Research Policy 48 (2019) 1855–1865

Contents lists available at ScienceDirect

Research Policy

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

Methods to account for citation inflation in research evaluation

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

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

ARTICLE INFO

Keywords:
Measurement error
Career evaluation
Citation analysis
Science policy

ABSTRACT

Quantitative research evaluation requires measures that are transparent, relatively simple, and free of disciplinary and temporal bias. We document and provide a solution to a hitherto unaddressed temporal bias – citation inflation – which arises from the basic fact that scientific publication is steadily growing at roughly 4% per year. Moreover, because the total production of citations grows by a factor of 2 every 12 years, this means that the real value of a citation depends on when it was produced. Consequently, failing to convert nominal citation values into real citation values produces significant mis-measurement of scientific impact. To address this problem, we develop a citation deflator method, outline the steps to generalize and implement it using the Web of Science portal, and analyze a large set of researchers from biology and physics to demonstrate how two common evaluation metrics – total citations and *h*-index – can differ by a remarkable amount depending on whether the underlying citation counts are deflated or not. In particular, our results show that the scientific impact of prior generations is likely to be significantly underestimated when citations are not deflated, often by 100% or more of the nominal value. Thus, our study points to the need for a systemic overhaul of the counting methods used evaluating citation impact – especially in the case of researchers, journals, and institutions – which can span several decades and thus several doubling periods.

1. Introduction

Whether for merit review, tenure and promotion of academics or for the assessment of national research systems, the evaluation of scientific productivity and impact increasingly relies on quantitative measures (Moed et al., 1985; Luukkonen, 1991; Moed, 2006; Vinkler, 2010; Hicks et al., 2015; Wildson, 2015; Wilsdon et al., 2015). In particular, the use of bibliometrics has been a boon for objective evaluation, but nevertheless requires statistical normalization, astute application, and careful inference (Bornmann and Marx, 2015; Stephan et al., 2017). That is, despite the improved quality, quantity and diversity of quantitative evaluation measures – in particular, a proliferation of publication and patent-based citation measures – many have made the case for caution in order to address the risk of promoting bad practice in commonplace evaluation and decision-making scenarios (Hicks et al., 2015; Wildson, 2015; Wilsdon et al., 2015; Stephan et al., 2017). By way of example, recent efforts to predict researchers' future bibliometric impact (Acuna et al., 2012) have been shown to suffer from cohort and autocorrelation bias (Penner et al., 2013a,b), making the proposed predictive methods unreliable for quantitative faculty evaluation.

Against this backdrop, here we address a more fundamental and under-appreciated bias in bibliometric evaluation – 'citation inflation' – a systematic measurement problem that arises from the persistent secular growth of the scientific system (Lariviere et al., 2008; Althouse et al., 2009). While the extant literature has primarily focused on developing methods to overcome field-normalization bias (Radicchi et al., 2008; Radicchi and Castellano, 2012b; Waltman and van Eck, 2013; Bornmann and Marx, 2015), the problem of citation inflation may indeed be even more fundamental. The problem is rather simple – when citations are produced in distinct historical periods their 'nominal values' are inconsistent and thus cannot simply be added together. The ramifications of this problem have been noted in recent work analyzing the citation dynamics of individual careers and publications across the citation life-cycle (Petersen et al., 2014a; Parolo et al., 2015; Yin and

* Corresponding authors.
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.respol.2019.04.009
Received 4 April 2016; Received in revised form 30 March 2019; Accepted 18 April 2019
Available online 30 April 2019

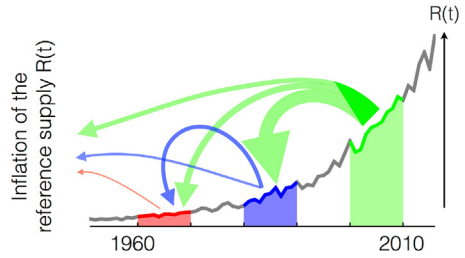

Fig. 1. Intergenerational flow of citations. The number of backward references $R(t)$ produced by scientific literature is growing steadily, due to growing publication rates and longer reference lists. Combined, these two effects correspond to a retroactive intergenerational effect (i.e. represented by the variable citation flows indicated by the arrows), the result of which is 'citation inflation'. Thus, when combining citations into tallies for scientific evaluation, a "deflator index" is needed in order to adjust for the growth of $R(t)$, thereby standardizing citations from different time periods to a common 'real' value. *Source*: Authors' elaboration (Color online).

Wang, 2017; Pan et al., 2018). To the extent that citation inflation is evident even over such short time periods, aggregating citation counts across several decades even further exacerbates this statistical measurement error. As such, it is likely to affect quantitative evaluation in hiring and promotion decisions as well as the science of the scientific endeavor (Fealing, 2011; Fortunato et al., 2018) when underling methods draw on cumulative citation counts or longitudinal citation analysis. As an example of the widespread use of citation counts in research evaluation studies, we used article-level keywords to estimate that roughly 6% (119 articles) of *Research Policy*'s publications over the period 2000–2018 have used research article or patent citations as a substantive measure. In addition to the increasing use of citation measures in science policy evaluation, another significant factor is the proliferation of rapid-publication online-only megajournals in the last decade. This paradigm shift in the production of scientific research (Solomon and Bjork, 2012; Binfield, 2013; Solomon, 2014; Bjork, 2015; Wakeling et al., 2016; Petersen, 2018a) is sustaining the 3–4% annual growth of science output, the net result of which is a substantial intergenerational flow of references made to prior literature, as illustrated in Fig. 1. When also taking into account the growth of reference lists, the growth rate of the reference supply is growing even faster at roughly 5.6% annual growth rate, which has implications for the structure and function of the science citation network (Pan et al., 2018). At this growth rate, the total number of references doubles roughly every 12 years; as such, it is possible that newer citations rather quickly outnumber the citations received by articles decades in the past. Thus, it is possible that a seminal research article from decades ago may feel a second wind – gaining more contemporary citations than received in the initial decade after its publication – not necessarily because of a delayed appreciation as in the case of "sleeping beauties" (Ke et al., 2015), but merely due to the growth of science.

In what follows, we start with a review of the literature on citation normalization methods for citation-based research evaluation, illustrating why various methods do not account for citation inflation. We proceed to identify and measure secular growth in science by analyzing the annual growth rates for both inputs and outputs of scientific research. We then derive a straight-forward method for defining a citation deflator index, drawing on the simple fact that an article can only cite another article once in its reference list, and demonstrate how to obtain a generic citation deflator using the public front end of the Clarivate Analytics Web of Science portal.

To assess the degree to which citation inflation might impact real-world evaluation, we apply our method to three typical units of analysis – individual researchers, journals, and institutions. In the case of the former, we apply our method to the publication profiles of 551 researchers, indexed by i, calculating each researcher's total citations C_i and Hirsch-index h_i (Hirsch, 2005) – with and without the citation deflator. Our results show that measurement errors upwards of 100% of the traditional nominal citation value can arise when citations are not deflated properly, which is especially exacerbated when comparing researchers from different age cohorts, as is customary in "all-time" lists (ACUMEN, 2018). We conclude with a discussion of our results and policy recommendations.

2. Literature review

Contemporary studies conceptualizing science as a growing and evolving complex system (Fortunato et al., 2018) owe much to Eugene Garfield's entrepreneurial efforts in developing the Science Citation Index for bibliometric data management, and Derek de Solla Price's theoretical efforts in formalizing the bibliometric citation network (Garfield, 1955; de Solla Price, 1965). In seminal work studying the growth of scientific production, De Solla Price used publication data collected over the 100-year period 1862–1961 to calculate a doubling time of $\tau_{2\times} = 13.5$ years for the scientific corpus, corresponding to a 5.1% annual growth rate, $g_n = \ln(2)/\tau_{2\times} = 0.051$ (de Solla Price, 1965).

This persistent doubling of the total volume of scientific output every 13.5 years poses institutional, technological, and cognitive challenges. While advances in information technology may have dramatically increased the accessibility of knowledge, the sheer volume of scientific knowledge production may have at the same time stretched individuals' cognitive abilities to browse, search, read, and re-use the information contained in scientific literature (Pan et al., 2018). Despite advances in search and retrieval algorithms, this represents a fundamental 'myopia problem' that occurs when search algorithms use as inputs, or sort outputs, primarily according to cumulative centrality or popularity measures (Mariani et al., 2015; Liao et al., 2017; Vaccario et al., 2017). In the particular context of citation-based document search, e.g. Google Scholar, older database items may appear to be less relevant if the search algorithm does not account for the secular growth of the system – i.e. does not account for the generic inflation of popularity measures as the population of both users and entities increases.

Moreover, De Solla Price's calculation does not take into account the increasing length of reference-lists, and the subtle but important implications this has for the connectivity of the citation network. Indeed, scientific articles are becoming longer, partly as a result of the online-only e-journal paradigm, whereby page-length limits are less stringent; it is also commonplace for print journals to allow online-only "supplementary materials" to accompany articles, thereby adjusting to the demand for more flexible research article lengths. Other explanations for reference list growth, e.g. self-citing behaviors, are more subtle and difficult to quantify, as traditional reasons to cite prior literature – i.e. attributing priority and explicitly demonstrating how one's research builds on prior knowledge – have become offset by the citation-based attention economy in science which incentivizes self-citation (Hellsten et al., 2007; Fowler and Aksnes, 2007; Costas et al., 2010; Zaggl, 2017; Petersen, 2018a; Biagioli et al., 2019; Seeber et al., 2019). In all, the increase in reference lists accounts for one-third of the growth rate g_R for the total number $R(t)$ of references produced by the scientific literature published in any given year t.

