistical noise in this part of the study, we only consider papers which have received more than ten citations in total. Moreover, to avoid getting incomplete cycles, we only focus on papers published before 2006.

We observe in Fig. 2B that the average time to reach the popularity peak, for all disciplines, increased by 1,5 years from the '80s to the 90's, stabilizing afterwards to a specific value typical of each discipline. The frequency at the peak, instead, does not show a significant variation in time (Fig. 2C). Finally, measuring the attention decay by the half-life time from the peak (Fig. 2D), we observe that the average half-life of papers increases with time in all disciplines. The combination of these measures indicates that the citation cycles of papers exhibit a more durable persistence in the scientific ecosystem.

In summary, while we observed an increase in inequality patterns, we also noted a lengthening of the attention period. This increased attention appears especially true for the most important papers and denotes with the "junkisation" phenomenon observed in online content.

2.3 Canons get more stable and "elites" get older

The higher concentration of citations on a small set of papers, together with the longer persistence of attention on papers observed in the previous section, has a consequence on the structuring of the pillars of the literature. In fact, we find that the ranking of such pillars becomes more and more stable. We calculate the similarity between the papers in the top-k list (the papers occupying the first k positions of the ranking) between two subsequent years, using the "ranked Jaccard", a measure to compare rankings previously introduced in Ref. [32]: From each top-k list we build a new list, that we name expanded-top-k, where each element, at position r of the top-k, is identically repeated $k - r$ times. Then the Jaccard similarity is calculated between the expanded lists. This measure allows to keep in consideration both the position in the ranking (differently from the traditional Jaccard similarity) and the new entries and the exits in the top-k list (differently from the traditional methods to compare the top-k lists, like the Kendall-Tau index).

In Fig. 3A, we show the ranked Jaccard between the top-k lists at two subsequent years for all disciplines (and for $k = 50$). Similar results are obtained for different values of k. Notice the significant increase of the ranked Jaccard for all the disciplines, meaning that the top-50 cited papers become more and more similar among them over time. The same results can be observed, in Fig. 3B, if we consider the top-50 papers according to the page-rank measure in the temporal citation networks, namely the directed graph whose nodes are the papers published in year t together with the papers, published in different years, citing and cited by the year t papers. The edges are the citations among the nodes-papers.

These results show that the scientific pillars stabilize even though there is an exponential growth of literature, giving origin to a lack of renewal of the canons.

To better analyze the concentration phenomenon of citations in a restricted number of papers, we analyze the structure of the "Elite" papers, $\mathcal{E}(t)$, defined as the set of papers that together get the 80% of the total number citations, given at year t.

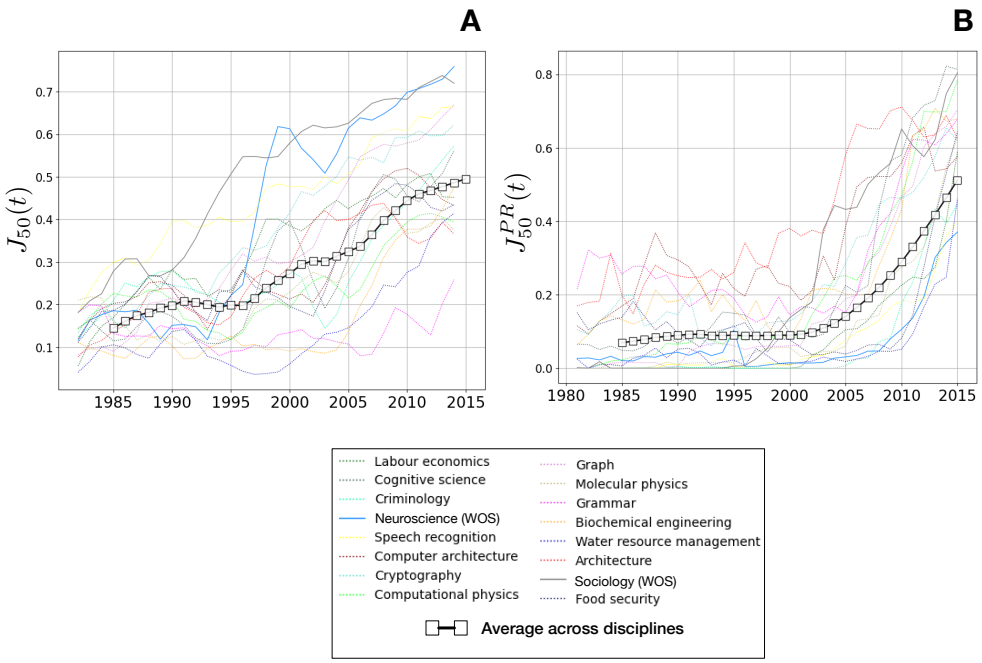

Fig. 3 **(A)** Ranked Jaccard similarity between the top 50 papers in the citation ranking at year Y and the ones at $Y-1$. **(B)** Similarity of the top 50 papers according to Page Rank centrality in the citation network at Y and those at $Y-1$s.

