Petersen and Penner *EPJ Data Science* 2014, **3**:24
http://www.epjdatascience.com/content/3/1/24

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

Inequality and cumulative advantage in science careers: a case study of high-impact journals

Alexander M. Petersen[1*] and Orion Penner[2]

*Correspondence:
petersen.xander@gmail.com
[1]Laboratory for the Analysis of Complex Economic Systems, IMT Lucca Institute for Advanced Studies, Piazza San Francesco 19, Lucca, IT-55100, Italy
Full list of author information is available at the end of the article

Abstract

Analyzing a large data set of publications drawn from the most competitive journals in the natural and social sciences we show that research careers exhibit the broad distributions of individual achievement characteristic of systems in which cumulative advantage plays a key role. While most researchers are personally aware of the competition implicit in the publication process, little is known about the levels of inequality at the level of individual researchers. Here we analyzed both productivity and impact measures for a large set of researchers publishing in high-impact journals, accounting for censoring biases in the publication data by using distinct researcher cohorts defined over non-overlapping time periods. For each researcher cohort we calculated Gini inequality coefficients, with average Gini values around 0.48 for total publications and 0.73 for total citations. For perspective, these observed values are well in excess of the inequality levels observed for personal income in developing countries. Investigating possible sources of this inequality, we identify two potential mechanisms that act at the level of the individual that may play defining roles in the emergence of the broad productivity and impact distributions found in science. First, we show that the average time interval between a researcher's successive publications in top journals decreases with each subsequent publication. Second, after controlling for the time dependent features of citation distributions, we compare the citation impact of subsequent publications within a researcher's publication record. We find that as researchers continue to publish in top journals, there is more likely to be a decreasing trend in the relative citation impact with each subsequent publication. This pattern highlights the difficulty of repeatedly producing research findings in the highest citation-impact echelon, as well as the role played by finite career and knowledge life-cycles, and the intriguing possibility that confirmation bias plays a role in the evaluation of scientific careers.

Keywords: science of science; computational sociology; Matthew effect; career growth; citation analysis; reputation; success premium

1 Introduction

The business of science is constantly evolving, on multiple levels and time scales, and this evolution has a profound impact on the institutions and individuals engaged in the production of scientific research. Competition plays a central role in pushing science forward, from the winner-takes-all race for the priority of discovery, to the awarding of research funds, and the challenge in obtaining a tenure-track faculty position [1–3]. However, high

levels of competition and inequality can be detrimental to the overall functioning of the science system [4–8], for example by affecting scientists' decision processes and sentiments of ethical responsibility [3, 9–14], and by altering the entry rate, the exit rate, and the overall appeal of careers in science [2, 15–18].

Ideally, academia should provide a science career path that is sustainable yet competitive and efficient [1, 18–21]. However, the improvement of the current career system in science requires a better understanding of how various complex social ingredients - reputation, cooperation, competition, risk-taking, and creativity - fit together. To begin with, two hallmarks of complex systems stand out as fundamental to improving our understanding of the complex science system:

(i) correlated behavior between individuals, due to the competition for finite resources, the increasing role of collaborative teams in science [3, 22], and ideation process arising from the combination of novel versus grounded ideas [23],

(ii) systemic memory, whereby cumulative advantage and reputation are known to play a strong role when integrated across the career [18, 24–28].

Here we investigate the high levels of inequality across researcher careers, and then quantify the role of cumulative advantage by analyzing longitudinal patterns of productivity and impact. Our focal unit throughout the analysis is the scientific career, even though we use publication and citation counts as the central quantitative measure. Our data comprises 412,498 publications drawn from 23 individual high-impact journals indexed by Thompson Reuters Web of Knowledge (TRWOK). From these data we extracted the publication profile of 258,626 individual scientists, where each trajectory is defined *within a set of journals*.

By analyzing researcher profiles within prestigious journals, we gather insights into the ascent of top scientists and the operational value of these highly-selective 'competitive arenas'. We focus most of the analysis on a case study of two journal sets in parallel, one representing the natural sciences and the other the economic sciences, each comprised of the highest impact journals in each domain. For the natural sciences we aggregated *Nature*, *the Proceedings of the National Academy of Science (PNAS)*, and *Science*. For the economic sciences we aggregated 14 highly cited journals (e.g. *American Economic Review*, *Quarterly Journal of Economics*, etc.), selected based on a page-ranking algorithm applied to journal citation data performed by *SCImago Journal & Country Rank* (http://www.scimagojr.com/index.php). Table 1 lists the journals comprising each journal set j.

In what follows, we explore at length and depth the statistical patterns that reflect the complex social processes underlying cumulative advantage in science. Our data are lim-

Table 1 Summary of journal set datasets

Journal set j	Years	Articles	Authors, R^j
Cell	1974-2012	12,349	20,521 (1,006)
Economics (top 14 journals)	1899-2012	44,571	11,882 (1,791)
Management Sci. (top 3 journals)	1954-2012	18,836	6,801 (479)
Nat./PNAS/Sci.	1958-2012	219,656	123,165 (10,317)
New England J. Med. (NEJM)	1958-2012	18,347	34,828 (916)
Phys. Rev. Lett. (PRL)	1958-2012	98,739	61,429 (13,085)

R^j is the number of 'sufficiently rare' surnames (see the Data & Methods section) we were able to identify in each journal set j over the denoted period. The R^j value in parentheses denotes the number of researcher profiles with $L_i \geq 5$, $N_p \geq 5$, and $y^j_{i,0} \geq 1960$ (Econ.) and $y^j_{i,0} \geq 1970$ (other).

ited in the sense that we are not able to pinpoint the specific covariates associated with cumulative advantage at the individual level (e.g. the emergence of individual reputation [26, 28], access to financial and human capital resources [29], refinement of talent and efficiency, collaboration spillovers [18, 27, 30], etc.). For an in-depth study using a control versus treated regression analysis approach, which astutely pinpoints specific covariates underlying the Matthew effect in science, see [28]. Here we take an alternative data-science approach, using longitudinal trends at the individual career level to provide novel insight into the emergence of cumulative advantage in the context of large number of scientists competing for limited publication space in prestigious journals.

To this end, we begin in Section 2.1 with a visualization of the historical publication patterns of highly-cited scientists in the natural and economic sciences. Following that, we present our analysis of the aggregate citation distribution of individual researchers and observe remarkable statistical regularities in the broad distribution of total citations within each publication 'arena'. We then compare these results with the distribution of longevity and productivity, finding that the skewed productivity distributions persist even among the scientists with the greatest longevity in each journal set. We also calculate the Gini inequality indices for both publications and cumulative citations. These initial descriptive analyses beg the question: How might these skewed distributions, representing relatively high levels of inequality in science, emerge at the micro level of individual careers?

To address this basic question, we used the longitudinal data for individual researchers in two complementary analyses to provide evidence for the manifestation of cumulative advantage. First, in Section 2.2 we analyze the waiting times between successive publications in these highly competitive journals. By analyzing the research profiles of prolific scientists within elite journals, our quantitative method shows how cumulative advantage manifests as an increasing publication rate. In Section 2.3 we present our second main result, showing that the relative citation impact of these researchers tends, on average, to decrease with each subsequent publication.

2 Results

2.1 General evidence of cumulative advantage in scientific careers

Given the complex institutional, economic, and behavioral factors at play in the academic career system it is no surprise that careers in science demonstrate two of the hallmark features of complex systems: strong correlations and long-term memory. For evidence of strong correlations one needs not look further than the collaboration and citation networks, which together serve as a backbone for the flow of reputation [26, 27]. Long-term systemic memory plays a role in the emergence of researcher reputation, and likely plays a strong role in social stratification [31–33]. Consequently, non-linear feedback can amplify small, early career, differences into large differences in successful outcomes over the course of scientific careers, a divergence which follows from integrating the 'Matthew effect' across time [24, 25, 34].

In this section we provide a descriptive analysis of research careers defined within two distinct sets of high-impact journals. The first set of economic researcher profiles are drawn from 14 highly-cited journals in political, financial, theoretical, and empirical economics. The second set of natural science researchers are drawn from the multidisciplinary journals Nature, PNAS, and Science. While we also analyzed other high-impact journal sets in the management science, cell biology, medicine, and physics domains, in

the interest of doing a side-by-side comparison, we focus mainly on the economics and multidisciplinary natural science journals sets. Within each journal set dataset we performed a name disambiguation estimation by analyzing only the research profiles of the sufficiently 'rare' surname + given-name combinations that we aggregated from the author lists. This disambiguation strategy was recently benchmarked on datasets of similar size to ours, demonstrating a remarkably high precision given its basic approach [35]. We defer our in-depth description of our disambiguation approach to the Appendix.

We start with two motivational questions to help guide our intuition on the path researchers take to success: Are the citation trajectories of top-cited scientists similar? Are the growth patterns smooth or marked by singular events? To answer these questions we first calculate the cumulative citation impact achieved by a given researcher, i, via his or her publications in a given journal set, j. It is important to note that citation counts are time and discipline dependent, and so we standardized our citation measures by normalizing each publication's net citation count by the average total citation count of all publications published in the same year y in j. This method effectively suppresses the time and discipline dependence [36, 37].

Hence, the normalized citations of a paper, p, published in a journal belonging to the journal set j in year y is given by

$$\tilde{c}^{j}_{i,p}(y) = c^{j}_{i,p,Y}(y)/\left\langle c^{j}_{Y}(y)\right\rangle, \tag{1}$$

where $c^{j}_{i,p,Y}(y)$ is the total number of citations in census year Y to publication p published in j in year y, and $\langle c^{j}_{Y}(y)\rangle$ is the average citations calculated over all publications in j from the same year. Y is the year when the citation data was collected from TRWOK (corresponding to 2009 for Nat./PNAS/Sci. and 2012 for the economics journals, see the Appendix for further explanation). It is worth mentioning that, despite the fact that Nature, PNAS, and Science are multidisciplinary journals, for the sake of our analysis, controlling for the base citation rate is the most important reason for the normalization in Eq. (1). Hence, in this regard, PNAS, Science and Nature are comparable since they each have roughly the same order of magnitude in their base citations rates (i.e. the total number of times their articles are cited per year).