Yet citation-based measurement studies continue to treat the objective significance of the total citation count c_p of a given publication p as though the nominal value of individual citations are steady over time, despite the steady growth of the production system. Failing to account for citation inflation results in significant measurement error $\delta(c_p)$, and thus a problematic noise-to-signal ratio $\delta(c_p)/c_p$.

This class of measurement problem is the motivation for various field-normalization procedures that address the problem of comparing citation counts for articles from different fields and subdisciplines (Radicchi et al., 2008; Radicchi and Castellano, 2012a,b; Waltman and van Eck, 2013, 2018; Bornmann et al., 2013; Bornmann and Marx, 2015; Wilsdon et al., 2015; Waltman, 2016; Vaccario et al., 2017).

Indeed, the problem of field normalization is quite similar to the problem of citation inflation, because different fields vary in the volume of literature produced by the researchers within the community, and so the citation supply also varies across fields. In order to appropriately compare research articles across fields according to their citation tallies, one must first establish a common baseline for citation measurement. By way of example, consider the task of comparing (or adding together) the citation count $c_{p,y}$ for two articles from two different fields but same publication year y. A straightforward yet powerful field-normalization method proceeds by rescaling each $c_{p,y}$ count by an appropriate field-specific baseline given by the mean citation impact $\bar{c_y}$ calculated by averaging over all articles published in that given field and year. The rescaled quantity is thus given by the ratio $c_f \equiv c_{p,y}/\bar{c_y}$. Using this normalization method, Radicchi et al. (2008) show that the distribution $P(c_f)$ of field-normalized citation values follows a log-normal probability distribution, independent of discipline f, indicating that c_f is net of discipline-specific factors.

Yet such field-normalization methods are limited to comparing articles from similar publication cohorts. There still remains the problem of right-censoring bias, which affects the comparison of two articles when one is significantly older than the other. One common solution to this problem is to only count citations within a fixed observation window so that both articles being compared have had the same time to accrue citations. For example, consider the balanced citation count $c_{p,y,\Delta t}$ calculated by only tallying citations arriving in a fixed Δt-year window after y. While this method partially facilitates comparing articles from different y, it is still susceptible to citation inflation because the doubling period for $R(t)$ is only 12 years (Pan et al., 2018).

To demonstrate the magnitude of citation inflation even after using a fixed citation window, consider c_q, the citation value corresponding to a given percentile $q = \{50, 25, 10, 1, 0.1\}$ of the citation distribution, calculated using the $c_{p,y,\Delta t=10}$ values for all articles from a given year. Fig. 2 shows that the cutoff for the top 1% of "Science" articles published in the year 2000 corresponds to 200 citations, whereas the top 1% cutoff for publications from 1965 was just under 100 citations. Similarly, the cutoff for the top 10% of "Social Science" publications from 1965 is around 10 citations, whereas in 2000 the threshold had risen to just over 40 citations. Each $c_q(t)$ curve is growing at a slow but nevertheless substantial rate that compounds across decades.

The steady growth of $c_q(t)$ illustrates how the relative value of a single citation is systematically decreasing over time, reflected by the fact that the entire citation distribution is systematically shifting toward higher values. In other words, more recent publications need increasingly more citations to be within the top x% (of publications from the same year) than do older publications with the same percentile score within their respective publication cohort. Thus, these arguments illustrate why neither field-normalization nor a fixed citation window can overcome the temporal bias induced by citation inflation. Indeed, in a recent review of citation impact indicators by Waltman (2016), while the merits and challenges of various citation normalization strategies are discussed at length, none address the systematic feature identified here of how to deal with the non-stationary inflation of nominal citation values.

There is one additional measurement issue relating to intensive versus extensive quantities. Evaluation of aggregate units of analysis (e.g. researchers, journals, institutions) typically calls for quantities that can be added together – i.e. extensive quantities that double if for example the underlying system doubles in size. This requirement follows because total citation impact (e.g. calculated by adding $c_{p,y}$ across individual articles), should have a clear and intuitive relation to the overall size of the aggregate unit. By way of contradistinction, there do exist standardized 'intensive' citation measures, such as the citation percentile $q(c_{p,y,\Delta t})$ and the normally-distributed z-statistic (obtained by standardizing the location and scale of the log-normal distribution) (Petersen, 2015; Petersen et al., 2018), which both facilitate standardized comparisons. While these methods produce citation measures that are time-independent, and thus robust to citation inflation, they are not amenable to aggregation because these quantities lack intuitive meaning when they are tallied across articles. Additionally, such intensive quantities do not readily decompose into increments, such as the annual citation rate $\Delta c_{p,y,t}$ corresponding to the number of new citations an article received in year t alone. While $\Delta c_{p,y,t}$ may appear suitable for time-series analysis, we show that citation inflation is still a burden, because a single citation arriving in one year is not necessarily equivalent in measure to a citation arriving in a different year.

3. Material and methods

We analyze all English publications (articles and reviews) in the Clarivate Analytics Web of Science (WOS) database that were published between 1965 till the end of 2012. We use the WOS journal classification system to group individual publications into broad research domains according to Web of Science Categories (denoted by the WC field in WOS publication records). In all there are 252 distinct WC, which we use to aggregate publications into 3 broad domains: "Science", "Social Science", and "Art & Humanities". The list of WC that define the "Science" domain are available at http://mjl.clarivate.com/scope/scope_scie/, and together comprise the "Science Citation Index Expanded"; similarly, the list of WC that define the "Social Science" domain are available at http://mjl.clarivate.com/scope/scope_ssci/,

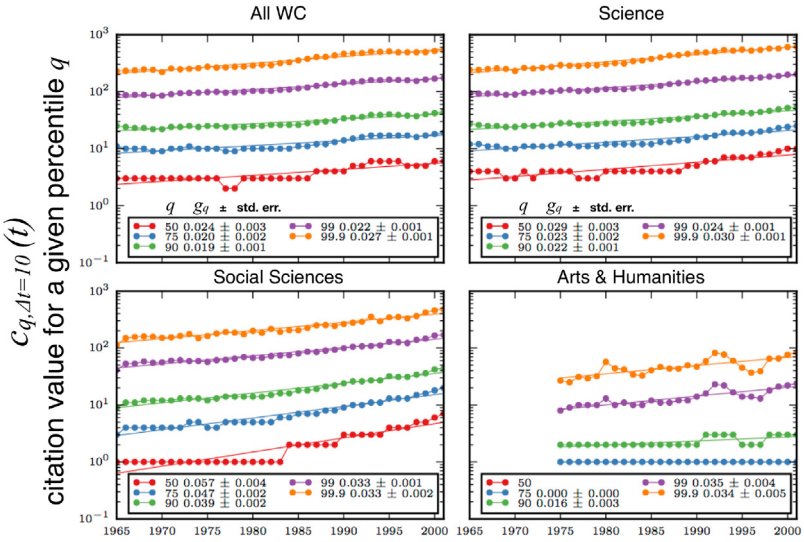

Fig. 2. Inflation of the number of citations received. Evolution of the citation value $c_q(t)$ corresponding to a given percentile value (q) of the citation distribution $P(c_{p,y,\Delta t=10})$. Balanced citation counts are calculated for all publications by using a $\Delta t = 10$-year citation window. The percentile values (q) are shown in each panel legend along with the best-fit exponential growth parameter for each curve. *Source*: Authors' elaboration using WOS data (Color online).

Box 1
Procedure for specifying a research area and obtaining citation deflator data from the Web of Science database portal.

1. **Use the Web of Science search portal to define the target subject.** Go to www.webofscience.com and select "Advanced Search". One can search for all publications by a particular journal(s) using the SO field, or for all publications corresponding to a particular subject category by using the WC field. For example, searching for "SO= RESEARCH POLICY" within the timespan 1900–2016 results in 3063 results within the *Web of Science Core Collection*.
2. **Obtain list of all records that have cited articles within the target subject.** On the "Results" page, select the "Create Citation Report" option toward the upper right-hand corner. Then, in the subsequent "Citation report" page, click on the "Citing Articles" to obtain the set of nearly 60,000 articles that cite the target subject (i.e. the collection of *RP* publications).
3. **Analyze results.** Select the "Analyze Results" option toward the upper right-hand corner of the search results webpage.
4. **Tabulate results by relevant field.** On the "Results Analysis" page, group the results according to a select data field indicated by the 16 tabs on the left, e.g. "Publication Years". In order to circumscribe the set of target subject articles, we select the "Web of Science Categories" tab, and then set the "Show" tab to 100; then proceed to select "Update Table" which tabulates the 58,431 articles by their WC values. We then take the top-*N* WC that most efficiently circumscribe the target subject. For example, in the case of *RP*, the WC fields "Management", "Business", and "Economics" cover 59% of the 58,431 articles citing *RP* literature.
5. **Obtain list of all records that belong to the circumscribing WC.** Go back to Step 1 and use the WC identified in Step 4 to modify the search query: "$WC = (WC_1$ OR WC_2 OR... OR $WC_N)$".[5] For example, using the top-3 "WC =(MANAGEMENT OR BUSINESS OR ECONOMICS)" returns roughly 1.67 million results over the 45-year period 1972–2016 that are within the WC scope of RP. Section 4.2 demonstrates that the method is robust to the number of WC used to define the scope, since it is the relative growth over time, and not the total number of articles identified in this step, that is captured by the deflator index defined in Eq. (2).
6. **Obtain the deflator time series $n_a(t)$ for the circumscribing WC.** Following from the previous step, the new search query will identify a new set of articles from journals belonging to the set of circumscribing WC. Select "Analyze Results" to again go to the "Results Analysis" page. Then select the "Publication Years" tab on the left, and again select "Show 250" and then "Update Table". This yields the set of circumscribing articles tabulated by their publication year – i.e. $n_a(t)$ – which can be downloaded to text file by selecting "Download".

and together comprise the "Social Science Citation Index". Together, the "Science" and "Social Science" domains account for more than 95% of the WOS database we analyzed.

