Journal of Informetrics 12 (2018) 656–678

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Regular article

The memory of science: Inflation, myopia, and the knowledge network

Raj K. Pan[a,1], Alexander M. Petersen[b,c,*,1], Fabio Pammolli[d,e], Santo Fortunato[f,g,**]

[a] Department of Computer Science, Aalto University School of Science, P.O. Box 15400, FI-00076, Finland
[b] Ernest and Julio Gallo Management Program, University of California, Merced, CA 95343, United States
[c] Department of Management of Complex Systems, School of Engineering, University of California, Merced, CA 95343, United States
[d] Department of Management, Economics, and Industrial Engineering, Politecnico di Milano, Milan 20156, Italy
[e] CADS, Center for Analysis, Decisions, and Society, Human Technopole, Milan 20157, Italy
[f] Center for Complex Networks and Systems Research, School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, USA
[g] Indiana University Network Science Institute (IUNI), Indiana University, Bloomington, IN, USA

ARTICLE INFO

Article history:
Received 12 September 2017
Received in revised form 12 June 2018
Accepted 12 June 2018
Available online 22 June 2018

Keywords:
Citation network
Reference distance
Models of science
Attention economy
Monte Carlo simulation
Citation inflation

ABSTRACT

Scientific production is steadily growing, exhibiting 4% annual growth in publications and 1.8% annual growth in the number of references per publication, together producing a 12-year doubling period in the total supply of references, i.e. links in the science citation network. This growth has far-reaching implications for how academic knowledge is connected, accessed and evaluated. Against this background, we analyzed a citation network comprised of 837 million references produced by 32.6 million publications over the period 1965–2012, allowing for a detailed analysis of the 'attention economy' in science. Our results show how growth relates to 'citation inflation', increased connectivity in the citation network resulting from decreased levels of uncitedness, and a narrowing range of attention – as both very classic and very recent literature are being cited increasingly less. The decreasing attention to recent literature published within the last 6 years suggests that science has become stifled by a publication deluge destabilizing the balance between production and consumption. To better understand these patterns together, we developed a generative model of the citation network, featuring exponential growth, the redirection of scientific attention via publications' reference lists, and the crowding out of old literature by the new. We validate our model against several empirical benchmarks, and then use perturbation analysis to measure the impact of shifts in citing behavior on the synthetic system's properties, thereby providing insights into the functionality of the science citation network as an infrastructure supporting the memory of science.

* Corresponding author at: Ernest and Julio Gallo Management Program, University of California, Merced, CA 95343, United States.
** Corresponding author at: Center for Complex Networks and Systems Research, School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, USA.
E-mail addresses: petersen.xander@gmail.com (A.M. Petersen), santo.fortunato@gmail.com (S. Fortunato).
[1] These authors contributed equally.

https://doi.org/10.1016/j.joi.2018.06.005

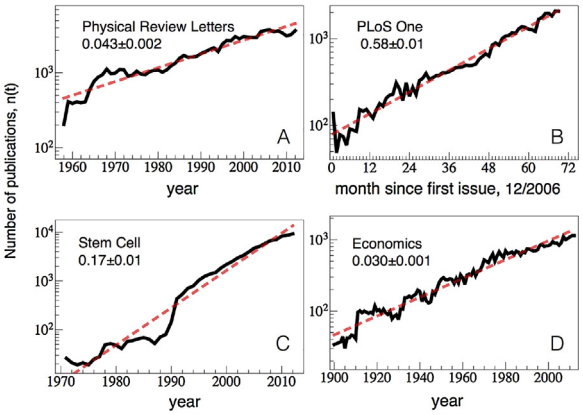

Fig. 1. Empirical growth of scientific output: journals, fields (A, B) Growth in two specific journals. Physical Review Letters (physics), and PLOS ONE (multidisciplinary open access). The remarkable growth rate for PLOS ONE – representative of the shift in scientific publication towards large online-only journals – dwarfs the others by an entire order of magnitude, with the exponential growth rate $g_n = 0.58$ corresponding to an annual growth rate $100(\mathrm{Exp}[0.58]-1) = 77.6\%$ and a publication doubling rate of $\ln 2/0.58 = 1.2$ years. (C, D) Growth in two specific research domains, Stem Cell research and Economics.

1. Introduction

Driven by public and private sector investment into people and projects (Stephan, 2012; Stokes, 1997), the rate of scientific production has exhibited persistent growth over the last century (Althouse, West, Bergstrom, & Bergstrom, 2009; Lariviere, Archambault, & Gingras, 2008). However, the extant empirical and theoretical literature provides little guidance for understanding the impact of these long-term growth trends on the structural properties of the science citation network. Only recently have studies shown that accounting for the growth of science can drastically change measured trends, e.g. in the decay rate of the citation life cycle of individual publications (Parolo et al., 2015). Insights such as this call for a better understanding of how the scientific attention economy (Franck, 1999; Klamer & van Dalen, 2002) is impacted by growth of the scientific system. Moreover, the recent proliferation of a new ecosystem of rapid-publication online-only "mega-journals" (Bjork, 2015; Petersen, 2018; Solomon, 2014; Solomon & Bjork, 2012; Wakeling et al., 2016) has further tipped the balance towards production over consumption of new research, making it a relevant and pressing issue.

To grasp the impact of this new publishing paradigm, consider the sole contribution by the online-only journal PLOS ONE, which grew over its first 6 years at an annual growth rate of 78.6%, corresponding to a doubling rate of only 1.2 years. To place this growth in real terms, after just 5 years since its inception, publications by PLOS ONE in 2012 (more than 23,000 articles) accounted for 1.4% of the entire volume of 2012 items indexed by the Web of Science (WOS) *Science Citation Index Expanded*; and calculated over the entire period 1900–2012, PLOS ONE publications represented 0.12% of all articles in this WOS index. Compared with other journals and fields in Fig. 1, which show typical growth around 3–4% annually, and even breakthrough fields such as Stem Cell research, which has grown at a thriving annual rate of 18.5%, the emergence of mega-journals appear ready to sustain long-term trends in the growth of scientific production well into the 21st century.

Applying methods from network science, complex systems, and data analytics, the 'science of science' (Fortunato et al., 2018) uses the millions of new research outputs produced each year by scientists around the world to illuminate knowledge production and innovation processes, thereby aiming to provide valuable insights for science policy guidance (Fealing, 2011). One particular link to the past that is preserved within each publication is the bibliographic list of references, which provide a means to measure how much today's research builds upon yesteryear's. As such, the citation network – where nodes are publications and links are the references within a publication to prior literature – has been used to conceptualize and measure the processes underlying the evolution of the scientific enterprise for more than half a century (de Solla Price, 1965; Garfield, 1955), and continues to be useful for making important insights into the long-term evolution of the scientific enterprise (Fortunato et al., 2018; Sinatra, Deville, Szell, Wang, & Barabasi, 2015).

Against this background we analyze the interplay between publication output growth and the attention to prior literature captured by the citation network. More specifically, we simultaneously identify and model three key features of this complex adaptive system:

1. The steady growth of the total number of references produced each year, arising from increasing publication output and reference-list lengths, and its relation to *citation inflation*.
2. The subsequent shifts in the concentration of citations received by publications at the lower and upper extremes of the citation distribution, providing perspective on *citation inequality*.
3. The distribution of references backwards in time which captures the historical breadth of *scientific attention*.

These considerations are important in the measurement, interpretation, and modeling of science for three fundamental reasons. First, citation inflation, which arises from the exponential growth in the production of references, affects the relative value of citations, thereby impacting the comparative evaluation of careers, institutions, and country output across different periods. For this reason, the bibliometrics community has put much effort into developing normalization strategies for comparing citation measures between varying time periods and disciplines (Waltman, 2016; Waltman & van Eck, 2018).

Second, this increasing supply of references has dramatically diminished the fraction of publications that go uncited. While this shift may at first appear to be an incremental change in the lower tail of the citation distribution, it has an enormous impact on the overall connectivity of the citation network, thereby affecting the scalability of "random walk" algorithms developed for the purpose of search, browsing, and recommendation in large yet sparse networks.

Third, there is the question of whether, and to what extent, a diminishing breadth of attention is related to the system's growth, and whether a decreasing attention to older literature may negatively impact the efficiency of knowledge production. Here we also uncover a decreasing attention to very recent literature – published within the most recent 6 years – that further indicates that scientists are not able to keep up with the deluge of new research, which may also reduce innovative capacity and efficiency. How innovation in information and communications technology (ICT) will address this information overflow by improving researchers' ability to browse, search, retrieve, store, analyze, and recall knowledge, thereby augmenting human memory capacity (Sparrow, Liu, & Wegner, 2011), is a question of particular relevance and importance.

By considering the three features simultaneously – each of which has been addressed separately in the literature, however with some notable disagreement – we aim to show how they are all related to the growth of science. We begin by outlining the fragmentation of the existing literature pertaining to these three features, and proceed to develop a framework and theoretical model for analyzing them together.

2. Background and framework

The general consensus is that the rate of uncited publications is declining (Lariviere, Gingras, & Archambault, 2009; Schwartz, 1997; Wallace, Lariviere, & Gingras, 2009). However, the level of inequality in the citation distribution has been shown to either decline (Acharya et al., 2014; Lariviere et al., 2009; Petersen & Penner, 2014) or increase (Barabasi, Song, & Wang, 2012; Evans, 2008), depending on the method and the perspective. There is also inconsistency concerning the obsolescence rate of scientific literature – used as a quantitative proxy to estimate the life-cycle of knowledge. For example, Evans (2008) shows that journals with more availability of online back-issues tend to have reference lists that are more focused on recent literature (more myopic), explained as the result of the availability of efficient online hyperlinks that mediate the browsing of the publications listed in reference lists. Meanwhile, two recent studies report that older publications are being cited (as a percent) more and more over time (Verstak et al., 2014; Wallace, Lariviere, & Gingras, 2012), which is consistent with an increasing mean reference distance, demonstrated empirically (Lariviere et al., 2008), and further shown analytically to follow directly from the growth of science (Egghe, 2010). These discrepancies demonstrate the need for a methodological framework that accounts for the systematic bias introduced by the exponential increase of scientific output. Indeed, although the inflation of scientific output has been considered previously (Broad, 1981; de Solla Price, 1965), only recently have analogs of inflation indices been used to normalize impact factors (Althouse et al., 2009) and individual citation counts (Petersen, Fortunato, et al., 2014; Petersen, Pan, Pammolli, & Fortunato, 2018).

To address these issues we analyzed the Clarivate Analytics Web of Science (WOS) publication index from 1965 to 2012, comprising 32.6 million publications and 838 million references made (or from the alternative perspective, citations received). That is, in our citation network analysis we consider the obsolescence problem from both the prospective (forward looking or diachronous) as well as the retrospective (backward looking or synchronous) perspectives (Glänzel, 2004; Nakamoto, 1988). We control for disciplinary variation by grouping the publication data using the three major subject area categories defined by WOS: (Natural) Science, Social Sciences, and Arts & Humanities (A&H). For each subject area, we analyzed the impact of the exponentially growing system (inflation) on the concentration of citations within the citation network (inequality) and the subsequent impact on the shifting breadth of attention, together illustrated in Fig. 2.

We start by analyzing how the inflation of the supply of references affects the distribution of citations received across different publication year cohorts. For example, we measured a 5.6% growth rate in $R(t)$, the total number of references produced (see Fig. 3), meaning that the total number of citations (links) in the WOS citation network doubles roughly every 12 years! These basic considerations then lead naturally to the question: is the concentration of incoming citations received increasing or decreasing? We find that the answer to this question is intrinsically linked to the decreasing share of uncited publications, which is inherently related to the increasing supply of references produced.

We then move to the question of whether attention (citations) to new, medium, and old literature is changing over time. To this end, we analyze the temporal distances between a referencing publication and the cited publication, denoted as the reference distance Δ_r. Instead of drawing conclusions from shifts in mean values, we focus this analysis on the shifts in the entire probability distribution $P(\Delta_r)$ across 10-year intervals from 1970 to 2010.

One of our main findings is demonstrating a narrowing attention consisting of a shift towards the middle range of the $P(\Delta_r)$ distribution – towards literature older than ≈6 y (Δ_r^-) and less than ≈50 y old (Δ_r^+) – which has increased on average by 25% over the study period for the three subject areas. These trends follow from an unexpected "fixed point" in the lower end of $P(\Delta_r)$, around $\Delta_r^- \approx 6$ years, which serves as a useful benchmark for classifying academic literature as contemporary (<6 y) or established (≥6 y).

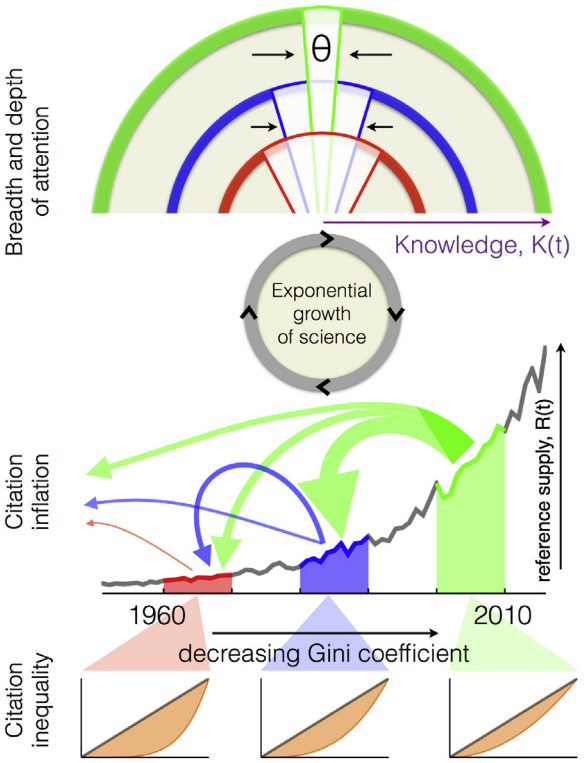

Fig. 2. Complementary perspectives on the growth of the scientific attention economy. The exponential increase in publications and reference-list lengths means that more citations are produced today than in the past. (Middle panel) As a result, the citation of prior literature is also growing with time, corresponding to a nonlinear inter-generational effect (i.e. the variable citation flows represented by the arrows). (Bottom panel) An increasing reference supply $R(t)$ impacts the concentration of the citations distribution as well as the real value of citations aggregated over different time periods, which are commonly used in research evaluation. (Top panel) From a researcher perspective, the increasing corpus of knowledge, indicated by $K(t)$, stresses the cognitive limits of researchers attention and expertise, as represented by the circular sector. Assuming a finite capacity for knowledge (fixed sector area), individuals must adapt by narrowing their breadth (as indicated by the range of expertise (θ) at the knowledge frontier) – and depth (as indicated by the decreasing proportion of the sector area reaching deep knowledge).

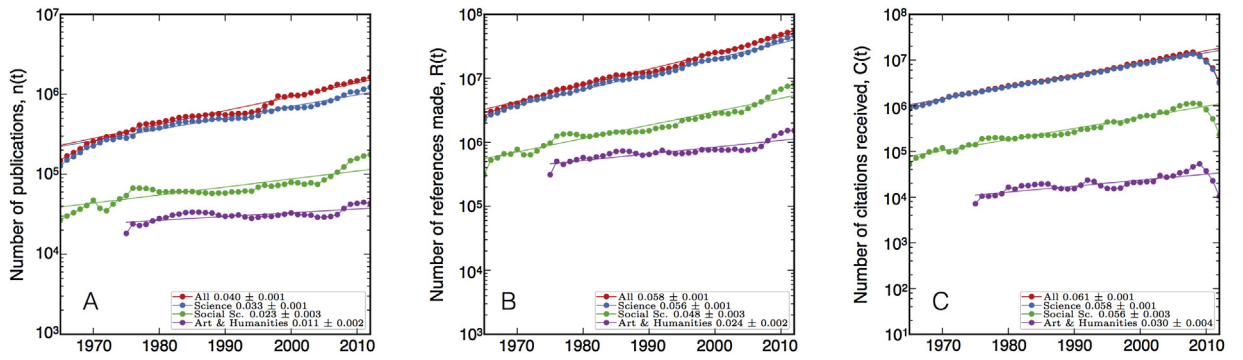

Fig. 3. Outputs of scientific R&D: empirical growth trends. Empirical growth trends of scientific output: publications and references. (A) Growth in the number of publications per year, $n(t)$. (B) Growth in the total supply of (outgoing) references per year, $R(t)$. (C) Growth in the total number of incoming citations per year, $C(t)$ (as measured in our citation census year $Y = 2012$).