Using the normalized citation count $\tilde{c}$, we define a scientist's net citation count $\tilde{C}^{j}_{i}(y)$ as the sum,

$$\tilde{C}^{j}_{i}(y) = \sum_{p=1}^{N^{j}_{p}(y)} \tilde{c}^{j}_{i,p}(y). \tag{2}$$

Here $N^{j}_{p}(y)$ represents the scientist's total publications up to year y. The measure is the scientist's cumulative citations measured in units of the mean citation baseline $\langle c^{j}_{Y}(y)\rangle$. For a given researcher, i, the time variable y runs from the first year $y^{j}_{i,0}$ he/she published in j to the arbitrary census year Y. Due to the finite citation life cycle of most publications [26], as long as the difference between Y and y is sufficiently long, then the publication p should have a relative stable ranking amongst the publications from its journal-year cohort. In our citation analyses we require the difference $Y - y$ to be at least 7 years. As such, $\tilde{C}^{j}_{i}(y)$ is a robust measure of cumulative citation impact. Additional methods have

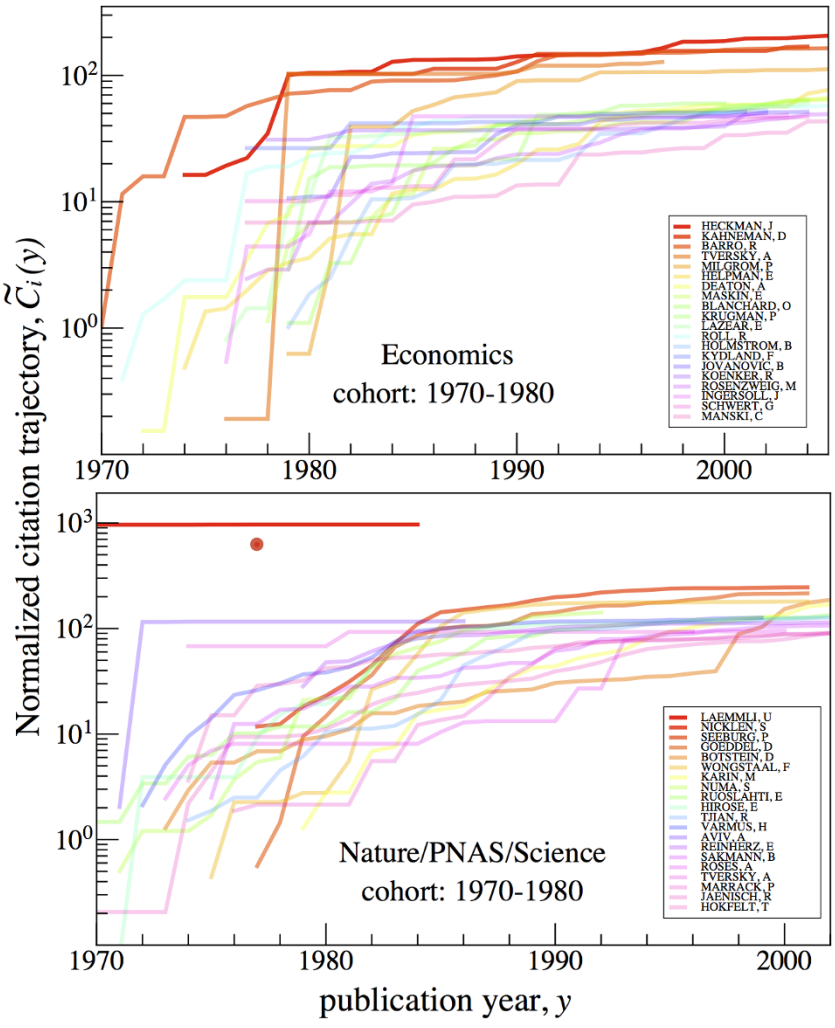

Figure 1 Top-cited scientists within the economists and natural sciences journal sets. Top-20 researchers, ranked by $\tilde{C}$, who had their first publication in the time-period cohort 1970-1980. A trajectory terminates in the last observed year in which there was a publication within the journal set, and so a single dot represents a researcher profile with only a single publication. The citation census year Y was 2012 (economics) and 2009 (Nat./PNAS/Sci.). Hence using a 7 year window to allow the citations to properly accrue, only publications published prior to 2005 for economics and 2002 for Nat./PNAS/Sci. are shown.

also been developed to account for variable team size by further normalizing by coauthor number, thus providing a way to aggregate scientists from varying time, discipline, and even sub-disciplines [38, 39]. In a very general sense, this detrending approach can be easily applied to other competitive arenas, such as professional sports, where success rates can be explicitly era dependent [40].

Figure 1 shows $\tilde{C}_i^j(y)$ trajectories for top-ranked researchers entering the journal sets over the decade 1970-1980 (see Figures S1 and S2 in Additional file 1 for researcher rankings using more recent time windows). In the case of economics there appears to be a greater level of separation (divergence) among the top ranked researches as qualitatively indicated by the gap between the highest-cited scientists (red curves) and the others. Each citation trajectory terminates at the year of the final publication within the journal set. In this way, a single dot corresponds to a scientist with a single publication. Figure 1 begins to

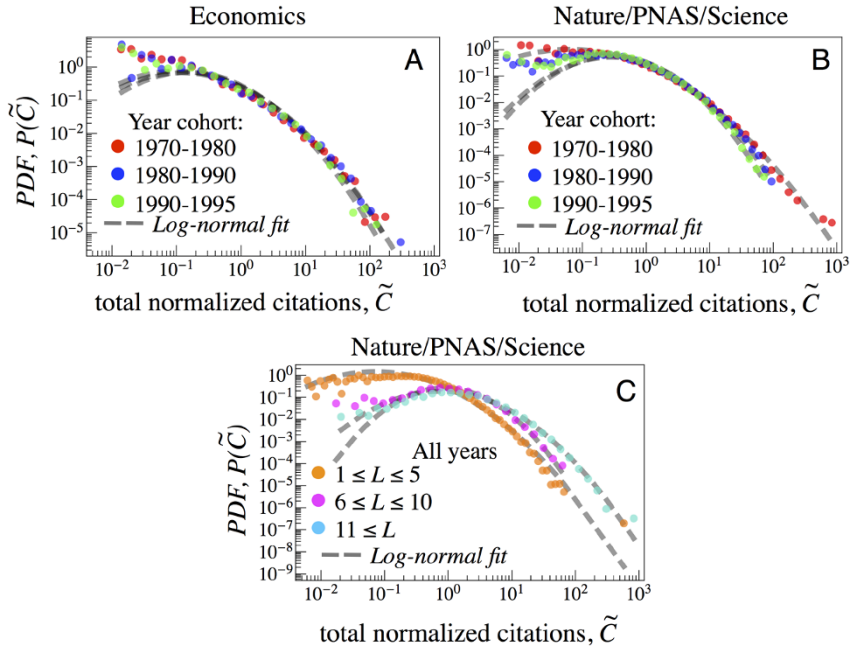

Figure 2 Skewed citation impact distributions. The right-skewed $P(\tilde{C})$, which range across more than 4 orders of magnitude, illustrate the wide range of outcomes in scientific winner-takes-all prestige competitions [1, 41]. To account for censoring bias we show the $P(\tilde{C})$ for the economics arena **(A)** and for the Nat./PNAS/Sci. arena **(B)** by considering non-overlapping scientist cohorts according to when each scientist entered a given arena by conditioning on his/her first publication being in the indicated time window. In **(C)** we show the $P(\tilde{C})$ conditional on the career length within the arena. The dashed grey best-fit curves are calculated using the log-normal distribution maximum likelihood estimator.

provide an answer to our preliminary questions, showing that the group of highest-cited scientists are a mixture of individuals whose accomplishments range from a single, monumental, contribution to persistent stream of high-impact publications, and everything between. However, as we will see below, despite this variability in the paths of ascent, there are remarkable statistical regularities in the distribution of $\tilde{C}_i^j$ across all researchers in each j.

To better understand the relative frequency of 'superstars' we calculated the distribution of normalized career citation counts $P(\tilde{C})$ using logarithmically sized bins to account for the broad distribution of $\tilde{C}_i^j$ values. Because $\tilde{C}_i^j$ controls for the average citation count of papers published within a specific year cohort, it is particularly well-suited for comparing achievements which occurred across a broad time range. Figures 2(A, B) each show three $P(\tilde{C})$ distributions, one for each of the cohorts indicated in the legend. Figure 2(C) shows conditional distributions $P(\tilde{C}|L)$, where L is the length of time between the first and last publication of author i in the journal set j,

$$L_i^j \equiv y_{i,f}^j - y_{i,0}^j + 1. \tag{3}$$

Interestingly, in each panel the aggregate success distribution is well-described by a log-normal distribution,

$$P(\tilde{C}) \propto \tilde{C}^{-1} \exp\left[-(\ln \tilde{C} - \mu)^2 / 2\sigma_{LN}^2\right], \tag{4}$$

Table 2 Summary of the Gini index (G) and top-1% share ($f_{1\%}$)

Journal set j	Cohort entry years	$G(\tilde{C})$	$f_{1\%}(\tilde{C})$	$G(N_p)$	$f_{1\%}(N_p)$
Economics	1970-1995	0.80	0.23	0.54	0.09
	1970-1980	0.83	0.26	0.56	0.10
	1980-1990	0.79	0.21	0.55	0.09
	1990-1995	0.74	0.19	0.47	0.07
Nat./PNAS/Sci.	1970-1995	0.69	0.18	0.46	0.10
	1970-1980	0.74	0.22	0.53	0.12
	1980-1990	0.67	0.15	0.45	0.08
	1990-1995	0.63	0.12	0.35	0.06

Inequality measures are calculated from the distribution of citation impact, $P(\tilde{C})$, and from the distribution of productivity, $P(N_p)$, for the cohorts of scientists whose first publication occurred in the indicated time intervals.

with varying location parameter μ and shape parameter σ, estimated using the log-normal distribution maximum likelihood estimator method. For small $\tilde{C}$ the log-normal fit has larger deviations from the empirical data due to fluctuations in the lower bound of $\tilde{C}$ arising from variability in the value of $\langle c_Y^j(y)\rangle$. Moreover, the poor fit for small $\tilde{C}$ further indicates that the aggregate empirical distributions are likely mixtures of underlying log-normal distributions with slightly varying shape and location parameters.

For example, in the 1980-1990 Economics cohort in Figure 2(A) we calculate $\mu = 0.23$ and $\sigma_{LN} = 1.53$ and for the 1980-1990 Nat./PNAS/Sci. cohort in Figure 2(B) we calculate $\mu = 0.30$ and $\sigma_{LN} = 1.25$. For contrast, the subset of Nat./PNAS/Sci. scientists in Figure 2(C) with $L \geq 11$ (with $\langle L\rangle = 20$, $\langle N_p\rangle = 6.8$ and $\langle\tilde{C}\rangle = 8.3$) have parameters $\mu = 1.31$ and $\sigma_{LN} = 1.26$. These values can be used to model the growth of $\tilde{C}$ using Gibrat's stochastic (proportional) growth model, $\Delta\tilde{C}_t = \tilde{C}_{t-1}(1 + \eta)$, where η is white noise with mean and standard deviation depending on the log-normal counterparts, μ and σ_{LN}. The limiting distribution of this multiplicative process is the log normal distribution (see [42] for recent empirical and theoretical results on firm growth that provides an appropriate starting point for the modeling of researchers' publication portfolios as companies in the small size limit).