In what follows, we provide basic steps for obtaining citation deflator data (see Box 1) that does not require purchasing a hard copy of the entire WOS database from Clarivate Analytics. Instead, this method just requires site-license access to the Web of Science front-end search portal located at www.webofscience.com. In effect, our method queries the entire Web of Science Core Collection comprised of the union of the Science Citation Index Expanded, Social Sciences Citation Index, Arts & Humanities Citation Index, Conference Proceedings Citation Index, Book Citation Index, and the Emerging Sources Citation Index (i.e. Indexes = SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC) and returns summary count data that is readily downloadable. Users may wish to refine their WOS query to just a single or smaller combination of indices as they see appropriate.

Upon obtaining deflator data using this method, we then apply our deflator method to the career profiles of 551 researchers. The researchers in this set started their careers over the broad interval spanning the 1940s to the 2000s. We demonstrate the impact of citation inflation on careers that span decades, and also illustrate the challenge of comparing researchers from significantly different age cohorts. See Penner et al. (2013a) for more details on these researcher profiles and the corresponding author name disambiguation method, which in short, leverages the WOS "distinct author" search query option, in addition to obtaining career profile data from *ResearcherID.com*.

To provide context for the growth of scientific publication output, which is the basis for citation inflation, we draw upon country-level research & development (R&D) data obtained from the World Bank (2019): Researchers in R&D (per million people) and Research and development expenditure (% of GDP) data combined with GDP (constant 2010 US$) and Total Population (in number of people) data. All dollar amounts are deflated to constant 2010 US$.

And finally, to facilitate the detail-oriented reader, we provide the following list of notation used in the remainder of the text:

(i) $c_{p,y,t}$ is the total number of citations tallied through year t, received by an article p that was published in year y.
(ii) $\Delta c_{p,y,t}$ is the citation rate, the number of new citations received in year t, corresponding to the 1-year increment given by $\Delta c_{p,y,t} \equiv c_{p,y,t} - c_{p,y,t-1}$. In what follows, because the publication year is either not relevant or redundant with respect to the generic time variable t, we suppress the index y from our notation.
(iii) $C = \Sigma c_p$ denotes the total number of citations, aggregated across a particular set of articles.
(iv) $n_a(t)$ is the total number of articles published in a given research area a in year t.
(v) g_x is the annual exponential growth rate associated with the generic time-dependent quantity $X(t)$. This growth rate is calculated by applying ordinary least squares regression to fit the exponential model: $\ln X(t) = \ln X(0) + g_x t$. For $g_x \ll 1$, then the 1-year growth in $X(t)$ is roughly $100 g_x$ percent. Quantities modeled in this way are the publication volume, $n(t)$; the reference supply, $R(t)$; the number of researchers in a given country that are active in R&D, $S(t)$; the total national expenditure on R&D, $E(t)$, measured in constant 2010 US dollars; and the citation value corresponding to a given percentile q of the citation distribution, $c_q(t)$.
(vi) $\tau_{2\times} = \ln(2)/g_x$ is the doubling period associated with g_x, which is the amount of time it takes for $X(t)$ to double in amplitude, i.e. $X(t + \tau_{2\times}) = 2X(t)$.
(vii) WC is the Subject Category assigned by WOS to each article.
(viii) $y_{0,i}$ is the year of first publication for a given researcher (indicated by the index i), which is useful for separating researcher profiles into age cohorts.

4. Results

4.1. Growth of scientific production

"*The beauty of science may be pure and eternal, but the practice of science costs money.*" – Dr. Paula Stephan.

Economic scholars have developed various lines of inquiry aimed at illuminating the role of incentives, allocation mechanisms, and markets for knowledge and innovation (Stephan, 2012). One particularly transcending theme is the nature of secular growth and its sources. The

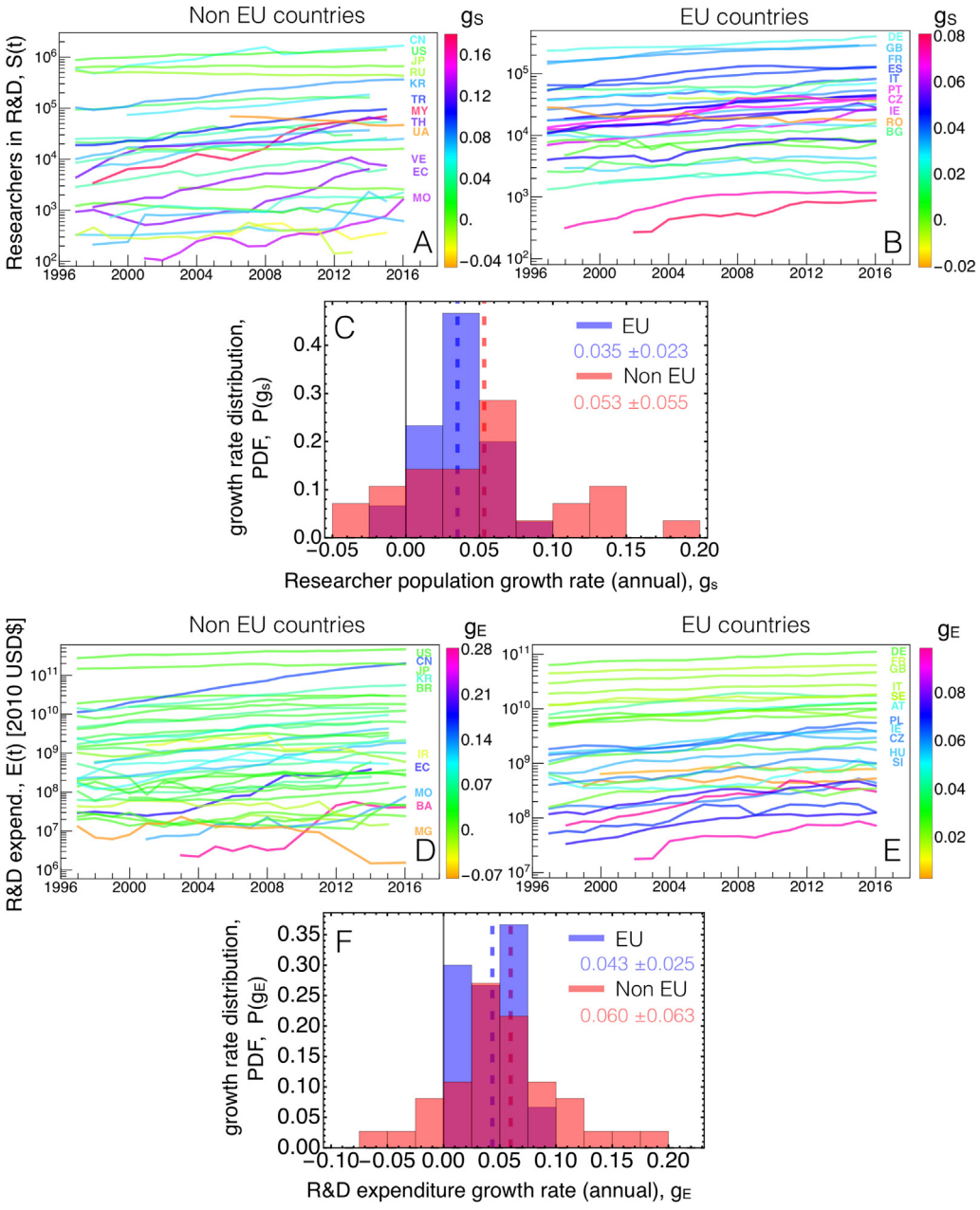

Fig. 3. Inputs of scientific R&D efforts: empirical growth trends. Country-level growth trends in researcher population and R&D funding over the 20-year period 1997–2016. (A–C) Growth in the number of researchers in R&D, $S(t)$, by country. (D–F) Growth in the total R &D expenditure, $E(t)$, by country (reported in constant 2010 US$). Only countries with more than 10 data points are analyzed. (A) $S(t)$ for 38 non-European countries. (B) $S(t)$ for 30 European Union and EFTA countries. (C) Frequency distribution of the exponential growth rate, g_S, estimated for each $S(t)$ time series. The mean (dashed vertical line) and std. dev. for each country subgroup are shown in each panel. (D) $E(t)$ for 28 non-European countries. (E) $E(t)$ for 30 European Union and EFTA countries. (F) Frequency distribution of the exponential growth rate, g_E. (A, B, D, E) Each color legend indicates the growth value corresponding to each individual time series. *Source*: Authors' elaboration using World Bank data (Color online).

main contemporary source of growth in science is national investment in R&D, with leading countries typically funding R&D activities as a targeted percentage of their GDP – i.e. strategic (re)investment into scientific industry. Aside from R&D infrastructure, much of this investment is placed in researchers and the resources that they need to prosecute their research programs (Stephan, 2012).