Interestingly, the decline in the percentage of references for $\Delta_r < \Delta_r^-$ may follow from the rapid expansion of the system, whereby the trade-off between short-term and long-term memory are stressing cognitive limits. There may also be social cohort effects within scientific communities, wherein the incentives to follow established leaders may reinforce the concentration of attention away from very recent as well as very distant research. Indeed, scientific reputation is a social mechanism that serves as an identification device to aid with the information overload problem, and its role may be becoming stronger as quantitative measures become increasingly prevalent in science (Petersen, Fortunato, et al., 2014).

To better understand these trends within a single framework, we develop a network-based citation model which incorporates four key ingredients that capture both real features of the academic citation system and the process that authors follow while searching within and traversing across the knowledge network:

(i) exponential growth in the number of publications published each year, $n(t)$, and the number of references per publication, $r(t)$,
(ii) crowding out of old literature by new literature, which we impose using an attention bias operationalized by $n(t)$ itself,
(iii) a citation mechanism (link-dynamics) capturing the orientation of scientific attention towards high-impact literature (a positive feedback mechanism), and
(iv) a redirection link-formation process that is inspired by the now common behavior of finding related knowledge by following the reference list of a source article.

Because (iv) is implemented via a tunable parameter controlling the rate of triadic closure in the citation network, this model falls into the class of network growth models incorporating redirection mechanisms (Gabel, Krapivsky, & Redner, 2013, 2014; Holme & Kim, 2002; Krapivsky & Redner, 2001, 2005; Simkin & Roychowdhury, 2007; Vazquez, 2003). We show that our generative model reproduces various stylized facts observed in empirical citation networks. Moreover, our model provides the opportunity to accurately explore the causal impact of shifts in citing behavior – increased reference list lengths and increased ability to follow the reference trail – to provide informative insights for researchers and policy makers interested in the impact of growth trends on the evolution of science.

3. Material and methods

We analyzed all publications (articles and reviews) written in English from 1965 till the end of 2012 included in the Clarivate Analytics Web of Science (WOS) database (not including the Emerging Sources Citation Index, the Conference Proceedings Citation Index, and the Book Citation Index). For each item we extracted its publication year, the journal in which it is published (for classifying according to subject area), the list of references in its bibliography, and the citations received by other publications.

We used the WOS journal classification (denoted by the WOS publication record field WC which associates the journal with one or more of the 252 Web of Science Categories) to separate the publication data into 3 broad subject area domains: (natural) "Science" (corresponding to the "Science Citation Index Expanded"), "Social Sciences" (corresponding to the "Social Sciences Citation Index"), and "Art & Humanities" (corresponding to the "Arts & Humanities Citation Index"). In total, our analysis comprises of 32,611,052 publications and 837,596,576 references. This latter number includes references to articles both indexed and not indexed within our WOS database, since we are still able to determine the publication year of articles not indexed in the WOS by the information in the reference lists. We focus our analysis on the science and social science domains, which account for more than 95% of the data. Note that there are many journals in WOS that are included in the Science, Social Sciences, and Art & Humanities citation indices. The articles published in these journals are, however, included in the category "All". As the unclassified journals are typically of relatively low citation impact, when included in our analysis the results may differ slightly from other well classified journals. We obtained the publication rates in Fig. 1 from the WOS using specific queries: panels (A, B) correspond to two specific journals; panel (C) is derived from the search term "Stem Cell"; and (D) corresponds to an aggregation of 14 high-impact economics journals (American Economic Review, Econometrica, Journal of Political Economy, Journal of Economic Theory, Journal of Econometrics, Journal of Financial Economics, Journal of Finance, Journal of Economic Growth, Journal of Economic Perspectives, Journal of Economic Literature, Quarterly Journal of Economics, Review of Economic Studies, Review of Financial Studies, Review of Economics and Statistics).

Further, we restrict our study to papers published after 1965, which helps to reduce one of the principal weaknesses of the WOS dataset, namely its dynamic coverage of journals. Fig. 11(A) shows the fraction of references in our database citing articles contained within the WOS dataset, which has increased steadily over time. For example, in 1965 roughly 65% (20%) of the references from Science (Social Sciences) were contained within the WOS database; and in the 2000s, roughly 90% (50%) of the references from Science (Social Sciences) were contained within the WOS database. It is important to note that this coverage issue primarily affects the analysis of incoming citations, since the citing (source) article must by construction be contained within the WOS dataset. This is not the case in our analysis of outgoing references, which contribute to our analysis even if the cited (destination) article is not contained within the WOS.

In summary, the dynamic coverage of WOS impacts the interpretation of our results in three particular ways. First, our estimates of the growth of science, based here on article counts and total reference counts, would be different if the WOS indexed every journal in production; instead, WOS has strict inclusion standards, and so our analysis of scientific production is subject to their criteria and conditions. Second, in terms of incoming citations received by publications, we assume that the set of articles outside of the WOS feature the same citation trends (e.g. citation inflation) as those contained in the WOS, independent of whether the excluded journals are less prominent and/or obscure. Third, in terms of outgoing references made to prior literature, we assume that references made outside of the WOS are generated according to the same processes as those existing within the WOS sample. In terms of citing behavior, this means that we assume that researchers cite literature independent of whether it is contained in the WOS index. To investigate the impact of these issues, we performed a comparative analysis based upon a subset of 4 high-impact journals which have been indexed by the WOS since the 1970s

(CELL, Nature, PNAS, and Science), and for which we can assume that: (a) most outgoing citations made by their articles are to other publications indexed by WOS; and (b) most citations made to their articles are by other publications indexed by the WOS. Our results show that the patterns for this stable subset are qualitatively aligned with the patterns for the entire dataset, and so the effects due to the WOS coverage appear to be marginal; regardless, the impact of dynamic WOS coverage is a topic that warrants further research.

And finally, we summarize the notation used in what follows. For each publication p published in given year t, we collected the set of references made by p. Then, for a given reference r with publication year t_r, we define the reference distance as $\Delta_r \equiv t - t_r$, i.e. the distance in time between p and r. Again, we calculate Δ_r for all references contained in the reference list of p that indicate a t_r, even if r is not contained in the WOS dataset. Using the subset of references contained within the WOS dataset, we calculated the total citation count $c_{p,t}$ up to year t, and the citation rate $\Delta c_{p,t} \equiv c_{p,t} - c_{p,t-1}$, corresponding to the total number of new citations received in year t. In Sections 4.2 and 4.3 we address the right-censoring citation bias by implementing a fixed citation window such that $c_{p,t,\Delta t}$ represents the total citation count tallied for just the first Δt years, choosing $\Delta t = 5$. In all other sections, we use an unlimited citation window, denoted by $c_{p,t} \equiv c_{p,t,\Delta t=\infty}$.

4. Empirical results

4.1. Growth of the science citation network

The growth of science is evident in its increasing funding, workforce (Petersen et al., 2018), article publication rate, and total knowledge production (Fortunato et al., 2018). In this study the units of analysis are primarily publication-to-publication associations – termed "references" from the outgoing perspective (i.e. reference list) and "citations" from the incoming perspective (i.e. citation count c_p of a given publication p). We calculated the growth of the reference supply drawing from two sources: (i) the increasing number of publications $n(t)$ produced per year t, and (ii) the increasing (average) number of references per publication, $r(t)$. For example, Fig. 3(A) shows exponential growth for the Science domain, where we measure an annual exponential growth rate $g_n = 0.033$ for publications and $g_r = 0.018$ for the references per publication. As a result, the net annual reference supply, $R(t)$, is also increasing exponentially with annual growth rate $g_R \approx g_n + g_r$. As a consistency check, we verified this exponential approximation by using the time series for the total references produced in a given year, $R(t)$, calculating $g_R = 0.056 \pm 0.001$ for Science. This increase in the reference supply (and thus the citation credit supply) is analogous to monetary inflation in economics and gives rise to "citation inflation" (Petersen et al., 2018).

De Solla Price estimated a publication doubling time of 13.5 years, corresponding to $g_n = \ln(2)/13.5y = 0.05\,y^{-1}$ in 1965 using publication data for the 100-year period 1862–1961 (de Solla Price, 1965). For comparison, here we use $R(t)$ to estimate the rate of growth of the number of links in the knowledge network, calculating a doubling period of roughly $\ln(2)/0.056 \approx 12$ y. That is, every 12 years, both $R(t)$ and the cumulative number of references in the citation network up to t, $\sum_{t'<t} R(t')$, roughly double in magnitude. In what follows, we use the empirical growth rate g_n as a key "inflation" parameter in our generative citation network model.

4.2. Growth and citation inflation

One challenge in using citations as a measure of scientific impact is the fact that references produced later in time can impact the citation tallies of publications earlier in time – a backwards inter-generational flow of the reference supply (Fig. 2). Because the doubling period for the reference supply is only 12 years, it is thus conceivable that some publications have an extended citation lifecycle across several decades merely due to the underlying inflation, as recently indicated in an analysis of the right tails of the citation life-cycle (Yin & Wang, 2017), and possibly serving as an explanation for "sleeping beauties" in science (Ke, Ferrara, Radicchi, & Flammini, 2015). Thus, in order to control for this retroactive effect, we tallied the citation counts $c_{p,t,\Delta t}$ for each publication p from year t using only the citations arriving in the fixed 5-year ($\Delta t = 5$) window $[t, t + \Delta t]$. Using a citation window addresses the right-censoring bias problem, and it also limits the impact of citation inflation resulting from the long-term increase in the supply of references across time.

However, 'citation inflation' is still significant even when using a fixed citation window. To demonstrate the this we calculated $C(q|t)$, the citation value corresponding to a given percentile q of the citation distribution, for each year t. For example, Fig. 4(A) and (B) shows that the top 1% of Science publications from $t = 2000$ had more than 100 citations as of 2005, whereas the top 1% of publications from 1965 had only 50 citations as of 1970. Each $C(q|t)$ is growing at a slow exponential rate $\lesssim g_n$, which is larger for smaller q, pointing to a decreasing concentration of citations which we will address in the next section. Because we only use citations within the 5-year window, these trends are not sensitive to long-term trends in the obsolescence rate of scientific publications (Parolo et al., 2015). In all, the steady growth of $C(q|t)$ illustrates how the *real* "value" of citations is systematically decreasing over time due to citation inflation, which is clearly evident in the systematic shift in the entire citation distribution towards higher values. As a result, more recent publications need increasingly more citations to be within the top 5% (of publications from the same year) than do older publications in order to achieve the same percentile value within their respective publication cohort. This effect has implications for the evaluation of publications from different periods, which is often the case when citation tallies are calculated across a subset of publications

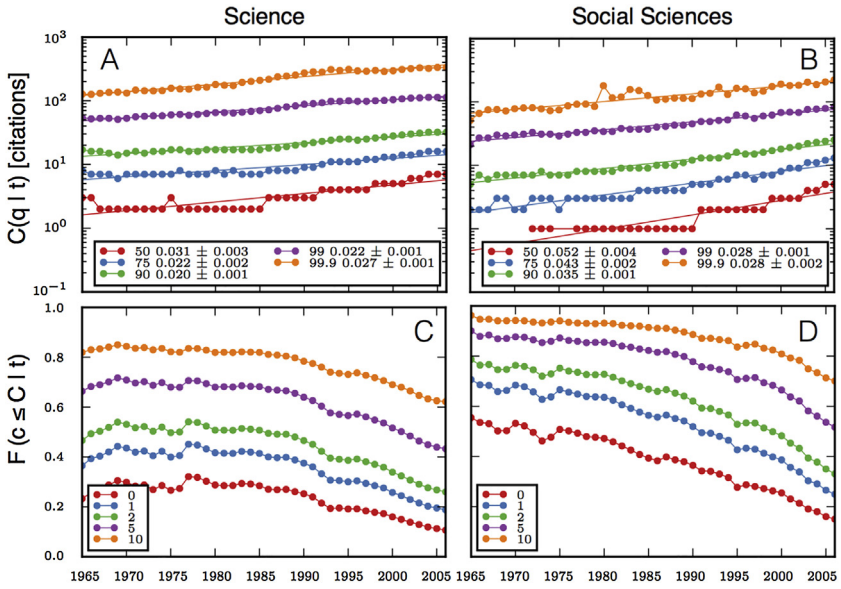

Fig. 4. Inflation & uncitedness – empirical trends in the upper and lower tails of the citation distribution. (A, B) Inflation: Increase in the broadness of the citation distribution. The citation value $C(q|t)$ corresponding to a given percentile $q = \{50, 75, 90, 99, 99.9\}/100$ of the citation distribution $P(c_{p,t,5})$. Each line of the legend shows two numbers: the percentile value $100 \times q$ and the best-fit exponential growth parameter calculated for each curve. (C, D) Uncitedness: Decrease in the fraction of uncited ($C = 0$) and less-cited ($C = 1, 2, 5, 10$) publications. Each curve represents the fraction of publications with $c_{p,t,5} \leq C$ citations received, for each threshold C and for each year.

that span a significant time period, e.g. researcher careers that span several inflationary doubling periods (Petersen et al., 2018).

4.3. Growth, uncitedness and inequality of the citation distribution

Recent work has shown that the rate of uncited articles is inversely related to the average number of references per article (Wallace et al., 2009). To further explore trends in the lower tail of the citation distribution, we calculated the fraction $F(c \leq C|t)$ of publications with $c_{p,t,5} \leq C$ citations received, for the range $0 \leq C \leq 10$ (e.g. the threshold $C = 0$ corresponds to uncited publications). Fig. 4(C) and (D) shows that the $F(c \leq C|t)$ are all decreasing, pointing to the relation between growth of the references supply $R(t)$ and decreasing likelihood of a publication being uncited in its first 5 years. For example, in 1980, roughly 30% of Science publications remained uncited 5 years after publication. By 2005, this percentage decreased to roughly 10%. Moreover, roughly 60% of Science publications from 2005 have 10 or less citations after 5 years. This decreasing trend has occurred in Science since the 1980s and in Social Sciences since at least the mid 1960s. Note that the gap between the curves for $C = 10, 5, 2$ is approximately constant over the last 20–30 years, indicating that the share of references *within* these ranges is not changing dramatically. Thus, the largest decrease over time is for the fraction of publications with $C = 0$, 2 citations, in that order.

While appearing to be just subtle changes in the fraction of uncited publications, the addition of this loosely-connected layer impacts the overall connectivity of the citation network. In practice, this topological alteration could impact the functionality of network-based search and retrieval algorithms, such as *Google Inc.'s* PageRank method (Page, Brin, Motwani, & Winograd, 1998), which is based on the principal of random walkers traversing the underlying information network, and serves as a common centrality measure in network science. As such, since citation networks also serve as the backbone for search and recommendation, the additional layer of loosely connected publications may affect the scalability and effectiveness of search algorithms designed to recommend and rank research (Vaccario, Medo, Wider, & Mariani, 2017).