While different between disciplines, the size of the elite,$|\mathcal{E}(t)|$, is small and, above all, it displays a clear linear decrease in time (Fig. 4A).

We define the age of the elite at a certain year t as the sum of the age of the papers belonging to this set (being the age of a paper the number of years from its publication). For all the disciplines, the age of the elite (Fig. 4B) increased of almost 5 years during the observation period (with the exception of neuroscience that displays an initial noisy decrease until the year 2000, before starting to increase).

We also analyzed the temporal co-citation networks, namely the networks where the nodes are the papers cited at year t and the nodes are linked if they appear together in the reference list of a paper published at year t. Inspired by rich club measures in networks [33] we finally investigate if there is an emerging tendency to cite together papers that are part of the elite or if the elite papers are disconnected among them, each having its own influence on an isolate part of the scientific literature (i.e. scientific sub-communities are disconnected and each of them has his own important reference). We compared the density of the sub–graph containing only the elite nodes with the density of the whole graph. We observe that the relationships between elite papers increase in time, creating a more and more marked core-periphery structure, where the elite papers are not only over-cited but also over-cited together (Fig. 4C).

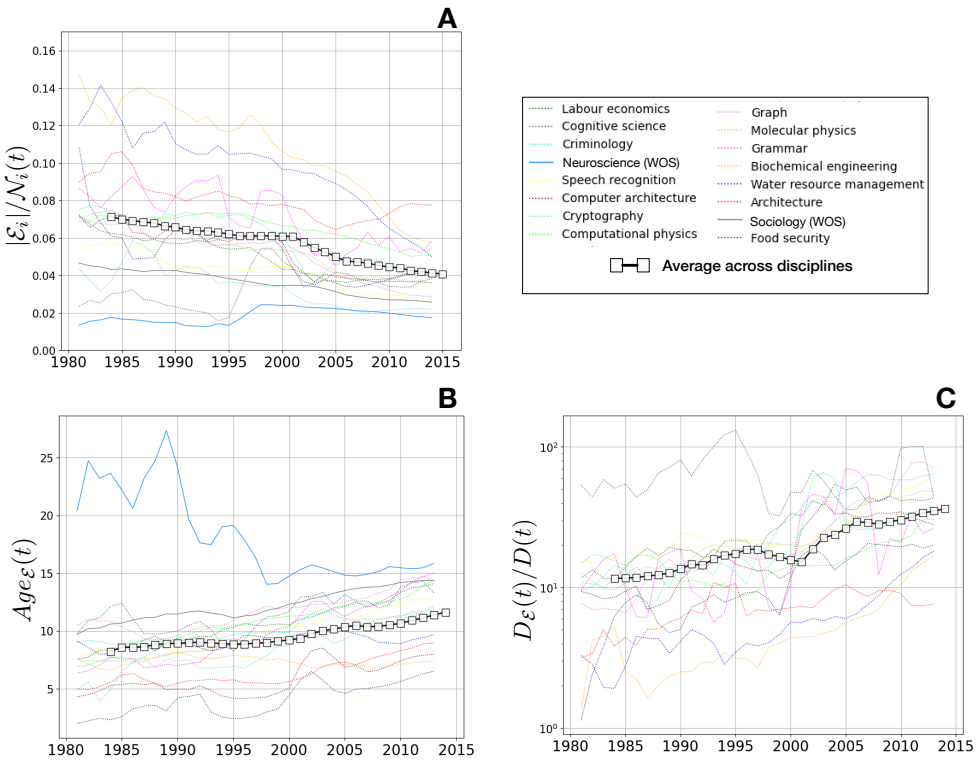

Fig. 4 **(A)** Elite relative size in time. **(B)** Average age of elite. **(C)** Co-citation density of the elite (respect to the whole co-citation network).

To summarize, the peculiar attention regime observed in the scientific ecosystem, characterized by the increase of the inequality of attention shares and by longer attention cycles on papers, is associated to a general consolidation of the canonical references and the emergence of an elite structure whose relative size is smaller and smaller and whose average age increases in time. Moreover this elite becomes more and more interconnected respect to the rest of the literature, meaning that elites become more and more responsible of the bridging process of the knowledge space. Based on these observations we can speak of a "gerontocracratic" re-organization of science.