To provide context for the underlying secular growth that drives research output, in particular publication growth, Fig. 3 shows the growth of the principal inputs of the scientific enterprise – people and money. Country-level data are drawn from the World Bank (2019) for the 20-year period 1997–2016. Because the European Union (EU) has its own integrated funding system, we separated the countries into two groups, non-EU and EU.

In terms of researcher population growth, measured by $S(t)$ for each country, non-EU (EU) countries are growing on average at a 5.3% (3.5%) annual rate. Extensive research shows that growth rates are typically inversely proportional to the enterprise size (Riccaboni et al., 2008). As such, the higher average value for non-EU countries largely reflects the emerging economies that have entered the scientific enterprise at the global scale only as of relatively recently. For comparison, the growth rate of the population size of post doctorates and graduate students in U.S. STEM fields shows roughly 2–4% annual growth over the period 1972–2010 (Petersen et al., 2014b). To put this growth in perspective, a 4% annual growth rate corresponds to a doubling period of 17 years.

In terms of total R&D expenditure, $E(t)$, growth rates are slightly larger. World Bank data shows that non-EU (EU) countries are on average growing at a 6.0% (4.3%) annual rate. The trend-breaker in this regard is China, emerging in the last 20 years as a global leader in the production of scientific knowledge (Zhou and Leydesdorff, 2006). For perspective, Fig. 3 highlights China's persistent 6.1% annual growth in researcher population accompanied by rapid 15.7% growth in R&D expenditure. South Korea also demonstrates significant growth in both quantities as well.

The growth of these fundamental R&D inputs sets the scale for output growth, such as the total number of publications $n(t)$. In

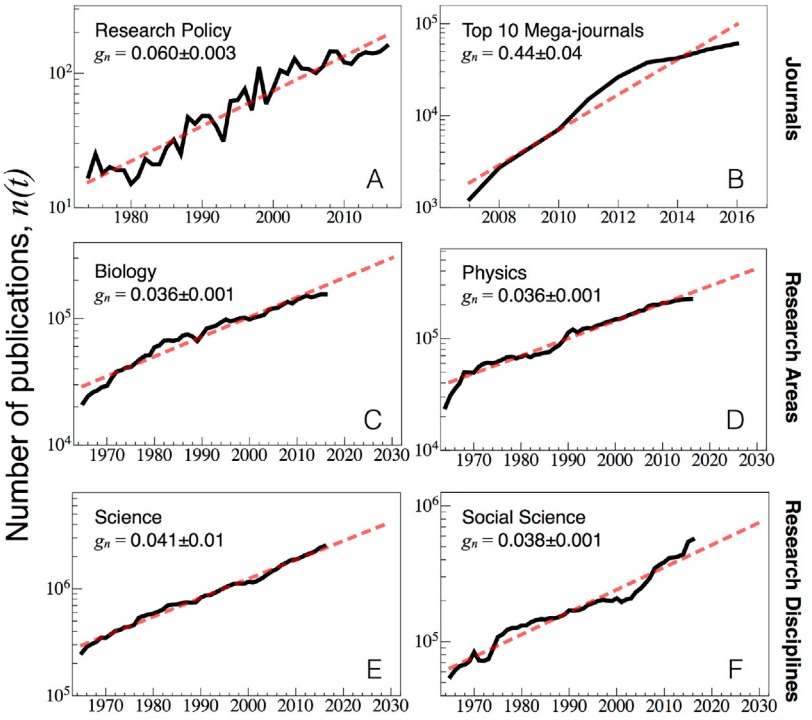

Fig. 4. Outputs of scientific R&D efforts: empirical publication rate growth for journals, research areas and disciplines. (A, B) Growth at the aggregation level of journals: *Research Policy* shows an annual growth rate of roughly 6%, while the 10 largest mega-journals combined show an exponential growth rate $g_n = 0.44$, which corresponds to an annual growth rate of 55%. (C, D) Growth of two prominent research areas in the Web of Science – biology and physics. Both areas are defined using the set of Web of Science "subject category" (WC) that contain either "biology" (or "physics") in the category name. Empirical growth trends demonstrate remarkably similar annual growth rates of 3.6%. (E, F) Growth of "Science" articles, showing a 4.1% annual growth rate, and Social Science articles, showing a 3.8% annual growth rate. (C–F) Growth trends extrapolated to 2030 using empirical $n(t)$ data from 1965 through 2016. *Source*: Authors' elaboration using WOS data (Color online).

particular, the quantity $n_a(t)$, calculated for a particular research area a, is a standardized unit of measurement for scientific productivity, e.g. because it accounts for common underlying trends such as globalization and innovation in scientific publishing. To this end, Fig. 4 shows the growth of $n(t)$ at three levels of aggregation: journals, research areas, and research domains. For example, at the level of journals, *Research Policy* shows an annual growth rate of 6% in the 46 years since its first issue in 1972, which is significantly larger than the 3.8% annual growth rate calculated for all "Social Science" research. However, by way of contrast, both these growth rates are relatively small compared to the 10 most prominent "mega-journals"[2] – journals characterized by the APC ("pay-to-publish"), open-access, continuous publication model – for which we calculate an exponential growth rate $g_n = 0.44$, corresponding to a 55% annual growth rate and a publication doubling period of just $\ln 2/0.44 = 1.6$ years.

For comparison, we also analyzed the growth of scientific output in two large subdomains, "Physics" and "Biology",[3] which show a remarkably similar annual growth rate of 3.6% over the period 1965–present, $g_n = 0.036$. Extrapolation of their growth trends to 2030 indicates productivity around 300,000 publications per year. And at an even higher level of aggregation, "Science" and "Social Science" appear to be growing even faster, $g_n = 0.041$ and $g_n = 0.038$, respectively, despite significant differences in the amplitude of $n(t)$. Science is the larger of the domains, with current productivity (estimated by counting the number of articles indexed by WOS) close to 2.5 million articles per year, whereas Social Science produces around 0.56 million articles per year.

The total number of references produced each year is growing even faster. In the case of "Science" ($g_n = 0.036$), we estimate the growth rate in the average number of references per publication per year to be $g_r = 0.018$. Combined, the net annual reference supply is thus increasing exponentially with $g_R \approx g_n + g_r = 0.054$. As a double check, we also estimate g_R from the time series $R(t)$, which is the total references produced in a given year, obtaining $g_R = 0.056 \pm 0.001$. From the growth of both scientific publication and reference list length, it follows that the total reference supply has a doubling period of $\tau_{2\times} = \ln(2)/0.056 \approx 12$ y. That is, $R(t + 12) \approx 2R(t)$, which also holds for integrated totals over time, i.e. $\sum_{t' \le t+12} R(t') \approx 2 \sum_{t' \le t} R(t')$, or equivalently, $\sum_{t' \in [0,t-13]} R(t') \approx \sum_{t' \in [t-12,t]} R(t')$, due to the properties of exponential growth. This growth is significant, as many careers or institutions span multiple 12-year doubling periods, which is the fundamental source of the measurement error associated with citation inflation.

4.2. Defining and obtaining a citation deflator

The method to account for citation inflation we propose is analogous to the method for adjusting real prices for monetary inflation. Instead of using a consumer price index as a price deflator, we use the publication rate $n(t)$ to convert 'nominal' citation rates (the raw citation counts one obtains from the likes of Web of Science, Google Scholar, and Scopus) into 'real' citation rates, so that they have common units and are comparable across time. This choice follows from the simple fact that a new article can only cite a prior article once – thus, $n(t)$ sets the upper limit as to how many new citations, $\Delta c_{p,t}$, a published article could conceivably receive from the entire set of new articles published in any given year.

[2] We used the list of mega-journals compiled at https://megajournals.info/. We then obtained WOS data using the following query to obtain records for only the top-10 journals in terms of publication volume: SO = ("PLOS ONE" OR "SCIENTIFIC REPORTS" OR "PEERJ" OR "ELIFE" OR "NATURE COMMUNICATIONS" OR "BMJ OPEN" OR "SAGE OPEN" OR "CHEMICAL SCIENCE" OR "IEEE ACCESS" OR "ONCOTARGET"). The smallest of these 10 journals is Sage Open, which published 302 articles in 2017; the largest is "Scientific Reports", which published 25,342 articles in 2017, or roughly 469 articles each week.

[3] We defined "Physics" and "Biology" using Web of Science Subject Categories (denoted by the WC field). We identified "Physics" articles by entering the WOS query: WC = ("Physics, Applied" OR "Physics, Atomic, Molecular & Chemical" OR "Physics, Condensed Matter" OR "Physics, Fluids & Plasmas" OR "Physics, Mathematical" OR "Physics, Multidisciplinary" OR "Physics, Nuclear" OR "Physics, Particles & Fields" OR "Astronomy & Astrophysics" OR "Biophysics" OR "Geochemistry & Geophysics"). Similarly, we identified "Biology" articles using the WOS query: WC = ("Biochemistry & Molecular Biology" OR "Cell Biology" OR "Biology" OR "Developmental Biology" OR "Evolutionary Biology" OR "Marine & Freshwater Biology" OR "Mathematical & Computational Biology" OR "Reproductive Biology").