To provide additional intuition regarding the level of 'inequality' within these citation distributions, we calculated the Gini index G as well as the citation share $f_{1\%}$ of the top 1% of researchers in each $P(\tilde{C})$. For example, for the 1970-1980 cohort we observe $G = 0.83$ (economics) and $G = 0.74$ (Nat./PNAS/Sci.) and found that the top 1% of researchers (comprised of 17 and 139 researchers, respectively) held a significantly disproportionate share of 26% and 22% of the total $\tilde{C}$ aggregated across all researchers in each distribution. Table 2 shows the $G(\tilde{C})$ and $f_{1\%}(\tilde{C})$ for each cohort group, which indicate for both journal sets a decreasing trend in the citation inequality over time. We note that our calculations do not control for the increasing prevalence of large collaborations in science [3]. Therefore, because there are correlations between the number of coauthors and the average citations a publication receives [22], and because we did not control for multiple counting of single publications in the calculation of the total $\tilde{C}$, it is difficult to assess whether the difference between the inequality values calculated for economics (where coauthorship effect is weak because the number of coauthors is typically small) and for natural sciences is attributable to this feature of the data.

For comparison, a recent analysis of US research funding at the institutional level provides a different picture, indicating a slow but steady increase in the Gini index across U.S. universities over the last 20 years, with current estimates of the Gini inequality index for

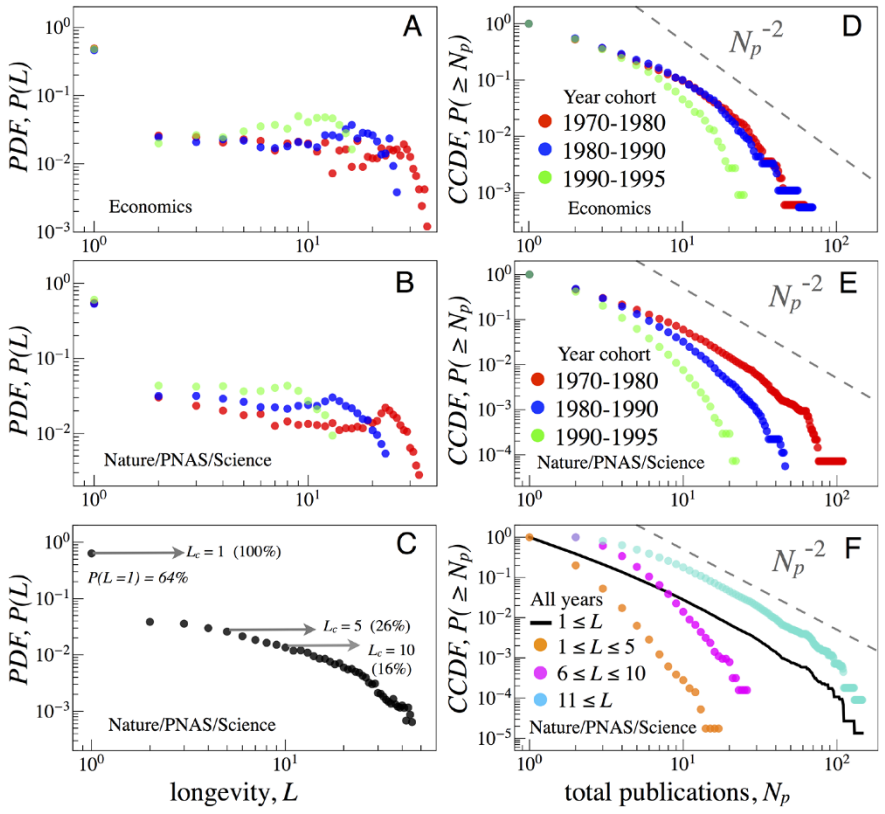

Figure 3 Competitive arenas in science. The right-skewed distributions of scientist longevity L and productivity N_p are the starting point for understanding the skewed success distributions in science. Panels **A** (Economics) and **B** (Nat./PNAS/Sci.) address censoring bias in the longevity distributions by considering separately the cohorts of scientists who entered the Nat./PNAS/Sci. arena in the indicated time window. Panel **C** (Nat./PNAS/Sci.) shows the unconditional aggregate longevity distribution. Panels **D/E/F** show $P(\geq N_p)$, the complementary cumulative distribution function (CCDF) of total publications N_p. Panels D/E are analogs of panels A/B, showing the fixed cohorts of researchers who entered the respective journal arena with their first publication in the indicated time window. Panel F shows the $P(\geq N_p|L)$ distributions conditioned on L, which indicate that even within the set of 'iron horse' researchers with longevity $L_i \geq 11$ the distribution is still extremely broad. To guide the eye and emphasize the curvature in $P(\geq N_p)$, which is not consistent with a power-law distribution, we plotted a curve proportional to N_p^{-2}.

university expenditure around $G \approx 0.8$ [7]. This increasing trend has also been noted in data measuring the share of the top 1% individuals in terms of U.S. income, which has increased from roughly 10% to 20% over the last half century; nevertheless, the 2010 U.S. income Gini coefficient reported was $G = 0.4$ [43], significantly less than what we observed for these citation distributions.

Success is typically assumed to be strongly correlated with career longevity, but to what degree does this assumption hold? In Figure 2, we conditioned the distributions on L_i and find that $P(\tilde{C}|L)$ is still well-described by a log-normal distribution, even after controlling for censoring and survivor bias. Hence, the correlation is somewhat weak, because even among researchers publishing in Nat./PNAS/Sci. for $L \geq 11$ years, the citation distributions still span a huge range, from $\tilde{C} \sim 10^{-1}$ to $\tilde{C} \sim 10^{3}$, with the maximum value being roughly 100 times larger than the characteristic mean value $\langle \tilde{C} \rangle \sim 10^{1}$.

Figures 3(A, B) show the longevity distributions $P(L)$ conditioned on the first publication being within a specified time window. Remarkably, roughly half the scientists enter

and exit the arena in a single year ($L_i^j = 1$), likely with a single publication. At the other end of the distribution, as indicated by the systemic shift in the tail across cohorts, a relatively small set of prolific scientists steadily publish within the arena throughout their scientific careers. The tail of the distribution, beginning around the peak in the far right of the distribution, consists of scientists sustained activity in j for longer than a decade, representing roughly 15-20% of the researchers analyzed. Aggregating across cohorts, Figure 3(C) shows that roughly 64% of authors enter this arena for the minimum time span of 1 year, with only 16% of the entrants publishing over a period $L_i \geq 10$ years.

While Figures 3(A-C) illustrate how long scientists stay active these high-impact arenas, Figures 3(D-F) show the productivity distributions $P(\geq N_p)$ for the same datasets shown in (A-C). The top 20% of the distribution corresponds to individuals publishing roughly five publications or more, signifying a rather broad productivity distribution even amongst the researchers with $L \geq 11$. Indeed, comparing $P(\geq N_p|L)$ conditioned on career length in Figure 3(F), there is a rather large range in N_p, e.g. from 3 to more than 100 publications for the subset with $L \geq 11$. We also note that none of the productivity distributions are consistent with Lotka's productivity law, $P(\geq N_p) \sim N_p^{-\lambda}$, for any value λ.

In order to compare the inequality levels for citation impact to productivity, we also calculated G and $f_{1\%}$ for each productivity distribution $P(\tilde{N_p})$. For example, for the 1970-1980 cohort we calculated $G = 0.56$ (economics) and $G = 0.53$ (Nat./PNAS/Sci.), finding that the top 1% of researchers (comprised of 17 and 139 researchers, respectively) had a share of 10% and 12% of the total publications. Table 2 shows the $G(N_p)$ and $f_{1\%}(N_p)$ for each cohort group, which like the citation inequality counterparts $G(\tilde{C})$ and $f_{1\%}(\tilde{C})$, suggests that productivity inequality is also becoming more equitable over time. However, it is worth noting that citation inequality is substantially larger than publication inequality for each cohort group, arising from the fact that all publications are measured equally and their value does not increase over time, in contrast to citations which accrue over time.

We conclude this section by noting the similarity and differences between the analysis performed in ref. [38]. First, the career citation share and paper share measures defined in [38] normalizes by the number of coauthors (dividing the credit among them equally). Also, a statistical method to eliminate 'unfinished' careers was implemented in [38] but was not used here. Hence, the results in this section, which represent finished and unfinished careers pooled together, neglect the censoring bias arising from including unfinished careers.

2.2 Decreasing waiting times as quantitative evidence of cumulative advantage

In the previous section we showed that the distributions of impact, productivity, and longevity are consistent with a highly competitive 'winner takes all' system. In this section we shift to the longitudinal perspective of researcher trajectories. The schematic in Figure 4(A) emphasizes the sequence of accomplishments as they might occur across a scientist's complex backdrop of career phases (grad student/postdoctoral fellow → assistant professor → tenured faculty). These career phases are characterized by varying roles in the research process, shifts in research interests, and the accumulation of various institutional responsibilities.

Our approach is to measure the longitudinal patterns in the sequence of inter-publication waiting times of individual researchers. It is important to note that we are not analyzing the complete publication profile of each researcher, but rather, just the set