Moreover, subtle shifts in the rate of uncited publications can have pronounced effect on the concentration of citations in the citation distribution. In order to demonstrate this relation, we calculated the Gini inequality coefficient

$$G(t) = (2n^2(t)\langle c(t)\rangle)^{-1} \sum_{i=1}^{n(t)} \sum_{j=1}^{n(t)} |c_{i,t} - c_{j,t}| \tag{1}$$

calculated for each year t, where i and j are indices running over the set of $n(t)$ papers from each publication cohort t and $\langle c(t)\rangle$ is the mean number of citations among p from t (Dixon, Weiner, Mitchell-Olds, & Woodley, 1987). Importantly, $G(t)$ is a relative (intensive) measure of the pairwise difference between all data values in the sample normalized by the value expected of this quantify for a uniform distribution. As such, $G(t)$ outperforms other extensive inequality measures, such as the Herfindahl–Hirschman Index $HHI(t)$, which can be significantly biased by temporal trends, e.g. if the number of entities

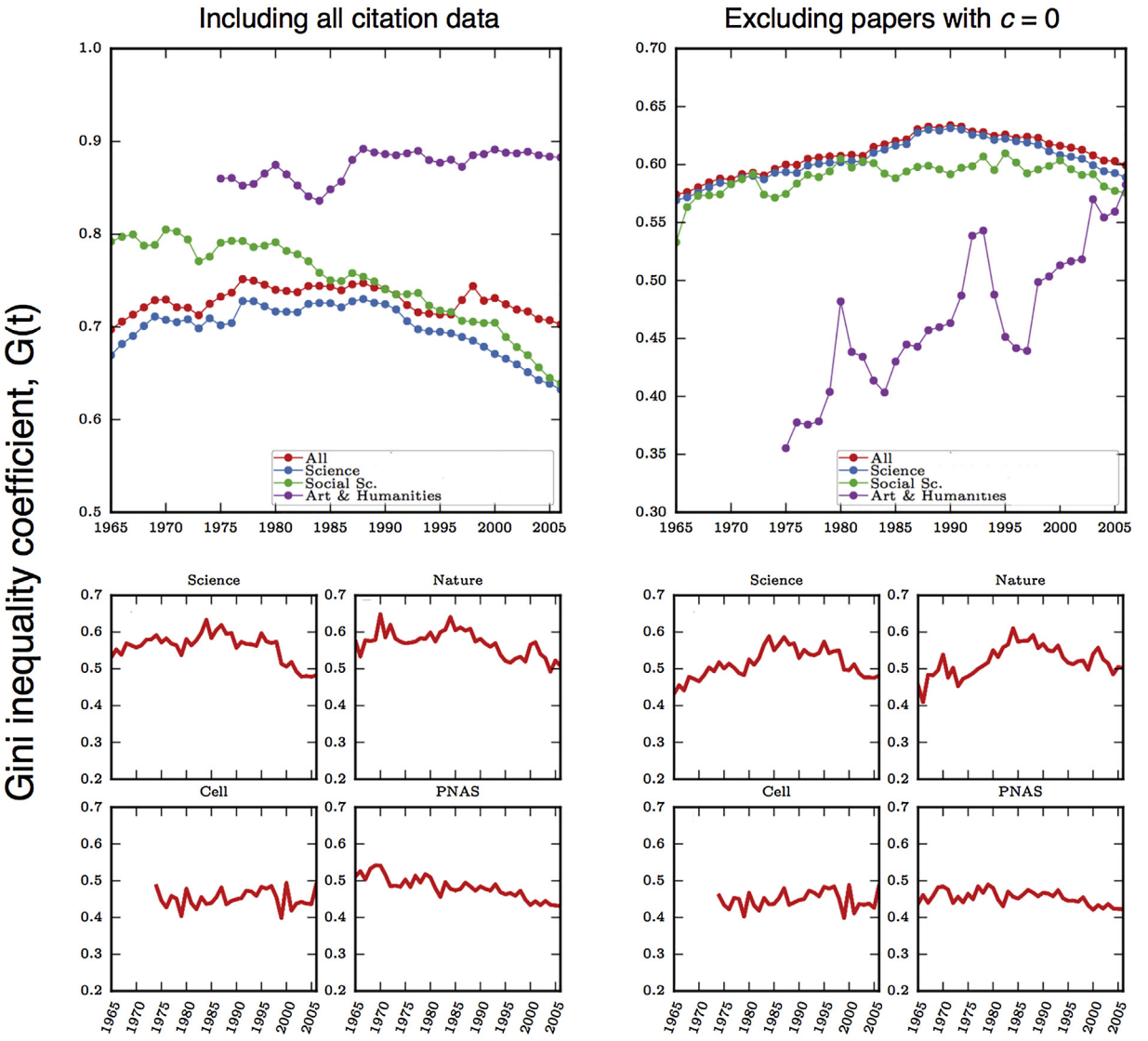

Fig. 5. Inequality – empirical analysis of the concentration of the citation distribution. The Gini inequality coefficient $G(t)$ is a standardized distribution measure which captures the relative concentration of citations with values ranging from 0 (all publications have the same number of citations) to 1 (extreme inequality, all publications have $c_{p,t,\Delta t=5} = 0$ except for one with $c_{1,t,\Delta t=5} > 0$). (Top row) Gini index by subject area. (Bottom row) Gini index by high impact journal. (Left column) Gini coefficient calculated using all publications ($c_{p,t,\Delta t=5} \geq 0$). (Right column) Gini coefficient calculated using only the publications with $c_{p,t,\Delta t=5} > 0$.

$n(t)$ is not reasonably constant.[2] As such, $G(t)$ is less sensitive to fluctuations and to system size bias because it incorporates information from the entire citation distribution (all moments instead of just the second moment, $\langle c_t^2 \rangle$). Moreover, it is a standardized distribution measure with values ranging from 0 ($c_{p,t,5} = \text{const.}\ \forall p$) to 1 (extreme inequality, i.e. all publications have $c_{p,t,5} = 0$ except for one).

Our analysis reported in Fig. 5 reveals a slow but substantial decrease in $G(t)$. To test whether the decreasing $G(t)$ is related to the decreasing trend in $F(c = 0|t)$, we recalculated $G(t)$ ignoring the uncited publications, finding the negative trend

[2] Without loss of generality, the HHI index in terms of the citation distribution can be written as $HHI(t) \equiv (C_t)^{-2} \sum_{i=1}^{n(t)} c_{i,t}^2 = n(t)\langle c_t^2 \rangle / C_t^2(t)$, with $C_t = \sum_{i=1}^{n(t)} c_{i,t}$. Thus, when considered in terms of the underlying distribution $P(c)$ instead of the unit-level "share" $c_{i,t}/C_t$, it becomes more clear that the HHI index is the product of the sample size, $n(t)$, the second moment of the distribution, $\langle c_t^2 \rangle$, divided by the net number of citations received by p from t, $C(t)$, squared. Thus, unlike $G(t)$ which is an intensive measure of the mean differences, the HHI is extensive (its value depends on the system size) as it confounds sample size bias due to the growing system with sensitivity to extreme values which are typical of citation distributions.

in $G(t)$ to be subsequently reduced. Moreover, the level of inequality also decreased when ignoring uncited publications in the calculation of $G(t)$.

Thus, the vanishing of uncitedness, in response to increasing $R(t)$, is sufficient to explain why citation inequality has decreased. It is possible that this trend merely follows from the expanding coverage of WOS, resulting in an artificial decrease in $G(t)$ following the indexing of more journals with on average lower impact factors. Our analysis of a select set of high-impact journals, which should be less sensitive to the expanding coverage of WOS, exhibits the same trends in $G(t)$ as for all Science, including and excluding uncited publications, (see Fig. 5), suggesting that the effect of WOS coverage expansion is marginal in this regard.

4.4. Growth and attention to scientific literature

As scientific production continues to grow, and because older publications (nodes) have shorter reference lists (fewer outgoing links), the new layers of knowledge added at the frontier subsequently diminish the accessibility of prior layers. The combined result is an abundance of short pathways between new knowledge and recent knowledge as apposed to the relatively fewer, and likely longer, pathways between new knowledge and older knowledge. Fig. 2 schematically illustrates this crowding out of the old by the new, and its implications on the breadth and depth of researcher attention and expertise.

More concretely, in this conceptual model the corpus of knowledge is a d-dimensional sphere of radius $K(t)$, with new contributions being added to its surface, i.e. the knowledge frontier. The volume V of generic spherical sector is defined by the angle θ (representing the breadth of an individual's attention at the frontier as well as at deeper layers), and the radius $K(t)$ (representing the depth of attention). In this way V represents the attention capacity of scientific agents to prior knowledge – and when aggregated over all agents, this corresponds to the system's memory capacity. Consider the average volume $\bar{V}$ as representative of scientists, which for simplicity of argument we assume to be relatively constant over time (i.e. we do not consider evolutionary factors or technological innovations that could increase $\bar{V}$). If we approximate the knowledge radius $K(t) \approx N(t) \sim \exp(g_n t)$, the total number of publications up to t, then constant capacity $\bar{V}_1 = \bar{V}_2$ for two average agents at different times $t_2 > t_1$ can be rewritten as $\theta_1^{d-1} \exp(dg_n t_1) = \theta_2^{d-1} \exp(dg_n t_2)$. Consequently, if agents are to maintain the same depth of knowledge, then $\theta_2/\theta_1 \approx \exp[-g_n(t_2 - t_1)d/(d-1)] < 1$, independent of d, implying a narrowing breadth of attention at the frontier. As such, the only compromise to narrowing θ is to reduce attention to deep foundational knowledge, i.e. decreasing the depth of the radius $K(t)$ so that it does not extend to the sphere's absolute core.

We provide evidence consistent with this crude attention model by analyzing trends in the reference distance, Δ_r, which is the number of years between the publication date of the referencing publication and the cited publication.[3] In total we analyzed 837,596,576 references over the period 1965–2012.[3]

We start by following the method used by Verstak et al. (2014), which analyzed articles indexed by Google Scholar between the years 1990 and 2013. More specifically, they calculate the fraction of references with distance Δ_r above a given threshold value, and track this fraction over time. Using the threshold values of 10, 15 and 20 years, Verstak et al. report an increasing fraction $F(\Delta_r \geq 10, 15, 20|t)$, indicating an increasing attention to older literature over time. However, 20 years is too short to investigate trends extending beyond the time period of any contemporary generation of scientists. In other words, in order to draw conclusions on attention to distant literature, one must reach further back in time to literature authored by researchers who are likely not still actively publishing – i.e. to investigate trends in generation-spanning attention.

With this in mind, we applied the Verstak et al. method to the WOS data, extending further back in time to 1965 and also exploring both smaller and larger threshold values, from 1 up to 50 years. The results of our reproduction are reported in Fig. 6, which are consistent with (Verstak et al., 2014) – namely exhibiting an increasing attention to literature older than 10 and 20 years. However, for larger threshold values, we observe a small yet significant increase in $F(\Delta_r \leq 50|t)$ for Science and Social Sciences and $F(\Delta_r \leq 30|t)$ for A&H. We also analyzed the small-Δ_r regime that was unexplored in Ref. (Verstak et al., 2014). In this case, for all subject areas analyzed, we observe a steady decline in the fraction of references less than 3 years old, $F(\Delta_r \leq 3|t)$. These results call for a more detailed analysis of the full distribution of reference distances.

For this reason, we analyzed the probability distribution $P(\Delta_r|t)$ and cumulative probability distribution $CDF(\geq \Delta_r|t)$ across the full range of values; note that $CDF(\geq \Delta_r|t) = 1 - F(< \Delta_r|t)$. Fig. 7(A) shows the results of our analysis for 10-year intervals over the period 1970–2010. The decline in attention to very recent literature is rather evident in the comparison of $P(\Delta_r|t)$ for small Δ_r values, with the more green curves (more distant t) peaking significantly above the more red curves (more recent t). Theoretical arguments by Egghe (2010) predict an increasing average and median reference distance as a results of the growing citation network. However, the distribution averages (shown as vertical dashed lines in Fig. 7(A)) do not display definitive increasing trends. This likely follows because the distribution of Δ_r is right-skewed and so the mean can be sensitive to extreme values.

Visual comparison of $P(\Delta_r|t)$ for varying t also reveals an unexpected intersection point Δ^- around the value $\Delta_r^- \approx 6$ years that is rather stable across t, despite the growth in $n(t)$ and $r(t)$. Upon closer inspection, we also observe a second, yet less-precise, intersection point Δ^+ in $CDF(\geq \Delta_r|t)$ around $\Delta_r^+ \approx 50$ (Science), 20 (Soc. Sciences), and 40 years (A&H). As

[3] Note that in Sections 4.3 and 5 we include all citations in our analysis, even those received after $\Delta_t = 5$ y. The measure $c_{p,t}$ departs from the fixed-window method used in Sections 4.2 and 4.3, in which $c_{p,t,\Delta t=5}$ only considered citations received through the first $\Delta t = 5$ years in order to account for right-censoring bias.

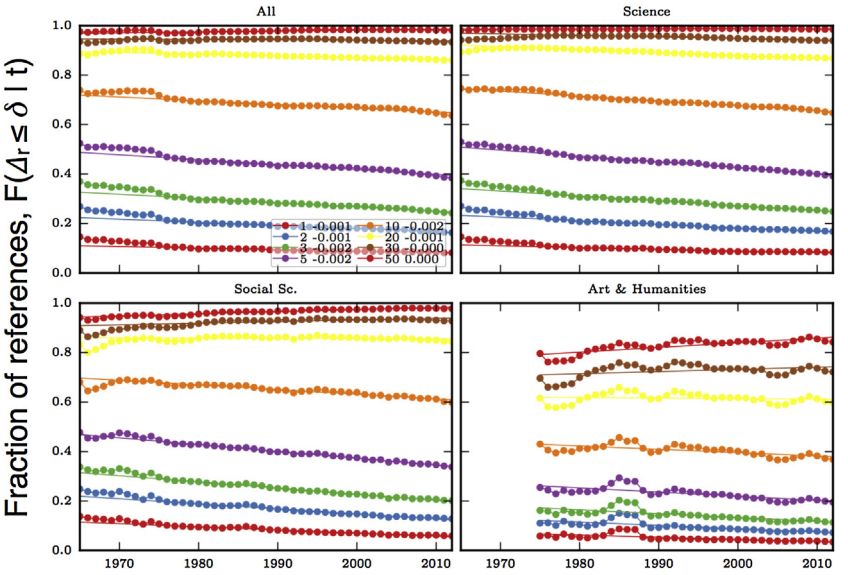

Fig. 6. Empirical trends in reference distances. The reference distance $\Delta_r = t - t_r$ is the number of years difference between the publication date t of p and the publication date t_r of any reference r appearing in its reference list. Shown is the fraction $F(\Delta_r \leq \delta|t)$ of references from year t with Δ_r value falling within the time window $[t-\delta, t]$. The values of δ are 50 years (top curve), 30, 20, 10, 5, 3, 2, 1 year (bottom curve).