2.4 Where does gerontocracy come from?

Is this gerontocratic re-oganization triggered by some exogenous factors (like the introduction of recommendation algorithms), is it due to some change of the scientific practices (publish or perish, proliferation of journals, etc.) or is it implicit in the hypergrowth regime of the scientific ecosystem?

A first hint in the direction of the last hypothesis comes from the Heaps' law of the citations (Fig. 5A). Heaps' laws [34] were originally introduced in lingustics to analyze the growing patterns of the vocabularies. According to the Heaps' law, the

number of different words (the dictionary, D) that we meet reading a text grows as a power-law function of the total number of words (N) used up to that moment: $D \sim N^{\nu}$. In our case, considering the time-ordered aggregation of papers' reference lists as a large "text" where the references are the words, we count the total number of different cited papers, N_{cited}, with respect to the total number of citations $N_{citations}$. This representation shows that all the disciplines strictly follow the same Heaps' law $N_{cited} \sim N_{citations}^{0.85}$, with a sub–linear exponent $\nu = 0.85$. In other words, as the system grows, less new papers are cited. The universality and the stationarity of this law suggest that the "aging" patterns, favoring the persistence of the attention share on few old contents, that we observed in the previous paragraph could be the direct result of a universal microscopic dynamics underlying the morphogenesis of citation networks.

In Fig. 5B, we show that the variation of the Gini index for each discipline, defined as $\Delta Gini = (Gini(2015) - Gini(1980))/Gini(1980)$, is positively correlated to the growing rate of the system, i.e., the exponent α of the growing curve for the fraction of papers in 2. This suggests that the process driving science to the increase of citations' concentration into a few "elite" papers could be connected to the exponential growing dynamics of the literature.

However other factors could also be identified as potential drivers, factors that, in turn are not related to the numerical composition of the ecosystem but rather to changes in the individual citation practices of the researchers. To dig in this direction, we studied, for example, the authors' propensity to cite papers that have not yet been cited, namely to discover new literature. In Fig. 5C we observe that the fraction of uncited papers in the reference list of the papers decreases over time, underlying a minor propensity of authors to cite literature that was not cited before. To perform this measure we had to had to drastically filter our dataset: in order to have the full citation history of each papers we limited our observation to the part of complete data in our datasets, namely the citations internal to the disciplines. We therefore considered only the part of the citation lists relative to papers that are also in the discipline (not in the extended dataset). The results of this measure did not change significantly from one discipline to the other, therefore the results shown in Fig.5C represent the aggregation of all disciplines. This last evidence is related to a change in the individual choices of the authors, reinforcing more and more the cumulative advantage mechanism. This phenomenon, could be also related to algorithms of the search engines of scientific literature (like Google Scholar), a potential effect will deserve a deeper analysis in future studies.

3 Hypergrowth drives the gerontocratization phenomenon

3.1 The model

In the previous paragraph we showed some hints suggesting that the hypergrowth phenomenon could be directly responsible of the emergence of the "gerontocratic" patterns observed in the scientific ecosystem, these lasts resulting from a microscopic