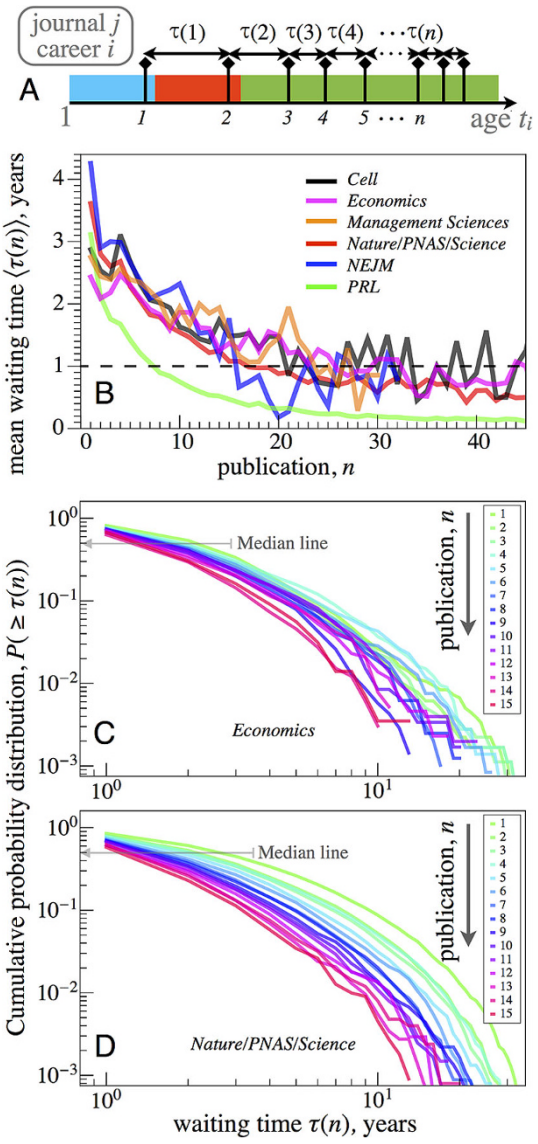

Figure 4 Decreasing inter-publication waiting time $\tau(n)$ is quantitative evidence for cumulative advantage in science. (A) Schematic of a science career, where major accomplishments sustain career growth. Specifically, publications in high-impact journals serve as a record of scientists capitalizing on opportunities for success, and the duration $\tau_i^j(n)$ between a scientist's success n and success $n+1$ provide a quantitative method for analyzing cumulative advantage. We search for quantitative evidence of self-reinforcing social mechanisms by analyzing productivity patterns in specific journal sets that are highly competitive and widely targeted. **(B)** The average waiting time $\langle \tau^j(n) \rangle$ between publication n and publication $n+1$ shows a significant decreasing trend as an author continues to publish in a given journal set. A decreasing $\tau^j(n)$ between publications suggests that an advanced publication career (larger n) facilitates future publications by leveraging reputation, expertise, seniority, and other cumulative resources. The values of $\langle \tau^j(1) \rangle$ are 2.9 yrs. (Cell), 2.4 yrs. (Econ.), 2.8 yrs. (Mgmt. Sci.), 3.6 yrs. (Nat./PNAS/Sci.), 4.3 yrs. (NEJM) and 3.1 yrs. (PRL). The journal PRL exhibits a more rapid decline in $\tau(n)$ because of possible rapidity in successive publications (often by large high-energy experiment collaborations that publish many publications together in a single issue). Only research profiles with $L \geq 5$ years and $N_p \geq 5$ are included in the calculation of these inter-event waiting-time curves. In order to reduce censoring bias arising form careers that started before the beginning of each data sample, we only included trajectories with the first publication year $y_{i,0}^j \geq 1970$ for the natural and management sciences and $y_{i,0}^j \geq 1960$ for the economic sciences. **(C, D)** Complementary cumulative probability distribution, $P(\geq \tau(n))$, for publications $n = 1, \ldots, 15$ in (C) the Economics and (D) Nat./PNAS/Sci. journal sets. The distributions are right-skewed, indicating the possibility of a relatively long waiting time $\tau(n)$ for all n. However, by $n = 10$ the observed likelihood of waiting 3 or more years, $P(\geq 3 | n = 10)$, falls to roughly 0.2 for both Econ. and Nat./PNAS/Sci.

of publications within each journal set j. Given the significant incentives for publishing in top journals, both in terms of prestige [1, 41] and financial benefits [44], we assume that 'if a researcher *could* publish in one of these journals, he/she *would*.' In this regard, the information contained in the waiting times between successive publications can provide quantitative insight into the workings of cumulative advantage.

For each i in j we define a sequence of waiting times, $\tau_i^j(n)$, for which the nth entry is the number of years between his/her publication n and publication $n+1$ in a given journal set j. For example, the average time $\langle\tau^j(1)\rangle$ between an author's first and second publication in both NEJM and Nat./PNAS/Sci. is roughly four years, whereas in the biology journal Cell and the physics journal PRL, the initial mean waiting time is closer to three years.

Figure 4(B) shows that the average $\langle\tau^j(n)\rangle$ decreases significantly with increasing n for each journal arena analyzed. Indeed, by the around the 10th publication the waiting time $\tau^j(10)$ has decreased to roughly 1/2 of the initial waiting time $\tau^j(1)$. Moreover the rate of publications becomes roughly one per year after the 30th publication in the economics journal set, and one per year after the 20th publication in the non-physics journal set, and on average one per year after the 10th publication in PRL.

In order to provide new insights beyond what was already shown in [38], we have extended the waiting-time analysis to the research domains of economics and management science, and have also analyzed the distribution of waiting times $P(\geq \tau^j(n))$ which are shown in Figures 4(C, D) for $n = 1,\ldots,15$. Notably, the systematic shift towards smaller $\tau^j(n)$ is not only reflected by the median and the mean $\tau^j(n)$ value, but is also visible across the entire distribution. Indeed, by $n = 10$ the observed likelihood $P(\geq 3|n = 10)$ of waiting 3 or more years until the next publication (3 years being a characteristic time scale associated with both a scientific project and a scientific collaboration), falls to roughly 0.2 for both Econ. and Nat./PNAS/Sci. A factor likely contributing to this systemic trend is the steady exponential growth in the total number of publications per year (recently measured for physics and cell biology to be around 5% growth per year [26]), as well as a slow but substantial 1% to 5% exponential growth in coauthorship size over time depending on the discipline [3], both of which could account for an overall decrease in publication waiting times.

The significant smaller values for the journal PRL largely reflects the large variations in team size as well as the type of research design - experimental and theoretical - occurring in physics. To elaborate, we ponder three basic pathways to publishing more than one publication in this high-impact journal per year. The first pathway involves a theoretical physicist with a very inspiring year - e.g. Albert Einstein's 1905 'Annus Mirabilis' - who is able to rapidly publish more than one (relatively short, $\leq$6 pages) letters in succession. This pathway, however, is likely unsustainable over the long run. The second pathway involves an experimental physicist working at a large particle collider or national laboratory, working in large teams that publish results with 500 or more coauthors. In this situation, a scientist in a top management position or involved with a critical experimental process may even be able to consistently publish multiple PRL articles per year; For a peculiar example consider L. Nodulman who has 388 PRL publications, but with on average 670 coauthors per publication! The third pathway, present to all scientists independent of discipline, reflects a mixture of the first two pathways, whereby a scientist is embedded in an efficient medium-sized team environment and capitalizes on collaboration spillovers, thereby consistently producing highly-cited publications. We should also mention that

PNAS offers a streamlined publication track ('contributed paper') for select US National Academy of Sciences members, an additional idiosyncratic and rare pathway, which nevertheless contributes to the surprisingly large number of scientists that have numerous publications in the Nat./PNAS/Sci. journal set.

Overall, Figure 4 provides evidence that cumulative advantage plays a strong role when it comes to publishing in elite journals. In fact, the mean waiting time, which can be empirically measured using publication data, also has a simple analytic relation to a position-dependent progress rate $g(n) = 1/\langle\tau(n)\rangle$ within a Poisson process framework, where $g(n)$ is the probability of moving from position n to $n+1$ in a unit time interval. This theoretical model has been tested on both scientific and sports career data, with the interesting feature that small modifications to the progress rate $g(n)$ for small n (early career transition rates) can lead to either a bimodal or a truncated power-law career longevity distributions [25], offering insight into the potential impact of career sustainability policies aimed at early-career researchers.

2.3 A decreasing longitudinal citation trend

In this section we investigate the longitudinal citation impact trends for the publications in each researcher profile. This analysis is related to the delicate topic of 'career predictability' [45–47], but is distinct in the sense that we focus exclusively on the citation impact *within the most prestigious journals* and *relative* to his/her own citation baseline. Hence, as a significant number of publications within each scientist's rank-citation profile [48, 49] are missing from our analysis, it is important to note that we do not contend that the citation trends within the high-impact journal set are representative of the trend within the scientist's entire publication portfolio.

We focus on the publication trajectory of individuals within select high-impact journals, acknowledging that it is likely to reflect factors beyond just the inherent citation impact of his/her average research output. One possibility is that there is no significant change in the citation impact of a researcher's publications over time. A second possibility is that there is an increase in the citation impact with each subsequent publication. This increasing trend is consistent with a researcher being able to leverage prior success to improve their research resources [29] and to leverage reputation within the community to increase their base citation rate [26]. A third scenario is a decrease in the citation impact over time. This negative trend is consistent with an opportunity premium that is provided to accomplished scientists via cumulative advantage, such that new opportunities arrive at effectively a 'lower cost' than the base 'entry cost'.

In order to investigate the longitudinal variation in the citation impact, we map the citation count $c^{j}_{i,p,y}(n)$ of the nth publication p of researcher i to a z-score,

$$z_i(n) \equiv \frac{\ln c^{j}_{i,p,y}(n) - \langle \ln c^{j}_{y} \rangle}{\sigma[\ln c^{j}_{y}]}, \tag{5}$$

which allows for a comparison of citation counts across time. The z-score of the log-citation count in Eq. (5) is measured relative to the mean ($\langle\cdots\rangle$) and standard deviation ($\sigma[\cdots]$) of the logarithm of the citations for a given journal set, j, in a given year, y. This follows naturally since the logarithm of a log-normally distributed variable is a normally distributed variable ($z \sim N(0,1)$), making the z-score an appropriate statistical measure.

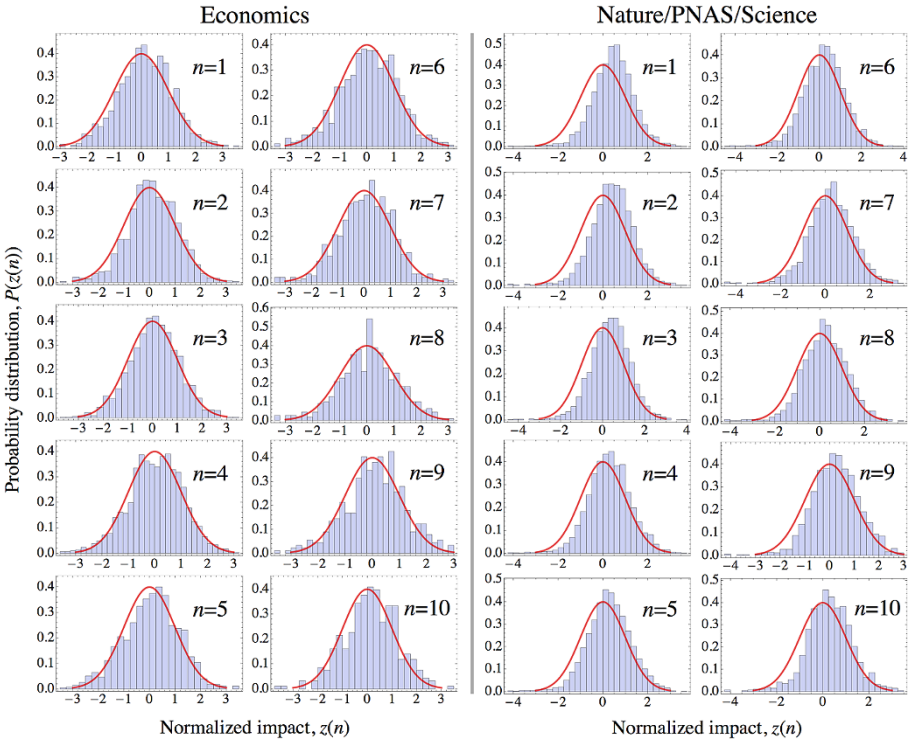

Figure 5 Empirical distribution of citation impact values conditioned on publication number *n*. We aggregate the normalized citation z values of researcher profiles with first publication year $y^j_{i,0} \geq 1970$ for Nat./PNAS/Sci. and $y^j_{i,0} \geq 1960$ for the economic sciences, and with $L_i \geq 5$ and $5 \leq N_p \leq 20$. Each panel shows the probability distribution $P(z(n))$ conditioned on publication number $n = 1, \ldots, 10$. The z-scores represented by each $P(z(n))$ represent a subset of the aggregate set of z values, independent of L_i and N_p. Because the unconditional distribution of z values is approximately normal with mean 0 and with units of the standard deviation ($\sigma_z = 1$), we also plot a normal distribution Normal(0, 1) in each panel for reference (red curve).