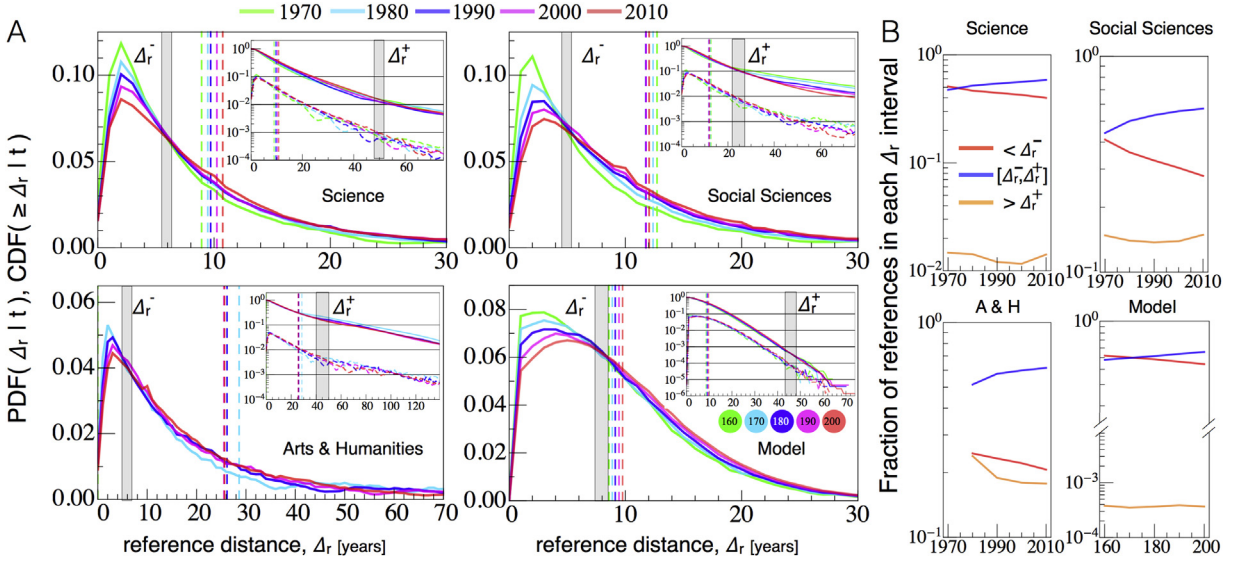

Fig. 7. Narrowing historical attention – reduction of scientific attention in the very near and very far fields. (A) Shown are the reference distance probability distributions, $P(\Delta_r|t)$, for select t indicated in the color legend; vertical dashed lines indicate distribution mean. (Inset) $CDF(\geq\Delta_r|t)$ (solid curve) and probability distributions $P(\Delta_r|t)$ (dashed curve) on log-linear scale to emphasize the shifts in the distributions for large Δ_r. Each panel shows a small Δ_r^- regime for which the $P(\Delta_r|t)$ cross – independent of t – signaling a fixed point in the reference distance distribution: $\Delta_r^- \approx 6$ years (Sci.), 5 y (Soc. Sci.), 6 y (A&H), and 8 y (Model). A second crossing point Δ_r^+ is indicated in the empirical $CDF(\geq\Delta_r|t)$, such that the fraction of citations going to publications with $\Delta_r > \Delta_r^+$ is decreasing for larger t: $\Delta_r^+ \approx 50$ years (Sci.), 20 y (Soc. Sci.), 40 y (A&H), and 45 y (Model). Interestingly, Science exhibits an increasing mean value with time, whereas Soc. Sci. and A&H indicate a decreasing mean value, demonstrating how single-value distribution measures can yield misleading comparisons. The lower-right panel shows the results of our model; Model parameters are listed in Fig. 9. (B) Temporal shifts in scientific attention towards the mid-field, demonstrated by tracking the fraction of references at 10-year intervals falling into 3 non-overlapping intervals: near-field, $\Delta_r < \Delta_r^-$; mid-field, $\Delta_r^- \leq \Delta_r \leq \Delta_r^+$; and far-field, $\Delta_r > \Delta_r^+$.

indicated in Fig. 6, the trends in $CDF(\geq\Delta_r|t)$ for $\Delta_r > \Delta^+$ more clearly show a decline in attention to the significantly older literature; this is most evident for Soc. Sci., where the green curve (corresponding to data for 1970) are above the red curve (corresponding to data for 2010). Combining these two general observations – the decreasing attention to very recent and very old literature – it follows by definition that there is a redistribution of Δ_r towards the reference distance interval $[\Delta_r^-, \Delta_r^+]$ years – i.e. a narrowing breadth of historical attention. This convergence corresponds to a narrowing because, if trends continue in the present direction, one can imagine the $P(\Delta_r|t)$ converging to a more normally distributed distribution

around a characteristic mean with characteristic width that is narrower than the width of the current $P(\Delta_r|t)$. As a robustness check, we also applied the same distribution analysis to data collected from a set of high-impact journals; Fig. 11(B) confirms that the narrowing trend continues to present day, as demonstrated by the most recent data for 2017.

Another result of our analysis of the full $P(\Delta_r|t)$ distribution is the identification of two crossing points, which appear to be relatively stable over time, thereby pointing to three specific attention windows:

(i) the myopic or near-field, $\Delta_r < \Delta_r^-$,
(ii) the mid-field, $\Delta_r \in [\Delta_r^-, \Delta_r^+]$,
(iii) the hyperopic or far-field, $\Delta_r > \Delta_r^+$.

Grouping data according to this data-driven classification, Fig. 7(B) shows that in all cases analyzed, including the results of our theoretical model developed in the next section, there is a decreasing attention to the near-field compensated by an increasing attention towards the mid-field. This redistribution towards the mid-field – for which the corresponding probability mass has grown by roughly 24% in Science, 30% Soc. Sci., and 19% in A&H over the last 50 years – constitutes a significant narrowing of attention. Interestingly, we observe a non-monotonic trend in attention to the far-field, which decreased for Science and Soc. Science over the period 1970–2000, but shows an uptick for 2010, as well as in more recent data for 2017 in Fig. 11(B). Because the amount of data representing the far-field is negligible with respect to the bulk of the distribution, the narrowing trend observed for the near- and mid-fields persists regardless of any observed trend in the far-field.

Altogether, these results are consistent with a narrowing θ and decreasing K described in our simple conceptual model of attention in science. However, they also signal nuances with respect to the far-field that may be related to the emergence of web-based information systems in the 2000s that augmented the way scientists access very distant literature. In Section 5 we develop a more elaborate model of the citation network to facilitate a mechanistic explanation for the narrowing of attention towards the mid-field; and in Section 6 we further discuss hypothetical behavioral and technological explanations for these trends.

5. Theoretical model

5.1. Network growth model featuring inflation, obsolescence, and attention redirection

Stochastic growth models can provide mechanistic insights into the evolution of competition and growth in various complex systems (Buldyrev, Growiec, Pammolli, Riccaboni, & Stanley, 2007; Golosovsky & Solomon, 2013, 2014; Petersen, Fortunato, et al., 2014; Petersen, Jung, Yang, & Stanley, 2011; Petersen, Riccaboni, Stanley, & Pammolli, 2012; Scharnhorst, Börner, & van den Besselaar, 2012). In order to gain such mechanistic insights into the impact of growth on citation inequality, uncitedness, and shifts in $P(\Delta_r|t)$, we developed a *generative model* of the science citation network, which we implement using Monte Carlo simulation. Because our model produces the entire network, it includes and leverages the information contained in the entire set of incoming citations and outgoing references. This is in contrast to similar "mean field" models which simulate the citation dynamics of individual publications, but neglect the degree-degree correlations embedded in the network with remarkable success (Golosovsky & Solomon, 2012, 2013, 2014; Peterson et al., 2010; Wang, Song, & Barabasi, 2013).

The initial conditions of our synthetic science system is a small batch of $n(0) \equiv 10$ nodes (publications), each with no outgoing references. Then, in each time step $t = 1, \ldots, T$ a cohort of $n(t)$ nodes (indexed by j) are added to the pre-existing system of nodes. We matched the slow exponential growth of $n(t)$ and $r(t)$ to empirical data, using $g_n^{model} \equiv 0.033$ and $g_r^{model} \equiv 0.018$, so that $g_R^{model} = g_n^{model} + g_r^{model} = 0.051$, thereby using g_n and g_r as the fundamental growth parameters for the simulation.

Our model is also designed to capture the behavioral aspects of synthesizing reference lists, wherein article browsing is facilitated by using the reference lists themselves as additional pathways to browse and find other articles. As such, this generative network growth model falls into the class of redirection models (Gabel et al., 2013, 2014; Holme & Kim, 2002; Krapivsky & Redner, 2001, 2005; Simkin & Roychowdhury, 2007; Vazquez, 2003), but is distinguished from other models by its relatively high rate of triadic closure (Holme & Kim, 2002; Ren, Shen, & Cheng, 2012; Wu & Holme, 2009).

More specifically, each new publication i from cohort t makes $r(t)$ references by two distinct processes illustrated in Fig. 8(A):

(a) global "browsing" leading to citation of a primary source publication j, and
(b) redirection of attention to other articles appearing in the reference list of j.

The relative rates of (a) and (b) are controlled by a parameter $\beta = \lambda/(\lambda + 1)$ used in step (b) to choose the random number x of references cited from the reference list of j, where x is drawn from a Binomial distribution with mean $\langle x \rangle = \lambda$. For any given simulation, let $r^b(t)$ be the total number of references occurring via process (b) in t, then on average $r^b(t) \approx \beta r(t)$, with the small discrepancy for small t arising from finite size effects due to fixed upper limit in the reference list length, $r(t)$. We

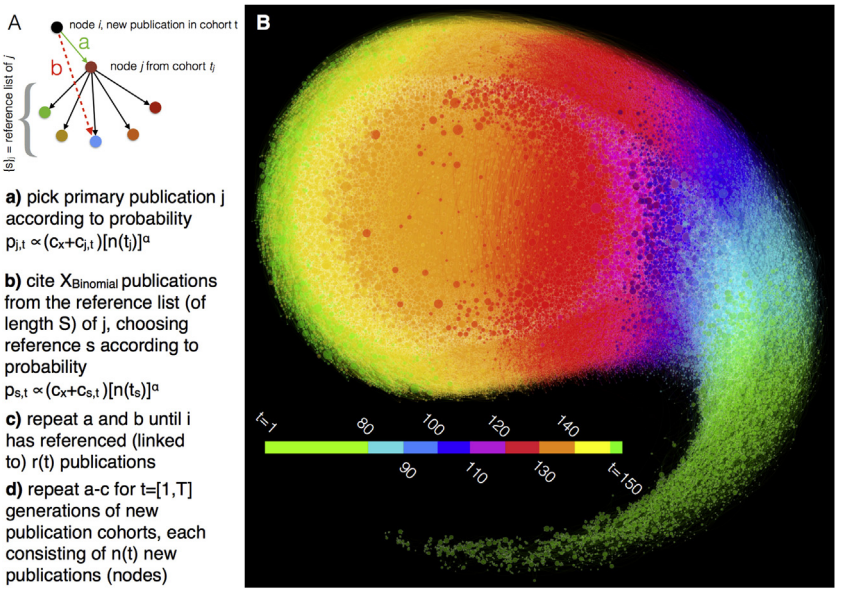

Fig. 8. Simulation of the generative network model featuring inflation, obsolescence, and attention redirection. (A) The model comprises three complementary mechanisms: preferential attachment (*PA*) link dynamics, crowding out of old nodes by new nodes induced by growth, and the redirection of citations via reference lists. The model parameter β controls the rate at which references occur via (b) for every initial reference from (a), thereby capturing the impact of shifts in "hyperlinking" citation behaviors. The entire citation network is generated by repeating (a–c) for many successive publications cohorts, with each Monte Carlo simulation period t representing a year. (B) Visualization of the citation network produced by the citation model, emphasizing the long-term expansion of literature juxtaposed by the relative thickening of the most recent layers. In order to emphasize nodes from every cohort, the node size is proportional to the normalized citation impact, $z_{p,t}$, which is a time-invariant citation measure suitable for cross-temporal analysis. The latest publication cohort represents a 'growth cone' which mediates the evolving connectivity of the growing network via the key redirection process included in our model, wherein more central and more recent publications are more likely to be referenced. The network was generated using the parameters given in Fig. 9, and is comprised of $N(T=150)=41,703$ nodes and 379, 454 links. Here we show all the publications (nodes) from periods $t=[1, 150]$ and assigned a color to each node according to its entry period (see legend), e.g. nodes from periods $t=[1, 79]$ are colored green, nodes afterwards are colored according to decade, except for the nodes from $t=150$ (the only nodes from this decade) which are colored lime green. The relatively large size of this last cohort, as compared to the first 149 periods, emphasizes the crowding out of the old by the new.

then repeat (a) and (b) for each new node until it has $r(t)$ links (outgoing references). The citation network is progressively grown by repeatedly adding new layers of $n(t)$ nodes, representing annual publication cohorts.

5.2. Results of Monte Carlo simulation

Fig. 8(B) provides a network visualization of a single realization of our model produced by Monte Carlo (MC) simulation of the link attachment dynamics. We colored nodes according to sequential time periods to illustrate and emphasize the exponential growth of the system. In this automated network layout, node colors are clustered into layers reflecting the fact that references primarily occur locally in time; however, some highly-cited nodes are pulled forward in time in the network as they maintain prominence (e.g. several red nodes are visible in the orange layer). Due to the exponential growth, the final layer of nodes (colored bright green) corresponding to the last "year" $t=T=150$ is already prominent with respect to the entire history of the model citation network. Specifically, the final network at $t=150$ has $N(T=150)=41,703$ nodes, $L=379,454$ links, modularity = 0.208, and a mean clustering coefficient = 0.018 indicating the relatively high rate of triadic closure (i.e. since $0.018 \gg L/(N(N-1)/2)=10^{-4}$, this means that the clustering coefficient is relatively large considering the number of edges and nodes).

Our model captures several important features of the science citation network. First, it incorporates the exponential growth of the system, both in $n(t)$ as well as $r(t)$, a feature which is not taken into account in prior citation models which assume that the citation sources produce a constant number of references per time unit (Golosovsky & Solomon, 2012, 2014; Peterson et al., 2010; Wang et al., 2013). As a result, the crowding out of the old by the new is readily apparent in the network visualization. Nodes introduced in older time periods are increasingly separated from the knowledge frontier, both in time as well as in network distance.

Second, while we implement classic linear preferential attachment (*PA*) (Barabasi et al., 2002; Jeong, Neda, & Barabasi, 2003; Peterson et al., 2010; Redner, 2005; Simon, 1955) in the link creation probabilities, we also include an additional obsolescence term that captures the crowding-out effect induced by the growth of the system. Combined, the attachment (citation) probability of node (publication) j from t_j is proportional to the weight

$$P_{j,t} = (c_\times + c_{j,t})[n(t_j)]^\alpha \tag{2}$$

where $c_{j,t}$ is the total number of citations received by j up to t, $n(t_j)$ is the number of new nodes entering in period t_j, $\alpha \equiv 5$ is a scaling parameter controlling the characteristic obsolescence rate, and $c_\times \equiv 7$ is a citation threshold, above which preferential attachment "turns on". A recent study has shown evidence for $c_\times$ on the order of 1 (Golosovsky & Solomon, 2013), and in a general analysis of network models, this offset parameter is supported against alternative models (Medo, 2014). Here we find that $c_\times = 7$ provides the best matching of the model and the empirical data with respect to the $P(\Delta_r)$ distributions shown in Fig. 7.

The obsolescence factor $[n(t_j)]^\alpha$ counteracts *PA*, because for two nodes with the same citation tally, the newer node will be preferentially selected – i.e. "crowding out" of the old by the new. The parameter α controls the rapidity of the obsolescence arising from temporal selection, as indicated by the ratio of the attachment rates between any two given publications with $t_{j'} \geq t_j$,

$$P_{j,t}/P_{j',t'} = \left(\frac{c_\times + c_{j,t}}{c_\times + c_{j',t}}\right) \exp[-\alpha g_n(t' - t)]. \tag{3}$$

As a result, the relative attachment rate between any two given publications with the same number of citations but with a difference in age, $t_{j'} \geq t_j$, is given by $P_{j,t}/P_{j',t'} = \exp[-\alpha g_n(t' - t)]$. This crowding out provides a mechanistic explanation for the obsolescence of knowledge. While preferential attachment will also produce obsolescence for most nodes due to "first mover advantage", it is not sufficient alone to reproduce stylized facts associated with obsolescence in real citation networks (Medo, Cimini, & Gualdi, 2011). Thus, while PA facilitates selection according to c_j, our model provides an additional selection according to t_j. In this way, our mechanism explains the exponential decay of the citation life-cycle $\Delta c(\tau|t)$ in terms of the exponential growth of the system, g_n; e.g. see Fig. 9(D) in Ref. Parolo et al. (2015).

Third, following the citation of j, a random number of publications from the reference list of j are also cited. This redirection step provides a "backdoor" to overcome the obsolescence induced by the growth of the system, since articles in the reference list are likely to be older.

In addition to reproducing the trends in $F(c \leq C|t)$, $G(t)$, and $P(\Delta_r)$, our model reproduces numerous other stylized facts representing both static and dynamic features of the empirical citation network: (i) the mean citation lifecycle is found to decay exponentially (Parolo et al., 2015); (ii) there is a relatively high mean clustering coefficient typical of real citation networks (Holme & Kim, 2002; Ren et al., 2012); (iii) we demonstrate that our model reproduces the increasing citation share of the top cited percentile of publications (Barabasi et al., 2012), shown in Fig. 9(G); and (iv) we demonstrate that our model's citation distribution is log-normally distributed (Radicchi, Fortunato, & Castellano, 2008), which we demonstrate by normalizing citation counts within age cohort according to the logarithmic transform, giving the normalized citation impact

$$z_{p,t} = (\log(c_{p,t}) - \mu_{LN,t})/\sigma_{LN,t} \tag{4}$$

where $\mu_{LN,t} = \langle \log(c_{p,t}) \rangle$ and $\sigma_{LN,t} = \sigma[\log(c_{p,t})]$ are the mean and the standard deviation of the logarithm of $c_{p,t}$ calculated across all p within each t. Fig. 9(F) shows that the citation distribution $P(z)$ is well-fit by the normal distribution, thereby confirming the log-normal distribution of $P(c_{p,t})$. Drawing on this feature, Fig. 8(B) uses a visualization scheme whereby each node size is proportional to the normalized citation impact z_p, and as a result, there is no visible temporal trend in the size distribution of the nodes. In this way, we demonstrate how an appropriate normalization that leverages the underlying statistical regularities of the data generating process can be useful for cross-temporal comparison, i.e. longitudinal panel regression (Petersen, 2015, 2018; Petersen & Penner, 2014).