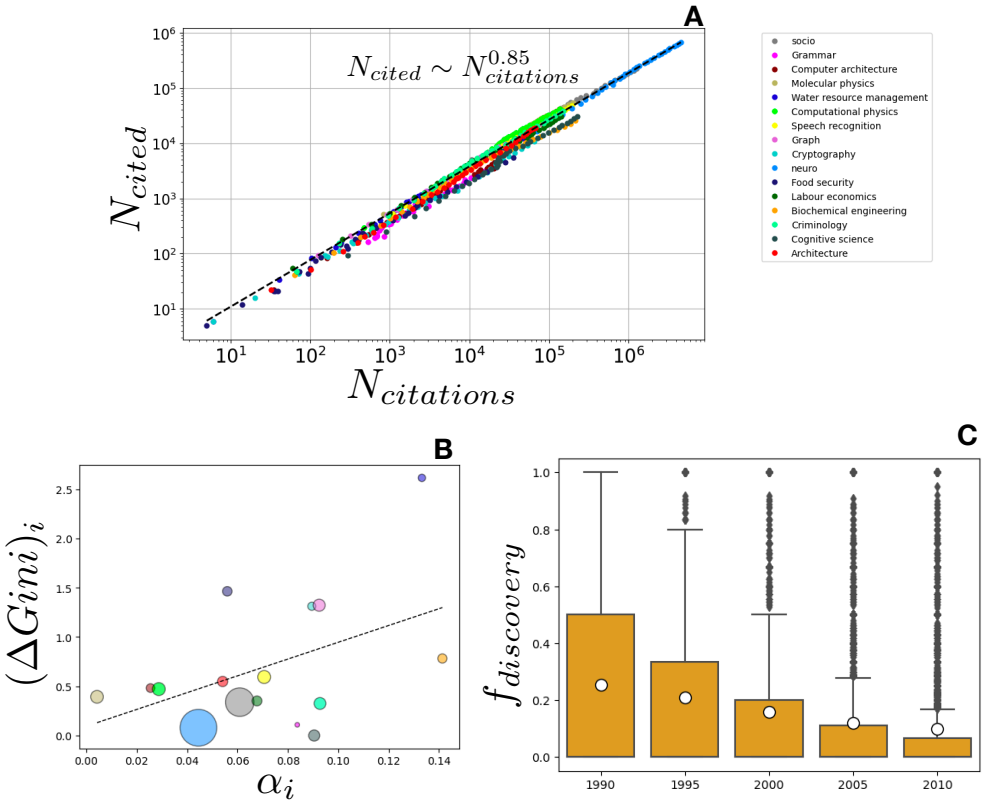

Fig. 5 **(A)** Heaps' law for the different disciplines. Number of different cited papers -vs- total number of citations. **(B)** Variation of the Gini index as a function of the exponent of the growing tate of the corpora' size. Each point is a corpus. The size of the points is proportional to the total size of the corpus. **(C)** Fraction papers with zero citations in papers' reference lists.

process where the basic morphogenetic rules do not change in time (at least not drastically) but where the system size expands exponentially.

We also identified other potential mechanisms that could determine the gerontocratic transition, like, for example, the decrease of the novelty discovery in papers bibliographies (Fig. 5C). In order to test the hypothesis of the direct emergence of gerontocratic patterns as a consequence of hypergrowth, we have to disentangle these different effects and, to do this, we build a simple generative model of the citations patterns where the only relevant variable is the slope of the litterature growing curve.

The model is inspired by Polya's urn model and Polya's urn model with triggering [34, 35], a modelling approach largely used for describing the dynamics of discovery processes.

The model simulates the selection process of references for new articles entering in the system. The number of new papers entering the system follows the exponential expansion rule $\mathcal{N}(t) = \mathcal{N}_0 + e^{\alpha t}$. The parameter α is therefore tuning the expansion rate of the system (with $\alpha = 0$ corresponding to the linear growth case).

The basic step of the model is explained in Fig. 6. At each time step, t, corresponding to a year, $\mathcal{N}(t)$ new papers enter into the system and each of them selects its references. To select the N_{ref} papers for its bibliography, each paper makes a choice of following the cumulative advantage mechanism, exploring the already cited papers, with probability p, or to discover uncited papers with probability $1 - p$.

Already cited papers are extracted from a first urn $\mathcal{S}$, where each paper is repeated as many times as the number of its previous citations (to reproduce the cumulative advantage process). Uncited papers are extracted with uniform probability from a second urn $\mathcal{U}$, containing all uncited published papers. According to the fact that almost 90% of the papers receive their first citation in the first two years from publication, we remove from $\mathcal{U}$ all papers with an age larger than 2 time steps.

The new entering papers, at time t, synchronously build their bibliographies on the base of the two urns $\mathcal{S}(t)$,$\mathcal{U}(t)$. After all the new papers have selected their bibliography, $\mathcal{C}_i$, the sum of all the papers' references constitutes the list of the year's citations $\mathcal{C}(t)$. The papers in $\mathcal{C}(t)$ are added, with all their repetitions, to the urn $\mathcal{S}$ and removed from $\mathcal{U}$ if previously uncited. The $\mathcal{N}(t)$ new papers enter in the $\mathcal{U}$ urn and the papers older than 2 years are removed from $\mathcal{U}$.

The parameter N_{ref} is fixed to 10 in the simulations, but this value does not impact the outcomes of the model. At the beginning of a simulation, the urn $\mathcal{S}$ is formed by a set of 200 papers (that is in the order of magnitude of the number of cited papers in 1980 in the data) with an initial number of citations uniformly varying between 1 and 3. The urn $\mathcal{U}$ contains 50% of the papers in the initial urn $\mathcal{S}$, because on average 50% of the papers is not cited. The ages of the papers are randomly assigned in a range between 1 and 2. The number of new papers at the first iteration is given by the sum of the sizes of the two urns divided by 2. The parameter p tunes the relative importance of the cumulative advantage with respect to the discovery process. We know from 5C that this parameter is slightly changing over time, indicating a higher preference toward cumulative advantage. However, to avoid confusing the effect caused by the variation of this parameter with the effect of the system growth, we fixed the value at 0.8 as to consider the "best scenario" observed in the data.(5C). However, we could consider the variations of this parameter in a second moment if interested to study the impact of the decreasing of the discovery process on the evolution. The process is iterated until the total number of papers reaches a final value of 50000. For each value of the parameter α we perform 100 replicas of the simulation.