We use the convention of replacing c_p by 1 for publications with zero citations; similarly, the mean $\langle \ln c^j_y \rangle$ and standard deviation $\sigma[\ln c^j_y]$ within each journal set are also calculated excluding publications with no citations. This method of dealing with the logarithm of zero has a negligible overall effect, since only 1.5% of publications over the time period 1970-2002 had 0 citations in the census year 2009 for the Nat./PNAS/Sci. journal set, and publications in the economics dataset had only twice this frequency.

Figure 5 shows the distributions of $z_i(n)$ conditioned on the publication number $n = 1, \ldots, 10$ and restricting to researchers with $L_i \geq 5$ and $5 \leq N_p \leq 20$. For example, $P(z(1))$ is the distribution of z-scores for the set of first publications, $P(z(2))$ is the distribution for the set of second publications, and so on. Each $P(z(n))$ is approximately normal, with a mean and standard deviation that deviates only slightly from the baseline Normal$(0, 1)$ distribution (red curve) shown for visual comparison.

Next, in order to account for author-specific heterogeneity before we aggregate citation trajectories across scientists, we centered the z-score around the mean value $\langle z_i \rangle \equiv N_p^{-1} \sum_{n=1} z_i(n)$ calculated for the N_p publications of a given scientist i. As a result, we obtain the relative citation impact trajectory,

$$\tilde{z}_i(n) \equiv z_i(n) - \langle z_i \rangle. \tag{6}$$

Table 3 Summary statistics for two aggregate regression models

Journal set	N_p	A	B	S	p-val.	N_{fit}	R^2
Economics	4-9	1,090	0.17(3)	−0.046(4)	$\mathbf{1 \times 10^{-5}}$	9	0.93
Shuffled	4-9	21,800	−0.003(6)	0.0001(1)	0.68	9	0.03
Economics	10-20	373	0.17(2)	−0.021(4)	$\mathbf{5 \times 10^{-4}}$	10	0.87
Shuffled	10-20	7,460	0.01(1)	−0.002(2)	0.23	10	0.17
Mgmt. Sci.	5-10	262	0.22(9)	−0.05(1)	$\mathbf{6 \times 10^{-3}}$	10	0.63
Shuffled	5-10	5,240	−0.01(3)	0.004(4)	0.40	10	0.09
Mgmt. Sci.	11-20	62	0.5(1)	−0.07(2)	$\mathbf{4 \times 10^{-3}}$	10	0.68
Shuffled	11-20	1,240	0.03(2)	0.005(4)	0.20	10	0.19
Nat./PNAS/Sci.	5-10	3,953	0.15(2)	−0.035(4)	$\mathbf{8 \times 10^{-6}}$	10	0.93
Shuffled	5-10	79,060	−0.006(8)	0.002(1)	0.28	10	0.16
Nat./PNAS/Sci.	11-20	847	0.23(3)	−0.032(4)	$\mathbf{10^{-4}}$	10	0.88
Shuffled	11-20	16,940	0.02(1)	−0.003(1)	0.05	10	0.36

Journal set	N_p	N_d	b	s	p-val.	A	R^2
Economics	4-9	6,183	0.19(3)	−0.053(7)	**0**	1,090	0.012
Economics	10-20	3,730	0.17(3)	−0.022(6)	$\mathbf{3 \times 10^{-4}}$	373	0.005
Mgmt. Sci.	5-10	1,710	0.26(4)	−0.07(1)	**0**	262	0.020
Mgmt. Sci.	11-20	620	0.48(9)	−0.07(2)	$\mathbf{10^{-4}}$	62	0.042
Nat./PNAS/Sci.	5-10	26,010	0.19(1)	−0.048(3)	**0**	3,953	0.013
Nat./PNAS/Sci.	11-20	8,470	0.23(2)	−0.032(4)	**0**	847	0.013

(Top) The regression model (ii) given by Eq.(8): A denotes the number of individual careers that were aggregated for each mean impact trajectory $\langle \tilde{z}(n) \rangle$. B and S are estimated using ordinary least squares, along with the F-test p-value, the number N_{fit} of data points, and the R^2 correlation value. The number in parentheses represents the standard error in the last digit shown. The 'shuffled' values correspond to the parameter estimations using our citation shuffling scheme (conserving the empirical citation distribution) that also allows for an increase in the sample size by a factor of 20). We also include the management science careers for comparison since the dataset contained a sufficient number of researcher profiles to analyze. Bold-faced p-values indicate the regressions with $p \leq 0.01$. (Bottom) The fixed-effects linear regression model (iii) (implemented by the function 'xtreg, vce(robust) fe' in STATA11) given by Eq. (9). We used the 'vce(robust)' Huber-White variance estimator to account for possible heteroscedasticity in the model errors. N_d denotes the number of observations, b and s are the coefficient estimates of the fixed-effects model (value in parenthesis is the robust standard error in the last significant digit), and p-val. corresponds to the model F-statistic $F(1, A-1)$.

This normalization also helps in controlling for latent effects arising from disciplinary variation within each j that can affect the citation potential of a paper over time. Using these standardized $\tilde{z}_i(n)$ trajectories, we pooled the data across scientists, noting that $\tilde{z}_i(n)$ is still measured in normalized units of the standard deviation $\sigma_{\ln c}$.

We also separated the researcher data into two sets of profiles, one with medium N_p and the other with relatively large N_p, requiring in both cases that $L_i \geq 5$ so that increasing n is more likely to correlate with increasing time. In order to reduce censoring bias arising from careers that started before the beginning of each data sample, we only analyzed trajectories with the first publication year $y^j_{i,0} \geq 1970$ for Nat./PNAS/Sci. and $y^j_{i,0} \geq 1960$ for the economic sciences.

For both disciplines and for each N_p subset we observed on average a negative trend in $\tilde{z}_i(n)$. We show this negative trend at two levels of aggregation outlined below, first at the individual level in method (i), and then at the systemic level in methods (ii) and (iii). Table 3 shows the summary statistics and parameter estimates for models (ii) and (iii).

(i) In order to analyze trends at the researcher level, we first analyzed each individual $\tilde{z}_i(n)$ separately by performing an ordinary least squares parameter estimation of the parameters of the basic linear model

$$\tilde{z}_i(n) = b_i + s_i n + \epsilon. \tag{7}$$

Figure 6 shows the cumulative distribution $P(\leq s_i)$ for four scientist subsets (see Fig. S3 in Additional file 1 for the analogous plots for the management science researcher profiles). In each case, the average value $\langle s_i \rangle$, indicated by the vertical blue line, is negative at the indicated p-value shown within each sub-panel (using the 1-sided z-statistic with the null hypothesis that $s = 0$). The $P(\leq 0)$ value, ranging between 60% to 70% across the four panels, indicates the excess proportion of the population with negative s_i. The asymmetry towards statistically significant negative s_i values is even more pronounced. For example, consider the asymmetry in the large N_p subsets: of the 373 economics profiles we analyzed, only 2 (0.5%) had p-val. < 0.01 and $s_i > 0$ whereas 18 (5%) had p-val. < 0.01 and $s_i < 0$; of the 847 Nat./PNAS/Sci. profiles we analyzed, only 8 (1%) had p-val. < 0.01 and $s_i > 0$ whereas 60 (7%) had p-val. < 0.01 and $s_i < 0$.

(ii) In the first aggregate method we calculated the mean citation impact z-score $\langle \tilde{z}(n) \rangle$ across all researcher profiles within j for a given n, and then performed the ordinary least squares parameter estimation of the analogous aggregate model,

$$\left\langle \tilde{z}(n) \right\rangle = B + Sn + \epsilon. \tag{8}$$

We plot $\langle \tilde{z}(n) \rangle$ (solid black curve) and the best-fit regression (dashed green line) for each researcher subset in Figure 6. To give an example, Figure 6(A), which refers to scientists in the Nat./PNAS/Sci. subset with between 5 and 10 publications, shows that the mean impact trajectory decreases by $S = 0.035 \pm 0.004$ - roughly 25% of $\langle \tilde{z}(1) \rangle$ - with each subsequent publication. This means that after the 4th publication, the relative impact typically is 'subpar' *with respect to a given scientist's mean* $\langle z_i \rangle$. Interestingly, for the cohort of scientists in Figure 6(B) with between 11 and 20 publications, the impact trajectory starts at a higher value, and since the slope is approximately equal to the slope in panel (A), the publications do not become subpar until after the 7th publication. We observe the analogous trends for the economics journal set. However, the S value for the relatively low-N_p economics subset in panel (C) is significantly more negative than the value estimated for the high-N_p researcher set in panel (D).

(iii) The previous model doesn't account for the fact that observations are not independent (since $\tilde{z}(n)$ values within each subset n also depend on i), and that the data are unbalanced (since N_p vary across researchers in each dataset). Hence, we apply a hierarchical approach in this second aggregate method by running an unbalanced fixed-effects regression with standard errors clustered by author i,

$$\tilde{z}_{i,p} = b + sn_{i,p} + \epsilon_{i,p}, \tag{9}$$

implemented using the STATA11 regression 'xtreg, vie(robust) fe'. We used the 'vce(robust)' option to implement the 'Huber/White/sandwich' estimate of the standard errors in order to account for possible heteroscedasticity in $\tilde{z}_{i,p}$. This approach also accounts for time-invariant characteristics of the authors. The parameter estimates in Table 3 of this hierarchical regression model show that the estimated coefficients B and b estimated in Eqs. (8) and (9) are consistent in value. The main difference is the explained variance provided by each method. Method (ii) indicates a large R^2 because it eliminates the variance in $\tilde{z}$ by representing only the systemic average, whereas the low R^2 value in method (iii) is a reminder that there are important hidden covariates affecting citation impact that are not captured by this simple model. Other covariates which have been shown