We discuss these empirical benchmarks in further detail in Appendix A, where we leverage the generative capacity of the citation network model to explore the effects of sudden perturbations of the system parameters during the network's evolution. In what follows we discuss the two most relevant perturbation scenarios.

First, we tested the synthetic citation network's response to a sudden increase in the parameter

$$\beta = r^b(t)/r(t) \tag{5}$$

which controls the rate of redirected referencing ("hyperlinking") via the redirection step (b) illustrated in Fig. 8. The sudden perturbation $\beta = 0.2 \rightarrow 0.4$ at $t = 165$ (see Fig. 10 – panel column 3) is meant to represent a shift from 1 in 5 citations to 2 in 5 citations occurring via mechanism (b), thereby simulating the sudden emergence of online journals that facilitate "reference hopping". We are unaware of any behavioral studies that would provide empirical guidance in the selection of β, and so we choose a sizable perturbation that corresponds to reasonable yet measurable shift in individual citing behavior. This particular perturbation is inspired by Evans' study (Evans, 2008), which used an econometric regression to estimate the impact of online journal archives on citation patterns. More specifically, Evans (2008) reports a negative relation between reference distance and online journal availability, concluding that the observed narrowing in historical attention was facilitated by the emergence of "hyperlinks" that allow one to more easily browse prior literature, arguing that the hyperlinks tend to lead to more recent rather than more distant literature. However, in contrast to Evans' conclusion (decreasing Δ_r with increased hyperlinking), our model shows the opposite – when we increase the rate of the redirection process (mimicking hyperlinking), we observe a decrease in *both* the frequency of $\Delta_r < \Delta_r^-$ and $\Delta_r > \Delta^+$, consistent with the empirical shifts we report in our analysis of $P(\Delta_r)$. Thus, our model suggests that the magnitude of the hyperlink effect reported in Evans (2008) is relatively small compared to the overall counter-effect of increased attention to the mid-field induced by the growth of the system. Also, this perturbation causes a decrease in $C(q|t)$ and an increase in $G(t)$, because more

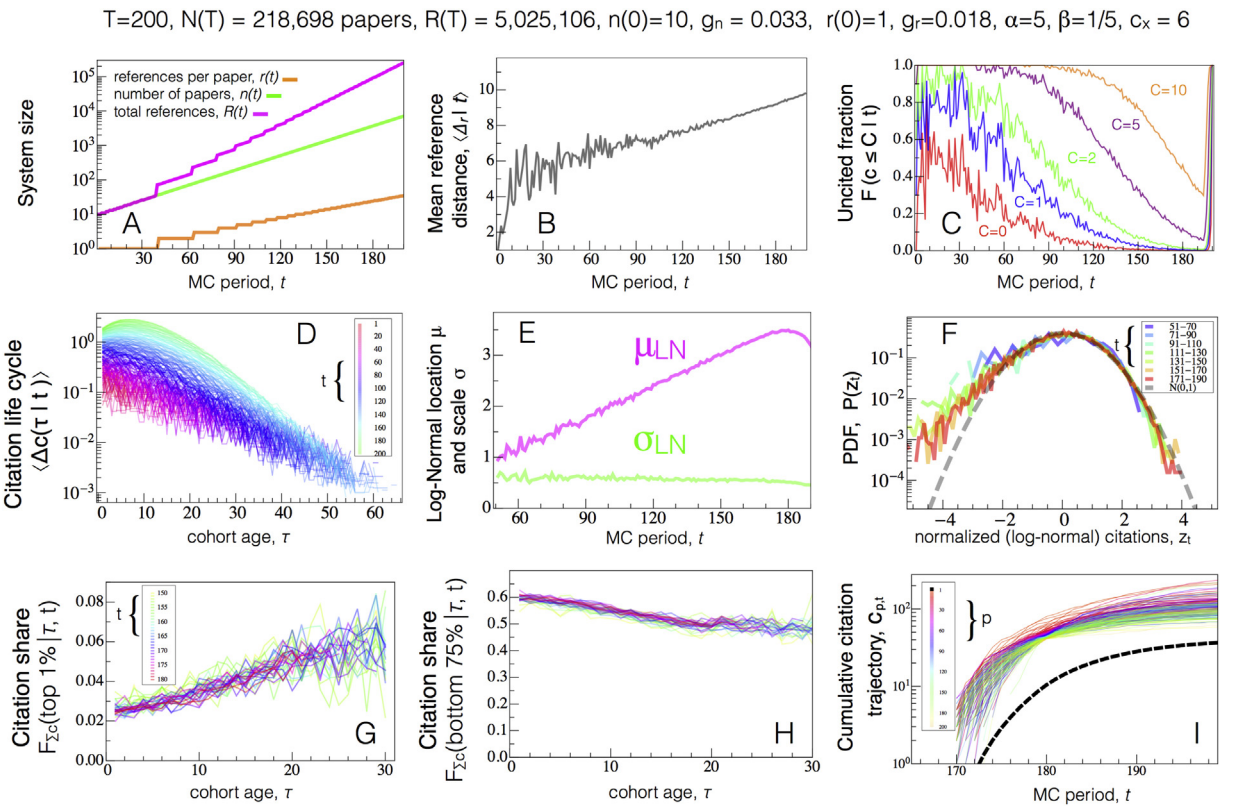

Fig. 9. Monte-Carlo growth parameters and benchmark validation. Shown are various properties of the synthetic citation network that can be compared with empirical trends. We evolved the simulation using the parameters: $T \equiv 200$ MC periods (~years), $n(0) \equiv 10$ initial publications, $r(0) \equiv 1$ initial references, exponential growth rates $g_n \equiv 0.033$ and $g_r \equiv 0.018$, secondary redirection parameter $\beta \equiv 1/5$ (corresponding to $\lambda = 1/4$), citation offset $C_\times \equiv 6$, and life-cycle decay factor $\alpha \equiv 5$. At the final period $t = T$, the final cohort has size $n(T) = 7112$ new publications, $r(T) = 35$ references per publication, and final citation network size $N(200) = 218{,}698$ publications (nodes) and $R(T) = 5{,}025{,}106$ total references/citations (links). (A) The size of the system in each MC period t. (B) Growth of the mean reference distance $\langle \Delta_r \rangle$. (C) The fraction $f_{c \le C}(t|\tau = 5)$ of publications which have C or less citations at cohort age $\tau = 5$. (D) The citation life cycle, measured here by the mean number of new citations τ periods after entry (publication). The different curves correspond to the publication cohort entry period t. For sufficiently large t the life cycle decays exponentially. (E) Growth of the logarithmic mean (location) value $\mu_{LN,t}$ and the relative stability of the logarithmic standard deviation (scale) value $\sigma_{LN,t}$. $\mu_{LN,t} = \langle \log(c_{p,t}) \rangle$ and $\sigma_{LN,t} = \sigma[\log(c_{p,t})]$ are the logarithmic mean and standard deviation calculated across all p within each age cohort t. (F) The distribution $P(z_{p,t})$ of the normalized citation impact $z_{p,t}$. For visual comparison we plot the Normal distribution $N(\mu = 0, \sigma = 1)$. (G) The increasing citation share $f_{\sum_c}$ – the fraction of the total citations received by all publications from cohort t – of the top 1% of publications from cohort t (ranked at cohort age $\tau = 10$). (H) The decreasing citation share $f_{\sum_c}$ of the bottom 75% of publications. (I) The cumulative citation count $c_p(t)$ of the top 200 publications (p) from the interval $t = [170, 179]$, ranked according to $c_p(t = 180)$. The dashed line represents the average citations for p from the same cohort over the same period.

references are redirected to older publications as demonstrated by the shifts in $F(c \le C|t)$, $P(\Delta_r|t)$ and $CDF(\ge \Delta_r|t)$ towards Δ_r in the mid-field range $[\Delta_r^-, \Delta_r^+] \approx [8, 45]$ years.

Second, we simulated a perturbation representing a sudden increase in the reference list growth rate, $g_r \to g_r + \delta g_r$, by suddenly increasing g_r from 0.013 to 0.019 at $t^* = 165$ (see Fig. 10 – panel column 4). This perturbation models the sudden influx of online-only "mega-journals", which are more flexible in terms of text length and also reference list length. For comparison, a similar perturbation is shown in the third panel column of Fig. 12 which is simulated using β=0. As far as the $P(\Delta_r)$ distribution, we observe that this second perturbation to g_r has the same qualitative impact as the first perturbation to β. Interestingly, however, the perturbation to β increased the citation inequality $G(t)$ whereas the perturbation to g_r decreased the citation inequality $G(t)$.

We direct the detail-oriented reader to additional description of the model and additional analysis of other perturbation scenarios in Appendix A.

6. Summary and discussion

The science citation network is a rich source of opportunities to model the structure and dynamics of knowledge creation. Yet the same dynamics also pose a challenge for those attempting to extract value from the citation network and other information networks. For example, information service providers must develop tools to store and facilitate searching such vast, sparsely connected, and growing information networks. In turn, the clients, in our case researchers, are also confronted

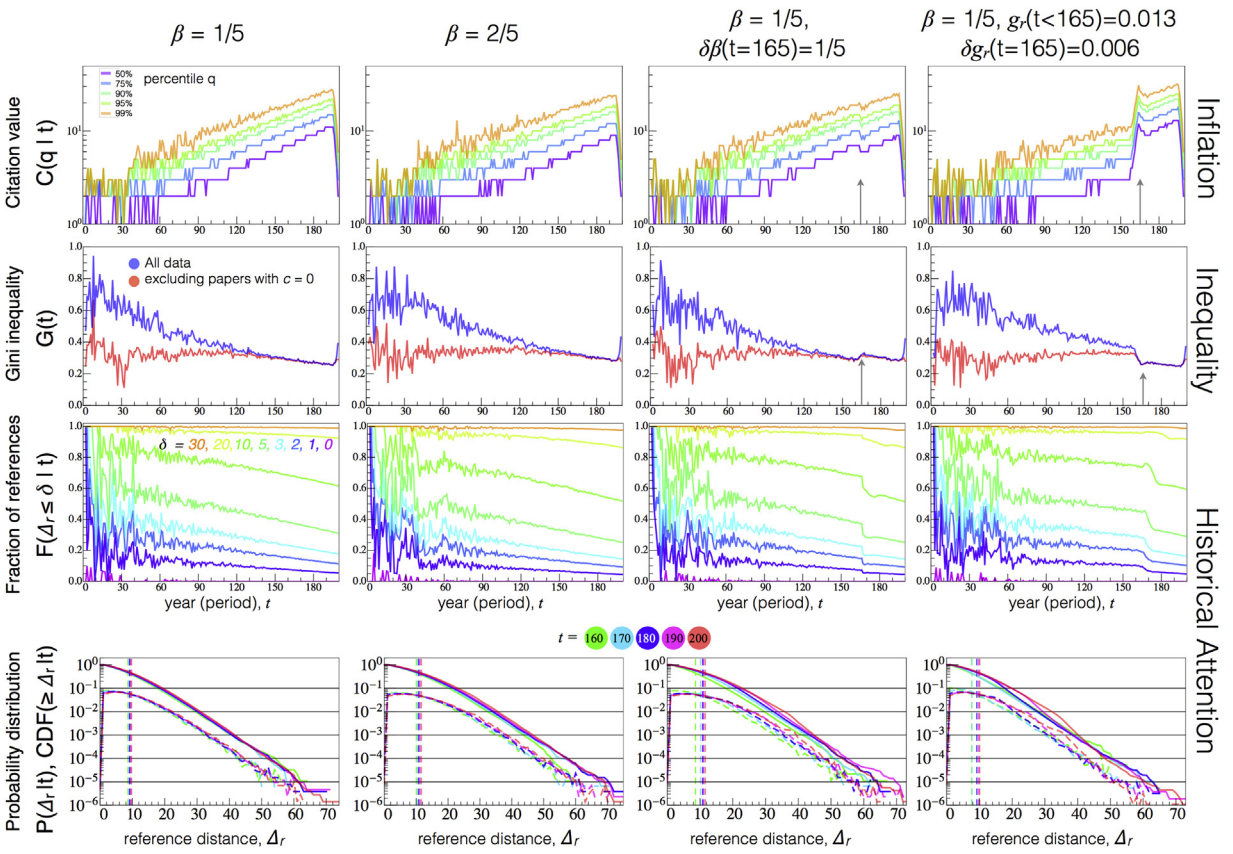

Fig. 10. Monte-Carlo simulation of the science citation network: with the redirection mechanism ($\beta > 0$). Each column represents a different modeling parameter set: the first and second columns differ only in the β value; the third column represents a perturbation at $t = 165$ from $\beta = 1/5$ to 2/5; and the final column represents the scenario with $\beta = 1/5$ where the reference list growth rate g_r is boosted from 0.013 to 0.019 at $t = 165$. (First row) Inflation demonstrated by the persistent exponential growth of the citation value $C(q|t)$ corresponding to the quantile q indicated in the plot legend. For example, the citation value corresponding to the 99th percentile grows from roughly 3 in $t = 30$ to 30 in $t = 195$ for the unperturbed simulations with $\beta = 1/5$ and $\beta = 2/5$. (Second row) The Gini index $G(t)$ of the total number of citations after 5 years from publication measures the citation inequality of the citation distribution. The model also indicates that the decreasing inequality is largely due to the decreasing proportion of uncited publications. Nevertheless, for large t the fraction of uncited publications is approximately zero, and so the decline in $G(t)$ is also induced by the growth of the system. (Third row) The fraction $F(\Delta_r \leq \delta|t)$ of references from year t going to publications within the interval $[t-\delta, t]$ shows nonlinear behavior in the perturbed systems, with sharp declines indicating that the perturbation causes a significant fraction of references to be directed back further than δ years in the past. These results capture the complex effects of growing sources of references and the subsequent crowding out of old publications by new publications. (Fourth row) The cumulative distribution $P(\geq \Delta_r|t)$ (solid lines) and the probability density function $P(\Delta_r|t)$ (dashed lines) of reference distance Δ_r, conditional on the publication cohort t. Vertical lines indicate the mean of each conditional distribution for varying t. To improve the data size, each $P(\Delta_r|t)$ and $CDF(\geq \Delta_r|t)$ are calculated by pooling the reference data from the 3-period interval $[t-2, t]$. For example, the scenario with constant $\beta = 1/5$ shows that medium Δ_r values becoming more frequent for $t > 160$. At the same time, recent publications corresponding to $\Delta_r \lesssim 4$ are being cited less and less. Other parameters used for each simulation are $T \equiv 200$ MC periods ($\sim$ years), $n(0) = 10$ initial publications, $r(0) \equiv 1$ initial references, exponential growth rates $g_n \equiv 0.033$ and $g_r \equiv 0.018$ ($g_R = g_n + g_r$), citation offset $C_\times = 6$, and life-cycle decay factor $\alpha \equiv 5$ so that $1/(\alpha g_R) \approx 4$ periods.

by the challenge of navigating the citation network in order to construct consistent theories, to validate empirical findings, and to aid in future problem selection, among other tasks. And as research assessment becomes increasingly quantitative, the citation network takes on additional relevance as the modus for inter-generational transfer of 'knowledge credit' that has become the focus of the citation economy in science – notwithstanding issues of 'citation inflation' which challenges the value of citations as a research evaluation metric (Petersen et al., 2018). These are just a few reasons why understanding the growth of scientific production and its implications merit increased attention.