Our intuitive approach thus corresponds to a simple rescaling of the 'nominal citation rate' $\Delta c_{p,t}$, thereby yielding a deflated citation rate $\Delta s_{p,t}$ given by

$$\Delta s_{p,t} \equiv \Delta c_{p,t} \times D_{a,t_b}(t), \tag{1}$$

where the deflator index,

$$D_{a,t_b}(t) = n_a(t_b)/n_a(t), \tag{2}$$

is defined using an arbitrary baseline year t_b, which we set to $t_b \equiv 2010$ for the remainder of our analysis. In practical terms, the deflator index $D_{a,2010}(t)$ is the real value of a single citation from year t in terms of 2010 citations. Here a refers to the research area that is broadly associated with the publication p. Thus, the first step to obtaining the deflator $D_{a,t_b}(t)$ is to define a in such a way that $n_a(t)$ can be estimated.

The Web of Science is a longstanding and major bibliometric indexing service provider, with databases that record natural science, social science, arts and humanities journals. As such, we shall demonstrate the steps to obtain $n_a(t)$ using their standardized and well-documented web portal accessible at www.webofscience.com. By way of example, we shall use *Research Policy (RP)* to illustrate our procedure based on WOS Subject Categories[4] (denoted by the WC field in WOS publication records). Our operating assumption is that one can approximately circumscribe the target subject's research area, denoted by the subscript a, using a single WC or combination of WC. Thus, in this example case, the target subject is a particular journal, although it could conceivably be any other unit of analysis, e.g. a researcher or institution. As we shall demonstrate, it is less important to be exhaustive in circumscribing a, and more important to just capture the growth trend for the research area over time.

We used the steps outlined in Box 1 to obtain $n_{RP}(t)$ for calculating $D_{RP,2010}(t)$, both of which are shown in Fig 5 . To demonstrate the robustness of our method, we used both the top-3 WC and the top-10 WC to define the circumscribing WC set in Step 5 of Box 1.[5] By construction, the amplitude of $n_{RP}(t)$ is larger for the top-10 WC, but the growth trends are similar enough that the corresponding deflators $D_{RP,2010}(t)$ are not substantially different. That is, despite the significant difference in the research volume circumscribed by the top-3 WC versus the top-10 WC, our method is robust as long as the WC capture the majority of the research area. For example, the top-3 (top-10) WC cover 59% (76%) of the 58,431 articles citing RP literature. This follows because the $n_a(t)$ amplitudes cancel out in Eq. (2). What remains is just the growth trend underlying $n_{RP}(t)$, which in the case of the top-3 WC corresponds to a doubling period of $\tau_{2\times} = \ln(2)/0.043 \approx 16$ y.

Accordingly, this doubling period indicates that a citation produced in year t (or conversely, in year $t - 16$) is worth roughly half (twice) as much as a citation produced in year $t - 16$ (t); and a citation produced in year t is worth roughly one quarter as a citation produced in year $t - 32$. Indeed, $D_{RP,2010}(1978) \approx 4.3$, consistent with our two doubling-period estimate. To see how this would affect the comparison of two individual articles, consider two hypothetical publications $p = 1$ and $p = 2$, both published in *RP* in the inaugural year 1974. Imagine the first gained a single citation every year for the first 20 years, thus its nominal citation count is $c_1 = 20$; the second also gained a single citation every year for 20 years, but over the period 1994–2013 instead, more characteristic of a "sleeping beauty" citation life-cycle (Ke et al., 2015). Despite both p being from the same journal and publication year and having equal nominal citation counts $c_1 = c_2$, after adjusting the citations using $D_{RP,2010}(t)$, the difference between the two p should be more clear: in terms of 2010 citations, the first gained 78 deflated citations whereas the second gained 39 deflated citations – a factor of two (or 100%) difference. This simple example highlights how it is important to account for the different timing of citations – in addition to discipline and publication year.

To demonstrate the magnitude of the citation inflation effect, we apply the method using real-world cases at three levels of aggregation – journal, institute and individual researcher. Fig. 6 shows the annual citation rate, the total number of citations received in a given year by all WOS articles published by each unit, for 4 entities: the journal *Research Policy*; the bio-medical research institution, *Cold Spring Harbor Laboratory*; the author of the most highly-cited author article from *RP*, D. J. Teece; and the most prolific author in *RP*, R. R. Nelson. The solid black curves show the 'nominal' or raw citation rates reported by WOS, and the red dashed curves show the real or deflated citation rates calculated using $D_{RP,2010}(t)$ (using the top-3 WC), except for *Cold Spring Harbor Laboratory* for which we used $D_{\text{Biology},2010}(t)$. The difference between the nominal and deflated time series is rather pronounced, demonstrating the measurement error incurred when citation inflation is neglected – ranging from 16% difference in the case of *RP* to 82% for R. R. Nelson.

4.3. Inflation-corrected productivity and impact measures – a researcher cohort study

The growth of science affects different units of analysis in different ways. For publications, following from the doubling period of roughly 2 decades, the growth significantly reduces the visibility of previous publication cohorts relative to the most recent publication cohorts (Pan et al., 2018). In the case of research careers, it affects the estimation of citation impact when the citation counts are tallied over long periods, e.g. decades. In such a case, as illustrated in Fig. 6, the total citation impact, corresponding to the area under each citation rate curve, is likely to strongly depend on whether *nominal values* or *deflated values* are used.

To demonstrate the magnitude of this measurement error across a larger set of scientists, we calculated one productivity measure (the h-index) and one citation impact measure (total citations) for 190 biologists and 361 physicists, each with an h-index (Hirsch, 2005) of 10 or greater, and with first publication year denoted by $y_{0,i}$, representative of the researcher's age cohort. The index i indicates an individual researcher, which we suppress on the right-hand side of the following definitions in order to focus on the important variables.

While it is not the purpose of this study to condone either impact or productivity measure, we focus on them because they are well-known and demonstrate the impact of citation inflation in an intuitive way. To calculate each summary measure, we assembled the publication portfolio of each individual researcher through 2010, comprised of a total N_i publications, which required tallying the (nominal) citation count $\Delta c_{p,t}$ in each year t for each individual publication p. The cumulative citations $c_{p,T} = \sum_{t=0}^{T} \Delta c_{p,t}$ is simply the total number of citations up to a given census year $T \equiv 2010$ to a given article (p) belonging to a particular individual (i). Using these final citation tallies we then calculated the h-index h_i (a productivity measure) and the net citations $C_i = \Sigma_{p\in i} c_{p,T}$ (a net citation impact measure). For comparison, we also calculated the deflated net citations

$$C_i^D = \sum_{p\in i} s_{p,t} \quad \text{with} \quad s_{p,T} = \sum_{t=0}^{T} \Delta s_{p,t} \tag{3}$$

[4] The Subject Categories (WC) are assigned and maintained by the Web of Science, and are attributes of particular journals and books, i.e. they are used to annotate all publications belonging to a particular journal. This is not to be confused with WOS Research Areas (SU) which are assigned to individual publications, independent of the source journal. It is also possible to tabulate the "Results Analysis" around "Research Areas", which is one of the 16 options, in addition to "Web of Science Categories" and "Publication Years".

[5] The query for the (ranked) top-10 WC for articles citing *Research Policy* literature is "WC = (MANAGEMENT OR BUSINESS OR ECONOMICS OR "PLANNING DEVELOPMENT" OR "OPERATIONS RESEARCH MANAGEMENT SCIENCE" OR "INFORMATION SCIENCE LIBRARY SCIENCE" OR "ENGINEERING INDUSTRIAL" OR "ENVIRONMENTAL STUDIES" OR "COMPUTER SCIENCE INTERDISCIPLINARY APPLICATIONS" OR GEOGRAPHY)".

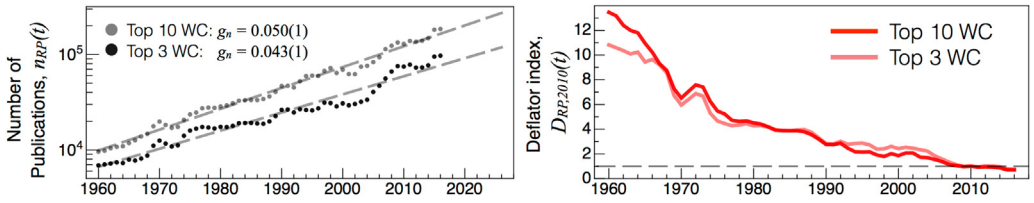

Fig. 5. Example of a citation deflator encompassing the broad research of *Research Policy*. (Left) Empirical growth of $n_a(t)$ for the research area defined by aggregating the top 3 and top 10 WOS subject categories (WC) of all publications in the WOS database that have cited *Research Policy* literature. For example, the top 3 WC are "Management", "Business", and "Economics"; the number of articles $n_a(t)$ published by journals classified by WOS as belonging to one of these three categories is growing at an annual rate of 4.3%. (Right) The corresponding deflator indices $D_{2010}(t)$ defined using $t_b \equiv 2010$ as the baseline year (i.e. $D_{2010}(2010) \equiv 1$). Despite the significant difference in $n_a(t)$ for the top 3 and top 10 WC, when normalized according to Eq. (2), the differences in the corresponding deflator indices become negligible. *Source*: Authors' elaboration using WOS data (Color online).