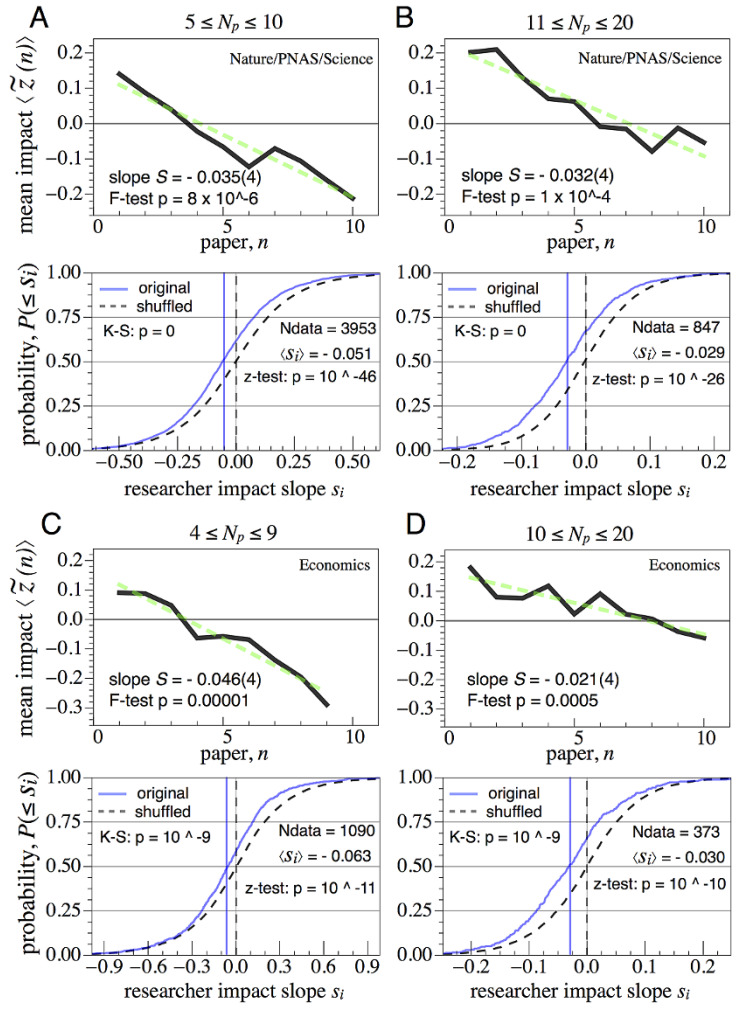

Figure 6 Evidence consistent with confirmation bias and a counter-effective role of cumulative advantage. We test whether the relative citation impact $\tilde{z}_i(n)$ decreases, increases, or is independent of n. While repeated publication in a highly competitive journals reflects the underlying quality of the researcher, it also indicates a strong role played by other factors such as author/institutional reputation and social ties with the journal editors and the referee base, and in the case of PNAS, membership in the US National Academy of Sciences. **(A)** Scientists with between 5 and 10 publications in the Nat./PNAS/Sci. arena. **(B)** Scientists with between 11 and 20 publications in the Nat./PNAS/Sci. arena. **(C)** Economists with between 4 and 9 publications in the top economics journal set arena. **(D)** Economists with between 10 and 20 publications in the top economics journal set arena. (A-D) For each cohort analyzed, the top panel shows a significant negative trend in $\langle\tilde{z}(n)\rangle$ (black curve) with each successive publication. Linear regression of each $\langle\tilde{z}(n)\rangle$ is shown by the dashed green line, with the best-fit slope and regression F-test p-value listed in each panel. In the lower half of each panel we show the empirical cumulative distribution $P(\leq s_i)$, and list the number of trajectories analyzed and the mean value $\langle s_i\rangle$ (indicated by the vertical solid blue line). For comparison, we also plot the $P(\leq s_i)$ for the shuffled data (dashed black curve), with the mean shuffled value (vertical dashed gray line). We apply the Kolmogorov-Smirnov test between the empirical and shuffled distributions, and for each panel we list the p-values that confirm that the underlying s_i values belong to different distributions. Only research profiles with $L \geq 5$ years were analyzed. In order to ensure that the relative citation impact z_p of a given publication had sufficient time to stabilize within the journal set dataset, only publications published prior to 2002 for Nat./PNAS/Sci. (since the publication citation counts used were current as of census year 2009) and 2005 for Economics (since citation counts used were current as of census year 2012) were analyzed. In order to reduce censoring bias arising form careers that started before the beginning of each data sample, we only included trajectories with the first publication year $y^j_{i,0} \geq 1970$ for the Nat./PNAS/Sci. and $y^j_{i,0} \geq 1960$ for the economic sciences.

to explain citation impact are team size [22], institutional prestige [32], conceptual novelty [23], and author reputation [26].

Additionally, in order to check that our results are not affected by systematic sampling bias, we analyzed the same sets of impact trajectories in panels (A-D) using a shuffling method to destroy the author-specific correlations across time. To be more specific, for a given scientist i we conserved his/her number of publications within the dataset. However, we randomly assigned a c_y^j to each of his/her publications, replacing the true citation value with a randomly drawn c_y^j value from the same year y and journal set j. Because in our shuffling algorithm we sampled without replacement, this technique conserves the overall probability distribution $P_y^j(c)$ of citations within a given journal set within a given year, and hence $\langle \ln c_y^j \rangle$ and $\sigma[\ln c_y^j]$ also remain unchanged, as do each $P(z(n))$. This shuffling technique also permits an increase in the number of trajectories analyzed within each subsample since we can reshuffle the data numerous times. Hence, for each journal set we increased the sample size by producing 20 shuffled synthetic datasets, thereby increasing the number of trajectories we analyzed by the same factor.

With respect to method (i), we tested the likelihood that the original s_i values and the shuffled s_i values arise from the same distribution by applying the Kolmogorov-Smirnov test between the original and shuffled cumulative distributions, $P(\leq s_i)$. In each case the p-value is less than 10^{-8}, rejecting the null hypothesis that the two sets of s_i values belong to the same distribution (values reported within each sub panel of Figure 6). With respect to method (ii), we also tested the model in Eq. (8) for each shuffled $\langle \tilde{z}_i(n) \rangle$, finding no significant positive or negative trend (see Table 3 for F-test p-values). Altogether, the comparison of the shuffled and empirical trajectories confirms that our estimates of S and s_i are not sensitive to systematic sampling artifacts.

Figure 7 shows a scatter plot which allows for the visual comparison of four descriptive variables for each researcher trajectory: the impact score of the 1st publication $z_i(1)$, the trajectory slope s_i, the mean impact value $\langle z_i \rangle$, and the total number of publications $N_{p,i}$. This scatter plot indicates an overall negative relation between $z_i(1)$ and s_i, indicative of the difficulty in sustaining high-impact research as well as the lack of predictive information contained in early achievement, $z_i(1)$. To further investigate their relation, for each journal set we estimated the coefficients of the linear regression model, $s_i = \beta_0 + \beta_1 z_i(1) + \beta_2 \log_{10} N_{p,i} + \beta_3 \langle z_i \rangle$. Consistent with the scatter plot, we observed the coefficient for $z_i(1)$ to be negative ($\beta_1 = -0.14$ for Econ. and Mgmt. Sci. and -0.17 for Nat./PNAS/Sci., each estimate statistically significant at the $p \approx 0$ level). Consistent with the S values for the medium versus large N_p subsets, we also observed a positive coefficient for $\ln N_{p,i}$ ($\beta_2 = 0.08$ for Econ., 0.17 for Mgmt. Sci., and 0.10 for Nat./PNAS/Sci., each estimate statistically significant at the $p = 0.01$ level). In each regression the coefficient for $\langle z_i \rangle$ was not statistically significant and the adjusted R^2 was roughly 0.35.

3 Conclusion

What can data science offer to the science of science? By leveraging the rich longitudinal, geographic, and cross-sectional aspects of large publication and patent datasets, new insights into career growth amidst the unabating competition for scientific credit [50] can provide institutions and policy makers important knowledge on how to assess and react to paradigm shifts in science.

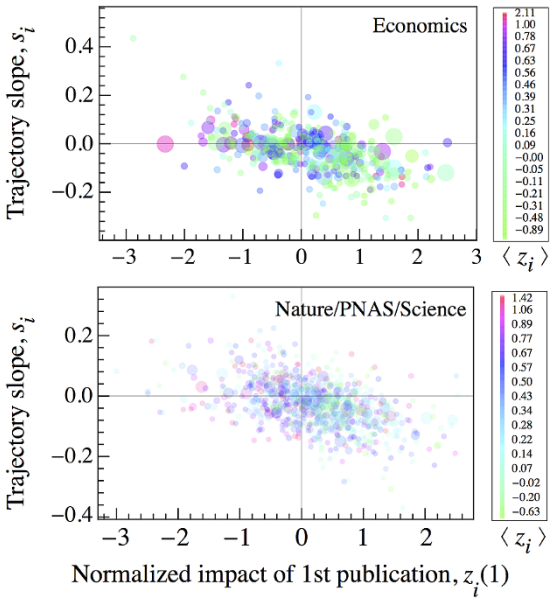

Figure 7 Visualizing the covariates underlying s_i. Scatterplot visualization of the impact score of the 1st publication $z_i(1)$, the trajectory slope s_i, the mean impact value $\langle z_i \rangle$ (color of the data point), and the total number of publications N_p (proportional to the size of the data point). These scatter plots demonstrate that there is an increased likelihood that a decreasing trend ($s_i < 0$) starts with a positive above-average citation value ($z_i(1) > 0$), indicative of how difficult it can be to consistently publish above-average research. A linear regression model indicates a negative relation between s_i and $z_i(1)$ and a positive relation between s_i and N_p (see discussion in text).

3.1 Success distributions in science

Here we have provided evidence that research careers exhibit the broad distributions of individual success characteristic of competitive systems in which cumulative advantage plays a key role. The inequality in research career activity in high-impact journals can be appreciated by considering the Gini coefficient calculated from the distribution of individual researcher productivity and impact. For example, pooling the Nat./PNAS/Sci. publication profiles that began within the period 1970-1995, we observed a Gini index $G = 0.46$ for publications and $G = 0.69$ for citations. For economics we observed even higher levels of inequality, with $G = 0.54$ for publications and $G = 0.80$ for citations. The fraction $f_{1\%}$ of the total output produced by the top 1% further demonstrates the disproportionate productivity levels even among scientists publishing in top ranked journals: $f_{1\%} = 0.09$ for publications and $f_{1\%} = 0.23$ for citations (economics), and $f_{1\%} = 0.10$ for publications and $f_{1\%} = 0.18$ for citations (Nat./PNAS/Sci.). Hence, it is important to note that the inequality amongst researchers is much greater when considering impact measures than for productivity measures. For perspective, the G values we calculated are larger than those observed for individual income in many developed nations of the world [43]. Nevertheless, with respect to individual achievement in science, we have provided evidence that the system became more equitable over the period 1970-1995.

3.2 On the role of cumulative advantage in academic career evaluation

The role played by the 'Matthew effect' is largely considered to be positive [24]. Indeed, cumulative advantage represents a 'positive' feedback mechanism that arises from the func-

tionally meritocratic system of science, which aptly rewards scientists who succeed in producing high-quality research [51].

Using a reasonably large and representative number of career profiles that satisfied our censoring bias criteria, we provided quantitative demonstration of how cumulative advantage in the publication process emerges, showing that the time between publications in top journals decreases as function of how many publications a researcher has published in those journals. This decrease is evident not only in the mean waiting time, but as a systematic shift in the distribution of waiting times towards smaller τ values.