3.2 Model's results

The model allows to analyze the direct effect of the system growth, namely the growth parameter α, on the evolution of the Gini index, of the ranked Jaccard and on the structure of the resulting Heaps' laws.

As we can see from Fig. 7A and Fig. 7B , the average value (on the 100 replicas) of the Gini and of the ranked Jaccard, have a direct dependency from the growth

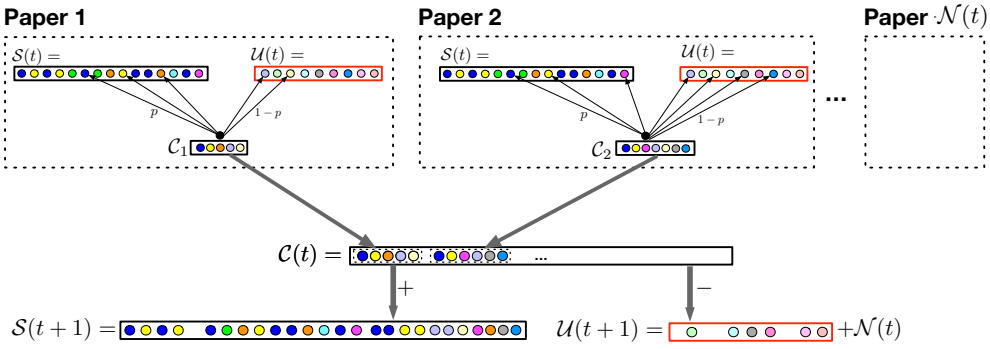

Fig. 6 Schematic representation of an iteration of the model.

mechanism. At the same population size, the value of the concentration is much lower if reached with a slow process (linear if $\alpha = 0$) respect to an exponential regime ($\alpha > 0$). Similar considerations can be done for the consolidation of canons, namely the ranked Jaccard.

In Fig 7C we plot the heaps laws for different values of the growth parameter. We see that the value of the exponent of the Heaps, ν, decreases for larger values of the growth parameter (Fig 7D) and starting from a certain value of α we observe the transition from a superlinear to a sublinear relationship between the number of different citations and the total number of citations.

In its simplicity this model explains that the presence of a too large number of new papers is compromising the equilibrium between the preferential attachment and the discovery process: when too many new papers enter in the urn $\mathcal{U}$, during their random discovery phase, these lasts are selected too few times to enter in the urn $\mathcal{S}$ with the capacity to compete with the pre-existing stars. This mechanism increases the first-mover advantage, and naturally suppresses the possibility for the newcomers to enter in competition.

4 Conclusions

Not unlike most other cultural sectors, science has been facing the problem of information overload.

In social media this overload is responsible of several communication troubles, among which the emergence of "junk news bubbles", namely ephemeral pieces of content that attract a large share of attention but are quickly forgotten and replaced by newer ones.

In scientific literature we do observe such acceleration of attention cycles. While a smaller and smaller portion of papers monopolizes most citations in each disciplinary arena, this does not produce the quick turnover observed in social media. Quite the contrary, popular scientific publications remains so for longer and longer periods of time, creating a growing "scientific gerontocracy", where smaller and older elites hold on to the top positions and scientific canons stagnate.