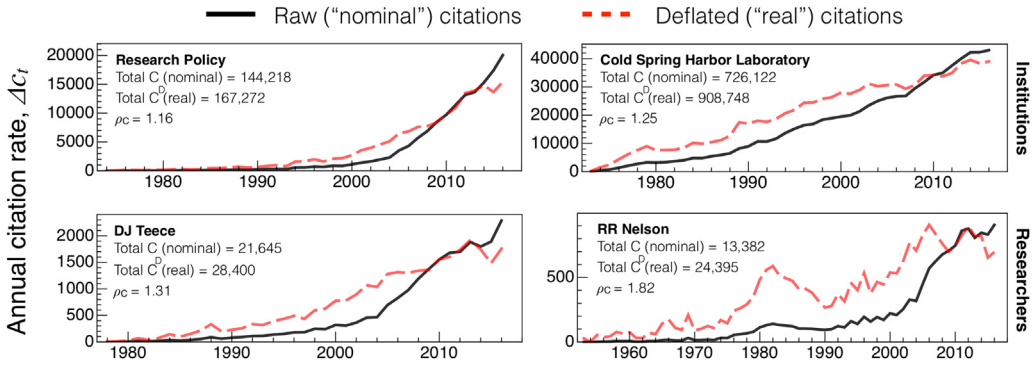

Fig. 6. The impact of citation inflation on the longitudinal evaluation of institutions and researchers. Shown is the annual citation rate, reported by WOS as nominal citations (solid black line): data collected using the WOS "Create Citation Report" service, for the journal "Research Policy", the "Cold Spring Harbor Laboratory" institute, the author of the highest-cited publication in *Research Policy*, D. J. Teece, and the most prolific publisher in *Research Policy*, R. R. Nelson. We calculated the deflated annual citation rate (dashed red line) for each case using the deflator index for the top 3 WC shown in Fig. 5 for each curve except for "Cold Spring Harbor Laboratory", for which we use the deflator time series $n_a(t)$ shown in Fig. 4 for "biology" to calculate the corresponding deflator index. For each example we report the total citations aggregated over time in terms of nominal citations C and deflated citations C^D, and report the deflated citations ratio $\rho_C \equiv C^D/C$. *Source*: Authors' elaboration using WOS data (Color online).

and the deflated h-index h_i^D using the deflated cumulative citations $s_{p,T}$ for each p. Because the $n_a(t)$ corresponding to physics and biology have different growth profiles, shown in Fig 4 (C, D), we used two distinct deflators, $D_{\text{Physics},2010}(t)$ and $D_{\text{Biology},2010}(t)$, respectively. We then separated the researcher profiles into age cohorts by grouping $y_{0,i}$ into 10-year non-overlapping intervals. Because each discipline includes 100 highly cited scientists, the range of h_i and C_i is rather broad, representing early career researchers with $h_i \sim 10$ up to eminent scientists with $h_i > 100$ (see Penner et al. (2013a) for more details on these researcher sets).

The difference between the traditional measures calculated using nominal citation rates, h_i and C_i, and their deflated counterparts, h_i^D and C_i^D, corresponds to the magnitude of the measurement error arising from citation inflation. The ratios

$$\rho_{H,i} \equiv h_i^D/h_i \quad \text{and} \quad \rho_{C,i} \equiv C_i^D/C_i \tag{4}$$

quantify this difference within each individual's research profile, with $100(\rho - 1)$ corresponding to the percent difference relative to the nominal value. Fig. 6 also serves as a visual guide in associating the differences in the nominal and real curves, and the summary difference quantified by ρ_C. For example, deriving from R. R. Nelson's contributions in the 1970s, there is a large deviation in the deflated and real curves in the 1980s once these works became highly cited. As such, his total deflated citations C_i^D are 82% larger than the nominal citations C_i^D, demonstrating the remarkable degree to which the impact of this early work is under-appreciated according to nominal citation values.

The mean ρ values, $\langle\rho_H\rangle = 1.08$ and $\langle\rho_C\rangle = 1.31$ are the same for both the physics and biology researchers we analyzed. Yet the growth trajectory of academic careers, even among prominent researchers in the same discipline, can vary considerably (Petersen et al., 2014a). As such, there can also be considerable variation in ρ_H and ρ_C within age cohorts.[6] Separating the $\rho_{H,i}$ and $\rho_{C,i}$ values by age cohort ($y_{0,i}$), Fig. 7 shows the wide range of error corresponding in this particular case to significant "underestimation" of scientific impact, especially for researchers from earlier cohorts. For example, the median ρ_H for researchers (either physics or biology) from the 1970s indicates a 10% error in the calculation of the h-index; similarly, the median ρ_C indicates a 35% error in C_i for the biologists and a 10% error for the physicists from the 1970s. The width of the ρ distributions also increases for older age cohorts, with some outliers from the 1970s having measurement error upwards of 100%, corresponding to $\rho = 2$.

The trend in $\rho_H(y_0)$ and $\rho_C(y_0)$ thus reflects the inflation rate of the science achievement measures themselves. Fig. 7 shows the mean value $\langle\rho(t)\rangle$ calculated for researchers from each 10-year group (e.g. from the 2000s to the 1940s). In order to estimate the 10-year growth rate g_{10}, we fit these data using the exponential form

$$\langle\rho(y_0)\rangle = \rho_0 \exp[g_{10}(2000 - y_0)/10]. \tag{5}$$

Based upon the average values $\langle\rho_H(y_0)\rangle$, we estimate $g_{10} \approx 0.061$ (biology) and $g_{10} \approx 0.076$ (physics), which means that for every 10 years in the past, ρ_H grows by roughly 6–7%. Similarly for $\langle\rho_C(y_0)\rangle$, we estimate a significantly larger growth rate, $g_{10} \approx 0.23$ (biology) and $g_{10} \approx 0.39$ (physics). Together these numbers quantify the extent to which a 10-year time difference can alter the relative values of productivity and impact measures – which could bias quantitative assessment in a senior-rank faculty search where candidates could differ significantly in age, for example. The above estimates follow from the average value, $\langle\rho(y_0)\rangle$, calculated across all individuals belonging to a particular age cohort. As a robustness check, we also disaggregated the averages and then pooled the individual $\rho_{H,i}$ ($\rho_{C,i}$) values along with each researcher's individual $y_{0,i}$. This disaggregation appropriately accounts for the varying number of data points (individuals) contributing

[6] Following from the same logic as our hypothetical example of two individual articles from the same year and journal in Section 4.2 that differ only in the timing of $\Delta c_{p,t}$. Thus, while nominal net citation counts are equal, $c_{1,T} = c_{2,T}$, the deflated net citations are not, $s_{1,T} \neq s_{2,T}$.

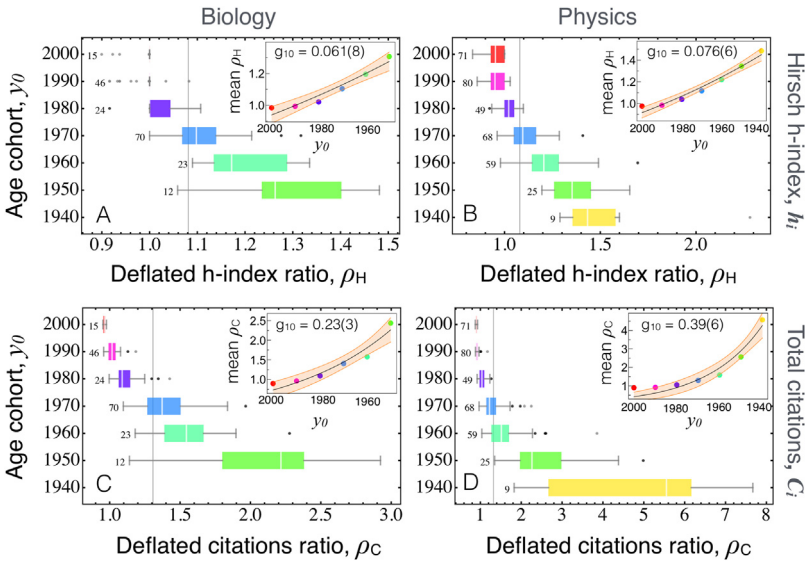

Fig. 7. Deflated productivity and impact measures by discipline and age cohort. The deflated h-index ratio $\rho_{H,i} \equiv h_i^D/h_i$ and the deflated total citations ratio $\rho_{C,i} \equiv C_i^D/C_i$ indicate the measurement error incurred when using nominal versus deflated citation values. Since we have used 2010 as the baseline year, researchers from the most recent cohorts have ratio values close to unity, whereas researchers from earlier cohorts are characterized by a real value boost ($r > 1$). Shown are box-whisker distributions of the deflated h-index ratio ρ_H (A, B) and deflated citations ratio ρ_C (C, D) by age cohort, with the midpoint of each box representing the median value; mean value across all data indicated by vertical black line. The mean deflated h-index ratio value is $\langle\rho_H\rangle = 1.08$ (for both biology and physics). The mean deflated citations ratio value is $\langle\rho_C\rangle = 1.31$ (biology) and 1.32 (physics). (insets) Progression of the mean $\rho_H(y_0)$ and $\rho_C(y_0)$ by each 10-year cohort determined by y_0; shaded regions indicate the 90% confidence interval. Also shown are the estimates of the 10-year exponential growth factor, g_{10}, as defined in Eq. (5); the standard error in the last digit is indicated in parenthesis. *Source*: Authors' elaboration using WOS data (Color online).

to the calculation of each $\langle\rho(y_0)\rangle$ value. Estimating g_{10} in this way resulted in nearly identical values. As such, the calculation of g_{10} is robust to the significant variation in cohort sizes in our researcher data sample.