Against this background, we studied the impact of growth on various properties of the science citation network. Our results show how growth in the number of new publications (nodes) and outgoing references (links), combined with an inhomogeneous distribution of reference distance, give way to the "crowding-out" of prior literature. In particular, this process compounds the natural obsolescence of knowledge, and may negatively impact the efficiency of search processes across the knowledge network and the efficiency of scientific progress on the knowledge frontier. The crowding out is rather evident in our model network visualization (Fig. 8), in which the thickening of each new layer significantly reduces the ability of would-be knowledge seekers to explore the more distant corners of the knowledge network given a finite number

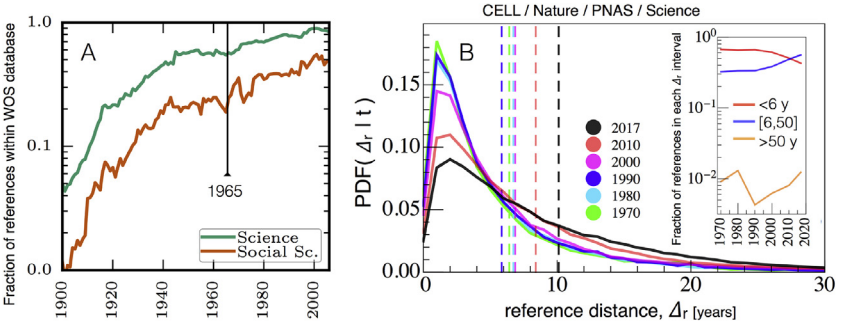

Fig. 11. Evolving coverage of the WOS database. (a) For each year and index, we calculated the fraction of references made by articles in the WOS database that cite articles also contained within the database. Since 1965 to present, the coverage has increased from roughly 65% to 90% for the natural sciences and roughly 20% to 50% for the social sciences. (b) Robustness of the reported trends in the reference distance distribution $P(\Delta_r)$ calculated using a balanced panel of high-impact journals (Nature, Science, PNAS, and CELL) which have been indexed by WOS since the 1970s; for comparison with Fig. 7. Shown are the probability distributions $P(\Delta_r|t)$ for select t indicated in the color legend; vertical dashed lines indicate distribution mean. (inset) The shifting range of scientific attention, demonstrated by the fraction of references in 3 non-overlapping intervals, $\Delta_r < 6$, $6 \leq \Delta_r \leq 50$, and $\Delta_r > 50$ years.

of jumps starting at the frontier. Thus, it is important to quantify and understand the trends in scientific attention in order to avoid, among other innovation inefficiencies, the syndrome of 'reinventing the wheel'.

One explanation for our empirical findings is the narrowing attention by scientific agents contributing to the knowledge frontier (see Fig. 2). This narrowing breadth and depth may be the result of various factors. First, in some areas scientific training may be increasingly focused on specialization during the doctoral training period, often to prepare for careers in large laboratory environments, together marking the end of the solo "renaissance" genius era (Simonton, 2013) and the steady emergence of the team science era (Milojevic, 2014; Pavlidis, Petersen, & Semendeferi, 2014; Petersen, Pavlidis, & Semendeferi, 2014; Wuchty, Jones, & Uzzi, 2007). Second, the deluge of new literature may push researchers to the limits of their individual cognitive abilities to browse and full digest new information. Whether scientific exploration occurs by jumping between reference lists or by using search engines, the channels to distant literature become increasingly narrow relative to the channels to more contemporary literature. Third, an increasing focus on production over consumption also means that researchers will spend more time writing and reviewing and less time reading and digesting the literature in the first place. Lastly, there may be narrowing scope due to the focusing power of circles of social influence which may affect the publication and referencing process. Indeed, the increasing prevalence of reputation-based systems in science, both individual organizational, may be the result of relying on reputation as a heuristic tool to aid in the process of filtering through the vast amounts of new literature (Petersen, Fortunato, et al., 2014).

Nevertheless, handling the overwhelming volume of knowledge required to make scientific advancement may, in part, be overcome by the division of labor. Indeed, the number of coauthors per publication, a proxy for team size, has also shown a persistent 4% annual growth over the last half century in the natural sciences (Pavlidis et al., 2014; Petersen, Pavlidis, et al., 2014). Moreover, new technologies for accessing, crowdsourcing (e.g. *Wikipedia.org*), searching, retrieving, exploring, storing, and organizing knowledge (Sparrow et al., 2011) could in principal fully counterbalance the aforementioned trend towards specialization, cognitive limits and social barriers.

Our study contributes to the literature by addressing two outstanding disagreements: (a) the level of citation inequality across publications (Acharya et al., 2014; Evans, 2008; Lariviere et al., 2009; Petersen & Penner, 2014) and (b) the obsolescence of knowledge (Evans, 2008; Verstak et al., 2014; Wallace et al., 2012). We provided clarity on these two issues by performing an in-depth analysis of (i) the entire citation distribution, as measured by the Gini coefficient $G(t)$, and (ii) the entire range of reference distances, quantified by the probability distribution $P(\Delta_r)$. In the case of $G(t)$, we found that the decreasing inequality in the number of citations received is largely due to the simultaneous decreasing trend in uncitedness. In the case of $P(\Delta_r)$, we found that the trends are rather nuanced, and thus susceptible to misinterpretation if simple yet biased summary statistics (e.g. mean values) are the focus. We found that the fraction of references to literature located in the intermediate or mid-field range of Δ_r is increasing, accompanied by a decline in the attention to both very recent and very distant literature.

This last point is distinguished as our most important finding. Unexpectedly, we found that the largest decline in attention was to the most recent literature – i.e. to articles less than 6 years old ($\Delta_r < 6$). One possible explanation is that researchers are not able to consume nearly as fast as they produce. This trend may reflect social factors associated with the rapid growth of $n(t)$. Instead of reading every new publication, researchers may increasingly depend on individual heuristics to determine an article's relevance and authority. Individual assessment is complemented by the "wisdom of the crowd" to collectively crowdsource the quality of research – i.e. as proxied by citation counts – which is a collective evaluation process that can take several years to accumulate, thereby slowing down the "digestion rate" of academic literature.

Accompanying this trend at the lower end of the reference distance distribution, we also found that the fraction of reference distances greater than Δ_r^+ to be decreasing for each subject area up until the 2000s. However, since then, we observe an upturn consistent with an increasing accessibility of older literature via efficient reference-list hyperlinking (see

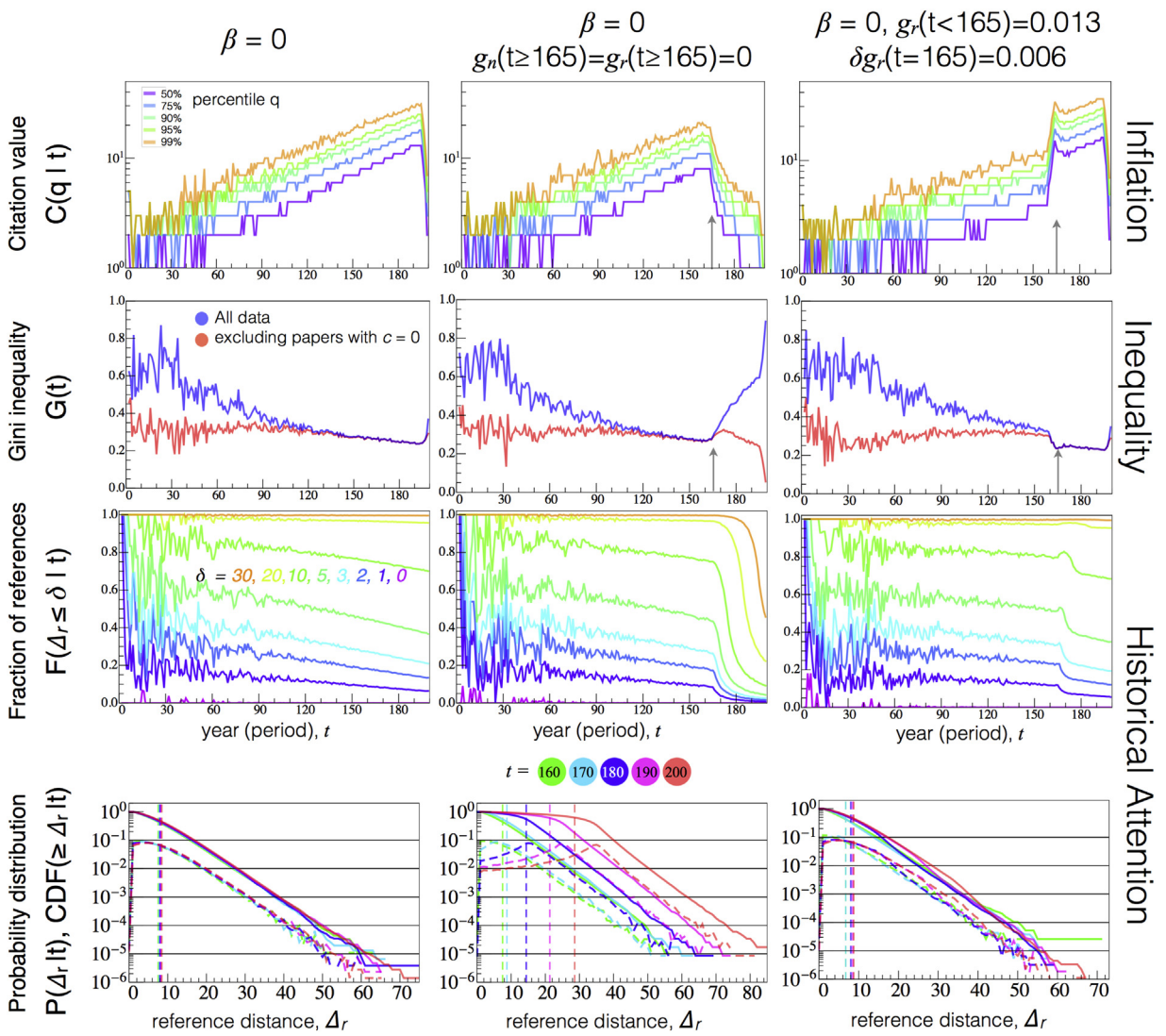

Fig. 12. Monte-Carlo simulation of the science citation network: without the redirection mechanism ($\beta = 0$). We benchmark the model using empirical trends observed in the real citation data. For each simulation, we use $\beta = 0$ meaning that there is no redirection mechanism. The only difference between the columns is in the growth of the system: in column 2 the growth rate of the system is quenched at $t = 165$ so that $n(t \geq 165) = n(164)$ since $g_n(t \geq 165) = 0$ and $r(t \geq 165) = r(164)$ since $g_r(t \geq 165) = 0$; in column 3 the growth rate of the reference lists is boosted at $t = 165$ so that $g_r(t < 165) = 0.013$ and $g_r(t \geq 165) = 0.019$. (First row) Inflation demonstrated by the persistent exponential growth of the citation value $C(q|t)$ corresponding to the quantile q indicated in the plot legend. For example, the citation value corresponding to the 99th percentile grows from roughly 10 in $t = 90$ to 30 in $t = 190$ for the unperturbed simulation. (Second row) The Gini index $G(t)$ of the total number of citations after 5 years from publication measures the citation inequality of the citation distribution. (Third row) The fraction $F(\Delta_r \leq \delta|t)$ of citations from year t going to publications within the interval $[t - \Delta_r, t]$ are all decreasing; the sharp decline for the perturbed growth scenario shows the importance of growth on sustaining attention to the recent literature base. (Fourth row) The cumulative distribution $CDF(\geq \Delta_r|t)$ (solid lines) and the probability density function $P(\Delta_r|t)$ (dashed lines) of reference distance Δ_r, conditional on the publication cohort t. Vertical lines indicate the mean of each conditional distribution for varying t. To improve the data size, each $P(\Delta_r|t)$ and $CDF(\geq \Delta_r|t)$ are calculated by pooling the reference data from the 3-period interval $[t - 2, t]$.

Fig. 7B). The crossover value varies by subject area with $\Delta_r^+ \approx 50$ y (Science), $\Delta_r^+ \approx 20$ y (Social Sciences) and $\Delta_r^+ \approx 40$ y (A&H). We incorporated this hyperlinking mechanism into our generative network model as a redirection process facilitating link formation.

Beyond Δ_r^+ is the very distant literature, the attention to which is captured by $CDF(\gg \Delta_r^+|t)$. Fig. 7(A) shows a decline in attention to very distant literature. This observation may appear to be in disagreement with results of refs. (Verstak et al., 2014; Wallace et al., 2012) which apply the same method. However the discrepancy is likely due to the fact that these analyses only investigated the citation trends for $\Delta_r \leq 10$, 15, and 20 years, and so they did not investigate sufficiently large "trans-generational" Δ_r to access the trends in the very classic literature – i.e. literature authored by scientists that are no longer active. Another explanation for the decline in fraction of references to $CDF(\gg \Delta_r^+|t)$ is likely due to technological turnover and the emergence of new disciplines which have relatively few foundational publications to reference – a branching process

of innovation that is not captured by our model. The content of academic literature also evolves differently in A&H versus Science: in the former, references are often historical or artifactual in context, referring to quasi-static representations, as opposed to the dynamic concepts and methods that can rapidly evolve in the latter.

Another interesting feature of $P(\Delta_r|t)$ is the fixed-point $P(\Delta_r^-|t)$, which appears to be stable across time (t). Thus, because of this stability, the reference distance $\Delta_r^- \approx 6$ y can be used to classify knowledge as recent ($\Delta_r \leq \Delta_r^-$) or contemporary ($\Delta_r^- \leq \Delta_r \leq \Delta_r^+$), representing a fundamental time scale characterizing the advancement of the scientific endeavor. Similarly, although less precise, a second crossing point Δ_r^+ in the cumulative distribution $CDF(\geq\Delta_r)$ distinguishes the classic literature ($\Delta_r \geq \Delta_r^+$). Complementary to the aforementioned trends, Fig. 7(B) shows that the fraction of references distances falling into the range $[\Delta_r^-, \Delta_r^+]$ has steadily increased, growing by roughly 24% in Science, 30% Soc. Sci., and 19% in A&H over the last 50 years, corresponding to a narrowing of historical attention around the mid-field.

Despite the consistency in the trends across the three broad domains analyzed, the dynamic coverage of the WOS database limits our ability to draw absolute conclusions from our empirical findings. This follows because over time journals enter and exit the WOS index, and as a result the set of references in a given year depend not only on the growth of science, but also on the coverage of the WOS. Fig. 11 illustrates this dynamic coverage, showing the fraction of outgoing references from articles indexed by the WOS that cite articles indexed by the WOS; this fraction has been increasing steadily since 1965 and is approaching 90% coverage for Science and 50% coverage for Social Sciences. Our analysis of outgoing citations are the least impacted by the dynamic coverage, since we were able to calculate Δ_r for all items in the reference list that indicate a publication year t_r, regardless of whether the cited publication is included in the WOS. In order to demonstrate the robustness of our results, we analyzed a select set of journals that have been indexed by the WOS since the 1970s (see Fig. 11(B)) whose outgoing citations are more likely to be contained within the WOS index rather than being directed at more obscure research not indexed by WOS. Using this balanced journal subset we analyzed the reference-age distribution $P(\Delta_r)$ and do not observe qualitatively different patterns as compared to the aggregate unbalanced set of "Science" articles. In the case of incoming citations, our estimate of c_p for each p indexed in the WOS uses only the citations arising from other sources contained in the WOS dataset. That is, the citation count c_p is a lower bound estimate of the true c_p arising from all articles every published, indexed or not by the WOS. Again, we tested the robustness of our results by analyzing a select set of high-impact journals, the logic being that the true c_p for such high-impact articles would only be a small marginal increase over the WOS estimate for c_p. Fig. 5 shows the results of our citation inequality analysis on this subset of high-impact journals, which again demonstrates the same qualitative patterns as the aggregate unbalanced data. Since it is beyond the scope of our focus, we leave it as an open problem to compare the citation distributions using citations derived only from a balanced set of journals to compare with the citation distribution calculated from the entire set of journals comprising the WOS index.

We also developed a generative citation model to facilitate exploring the impacts of technological (e.g. web-based article search and bibliography registry tools; web-based manuscript preparation and submission tools) and industrial (e.g. new layer of mega-journals) shifts in the production of scientific literature, and the subsequent impact they might have on researcher citing behavior. We implemented our model using Monte Carlo simulation and validated its results by reproducing numerous well-established features, both static and dynamic, of the empirical science citation network. Despite the model's apparent success in capturing the essential features of the science citation network, it has some clear limitations which are also worth discussing. First, we do not incorporate the intrinsic quality of new publications nor any other node features (e.g. journal, authors, author affiliations) meaning that our model lacks heterogeneity in the intrinsic fitness of each p, a factor which can explain the extremely wide variation in long-term citation impact of individual publications (Wang et al., 2013). And second, the model lacks social factors, such as author-specific effects such as reputation (Petersen, Fortunato, et al., 2014), collaboration (Barabasi et al., 2002; Petersen, 2015; Petersen, Pavlidis, et al., 2014), and self-citation (Wallace et al., 2012), which are inextricable features of the science system that lead to important correlations in the coevolutionary dynamics of the citation network. In particular, the tendency for authors to self-cite, in other words their heightened awareness of their own work, may affect the distribution of Δ_r in both the short and the long term. By incorporating a social layer into our model, one which captures self-citation, we might be able to better match the model predictions to the empirical $P(\Delta_r|t)$ distributions in the small Δ_r regime.