It is, perhaps, unsurprising to practicing researchers that as a researcher places more of his or her publications in a top journal that the preexisting publication barriers progressively decrease. There are a number of anecdotally well-accepted mechanisms that likely contribute to this phenomena, being as simple as an increase in research funding resulting from previous high profile publications, the ability to attract the best graduate students, election into a prestigious academy, or simply an editor spending five additional minutes evaluating a new submission by a prominent scientist before making the initial reject or review decision. Nonetheless, it is important that this phenomena be quantified using longitudinal researcher profiles from distinct research fields.

Our first quantitative observation of a decreasing waiting time between publications is consistent with the reasonable assumption that, given a researcher's history of publishing in high-impact journals, his/her next publication is likely to also be high-impact. However, this hypothesis is inconsistent with our second quantitative finding that on average there is a statistically significant decrease in the relative impact of each subsequent publication ($S < 0$) when conditioning on the publication number n. We also observed this imbalance at the individual level, finding more researcher trajectories with statistically significant decreasing trend ($s_i < 0$) than with statistically significant increasing trend ($s_i > 0$), although this asymmetry contributes less to the overall negative S value than the aggregate trend across all scientists. In other words, the decreasing trend is not attributable to individual scientists per se, but rather, is representative of a larger aggregate trend.

Nevertheless, it is important to consider how reputation arising from highly-cited papers may contribute to a detrimental false-positive rate due to the intrinsic noise associated with success outliers [52]. For example, a side-effect of a systematic type-II confirmation-bias error in the identification of high quality research(ers) may induce a 'crowding out' of young and inexperienced scientists. This is not to say that there are not enough opportunities to go around, but that in light of the broad distribution of N_p, it is important to know what role reputation plays in detecting signal from noise. Interestingly, in our analysis of $\langle \tilde{z}(n) \rangle$, we found that the set of researchers with larger N_p cross the zero baseline for a larger n value than the subset with smaller N_p, which was also supported by the positive value of the β_1 coefficient relating s_i and N_p. Together, these two observations indicate that cumulative advantage is functioning properly in the case of researchers with large N_p. It will be important in follow-up research to add more researcher covariates to further test the origin of the non-zero s_i.

So what do our results mean in the context of academic careers? It is difficult to interpret the decreasing impact trend ($S < 0$) as a desirable property of cumulative advantage in science. Since it is likely a researcher consistently publishing in high impact journals is also gaining access to greater resources, it is disappointing that the impact trend is not, at least, stable, if not increasing. But we also have to be careful in over-interpreting this result,

since we have shown that impact decreases relative only to the author's average citation impact $\langle z_i \rangle$. Additional explanations for the negative S value and the relative abundance of individual negative s_i values are the difficulty in sustaining high-impact research in the top citation percentile, aging across intrinsic creativity and career life-cycles [19, 53], and aging within knowledge life-cycles reflecting the difficulty in staying at the innovative front of science [54–57].

More generally it is important to discuss the impact of cumulative advantage upon how individual careers evolve and are evaluated. In a system with even a subtle feedback loop, small advantages at an early stage compound over time and can produce stratification at later stages. In the case of academic careers this stratification process can be accelerated by the fact that many careers leave academia at a relatively early stage. Recently that competition increased by the emergence of a 'PhD bubble' characterized by an unreasonably high market valuation of graduate education, resulting in an excessive supply of doctoral degrees. Evidence for this supply-demand imbalance in the US are evident in the number of PhDs awarded relative to tenure-track openings [1–3, 58].

It is important to keep in mind that a small advantage in the early stage can just as easily be due to noise as due to signal. To avoid type I and II errors in career evaluation, extra care should to be taken in evaluating the entire publication portfolio of early stage researchers, not just their high-impact factor publications, to reduce the possibility that early publication success is misinterpreted as a signal of high research potential. On the contrary, it is also important to avoid the scenario in which a scientist is eliminated merely because he/she failed to publish early and consistently in top journals. For early career researchers, especially those with relatively few (and recent) publications, quantitative citation metrics should be used mainly as an initial tool to reduce the candidate pool size [45, 46].

Furthermore, a decreasing barrier to publication in top journals with increasing achievement and reputation (here proxied by n) is important to consider for two reasons. First, one should consider the advantage an early stage researcher has in publishing in top journals via collaboration with a senior research possessing an outstanding track record. Second, the lowering of impact with continued publishing means that, perhaps, higher impact publications by less established researchers are being overlooked by the top journals in favor of lower impact publications by more established ones. In this sense, due to the implicit competition for the select publication slots in highly visible and reputable journals, the current system may be crowding out less established researchers, an inefficiency within the reward system of science suggesting that 'the cream may not always rise to the top.'

It is clear that research careers are multifaceted and complex and in studying them many aspects must be taken into account. Specifically, it is crucial to better understand the role that both social and knowledge networks play in the career growth process, and perhaps one day, understanding how they can be predicted in order to manipulate both research and career success strategies. The most readily available data source for producing insight on careers, and scientific progress in general, is publication metadata. However, this data is shaping how careers are both studied by the science of science community, as well as how academics ad hoc measure their impact and the impact of colleagues. As a result citations are pushed to the forefront, again both in terms of how careers are studied and how researchers view themselves and colleagues. In this regard, we are entering an era where the 'hunters become the hunted.'

3.3 The role of scientometric data science

Moving forward, what can scientometrics offer towards our understanding of careers in science against the backdrop of implicit competition and reward? On one hand, citation data are well-suited for developing testable models of longitudinal productivity and impact dynamics within and across research careers [18, 26]. On the other hand, it can be quite technically challenging (*ex.* overcoming author ambiguity [59, 60]) to extend these analyses beyond productivity and impact and into the social network even if we use the coarse proxy of co-authorship. In Merton's seminal paper 'The Matthew Effect in Science' [24] he outlines the various specific mechanisms by which the reputation premium (Matthew effect) is generated in academic careers. Those mechanisms, however, do not manifest themselves purely in the citation data. Thus it is also important that data outside publication metadata be accessed to shed further light on the role of cumulative advantage. For example, it is important to better understand the embedding of researchers in other advantageous social networks, ones which cannot be captured by co-authorship.

However complex a role cumulative advantage plays in research careers it is a key problem that must be addressed both by the community of researchers studying careers, as well as the gatekeepers of the academic profession, which are often researchers themselves. As with nearly all advances in scientometrics, data must play a critical role and this work represents a small example of how existing data can be exploited to better understand the vast issue of cumulative advantage, and raises the important question as to whether or not the cumulative advantage plays an overall positive roll in the scientific selection process.

Appendix: Data and methods

A.1 Our data-science approach

We defined researcher subsets using several thresholds to account for sources of censoring bias in the data. (a) We removed career profiles with relatively short longevity $L < 5$ years between the first and last publication. (b) We only analyzed profiles with first publication year $y^j_{i,0}$ at least a decade after the starting year of the dataset so that we could be reasonably confident that the first publication observed was actually the researchers first publication within the dataset. (c) We conditioned the careers on the number of publications N_p to ensure that there are sufficient statistics to quantify a trend in the citation impact trajectory $\tilde{z}_i(n)$. (d) In our analysis of the citation impact trajectory we only included publications that were published at least 7 years before the TRWOK citation census year Y (corresponding to the data download date which was $Y = 2012$ for the economics journals and $Y = 2009$ for Nat./PNAS/Sci.) to ensure that each publication had a sufficient time to accrue citations which we use as a proxy for research impact. With this time lag, the distribution $P^j_y(c)$ has time to converge to a log-normal distribution, and the ranking of publications within j is likely to become sufficiently stable that the z value is a robust measure of relative impact.

A.2 Name disambiguation

The 'disambiguation problem' is a major hurdle in the analysis of scientific careers as career profiles may be split or aggregated resulting in inaccurate portraits of productivity and impact. Recent methods have been proposed to solve this problem, ranging from relatively simple name disambiguation methods (as employed here) which provide sufficient

accuracy within a reasonably small dataset [35, 61], to more sophisticated network-based solutions that are more appropriate for comprehensive databases like *Thomson Reuters Web of Knowledge (TRWOK)* [60] and comprehensive patent office data (e.g. USPTO) [59].

From *TRWOK* we downloaded annual publication data for 3 high-impact multidisciplinary journals *Nature*, *Proceeding of the National Academy of Sciences USA*, and *Science*; 3 discipline-specific journals *Cell*, the *New England Journal of Medicine (NEJM)*, *Physical Review Letters (PRL)*; 14 top 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*; and 3 management science journals *Management Science*, *Operations Research*, *Organization Science*. For the natural science journals we restricted our analysis to publications denoted as 'Articles', which excludes reviews, letters to editor, corrections, and other content types. For the economics publications we restricted our analysis to the publication types: 'Articles,' 'Reviews' and 'Proceedings Paper'. Natural science journal data were downloaded and curated in 2009, meaning that the citation counts we analyze do not include citations arriving afterwards. Similarly, the economics and management science journal data were downloaded in 2012.

For a given journal set j we aggregate publications together and create a registry of surname and first/middle-initial pairs {*Surname, FM*} where *FM* can consist of one, two, or three alphabetic characters α, hence $FM = \alpha_1\alpha_2\alpha_3$. For a given journal set, we aggregate and analyze the publications associated with {*Surname, FM*} if it is sufficiently rare in the entire database using the following criteria: if there is only one instance of *FM* for a given {*Surname, FM*} then it is used; however, if there is more than one type of $\alpha_2\alpha_3$ for a given α_1, then this surname and first/middle-initial pairs is omitted from the analysis. For example, we consider *Smith, AM* and *Smith, BM* as not being in conflict, but treat {*Smith, AM*} and {*Smith, A*} as indeterminately distinct authors and so we exclude all profiles with {*Smith,* $A\alpha_2\alpha_3$} from our analysis.

For each {*Surname, FM*} that meets this criteria, we aggregate the corresponding publications together creating a profile which is assigned to author i in a given journal set j. This simple initials-based disambiguation method is well-suited for datasets of similar size to those analyzed here, with demonstrated precision (1-'contamination rate') ranging from 95-97% [35].

We use this method under the assumption that there is no intrinsic bias associated with selecting sufficiently rare {*Surname, FM*} pairs, and hence, the set of 'rare' surname profiles should provide a representative sample from the entire career distribution [61]. Indeed, there are some notable scientists with sufficiently common surnames that are omitted from our analysis, e.g. Stanley HE and Vogelstein B, but we maintain that the number of profiles analyzed is sufficiently large to include a representative proportion of these elite careers comprising the tail of the productivity and citation impact distributions. This assumption appears to be valid, as recent analysis comparing the aggregate h-index distribution $P(h)$ comprising all scientist profiles within the TRWOK dataset with the $P(h)$ comprising only the 'extremely rare' scientist profiles within the TRWOK dataset shows that the distributions are remarkably similar except in the extreme right tail, which is only a finite-size effect due to the difference in dataset sizes [60].