4.4. Additional sources of citation inflation

The evaluation of scientific careers using citation measures is confounded by another secular trend – the increasing number of coauthors per article (Wuchty et al., 2007; Petersen et al., 2014b; Milojevic, 2014). The annual growth rate in the number of authors per paper is roughly 3–4% in biology and medicine, roughly 4.5% in physics, and 1% in economics; and other collaborative domains, such as patenting, also demonstrate a 1–2% growth rate in the number of coinventors (Petersen et al., 2014b). In everyday terms, whereas the average article from Nature, Science and PNAS had 2 coauthors in 1960, nowadays the typical article from these multi-disciplinary journals has on average 10 coauthors (Pavlidis et al., 2014).

Consequently, in the context where researchers are the primary unit of analysis, the amount of "citation credit" introduced into the scientific system when a publication with a_p coauthors is published is for all intensive purposes compounded by the same factor a_p. That is, each time an article is cited, it introduces a_p citations across all the coauthor profiles, considered separately. At the aggregate level, this corresponds to $c_p \times a_p$ citations introduced into the system for an article with c_p citations. Most quantitative methods for evaluating scientific impact of individual researchers neglect to take into account the number of other authors sharing credit for a common research output, i.e. a single article or patent.

Recent efforts to develop quantitative credit allocation methods that account for a_p are promising (Stallings et al., 2013; Shen and Barabasi, 2014). However, accounting for team size in a fair, transparent, and universal manner remains challenging since it requires authors to accurately agree upon and report their individual contributions (Haeussler and Sauermann, 2013; Allen et al., 2014; Pavlidis et al., 2014; Sauermann and Haeussler, 2017). Using this contribution information to distribute shares of citation credit, so that the total impact of the article is conserved independent of a_p, similar to the approach developed by Stallings et al. (2013) who use mathematical arguments based upon author order, would be an additional challenge. See Waltman (2016) for a review on fractional counting methods that address authorship-related citation inflation, but not without additional challenges.

5. Discussion

Citation inflation is an under-appreciated statistical bias that affects the quantitative evaluation of science – from careers to institutions. Its source is the growth of the scientific endeavor. At the most fundamental level, the growth is the direct result of investment in the researcher population, which over the period 1997–2016 has been growing at an average annual rate of 3.5–5.3% (see Fig. 3). Technological innovations in the publication process have also contributed to the growth of $n(t)$, such as word processing advancements (e.g. layout and equation editing in LaTex), bibliographic management tools, streamlined submission, referee and editorial services, and the advent of online-only journals. Judging by the total volume of literature indexed by the WOS each year, $n(t)$, this quantity exhibits a 4% annual growth – in both the natural and the social sciences. Indeed, $n(t)$ is both a measure for and a solution to the citation inflation problem.

The emergence of 'rapid-publication' mega-journals is another important and relevant factor. To place their growth in real terms, just one decade after the emergence of PLOS ONE, the first mega-journal, the publication volume in 2016 by the 10 largest megajournals (61,000 articles) represented roughly 2.4% of the "Science" articles from the same year indexed in the Web of Science Core Collection (see Fig. 4). In 2016 alone, the journal PLOS ONE published 23,000 articles, roughly 145 times as many as *Research Policy*. These new mega-journals have transformed the industrial throughput of science and democratized the scientific publication process by altering traditional standards for the acceptance of research findings into the larger scientific canon (Petersen, 2018a).

Here we are primarily interested in the impact of scientific growth on the methods for evaluating scientific impact. To address this issue, we first analyzed several fundamental sources of growth to establish consistency, serving also as a robustness check. We then developed a statistical method that factors out this growth which amounts to deflating citations. In particular, our method accounts for the different real value of citations occuring at different times, and is readily applicable whenever an appropriate target research area a can be defined. As such, one drawback of our method is that it requires selecting a target research area a, which for interdisciplinary researchers or institutions may be challenging to define.

The application of this statistical method is necessary because gross citation rates, and thus the real value, of citations are not stationary in time. For example, because the doubling time of publications in the Management, Business, and Economics research domain of *Research Policy* is roughly $\tau_{2\times} = \ln(2)/0.043 = 16$ years, our method indicates that every 16 years the "real value" of a citation is halved in relative terms. This phenomena is quite general, e.g. when evaluating a researcher's funding it is standard practice to deflate nominal dollar values amounts to constant dollars. The concept of deflating performance metrics is also generalizable to quantitative evaluation scenarios outside of science. For example, in order to compare sports achievements and records across era, one must also account for inflation in the total number of possible player opportunities per season, in addition to other cultural and physiological trends that have altered athletic rates of achievement (Petersen et al., 2011). We show that measurement error associated with evaluating career citation-based measures, e.g. $\delta(C)/C$, can be upwards of 100%. For example, the total citation impact of Richard R. Nelson is significantly underestimated using nominal citation tallies, as our method yields a real net citation impact that is larger by 82%.

As such, these results show why citation deflators should be used in scientific evaluation whenever citation tallies are the basis for objective assessment. Indeed, current methods to normalize citations for cross-field or cross-cohort comparison are not sufficient (Vaccario et al., 2017), because they do not account for *when* the citations were received – Section 4.2 shows this to be a subtle yet crucial factor. Take for example the "citation window" method to address temporal right-censoring bias – i.e. when comparing the net citation count c_p of two articles published in different years (Waltman, 2016). The solution proposed is to count only the citations accrued in the first Δt years, $c_{p,\Delta t}$. In this way, both publications are effectively compared at the same age; nevertheless, we emphasize that this approach *is not sufficient* to eliminate the inflation bias because there are more citations produced in the later Δt-year period than the earlier Δt-year period. We demonstrated this in Section 2, where we use a 10-year window to calculate the distribution of citation tallies 10 year post-publication, $c_{p,\Delta t=10}$, and track the corresponding citation value of the top percentiles. Quite clearly, a Social Sciences article published in 2000 (1965) requires 40 (10) citations to be considered in the top 10% – clearly demonstrating the inflation of nominal citation values. Field normalization methods are also not satisfactory, as the measures they produce are not as intuitive (i.e. rescaled factors that are intensive rather than extensive quantities) and often implement a baseline average to factor out disciplinary and age-cohort differences, however such a baseline average is itself susceptible to citation inflation.

We further demonstrated the value of our method in a common research evaluation scenario by analyzing career productivity (h_i-index) and citation impact (C_i) measures for a broad sample of biologists and physicists. Our results in Section 4.3 show that the standard method of aggregating citation counts can lead to a significant underestimation of h_i and C_i, with particularly larger penalties on researchers from earlier age cohorts. This is particularly relevant to tenure and promotion considerations where researchers are being assessed relative to specific thresholds, to the extent that J. Hirsch wrote in his seminal work: "for [physics] faculty at major research universities, $h \approx 12$ might be a typical value for advancement to tenure (associate professor) and that $h \approx 18$ might be a typical value for advancement to full professor" (Hirsch, 2005). While these numbers may have been reasonable l 14 years ago, our analysis shows why they are undervalued by present day measure. Using the rate of change in ρ_H by each age cohort, our estimates indicate that these h-index thresholds should be increased by 6–7% every decade – i.e. each h-index value above should be increased by roughly 1. Similarly, if $C_i = 10{,}000$ citations were a benchmark one decade, our estimates indicate that the benchmark should be increased by 20–40% the following decade, depending on the discipline.

In order to facilitate the implementation of this citation deflator method, Box 1 provides the basic steps to obtain the time series $n_a(t)$ from the Web of Science portal. Indeed, the time series $n_a(t)$ is the only quantity needed to calculate the deflator index defined in Eq. (2). In order to demonstrate the utility of this deflator method in data-driven research, we refer the reader to three recent studies focusing on various units of analysis: individual careers (Petersen et al., 2014a; Petersen, 2018b) and individual publications (Parolo et al., 2015). As such, the contribution of the analysis reported here is to delineate the steps needed to obtain a generic deflator index $D_a(t)$ defined in Eq. (2), and to demonstrate the magnitude of the measurement error associated with quantitative research evaluation that neglects the impact of citation inflation. While in-house methods for normalizing citation measures may be part of the existing "best-practice" within select research evaluation groups, to the best of our knowledge, there is widespread negligence of this fundamental statistical bias.

Thus, a deflator index is needed to precisely correct for inflationary temporal bias, and could be readily incorporated into popular bibliometrics indices such as Web of Science and Google Scholar. Indeed, adjusting for the growth of science in various modeling frameworks is an essential ingredient for understanding the evolution of science as a complex socio-economic system (Stephan, 2012; Fortunato et al., 2018), the quantitative evaluation of scientific production (Wildson, 2015; Wilsdon et al., 2015) and the science of science policy (Fealing, 2011).

6. Conclusions

We conclude with two policy recommendations. First, researchers, scientific evaluators, and the major bibliometric index providers (Web of Science, Google Scholar, Scopus, Microsoft Academic Graph, etc.) should account for citation inflation when analyzing and reporting citation tallies. Second, because the supply of references is a source of inflation, journals should consider standardizing the maximum number of references per publication, depending on the article page length or type (letter, article, review, etc.). This would limit the growth in the total supply of references, and may discourage other bad habits such as self-citation with the intent to surgically enhance one's citation prominence (Hellsten et al., 2007; Fowler and Aksnes, 2007; Costas et al., 2010; Zaggl, 2017; Petersen, 2018a; Biagioli et al., 2019; Seeber et al., 2019).

Data availability statement

Science funding data is openly available from the World Bank (2019). Certain data included herein are derived from the Science Citation Index Expanded, Social Science Citation Index and Arts & Humanities Citation Index, owned by Clarivate Analytics Web of Science, which are not shareable due to binding contract agreement.