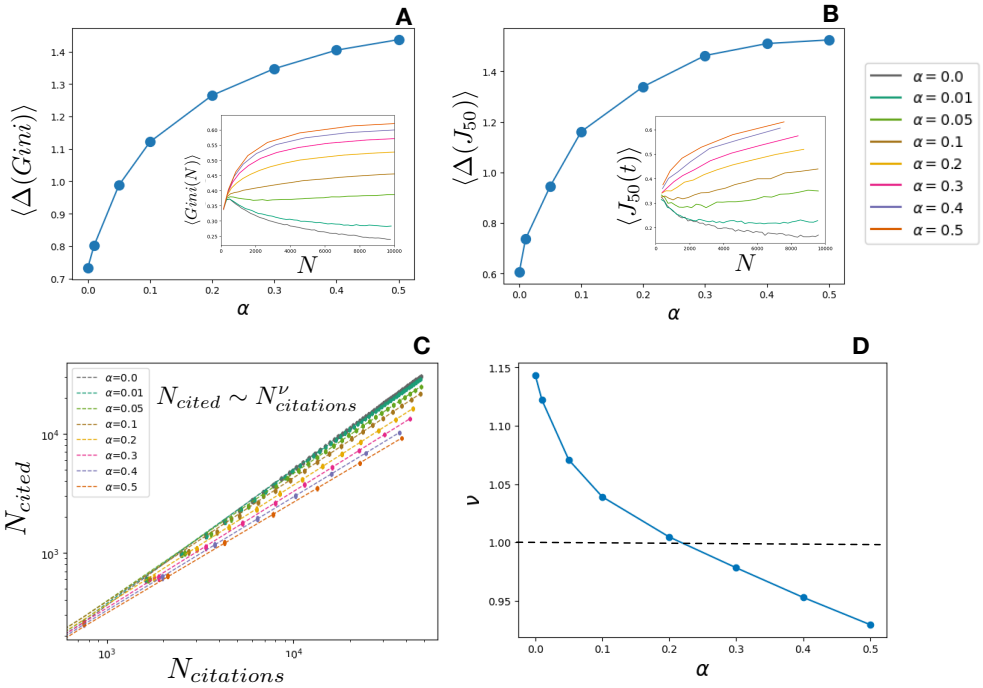

Fig. 7 **(A)** Variation of the Gini index when the population grows from 500 to 5000, as a function of the growth parameter α. In the inbox we plot the growth of the Gini index with respect to the population size. **(B)** Variation of the ranked Jaccard index when the population grows from 500 to 5000, respect to the population growth parameter. Inbox: growth of the ranked Jaccard index respect to the population size. **(C)** Heaps' laws for different values of α. Plot D: Exponent of the heaps law as a function of the parameter α. The results are obtained averaging on 100 replicas of the model for each value of the parameter.

This trend is surely due to many factors, including the transformations of the scientific and editorial practices, the platformization of science, the changing ways in which researchers explore the existing literature as well as the simple fact that, because of the exponantial its growth, disciplinary literatures are more and more difficult to consider in their entirety.

To separate the effects of the structural changes in scientific practices from the effects of the information overload, we proposed a simple model allowing to analyze different growth scenarios (while keeping scientific practices are costant).

The model shows that a "gerontocratic" regime emerges as a direct consequence of the exponential expansion of the system, because of the low probability of selection of innovative contents.

The model was formulated to disentangle the effects of hyper-growth from the other processes that could drive the same dynamics: the variation of the preference toward preferential attachment, an eventual lowering of the inner quality of papers due

to the publish or perish diktat, the platformization of science with the introduction of recommendation systems, etc. Overall, this paper highlights how science hypergrowth may have triggered a generalized gerontocratization of academic literature as a direct consequence of the information overload in the scientific system. Understanding of this mechanism, we hope, can help define strategies to reverse or at least mitigate this gerontocratization process and restore a healthy renewal of scientific canons.

5 Data

5.1 Data Sources

In this study, we mobilize the OpenAlex (OA) database; this choice was motivated by the availability of data and the extensive scope of the base. OA offers an extensive, unfiltered database, enabling us to include the entire corpus available to scientists when conducting research, including low-value or "junk" science [36] [37] typically excluded by providers such as Web of Science or Scopus. Additionally, OA provides a classification system, the "concept" , that allows fine-grained collection focused on specific disciplines sub-disciplines, methods or theoretical backgrounds [1]. This hierarchical classification, based on the documents abstracts, allowed to identify sub-disciplines (level 2) and extract corpus associated to 14 of them.

Finally, to validate the measurements obtained from the OA concepts, we utilized two corpora based on the Web Of Science (WOS) classification, previously compiled in the context of other projects [38].

Concept	Size	Size extended dataset
Neuroscience **(WOS)**	1,578,556	11,888,972
Sociology **(WOS)**	949,605	6,015,791
Cognitive science	148,027	2,661,581
Computer architecture	85,470	507,037
Food security	113,283	609,721
Labor economics	112,075	1,234,409
Molecular physics	160,524	2,318,947
Cryptography	80,314	318,990
Architecture	126,215	1,516,638
Biochemical engineering	110,452	3,695,296
Speech recognition	207,264	1,632,350
Water resource management	75,050	818,540
Computational physics	177,372	1,939,356
Criminology	178,532	1,533,333
Graph	177,914	1,666,413
Grammar	24,367	301,240

Table 1 Dataset size and associated concept or discipline.