With these limitations in mind, our primary goal was to use the model to determine the response of the synthetic science system to four scenarios operationalized as modifications and sudden perturbations of the model parameters:

(i) no redirection mechanism ($\beta = 0$), see Fig. 12;
(ii) a sudden perturbation increasing β after t^*, thereby increasing the frequency of the redirection process (b) relative to the direct process (a), representing the technological innovation of "hyperlinking" (Evans, 2008), see Fig. 10;
(iii) a sudden perturbation to g_r representing a sudden increasing of reference list lengths after t^*, see Fig. 10;
(iv) a sudden perturbation to g_n, causing either a decrease or increase in the growth rate of the system after t^* (see Fig. 12).

We discuss these *in silico* experiments in further detail in Appendix A.

In all, these MC simulations provide various insights into the evolution of the citation network that are not possible otherwise, since closed-form analytic methods are typically not tractable for such time-dependent network growth models. We summarize our four principal model-oriented findings as follows. First, our simulations indicate how $n(t)$, the "crowding-out" factor included in Eq. (2), can explain the exponential decay of the citation lifecycle. For a visual example, this "crowding out"

is indicated in Fig. 8(A) where the publications from the final period $t = 150$ (bright yellow) are already prominent compared to the entire corpus of the preceding 149 periods. Second, our model clarifies how the narrowing of attention towards the mid-field counteracts the positive feedback arising from the preferential attachment, together producing realistic citation (knowledge) life-cycles that peak and then decay exponentially after a time scale $\sim 1/(\alpha g_R) \approx 4$ periods (see Fig. 9). Third, redirected references represent a mechanism to overcome the implicit citation lifecycle induced by the growth of $n(t)$ and $r(t)$, because the references that arise from the redirection process cite older literature on average. Fourth, we also find that the growth of $n(t)$ is necessary for sustaining the citation of recent literature, as indicated by the perturbation $g_{n,t} \to 0$, which results in a sudden decline in the number of citations received and a sudden increase in the citation inequality $G(t)$, as demonstrated in Fig. 12.

In conclusion, our empirical analysis demonstrate various ways in which the growth of the scientific endeavor impacts the structure of the science citation network. Going forward, models of science (Fortunato et al., 2018; Scharnhorst et al., 2012) will be important tools for providing both mechanistic insights and practical guidance on research evaluation (Petersen et al., 2018; Wildson, 2015; Wilsdon et al., 2015) and science policy (Fealing, 2011). Because collective attention translates into long-term memory formation, it is important to understand the mechanisms underlying the attention economy in science. And more broadly speaking – an increasingly relevant question is how individuals, organizations, and societies will find the right balance between breadth and depth of (historical, cultural, and technical) knowledge in the age of information overload.

Author contributions

All authors conceived the research. RKP, AMP, and SF designed the analyses. RKP, AMP, and SF conducted the analyses. All authors wrote the manuscript.

Acknowledgements

The authors are grateful for expert comments and recommendations for improvement from two reviewers, and for helpful discussions with A.-L. Barabási, A. Bonaccorsi and O. Penner. AMP and FP acknowledge financial support from the Italian Ministry of Education, PNR Crisis Lab, www.crisislab.it. The authors also acknowledge the opportunity to receive feedback via COST Action TD1210 "KnowEscape." Certain data included herein are derived from the Science Citation Index Expanded, Social Sciences Citation Index and Arts & Humanities Citation Index, prepared by Clarivate Analytics.

Appendix A. Model description

Network growth model: Our model is implemented in a strict network framework. That is, a fixed number of references are produced each year which are then allocated as links between the $n(t)$ incoming and the $\sum_{t'=0}^{t-1} n(t')$ existing nodes up to year t. This is unlike other citation "network growth" models in which stochastic differential equations are estimated and integrated for the average "mean field" node, meaning that node-to-node correlations are not taken into account (Golosovsky & Solomon, 2012, 2013, 2014; Peterson et al., 2010; Wang et al., 2013). More importantly, we take the systemic route to simulating the citation network because it allows us to explicitly control $r(t)$ and $n(t)$, which is the starting point for determining the impact of the exponential growth of science on citation patterns. It also allows us to monitor the Δ_r produced by our model, as well as other dynamic features of the citation network which we are able to match to empirical benchmarks.

In particular, the model presented here is inspired by the Peterson et al. (2010) citation network model, but with an enhanced triadic closure mechanism similar to Wu and Holme (2009). The key development in our model is the fact that each new publication i adds not only j to its reference list, but also adds a random number of x_{Binomial} indirect references – thereby contributing x_{Binomial} triadic closures within the citation network. Thus, one crucial difference in our link-formation model is that the triadic closure rate is relatively high. This is measured by calculating the mean clustering coefficient $\bar{\psi}$: for the $N(T = 150) = 41{,}703$ nodes entering the network in the interval $t \in [1, 150]$ comprising $L(T = 150) = 379{,}454$ links, we calculate $\psi = 0.018$.

This ψ value is significantly larger than the baseline clustering coefficient value $q = 2N/(L^2 - L)$ expected of a random network. However, the $\bar{\psi}$ value produced by our model is significantly smaller than the values $\psi = 0.31$ (ArXiv) and $\psi = 0.17$ (PNAS) calculated for empirical citation networks (Ren et al., 2012). This discrepancy is resolved when taking into consideration the discrepancy in the network sizes and the data used, as only partial snapshots of the citation network from highly correlated publication subsets are used in (Ren et al., 2012). Thus, because our network is comprised of the entire system, the clustering coefficient is undoubtably expected to be considerably smaller, especially considering the limitations in the triadic closure due to small $r(t)$ for small t.

Another important, yet oftentimes overlooked, feature of our model is the explicit exponential growth of the system, controlled by the parameters g_n and g_r. And finally, by including $n(t)$ in the link dynamics governed by Eq. (2), we are able to reproduce non-monotonic attention life-cycles, even for the highly cited nodes (publications), which is not produced by

pure *PA* models in which the hubs grow uncontrollably. As a result, the node attachment rate (here representing citation rates $\Delta c_p(t)$) eventually decay exponentially in our model, as they typically also do in real systems.

We model the network growth using Monte Carlo (MC) simulation based around the birth (publication) of $n(t)$ nodes (publications) in time step $t \geq 0$. The scientific publication base starts from a tiny seed at $t = 0$ with $n(t = 0) \equiv 10$ disconnected nodes. The number of new publications in each MC period t is $n(t) = n(t = 0)\exp[g_n t]$, where $g_n \equiv 0.033$ is the node (publication) growth rate, using the approximate empirical values reported here for the Science. Similarly, the number of outgoing links (references) $r(t)$ per node (publication) also grows exponentially, $r(t) = r(0)\exp[g_r t]$, using the empirical growth rate value $g_r \equiv 0.018$ and initial reference list length $r(0) \equiv 1$ references. Using these growth parameters, we sequentially add cohorts of $n(t)$ nodes to the network over $t = 1, \ldots, T$ periods according to the following prescription:

Network growth rules:

1. **System growth:** In each period t, we introduce $n(t)$ new publications (nodes), each with a reference list of length $r(t)$. As such, the total number of references produced in t is $R(t) = n(t)r(t)$. Also, since the seed publications from period $t = 0$ are foundational, they have reference lists of length 0.
2. **Link dynamics:** For each new publication $i \in n(t)$:

   (a) Direct citation $i \rightarrow j$**:** Each new publication i starts by referencing 1 publication j from period $t_j \leq t_i$ (where $t_i = t$ by definition). The publication j is selected proportional to the weight $P_{j,t} \equiv (c_\times + c_{j,t})[n(t_j)]^\alpha$ given by Eq. (2). The factor $c_{j,t}$ is the total number of citations received by j, tallied at the end of period $t-1$, representing a linear preferential attachment link dynamics (Barabasi et al., 2002; Jeong et al., 2003; Peterson et al., 2010; Redner, 2005; Simon, 1955). The factor $n(t_j)$ is the number of new nodes introduced in cohort t_j of j, and represents the endogenous crowding out of old literature by new literature. The parameter $c_\times$ is a citation offset controlling for the citation threshold, above which preferential attachment "turns on" (Golosovsky & Solomon, 2012, 2013; Petersen et al., 2014). It is needed so that a sufficient number of references go to publications with $\Delta_r \lesssim 5$ in the citation distribution. In this way, the preferential attachment mechanism is less sensitive to p with for $c_p < c_\times$.

   (b) Redirection $i \rightarrow s(j)$**:** The referencing publication i then cites a random number x of the publications contained in the references list $\{s\}_j$ (of length S_j) of publication j. The probability of selecting x references is given by the binomial distribution

$$P(x = k) = \binom{S_j}{k} (q)^k (1-q)^{S_j - k} \tag{6}$$

   with success rate $q = \lambda/S_j$ and average value $\langle x \rangle = \lambda$. Thus, we select a random number $x_{\text{Binomial}(S_j, \lambda/S_j)}$ according to a prescribed rate of redirected citations per direct citation,

$$\lambda = r^{(2)}/r^{(1)} = \beta/(1-\beta) \tag{7}$$

   where $0 \leq \beta \leq 1$ is the prescribed fraction of total references $r(t)$ that are made according to step (b). In this way, on average, the total number of redirected references is $r^b(t) = \beta r(t)$ for any t. Once the sample size x is determined, the $s(j)$ are selected from the reference list proportional to the same weights used in step (a), $p_{s(j),t} = (c_\times + c_{s(j),t})[n(t_{s(j)})]^\alpha$, which again gives preference within this secondary redirection process to the references with larger $t_{s(j)}$. We do not allow i to add any given $s(j)$ more than once to its reference list. Note that the quantity $q = \lambda/S_j$ represents the Bernoulli trial success rate, which for small t can be greater than unity if $\lambda > S_j$; thus, for small t we use $q = Min(1, \lambda/S_j)$. For large t, encountering this scenario is not very likely, as most candidate j selected for large t have $S_j > \lambda$. In other words, this subtle modeling limitation is only important in the beginning of the simulation when $r(t) < \lambda$, and plays an insignificant role on the system evolution thereafter.

   (c) Stop according to fixed reference list length: The referencing process alternates between (a) and (b) until publication i has referenced $r(t)$ publications.
3. Repeat 2 for each new publication in period t. At the end of each t the weights $P_{j,t}$ are updated.
4. Perform steps (1–3) for $t = 1, \ldots, T$.

A.1 Model benchmarking and case studies of parametric perturbations

Testing the model against empirical benchmarks: The base parameter set we used are: $T \equiv 200$ MC periods (years), $n(0) = 10$ initial publications, $r(0) \equiv 1$ initial references, exponential growth rates $g_n \equiv 0.033$ and $g_r \equiv 0.018$, secondary redirection parameter $\beta \equiv 1/5$ (corresponding to $\lambda = 1/4$), citation offset $c_\times = 6$, and life-cycle decay factor $\alpha \equiv 5$. Each model realization was simulated to size $N(T) = 218,698$ nodes at the time of the stopping time $T = 200$, with final reference list length $r(T) = 35$. We manually explored the parameter space of $(\alpha, \beta, c_\times)$ to determine the combination which best reproduces various empirical regularities known for the structure of citation network and the dynamical citation patterns of individual publications.

We summarize a typical network produced by our model in Figs. 8–10. First, Fig. 9(A) shows the exponential growth of $n(t)$, $r(t)$, and $R(t)$ as determined by the empirical parameters g_n, g_r, and g_R. In Figs. 9(C)–(I) we visualize various quantities measured from the synthetic citation network, such as: the mean reference distance $\langle \Delta_r \rangle$, the fraction of uncited publications

$F(c \leq C|t)$, the mean citation lifecycle $\langle \Delta c(\tau|t)$ calculated across the p for the τ periods after the publication year t, the mean $\ln\mu$ and standard deviation σ_{LN} of the natural log of the number of citations received by p from each t, the log-normal distribution of the normalized citations z_t, the evolution of the citation share of the top and bottom percentile groups $F_{\sum_c}(Q|\tau, t)$, and some typical citation trajectories $c_{p,t}$ produced by the model. Second, Fig. 12 shows the results of a simulation without redirection ($\beta = 0$) while varying g_n and g_r And finally, Fig. 10 shows the model results while varying β and varying g_r.

Fig. 6 shows the fraction $F(\Delta_r \leq \delta|t)$ of references from year t falling within the time window $[t-\delta, t]$ years. The model also reproduces the decreasing $F(\Delta_r \leq \delta|t)$ for $\delta \leq 20$, with the fastest decay for $\delta = 5$, pointing to the fact that relatively new literature is being cited less and less, as a percentage, over time. This is also evident in the $P(\Delta_r|t)$ and $CDF(\Delta_r|t)$ panels shown in Figs. 12 and 10.

The log-normal distribution of citations – within cohort – is another statistical benchmark that our model reproduces. To show this, we took all the publications with $c_{p,t} > 0$ from a given cohort t and calculated the average $\mu_{LN,t} = \langle \log(c_{p,t}) \rangle$ and the standard deviation $\sigma_{LN,t} = \sigma[\log(c_{p,t})]$ of the log citation count. Fig. 9(E) shows the evolution of $\mu_{LN,t}$ and $\sigma_{LN,t}$. The normalized citation counts $z_{p,t}$ are appropriately scaled by these cohort-dependent location and scale parameters. Fig. 9(E) shows the probability distribution $P(z_t)$ for varying cohorts (grouped over 20-year intervals). Each $P(z_t)$ distribution collapses onto the baseline Normal distribution $N(\mu = 0, \sigma = 1)$, except for in the lower tail where the log transform of small citation counts does not behave well. The log-normal distribution is a fundamental benchmark, representing a universal pattern observed for the within-cohort normalized citation distribution (Radicchi et al., 2008).

We also show in Fig. 9(G) that the model reproduces the increasing dominance of the top 1% of publications from each cohort t, consistent with empirical findings within high-impact journals (Barabasi et al., 2012). To demonstrate this, we calculated $F_{\sum_C}(1\,\%|t, \tau)$ which is the fraction of the total citations $\sum_p c_{p,t}$ accrued by the top 1% of publications from cohort t in year τ (the top1% is determined by ranking the p according to $c_{p,t}$ after $\tau = 20$ periods), where τ is the number of years since publication. Fig. 9(H) shows how the complementary bottom 75% of publications loses its citation share over time, largely because the lower-cited publications have a shorter citation lifecycle. Fig. 9(I) shows a sample of the top-200 cited p (ranked at t=180) from the cohorts $t \in [170 - 180]$, demonstrating growth and mixing of citation trajectories, consistent with real citation curves (Petersen, Fortunato, et al., 2014).

Understanding the role of the key parameters. The size of each new cohort $n(t)$ plays a key role in mediating the intrinsic citation lifecycle induced by crowding out of old nodes by new node, as mediated by the parameter α, see Eqs. (2) and (3). This follows from the quantity $1/(\alpha g_R) \approx 4$ periods, which determines the natural decay time scale of a publications citation attractiveness and governs the balance between preferential attachment effect and the crowding-out effect. The parameter α cannot be too large or too small. If α is too small, then preferential attachment dominates and the share of references to recent publications is significantly smaller than what we observe empirically because there is no crowding-out effect. Contrariwise, if α is too large, then the citation lifecycle decays too quickly meaning too large of a concentration of references on recent literature, which also eliminates the increasing dominance of the top 1% because their citation rates also decay to zero for large τ. In this way, a smaller α can be seen as increasing the citation equality.

We used the value $\alpha \equiv 5$ which produces a peak citation timescale that is consistent with he empirical peak citation time scale of a few years found by Parolo et al. (Parolo et al., 2015). Furthermore, for $\tau \gg \tau_{peak}$ the mean citation rate $\langle \Delta c(\tau|t) \rangle$ of the typical p from t decays exponentially as demonstrated by the approximately linear decay when plotted on log-linear axes, see Fig. 9(D). The peak citation year also appears to decrease and the exponential decay rate appears to increase for larger t, signaling that obsolescence is increasing due to systemic 'crowding out'.