We also note that one source of selection bias arising from the selection of rare surnames is the bias against common Asian and Anglo-Saxon names and in favor of underrepresented nationalities in science. Correcting for this bias is difficult without information on the distribution of surnames in science; however, we assume that its affect is negligible since our simple method was able to extract a significant number of prolific profiles with $5 \leq N_p \leq 20$ within each journal set, providing ample statistics in order to analyze the overall longitudinal trends in citation impact. Future avenues of research in this general direction may benefit from additional covariates, including gender, nationality, and ethnic background, in order to better understand the possible sources of bias.

Additional material

Additional file 1: Supplementary additional figures.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
AMP downloaded, curated, and cleaned the data and performed the statistical analysis. AMP and OP designed research and wrote the manuscript.

Author details
[1]Laboratory for the Analysis of Complex Economic Systems, IMT Lucca Institute for Advanced Studies, Piazza San Francesco 19, Lucca, IT-55100, Italy. [2]Laboratory of Innovation Management and Economics, IMT Lucca Institute for Advanced Studies, Piazza San Francesco 19, Lucca, IT-55100, Italy.

Acknowledgements
The authors thank the four anonymous referees for their diligent service. We also thank the discussant P. Stephan and other participants of the 'The Organization, Economics and Policy of Scientific Research' 2013 BRICK conference for helpful comments on an early version of the manuscript, and A. Rungi and R. Metulini for guidance with the implementation of the fixed-effects regressions. The authors acknowledge support from Italian PNR project 'CRISIS Lab.' AMP acknowledges the opportunity to present the research and receive feedback via COST Action TD1210 'KnowEscape.' OP acknowledges funding from the Canadian SSHRC.

Received: 25 April 2014 Accepted: 19 September 2014 Published online: 22 October 2014

References

1. Stephan P (2012) How economics shapes science. Harvard University Press, Cambridge
2. Schillebeeckx M, Maricque B, Lewis C (2013) The missing piece to changing the university culture. Nat Biotechnol 31(10):938-941
3. Petersen AM, Pavlidis I, Semendeferi I (2014) A quantitative perspective on ethics in large team science. Sci Eng Ethics. doi:10.1007/s11948-014-9562-8
4. Weinberg BA, Owen-Smith J, Rosen RF, Schwarz L, Allen BM, Weiss RE, Lane J (2014) Science funding and short-term economic activity. Science 344(6179):41-43
5. Alberts B, Kirschner MW, Tilghman S, Varmus H (2014) Rescuing us biomedical research from its systemic flaws. Proc Natl Acad Sci USA 111:5773-5777
6. Schekman R (2013) How journals like Nature, Cell, and Science are damaging science. Guardian 12/09/2013
7. Xie Y (2014) 'Undemocracy': inequalities in science. Science 344:809-810
8. Couzin-Frankel J (2014) Chasing the money. Science 344:24-25
9. Maher B (2008) Poll results: look who's doping. Nature 452:674-675
10. Committee on the Conduct of Science (1989) On being a scientist. Proc Natl Acad Sci USA 86(23):9053-9074
11. Committee on Science, Engineering, and Public Policy, National Academy of Sciences, National Academy of Engineering, and Institute of Medicine (2009) On being a scientist: a guide to responsible conduct in research, 3rd edn. The National Academies Press, Washington
12. Fang FC, Steen RG, Casadevall A (2012) Misconduct accounts for the majority of retracted scientific publications. Proc Natl Acad Sci USA 109(42):17028-17033
13. Martinson BC, Anderson MS, De Vries R (2005) Scientists behaving badly. Nature 435:737-738
14. Martinson BC, Anderson MS, Crain AL, De Vries R (2006) Scientists' perceptions of organizational justice and self-reported misbehaviors. J Empir Res Hum Res Ethics 1:51-66
15. Anderson MS, Ronning EA, De Vries R, Martinson BC (2007) The perverse effects of competition on scientists' work and relationships. Sci Eng Ethics 13:437-461
16. Martinson BC (2011) The academic birth rate. EMBO Rep 12(8):758-762
17. Freeman R, Weinstein E, Marincola E, Rosenbaum J, Solomon F (2001) Competition and careers in biosciences. Science 294:2293-2294

18. Petersen AM, Riccaboni M, Stanley HE, Pammolli F (2012) Persistence and uncertainty in the academic career. Proc Natl Acad Sci USA 109:5213-5218
19. Committee on Science, Engineering, and Public Policy (2014) The arc of the academic research career issues and implications for U.S. science and engineering leadership, The National Academies Press, Washington
20. Janger J, Nowotny K (2013) Career choices in academia. Welfare wealth work for Europe working paper. http://www.foreurope.eu/fileadmin/documents/pdf/Workingpapers/WWWforEurope_WPS_no036_MS64.pdf
21. Janger J, Strauss A, Campbell DFJ (2013) Academic careers: a cross-country perspective. Welfare wealth work for Europe working paper. http://ideas.repec.org/p/feu/wfewop/y2013m8d0i37.html
22. Wuchty S, Jones BF, Uzzi B (2007) The increasing dominance of teams in production of knowledge. Science 316(5827):1036-1039
23. Uzzi B, Mukherjee S, Stringer M, Jones B (2013) Atypical combinations and scientific impact. Science 342:468-472
24. Merton RK (1968) The Matthew effect in science. Science 159:56-63
25. Petersen AM, Jung W-S, Yang J-S, Stanley HE (2011) Quantitative and empirical demonstration of the Matthew effect in a study of career longevity. Proc Natl Acad Sci USA 108(1):18-23
26. Petersen AM, Fortunato S, Pan RK, Kaski K, Penner O, Rungi A, Riccaboni M, Stanley HE, Pammolli F (2014) Reputation and impact in academic careers. Proc Natl Acad Sci USA. doi:10.1073/pnas.1323111111
27. Sarigl E, Pfitzner R, Scholtes I, Garas A, Schweitzer F (2014) Predicting scientific success based on coauthorship networks. EPJ Data Sci 2014:9
28. Azoulay P, Stuart T, Wang YM (2014) Effect or fable? Manag Sci 60:92-109
29. Duch J, Zeng XHT, Sales-Pardo M, Radicchi F, Otis S, Woodruff TK, Amaral LAN (2012) The possible role of resource requirements and academic career-choice risk on gender differences in publication rate and impact. PLoS ONE 125:51332
30. Azoulay P, Zivin JSG, Wang J (2010) Superstar extinction. Q J Econ 125:549-589
31. Cole JR (1981) Social stratification in science. The University of Chicago Press, Chicago
32. Jones BF, Wuchty S, Uzzi B (2008) Multi-university research teams: shifting impact, geography, and stratification in science. Science 322(5905):1259-1262
33. Deville P, Wang D, Sinatra R, Song C, Blondel VD, Barabasi A-L (2014) Career on the move: geography, stratification, and scientific impact. Sci Rep 4:4770
34. van de Rijt A, Kang SM, Restivo M, Patil A (2014) Field experiments of success-breeds-success dynamics. Proc Natl Acad Sci USA 111:6934-6939
35. Milojevic S (2013) Accuracy of simple, initials-based methods for author name disambiguation. J Informetr 7:767-773
36. 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
37. Radicchi F, Castellano C (2012) A reverse engineering approach to the suppression of citation biases reveals universal properties of citation distributions. PLoS ONE 7:33833
38. Petersen AM, Wang F, Stanley HE (2010) Methods for measuring the citations and productivity of scientists across time and discipline. Phys Rev E 81:036114
39. Radicchi F, Castellano C (2011) Rescaling citations of publications in physics. Phys Rev E 83:046116
40. Petersen AM, Penner O, Stanley HE (2011) Methods for detrending success metrics to account for inflationary and deflationary factors. Eur Phys J B 79:67-78
41. David PA (2008) The historical origins of 'open science': an essay on patronage, reputation and common agency contracting in the scientific revolution. Capital Soc 3:5
42. Fu D, Pammolli F, Buldyrev SV, Riccaboni M, Matia K, Yamasaki K, Stanley HE (2005) The growth of business firms: theoretical framework and empirical evidence. Proc Natl Acad Sci USA 102:18801-18806
43. Chin G, Culotta E (2014) The science of inequality: what the numbers tell us. Science 344:819-821
44. Franzoni C, Scellato G, Stephan P (2011) Changing incentives to publish. Science 333(6043):702-703
45. Penner O, Petersen AM, Pan RK, Fortunato S (2013) Commentary: the case for caution in predicting scientists' future impact. Phys Today 66(4):8-9
46. Penner O, Pan RK, Petersen AM, Fortunato S (2013) On the predictability of future impact in science. Sci Rep 3:3052
47. Acuna DE, Penner O, Orton CG (2013) Point/counterpoint: the future h-index is an excellent way to predict scientists' future impact. Med Phys 40(11):110601
48. Petersen AM, Succi S (2013) The Z-index: a geometric representation of productivity and impact which accounts for information in the entire rank-citation profile. J Informetr 7(4):823-832
49. Petersen AM, Stanley HE, Succi S (2011) Statistical regularities in the rank-citation profile of scientists. Sci Rep 1:181
50. Radicchi F, Fortunato S, Markines B, Vespignani A (2009) Diffusion of scientific credits and the ranking of scientists. Phys Rev E 80:056103
51. Kleinberg J, Oren S (2011) Mechanisms for (mis)allocating scientific credit. In: Proceedings of the forty-third annual ACM symposium on theory of computing. STOC '11. ACM, New York, pp 529-538
52. Denrell J, Liu C (2012) Top performers are not the most impressive when extreme performance indicates unreliability. Proc Natl Acad Sci USA 109:9331-9336
53. Levin SG, Stephan PE (1991) Research productivity over the life cycle: evidence for academic scientists. Am Econ Rev 81:114-132
54. Stephan PE, Levin SG (1993) Age and the Nobel prize revisited. Scientometrics 28:387-399
55. Jones BF, Weinberg BA (2011) Age dynamics in scientific creativity. Proc Natl Acad Sci USA 108:18910-18914
56. Horlings E, Gurney T (2013) Search strategies along the academic lifecycle. Scientometrics 94:1137-1160
57. Jones BF, Reedy EJ, Weinberg BA (2014) Age and scientific genius. NBER working paper, 19866
58. Cyranoski D, Gilbert N, Ledford H, Nayar A, Yahia M (2011) Education: the PhD factory. Nature 472:276-279
59. Lai R, D'Amour A, Yu A, Sun Y, Torvik V, Fleming L (2011) Disambiguation and co-authorship networks of the US patent inventor database, vol 2138. Harvard Institute for Quantitative Social Science, Cambridge
60. Schulz C, Mazloumian A, Petersen AM, Penner O, Helbing D (2014) Exploiting citation networks for large-scale author name disambiguation. EPJ Data Sci 2014:11
61. Mazloumian A (2012) Predicting scholars' scientific impact. PLoS ONE 7(11):49246

doi:10.1140/epjds/s13688-014-0024-y
Cite this article as: Petersen and Penner: Inequality and cumulative advantage in science careers: a case study of high-impact journals. *EPJ Data Science* 2014 **3**:24.