5.2 Data Collection

For our analysis we first identified a set of concepts representing a large panel of disciplines having different publishing practices, we then made a first query to know the number of work represented for each of the identified concepts and randomly selected a set of concept with constrained number of associated work ($\approx$ 100 000) to ensure that the concept represent a bounded sub-field.

To reconstruct the disciplines based on OpenAlex concepts we performed the following steps:

1. We retrieve all the papers that contain the selected concept (representing the papers of the discipline)
2. To get the whole citation history of the paper and its relationship with external disciplines, we also retrieve the information for all the papers that are cited and that are citing the papers retrieved during the first step (extended dataset).

To reconstruct the disciplinary corpora based on journals we performed the following steps:

1. We identify all the journals that in the WoS are identified in the selected disciplines and we build the correspondence between these journals and their OpenAlex identifier.
2. We retrieve, from OpenAlex, all the papers that appeared in the selected journals (papers of the discipline)
3. We retrieve the information for all the papers that are cited and that are citing the papers retrieved during the first step (extended dataset).

All collections were conducted from September 2023 to January 2024 and represent work published from 1980 to 2019. Additionally, we wish to mention that an initial filtering was made using the OpenAlex API in order to only consider published article (e.g including a DOI) and excluding retracted paper and document including no references.

In the following, when we refer to a discipline, we strictly indicate the papers that contain the disciplinary concept (or that appear in the disciplinary journals). On the contrary, the citations that a paper receives are counted on the extended datasets, including the papers that do not contain the selected concept (or journals).

Acknowledgements. The author(s) acknowledge(s) the support of the French Agence Nationale de la Recherche (ANR), under grant ANR-XXXXXXX (project ScientIA). The authors would like to thank A. Maddi for the useful discussions.

References

[1] Priem, J., Piwowar, H., Orr, R.: OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts (2022). https://arxiv.org/abs/2205.01833

[2] Ioannidis, J.P.A., Klavans, R., Boyack, K.W.: Thousands of scientists publish a paper every five days. Nature **561**(7722), 167–169 (2018) https://doi.org/10.1038/d41586-018-06185-8

[3] Solla Price, D.J.: Little Science, Big Science, (1963)

[4] Bornmann, L., Haunschild, R., Mutz, R.: Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. Humanities and Social Sciences Communications **8**(1) (2021) https://doi.org/10.1057/s41599-021-00903-w

[5] Larivière, V., Archambault, É., Gingras, Y.: Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900–2004). Journal of the American Society for Information Science and Technology **59**(2), 288–296 (2007) https://doi.org/10.1002/asi.20744

[6] Malthus, T.R.: An essay on the principle of population (1798). The Works of Thomas Robert Malthus, London, Pickering & Chatto Publishers **1**, 1–139 (1986)

[7] Verstak, A., Acharya, A., Suzuki, H., Henderson, S., Iakhiaev, M., Lin, C.C.Y., Shetty, N.: On the Shoulders of Giants: The Growing Impact of Older Articles. arXiv (2014). https://doi.org/10.48550/ARXIV.1411.0275 . https://arxiv.org/abs/1411.0275

[8] Spezi, V., Wakeling, S., Pinfield, S., Creaser, C., Fry, J., Willett, P.: Open-access mega-journals: The future of scholarly communication or academic dumping ground? a review. Journal of Documentation **73**(2), 263–283 (2017) https://doi.org/10.1108/jd-06-2016-0082

[9] Sarewitz, D.: The pressure to publish pushes down quality. Nature News **533**(7602), 147 (2016) https://doi.org/10.1038/533147a

[10] Merton, R.K.: The matthew effect in science. Science **159**(3810), 56–63 (1968) https://doi.org/10.1126/science.159.3810.56

[11] Collins, R.: Conflict Sociology. Academic Press, San Diego, CA (1977)

[12] Lorenz-Spreen, P., Mønsted, B.M., Hövel, P., Lehmann, S.: Accelerating dynamics of collective attention. Nature Communications **10**(1) (2019) https://doi.org/10.1038/s41467-019-09311-w

[13] Candia, C., Jara-Figueroa, C., Rodriguez-Sickert, C., Barabási, A.-L., Hidalgo, C.A.: The universal decay of collective memory and attention. Nature Human Behaviour **3**(1), 82–91 (2018) https://doi.org/10.1038/s41562-018-0474-5

[14] Rosa, H.: Social Acceleration: A New Theory of Modernity. Editions La Découverte, Paris (2010)

[15] Bergström, A., Jervelycke Belfrage, M.: News in social media: Incidental consumption and the role of opinion leaders. Digital Journalism **6**(5), 583–598 (2018) https://doi.org/10.1080/21670811.2018.1423625