The citation offset $c_\times$ is also important for shifting the peak in $P(\Delta_r|t)$ to smaller values, around $\Delta_r = 2$ to 3 periods. This occurs because $c_\times$ draws more citations to newer publications since it effectively diminishes the role of preferential attachment for the newer nodes with $c < c_\times$. We found that the fraction $F(c \leq C|t)$ of nodes with less than C citations decreases faster with t with larger $c_\times$ values. For the value $c_\times \equiv 6$ we observed a rapidly decaying uncited fraction, $F(c \leq 0|t)$ which is notably faster than the empirical curves shown in Fig. 4. This discrepancy likely arises from the fact that our model does not incorporate intrinsic variation in publication "quality" (i.e. fitness (Wang et al., 2013)), which is analogous to all the p being drawn from a single journal with the same impact factor. In other words, the empirical $F(c \leq C|t)$ do not decay as quickly as the model, most likely because of impact factor heterogeneity – there are many more journals with low impact factor than with relatively large impact factor in the aggregate WOS dataset.

The parameter β controls the rate at which references are made according to the redirection process (b) relative to the random (with PA) referencing process (a) (see Fig. 8). These two processes correspond to the two modes of "searching" described by Evans (Evans, 2008), the first being a "browsing" mode, the second being a "redirection" mode that is facilitated by online journals and indexes that have 1-click "hyperlinks". Thus, our model allows us to mechanistically test Evans' theory about the impact of electronic media on citation "inequality" and narrowing scholarship in science. Since redirected citations on average go to older papers, this parameter can be used to tune the mean reference distance $\langle \Delta_r|t \rangle$, which in our model grows linearly in time, from roughly 6 periods in $t = 60$ to 10 in T, see Fig. 9(B). This feature is also verified by noting that increasing β also shifts the distribution $P(\Delta_r)$ towards larger Δ_r values in the bulk of the distribution.

1 **Simulation without the redirection parameter: $\beta = 0$.** Fig. 12 shows the results of the simulation when we excluded the redirection mechanism (b) by setting $\beta = 0$. Comparing the benchmarks in the first column of Fig. 12 for $\beta = 0$ with the first two columns of Fig. 10 for $\beta = 1/5$ and $\beta = 2/5$, respectively, the citation inflation quantified by $C(q|t)$, the citation

inequality quantified by $G(t)$, and the distribution of reference distances quantified by $F(\Delta_r \leq \delta|t)$, $P(\Delta_r|t)$ and $CDF(\geq \Delta_r|t)$ are qualitatively similar. There are, however, differences in the decay of $G(t)$, which is faster for the case $\beta = 0$ and the decay of $F(\Delta_r \leq \delta|t)$, which is faster for larger β, mainly because the redirection mechanism facilitates the citation of older publications. Also, the bulk of the distribution $P(\Delta_r|t)$ is wider as β increases, representing the redirection of references into the range $[\Delta_r^-, \Delta_r^+] \approx [8, 45]$.

We also performed three sudden parametric perturbations to gain a better understanding of how the inflation of science, attention inequality, and the obsolescence of knowledge are related.

2 **The role of system growth: quenching the system growth, perturbations of g_n and $g_r \to 0$.** The second column of Fig. 12 shows the results of perturbing the system at $t = 165$ by suddenly quenching the system growth: $g_n \to 0$ and $g_r \to 0$ for $t \geq 165$. These two perturbations mean that a constant $n(t \geq 165) = 2168$ number of new nodes (publications) are added each period (year) thereafter, with each publication having a constant reference list length $r(t \geq 165) = 18$. This simulation demonstrates how the growth of the system sustains the citation of recent publications, because after $t = 165$ the citation values $C(q|t)$ suddenly drop as most of the references are drawn to the highly cited publications because the crowding out effect is not as strong as when there is sustained growth. Moreover, the attention of the system becomes "frozen" on the articles from the last cohort before the quenching, as evident in the peaks of the $P(\Delta_r|t)$ distributions for Δ_r values coinciding with the $t = 165$. This implies that the growth is necessary to sustain the attention to the more recent literature. From a different perspective, it suggests that slowing the growth of science could increase the attention to older literature.

3 **Increasing the rate of "hyperlinking" by perturbing the redirection parameter β.** The third column Fig. 10 shows the result of suddenly increasing the rate of redirection from $\beta = 1/5$ to $\beta = 2/5$ at $t = 165$. This perturbation causes a decrease in $C(q|t)$ and an increase in $G(t)$ because more references are redirected to older publications as demonstrated by the shifts in $F(c \leq C|t)$, $P(\Delta_r|t)$ and $CDF(\geq \Delta_r|t)$ towards Δ_r in the mid-field range $[\Delta_r^-, \Delta_r^+] \approx [8, 45]$.

4 **Increasing the length of reference lists: perturbation of the growth rate parameter g_r.** The third column of Fig. 12 (with β=0) and the fourth column of Fig. 10 (with β=1/5) show the result of increasing the reference supply by suddenly increasing g_r from 0.013 to 0.019 at $t = 165$. This perturbation causes a significant increase in $C(q|t)$ and a decrease in $G(t)$ as the increasing supply of references is distributed to a larger share of the publications, demonstrating the relation between growth and decreasing inequality. Nevertheless, as in the perturbation of β, the increasing supply of references going through the redirection process (b) means that more references are redistributed to the mid-field range $[\Delta_r^-, \Delta_r^+] \approx [8, 45]$ of $P(\Delta_r|t)$. This follows because those publications which introduced right after the increase δg_r, which gained a slight advantage over the cohort preceding them; this is evident in the subsequent $P(\Delta_r|t)$ distributions which show a slight increase for Δ_r values coinciding with $t = 165$.

References

Acharya, A., Verstak, A., Suzuki, H., Henderson, S., Iakhiaev, M., Lin, C. C. Y., et al. (2014). *Rise of the rest: The growing impact of non-elite journals.* arXiv:1410.2217

Althouse, B. M., West, J. D., Bergstrom, C. T., & Bergstrom, T. (2009). Differences in impact factor across fields and over time. *JASIST*, *60*, 27–34.

Barabasi, A.-L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A*, *311*, 590–614.

Barabasi, A. L., Song, C., & Wang, D. (2012). Publishing: Handful of papers dominates citation. *Nature*, *491*, 40.

Bjork, B.-C. (2015). Have the 'mega-journals' reached the limits to growth? *Peer Journal*, *3*, e981.

Broad, W. (1981). The publishing game: Getting more for less. *Science*, *211*, 1137–1139.

Buldyrev, S. V., Growiec, J., Pammolli, F., Riccaboni, M., & Stanley, H. E. (2007). The growth of business firms: Facts and theory. *Journal of the European Economic Association*, *5*, 574–584.

Dixon, P. M., Weiner, J., Mitchell-Olds, T., & Woodley, R. (1987). Bootstrapping the Gini coefficient of inequality. *Ecology*, *68*, 1548–1551.

Egghe, L. (2010). A model showing the increase in time of the average and median reference age and the decrease in time of the price index. *Scientometrics*, *82*, 243–248.

Evans, J. A. (2008). Electronic publication and the narrowing of science and scholarship. *Science*, *321*, 395–399.

Fealing, K. H. (Ed.). (2011). *The science of science policy: A handbook*. Stanford CA, USA: Stanford Business Books.

Fortunato, S., Bergsrom, C. T., Borner, K., Evans, J. A., Helbing, D., Milojevic, S., et al. (2018). Science of science. *Science*, *359*, eaao0185.

Franck, G. (1999). Scientific communication – A vanity fair? *Science*, *286*, 53–55.

Gabel, A., Krapivsky, P. L., & Redner, S. (2013). Highly dispersed networks by enhanced redirection. *Physical Review E*, *88*, 050802.

Gabel, A., Krapivsky, P. L., & Redner, S. (2014). Highly dispersed networks generated by enhanced redirection. *Journal of Statistical Mechanics: Theory and Experiment*, *2014*, P04009.

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, *122*, 108–111.

Glänzel, W. (2004). Towards a model for diachronous and synchronous citation analyses. *Scientometrics*, *60*, 511–522.

Golosovsky, M., & Solomon, S. (2012). Stochastic dynamical model of a growing citation network based on a self-exciting point process. *Physical Review Letters*, *109*, 098701.

Golosovsky, M., & Solomon, S. (2013). The transition towards immortality: Non-linear autocatalytic growth of citations to scientific papers. *Journal of Statistical Physics*, *151*, 340–354.

Golosovsky, M., & Solomon, S. (2014). *Uncovering the dynamics of citations of scientific papers.* arXiv:1410.0343v1

Holme, P., & Kim, B. J. (2002). Growing scale-free networks with tunable clustering. *Physical Review E*, *65*, 026107.

Jeong, H., Neda, Z., & Barabasi, A. L. (2003). Measuring preferential attachment in evolving networks. *EPL*, *61*, 567.

Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences*, *112*, 7426–7431.

Klamer, A., & van Dalen, H. P. (2002). Attention and the art of scientific publishing. *Journal of Economic Methodology*, *9*, 289–315.

Krapivsky, P. L., & Redner, S. (2001). Organization of growing random networks. *Physical Review E*, *63*, 066123.

Krapivsky, P. L., & Redner, S. (2005). Network growth by copying. *Physical Review E*, *71*, 036118.

Lariviere, V., Archambault, E., & Gingras, Y. (2008). Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900–2004). *JASIST*, *59*, 288–296.

Lariviere, V., Gingras, Y., & Archambault, E. (2009). The decline in the concentration of citations, 1900–2007. *JASIST*, *60*, 858–862.
Medo, M. (2014). Statistical validation of high-dimensional models of growing networks. *Physical Review E*, *89*, 032801.
Medo, M., Cimini, G., & Gualdi, S. (2011). Temporal effects in the growth of networks. *Physical Review Letters*, *107*, 238701.
Milojevic, S. (2014). Principles of scientific research team formation and evolution. *Proceedings of the National Academy of Sciences*, *111*, 3984–3989.
Nakamoto, H. (1988). Synchronous and diachronous citation distributions. In L. Egghe, & R. Rousseau (Eds.), *Informetrics 87/88: Select proceedings of the 1st international conference on bibliometrics and theoretical aspects of information retrieval* (pp. 157–163). New York: Elsevier.
Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). *The PageRank citation ranking: Bringing order to the Web*. Technical Report Standford University.
Parolo, P. D. B., Pan, R. K., Ghosh, R., Huberman, B. A., Kaski, K., & Fortunato, S. (2015). Attention decay in science. *Journal of Informetrics*, *9*, 734–745.
Pavlidis, I., Petersen, A. M., & Semendeferi, I. (2014). Together we stand. *Nature Physics*, *10*, 700–702.
Petersen, A. M. (2015). Quantifying the impact of weak, strong, and super ties in scientific careers. *Proceedings of the National Academy of Sciences*, *112*, E4671–E4680.
Petersen, A. M. (2018). Megajournal mismanagement: Manuscript decision bias and anomalous editor activity at PLOS ONE. *SSRN e-print:2901272*.
Petersen, A. M., Fortunato, S., Pan, R. K., Kaski, K., Penner, O., Rungi, A., et al. (2014). Reputation and impact in academic careers. *Proceedings of the National Academy of Sciences USA*, *111*, 1531–15321.
Petersen, A. M., Jung, W.-S., Yang, J.-S., & Stanley, H. E. (2011). Quantitative and empirical demonstration of the Matthew effect in a study of career longevity. *Proceedings of the National Academy of Sciences*, *108*, 18–23.
Petersen, A. M., Pan, R. K., Pammolli, F., & Fortunato, S. (2018). Methods to account for citation inflation in research evaluation. *SSRN e-print:3193712*.
Petersen, A. M., Pavlidis, I., & Semendeferi, I. (2014). A quantitative perspective on ethics in large team science. *Science and Engineering Ethics*, *20*, 923–945.
Petersen, A. M., & Penner, O. (2014). Inequality and cumulative advantage in science careers: A case study of high-impact journals. *EPJ Data Science*, *3*, 24.
Petersen, A. M., Riccaboni, M., Stanley, H. E., & Pammolli, F. (2012). Persistence and uncertainty in the academic career. *Proceedings of the National Academy of Sciences USA*, *109*, 5213–5218.
Peterson, G. J., Presse, S., & Dill, K. A. (2010). Nonuniversal power law scaling in the probability distribution of scientific citations. *Proceedings of the National Academy of Sciences USA*, *107*, 16023–16027.
Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences USA*, *105*, 17268–17272.
Redner, S. (2005). Citation statistics from 110 years of physical review. *Physics Today*, *58*, 49–54.
Ren, F.-X., Shen, H.-W., & Cheng, X.-Q. (2012). Modeling the clustering in citation networks. *Physica A*, *391*, 3533–3539.
Scharnhorst, A., Börner, K., & van den Besselaar, P. (Eds.). (2012). *Models of science dynamics*. Berlin: Springer-Verlag Berlin Heidelberg.
Schwartz, C. A. (1997). The Rise and Fall of Uncitedness. *College & Research Libraries*, *58*, 19–29.
Simkin, M. V., & Roychowdhury, V. P. (2007). A mathematical theory of citing. *JASIST*, *58*, 1661–1673.
Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika*, *42*, 425–440.
Simonton, D. K. (2013). After Einstein: Scientific genius is extinct. *Nature*, *493*, 602–602.
Sinatra, R., Deville, P., Szell, M., Wang, D., & Barabasi, A.-L. (2015). A century of physics. *Nature Physics*, *11*, 791–796.
de Solla Price, D. J. (1965). Networks of scientific papers. *Science*, *149*, 510–515.
Solomon, D. J. (2014). A survey of authors publishing in four megajournals. *PeerJ*, *2*, e365.
Solomon, D. J., & Bjork, B.-C. (2012). A study of open access journals using article processing charges. *Journal of the American Society for Information Science and Technology*, *63*, 1485–1495.
Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, *333*, 776–778.
Stephan, P. (2012). *How economics shapes science*. Cambridge MA, USA: Harvard University Press.
Stokes, D. E. (1997). *Pasteur's quadrant: Basic science and technological innovation*. Washington, D.C., USA: Brookings Institution Press.
Vaccario, G., Medo, M., Wider, N., & Mariani, M. S. (2017). Quantifying and suppressing ranking bias in a large citation network. *Journal of Informetrics*, *11*, 766–782.
Vazquez, A. (2003). Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Physical Review E*, *67*, 056104.
Verstak, A., Acharya, A., Suzuki, H., Henderson, S., Iakhiaev, M., Lin, C. C. Y., et al. (2014). *On the shoulders of giants: The growing impact of older articles*. arXiv:1411.0275
Wakeling, S., Willett, P., Creaser, C., Fry, J., Pinfield, S., & Spezi, V. (2016). Open-access mega-journals: A bibliometric profile. *PLOS ONE*, *11*, e0165359.
Wallace, M. L., Lariviere, V., & Gingras, Y. (2009). Modeling a century of citation distributions. *Journal of Informetrics*, *3*, 296–303.
Wallace, M. L., Lariviere, V., & Gingras, Y. (2012). A small world of citations? The influence of collaboration networks on citation practices. *PLoS ONE*, *7*, e33339.
Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, *10*, 365–391.
Waltman, L., & van Eck, N. J. (2018). *Field normalization of scientometric indicators*. ArXiv e-print:1801.09985
Wang, D., Song, C., & Barabasi, A.-L. (2013). Quantifying long-term scientific impact. *Science*, *342*, 127–132.
Wildson, J. (2015). We need a measured approach to metrics. *Nature*, *523*, 129.
Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., et al. (2015). The metric tide: Report of the independent review of the role of metrics in research assessment and management. *Technical Report Higher Education Funding Council for England (HEFCE)*.
Wu, Z.-X., & Holme, P. (2009). Modeling scientific-citation patterns and other triangle-rich acyclic networks. *Physical Review E*, *80*, 037101.
Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, *316*, 1036–1039.
Yin, Y., & Wang, D. (2017). The time dimension of science: Connecting the past to the future. *Journal of Informetrics*, *11*, 608–621.