Acknowledgements

The authors are grateful 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. AMP, SF, and FP also acknowledge support from EU FP project

"Multiplex". The authors also acknowledge the opportunity to receive feedback via COST Action TD1210 "KnowEscape."

References

ACUMEN, 2018. Highly cited researchers (h > 100) according to their Google Scholar citations public profiles. (Accessed 2018). http://www.webometrics.info/en/node/58.

Acuna, D.E., Allesina, S., Kording, K.P., 2012. Future impact: predicting scientific success. Nature 489, 201.

Allen, L., Brand, A., Scott, J., Altman, M., Hlava, M., et al., 2014. Credit where credit is due. Nature 508, 312–313.

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.

Biagioli, M., Kenney, M., Martin, B.R., Walsh, J.P., 2019. Academic misconduct, misrepresentation and gaming: a reassessment. Res. Policy 48, 401–413.

Binfield, P., 2013. Open access megajournals – have they changed everything? (Posted online 23 October 2013). http://creativecommons.org.nz/2013/10/open-access-megajournals-have-they-changed-everything/.

Bjork, B.C., 2015. Have the 'mega-journals' reached the limits to growth? PeerJ 3, e981.

Bornmann, L., Leydesdorff, L., Mutz, R., 2013. The use of percentiles and percentile rank classes in the analysis of bibliometric data: opportunities and limits. J. Informetr. 7, 158–165.

Bornmann, L., Marx, W., 2015. Methods for the generation of normalized citation impact scores in bibliometrics: which method best reflects the judgements of experts? J. Informetr. 9, 408–418.

Costas, R., van Leeuwen, T.N., Bordons, M., 2010. Self-citations at the meso and individual levels: effects of different calculation methods. Scientometrics 82, 517–537.

de Solla Price, D.J., 1965. Networks of scientific papers. Science 149, 510–515.

Fealing, K.H. (Ed.), 2011. The Science of Science Policy: A Handbook. Stanford Business Books, Stanford, CA, USA.

Fortunato, S., Bergsrom, C.T., Borner, K., Evans, J.A., Helbing, D., Milojevic, S., Petersen, A.M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., Barabasi, A.L., 2018. Science of science. Science 359, eaao0185.

Fowler, J.H., Aksnes, D.W., 2007. Does self-citation pay? Scientometrics 72, 427–437.

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

Haeussler, C., Sauermann, H., 2013. Credit where credit is due? The impact of project contributions and social factors on authorship and inventorship. Res. Policy 42, 688–703.

Hellsten, I., Lambiotte, R., Scharnhorst, A., Ausloos, M., 2007. Self-citations, co-authorships and keywords: a new approach to scientists' field mobility? Scientometrics 72, 469–486.

Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., Rafols, I., 2015. The Leiden Manifesto for research metrics. Nature 520, 429.

Hirsch, J., 2005. An index to quantify an individual's scientific research output. Proc. Natl. Acad. Sci. USA 102, 16569–16572.

Ke, Q., Ferrara, E., Radicchi, F., Flammini, A., 2015. Defining and identifying sleeping beauties in science. Proc. Natl. Acad. Sci. USA 112, 7426–7431.

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.

Liao, H., Mariani, M.S., Medo, M., Zhang, Y.C., Zhou, M.Y., 2017. Ranking in evolving complex networks. Phys. Rep. 689, 1–54.

Luukkonen, T., 1991. Citation indicators and peer review: their time-scales, criteria of evaluation, and biases. Res. Eval. 1, 21–30.

Mariani, M.S., Medo, M., Zhang, Y.C., 2015. Ranking nodes in growing networks: when pagerank fails. Sci. Rep. 5, 16181.

Milojevic, S., 2014. Principles of scientific research team formation and evolution. Proc. Natl. Acad. Sci. USA 111, 3984–3989.

Moed, H.F., 2006. Citation Analysis in Research Evaluation, vol. 9 Springer Science & Business Media.

Moed, H.F., Burger, W., Frankfort, J., Van Raan, A.F., 1985. The use of bibliometric data for the measurement of university research performance. Res. Policy 14, 131–149.

Pan, R.K., Petersen, A.M., Pammolli, F., Fortunato, S., 2018. The memory of science: inflation, myopia, and the knowledge network. J. Informetr. 12, 656–678.

Parolo, P.D.B., Pan, R.K., Ghosh, R., Huberman, B.A., Kaski, K., Fortunato, S., 2015. Attention decay in science. J. Informetr. 9, 734–745.

Pavlidis, I., Petersen, A.M., Semendeferi, I., 2014. Together we stand. Nat. Phys. 10, 700–702.

Penner, O., Pan, R.K., Petersen, A.M., Fortunato, S., 2013a. On the predictability of future impact in science. Sci. Rep. 3, 3052.

Penner, O., Petersen, A.M., Pan, R.K., Fortunato, S., 2013b. Commentary: The case for caution in predicting scientists' future impact. Phys. Today 66, 8–9.

Petersen, A.M., 2015. Quantifying the impact of weak, strong, and super ties in scientific careers. Proc. Natl. Acad. Sci. USA 112, E4671–E4680.

Petersen, A.M., 2018a. Megajournal mismanagement: manuscript decision bias and anomalous editor activity at PLOS ONE. SSRN: e-print:2901272.

Petersen, A.M., 2018b. Multiscale impact of researcher mobility. J. R. Soc. Interface 15, 20180580.

Petersen, A.M., Fortunato, S., Pan, R.K., Kaski, K., Penner, O., Rungi, A., Riccaboni, M., Stanley, H.E., Pammolli, F., 2014a. Reputation and impact in academic careers. Proc. Natl. Acad. Sci. USA 111, 15316–15321.

Petersen, A.M., Majeti, D., Kwon, K., Ahmed, M.E., Pavlidis, I., 2018. Cross-disciplinary evolution of the genomics revolution. Sci. Adv. 4, eaat4211.

Petersen, A.M., Pavlidis, I., Semendeferi, I., 2014b. A quantitative perspective on ethics in large team science. Sci. Eng. Ethics 20, 923–945.

Petersen, A.M., Penner, O., Stanley, H.E., 2011. Methods for detrending success metrics to account for inflationary and deflationary factors. Eur. Phys. J. B 79, 67–78.

Radicchi, F., Castellano, C., 2012a. A reverse engineering approach to the suppression of citation biases reveals universal properties of citation distributions. PLoS One 7, e33833.

Radicchi, F., Castellano, C., 2012b. Testing the fairness of citation indicators for comparison across scientific domains: the case of fractional citation counts. J. Informetr. 6, 121–130.

Radicchi, F., Fortunato, S., Castellano, C., 2008. Universality of citation distributions: toward an objective measure of scientific impact. Proc. Natl. Acad. Sci. USA 105, 17268–17272.

Riccaboni, M., Pammolli, F., Buldyrev, S.V., Ponta, L., Stanley, H.E., 2008. The size variance relationship of business firm growth rates. Proc. Natl. Acad. Sci. USA 105, 19595–19600.

Sauermann, H., Haeussler, C., 2017. Authorship and contribution disclosures. Sci. Adv. 3, e1700404.

Seeber, M., Cattaneo, M., Meoli, M., Malighetti, P., 2019. Self-citations as strategic response to the use of metrics for career decisions. Res. Policy 48, 478–491.

Shen, H.W., Barabasi, A.L., 2014. Collective credit allocation in science. Proc. Natl. Acad. Sci. USA 111, 12325–12330.

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. J. Am. Soc. Inf. Sci. Technol. 63, 1485–1495.

Stallings, J., Vance, E., Yang, J., Vannier, M.W., Liang, J., Pang, L., Dai, L., Ye, I., Wang, G., 2013. Determining scientific impact using a collaboration index. Proc. Natl. Acad. Sci. USA 110, 9680–9685.

Stephan, P., 2012. How Economics Shapes Science. Harvard University Press, Cambridge, MA, USA.

Stephan, P., Veugelers, R., Wang, J., et al., 2017. Blinkered by bibliometrics. Nature 544, 411–412.

Vaccario, G., Medo, M., Wider, N., Mariani, M.S., 2017. Quantifying and suppressing ranking bias in a large citation network. J. Informetr. 11, 766–782.

Vinkler, P., 2010. The Evaluation of Research by Scientometric Indicators. Elsevier.

Wakeling, S., Willett, P., Creaser, C., Fry, J., Pinfield, S., Spezi, V., 2016. Open-access mega-journals: a bibliometric profile. PLoS One 11, e0165359.

Waltman, L., 2016. A review of the literature on citation impact indicators. J. Informetr. 10, 365–391.

Waltman, L., van Eck, N.J., 2013. A systematic empirical comparison of different approaches for normalizing citation impact indicators. J. Informetr. 7, 833–849.

Waltman, L., van Eck, N.J., 2018. Field normalization of scientometric indicators. arXiv e-print:1801.09985.

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., Jones, R., Kain, R., Kerridge, S., Thelwall, M., Tinkler, J., Viney, I., Wouters, P., Hill, J., Johnson, B., 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).

World Bank, 2019. World Bank data sources. (Accessed March 2019). http://data.worldbank.org/indicator.

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. J. Informetr. 11, 608–621.

Zaggl, M.A., 2017. Manipulation of explicit reputation in innovation and knowledge exchange communities: the example of referencing in science. Res. Policy 46, 970–983.

Zhou, P., Leydesdorff, L., 2006. The emergence of China as a leading nation in science. Res. Policy 35, 83–104.