[16] Costera Meijer, I., Groot Kormelink, T.: Checking, sharing, clicking and linking: Changing patterns of news use between 2004 and 2014. Digital journalism **3**(5), 664–679 (2015)

[17] Kormelink, T.G., Meijer, I.C.: What clicks actually mean: Exploring digital news user practices. Journalism **19**(5), 668–683 (2018)

[18] Karlsen, R.: Followers are opinion leaders: The role of people in the flow of political communication on and beyond social networking sites. European Journal of Communication **30**(3), 301–318 (2015) https://doi.org/10.1177/0267323115577305

[19] Castaldo, M., Venturini, T., Frasca, P., Gargiulo, F.: Junk news bubbles modelling the rise and fall of attention in online arenas. New Media & Society **24**(9), 2027–2045 (2022)

[20] Bornmann, L., Daniel, H.: What do citation counts measure? a review of studies on citing behavior. Journal of Documentation **64**(1), 45–80 (2008) https://doi.org/10.1108/00220410810844150

[21] Klamer, A., Dalen, H.P.: Attention and the art of scientific publishing. Journal of Economic Methodology **9**(3), 289–315 (2002) https://doi.org/10.1080/1350178022000015104

[22] Bouabid, H., Larivière, V.: The lengthening of papers' life expectancy: a diachronous analysis. Scientometrics **97**(3), 695–717 (2013) https://doi.org/10.1007/s11192-013-0995-7

[23] Wallace, M.L., Larivière, V., Gingras, Y.: A small world of citations? the influence of collaboration networks on citation practices. PLoS ONE **7**(3), 33339 (2012) https://doi.org/10.1371/journal.pone.0033339

[24] Barabási, A.-L., Song, C., Wang, D.: Handful of papers dominates citation. Nature **491**(7422), 40–40 (2012) https://doi.org/10.1038/491040a

[25] Pan, R.K., Petersen, A.M., Pammolli, F., Fortunato, S.: The memory of science: Inflation, myopia, and the knowledge network. Journal of Informetrics **12**(3), 656–678 (2018) https://doi.org/10.1016/j.joi.2018.06.005

[26] Chu, J.S., Evans, J.A.: Slowed canonical progress in large fields of science. Proceedings of the National Academy of Sciences **118**(41), 2021636118 (2021)

[27] Meng, X., Varol, O., Barabási, A.-L.: Hidden citations obscure true impact in science. arXiv preprint arXiv:2310.16181 (2023)

[28] Petersen, A.M., Arroyave, F., Pammolli, F.: The disruption index suffers from citation inflation and is confounded by shifts in scholarly citation practice (2024). https://arxiv.org/abs/2406.15311

[29] Gini, C.: On the measure of concentration with special reference to income and statistics, colorado college publication. General series **208**(1) (1936)

[30] Kozlowski, D., Andersen, J.P., Larivière, V.: The decrease in uncited articles and its effect on the concentration of citations. Journal of the Association for Information Science and Technology **75**(2), 188–197 (2024)

[31] Parolo, P.D.B., Pan, R.K., Ghosh, R., Huberman, B.A., Kaski, K., Fortunato, S.: Attention decay in science. Journal of Informetrics **9**(4), 734–745 (2015)

[32] Gargiulo, F., Caen, A., Lambiotte, R., Carletti, T.: The classical origin of modern mathematics. EPJ Data Science **5**(1), 26 (2016)

[33] Colizza, V., Flammini, A., Serrano, M.A., Vespignani, A.: Detecting rich-club ordering in complex networks. Nature physics **2**(2), 110–115 (2006)

[34] Tria, F., Loreto, V., Servedio, V.D.: Zipf's, heaps' and taylor's laws are determined by the expansion into the adjacent possible. Entropy **20**(10), 752 (2018)

[35] Tria, F., Loreto, V., Servedio, V.D.P., Strogatz, S.H.: The dynamics of correlated novelties. Scientific reports **4**(1), 5890 (2014)

[36] Venturini, T.: From fake to junk news: The data politics of online virality. In: Ruppert, E., Isin, E., Bigo, D. (eds.) Data Politics: Worlds, Subjects, Rights, pp. 123–144. Routledge, London and New York (2019)

[37] Lattier, D.: Why professors are writing crap that nobody reads (2016)

[38] Fontaine, S., Gargiulo, F., Dubois, M., Tubaro, P.: Epistemic integration and social segregation of ai in neuroscience. arXiv preprint arXiv:2310.01046 (2023)