INTERFACE

royalsocietypublishing.org/journal/rsif

Research

Cite this article: Li J, Yin Y, Fortunato S, Wang D. 2020 Scientific elite revisited: patterns of productivity, collaboration, authorship and impact. *J. R. Soc. Interface* **17**: 20200135. http://dx.doi.org/10.1098/rsif.2020.0135

Received: 17 December 2019
Accepted: 31 March 2020

Subject Category:
Life Sciences–Mathematics interface

Subject Areas:
computational biology, biocomplexity

Keywords:
science of science, scientific careers, Nobel Prize, big data

Author for correspondence:
Dashun Wang
e-mail: dashun.wang@northwestern.edu

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.4938111.

THE ROYAL SOCIETY PUBLISHING

Scientific elite revisited: patterns of productivity, collaboration, authorship and impact

Jichao Li[1,2,3,4], Yian Yin[2,3,5], Santo Fortunato[6,7] and Dashun Wang[2,3,4,5]

[1]College of Systems Engineering, National University of Defense Technology, Changsha, People's Republic of China
[2]Center for Science of Science and Innovation, [3]Northwestern Institute on Complex Systems, [4]Kellogg School of Management, and [5]McCormick School of Engineering, Northwestern University, Evanston, IL, USA
[6]School of Informatics, Computing, and Engineering, and [7]Indiana University Network Science Institute (IUNI), Indiana University, Bloomington, IN, USA

DW, 0000-0002-7054-2206

Throughout history, a relatively small number of individuals have made a profound and lasting impact on science and society. Despite long-standing, multi-disciplinary interests in understanding careers of elite scientists, there have been limited attempts for a quantitative, career-level analysis. Here, we leverage a comprehensive dataset we assembled, allowing us to trace the entire career histories of nearly all Nobel laureates in physics, chemistry, and physiology or medicine over the past century. We find that, although Nobel laureates were energetic producers from the outset, producing works that garner unusually high impact, their careers before winning the prize follow relatively similar patterns to those of ordinary scientists, being characterized by hot streaks and increasing reliance on collaborations. We also uncovered notable variations along their careers, often associated with the Nobel Prize, including shifting coauthorship structure in the prize-winning work, and a significant but temporary dip in the impact of work they produce after winning the Nobel Prize. Together, these results document quantitative patterns governing the careers of scientific elites, offering an empirical basis for a deeper understanding of the hallmarks of exceptional careers in science.

1. Introduction

According to Zuckerman [1], scientific elites 'are worthy of our attention not merely because they have prestige and influence in science, but because their collective contributions have made a difference in the advance of scientific knowledge'. Indeed, across the broad spectrum of sciences, scientific elites are often pathbreakers and pacesetters in the science of their time [2–7]. Understanding patterns governing the careers of scientific elites helps us uncover insightful markers for exceptional scientific careers, useful for scientists and decision-makers who hope to identify and develop individual careers and institutions [8].

The Nobel Prize, widely regarded as the most prestigious award in science, offers a unique opportunity to systematically identify and trace many of the world's greatest scientists [1,3,8–15]. These scientific elites have attracted interest from a wide range of disciplines [1,3,8,11,12,15–27], spanning sociology, economics, psychology and physics. On the one hand, quantitative studies analysing publication and citation records have mainly focused on the prize-winning work alone, helping uncover a set of highly reproducible patterns ranging from understanding the link between age and creativity [3,16,17,28–30] to allocating credits and recognition [4,15,19,21]. On the other hand, Zuckerman's canonical work [1] probes into the *entire* career histories of Nobel laureates through qualitative methods [13,14,16,31–35]. The rich patterns articulated by Zuckerman vividly highlight the need to go beyond their prize-winning works, and put them in the context of the entire careers of

laureates. Together, the two strands of research call for a quantitative, career-level analysis relying on large-scale datasets to study patterns of productivity, collaboration, authorship and impact governing the careers of scientific elites.

Despite the recent surge of interest in the science of science [3,19,28,29,36–43] and efforts in constructing large-scale datasets of scholarly activities [3,44–46], large-scale studies of the career histories of Nobel laureates remained limited, largely owing to the difficulty in collecting systematic data for their scientific contributions. Here, by combining information collected from the Nobel Prize official websites, laureates' university websites, Wikipedia entries, publication and citation records from the Microsoft Academic Graph (MAG) (https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/), and extensive manual curations, we constructed a unique dataset capturing career histories of nearly all Nobel laureates in physics, chemistry, and physiology or medicine from 1900 to 2016 (545 out of 590, 92.4%) [47]. We cross-validated this dataset with four different approaches to ensure the reliability of our results. We deposited the derived dataset in a public data repository [48], and described our data collection and validation procedures in a data descriptor with great detail [47].

We further constructed a comparison dataset of scientific careers using data from the Web of Science (WOS) and Google Scholar (GS) [46], representing the kinds of 'ordinary' careers that tend to be studied in the science of science literature [29,49]. For each laureate who published the first paper after 1960, we randomly selected 20 scientists in the same discipline who started their careers in the same year (electronic supplementary material, S1). Note that the goal here is not to create a matching sample of Nobel-calibre scientists, but a comparison group consisting of scientists who are more similar to typical scientists in the field. One advantage of this comparison approach is that, by selecting individuals with long careers and well-maintained GS profiles, it covers scientists with relatively higher visibility and impact than typical scientists, indicating that our comparisons offer a conservative estimate of the difference between Nobel laureates and their contemporary peers.

2. Results

2.1. Early performance

Widely held is the belief that the great minds do their critical work early in their careers [3,16,17], prompting us to ask if there is any early signal that distinguishes Nobel laureates. Here, we focus on the first 5 years since their first publication and measure their productivity and impact at this early stage of their careers. Consistent with Zuckerman's observation [1], we find that Nobel laureates were energetic producers from the outset, publishing almost twice as many papers as scientists in our comparison group (figure 1*a*). Yet, compared with this productivity difference, more impressive is the gap in impact. Indeed, the future laureates had a more than sixfold increase over the comparison group in terms of the rate of publishing hit papers, defined as papers in the top 1% of rescaled 10 year citations (equation (4.1)) in the same year and field (electronic supplementary material, S3.1) (figure 1*b*). This difference is not simply driven by the early onset of prize-winning works. Indeed, we repeated our measurements by omitting the careers of laureates who published their prize-winning work in this period, finding that a substantial gap remained (electronic supplementary material, figure S1).

To conceptualize the observed difference in productivity and impact, we separated team- and solo-authored papers, finding that both types of work boost early performance, but they do so in different ways: most of the difference in early productivity is accounted for by team-authored papers, as solo-authored papers show meagre productivity difference between the laureates and their comparison group (figure 1*c*), documenting a greater propensity towards collaborations for scientific elites in their early careers [1]. The only exception is physics laureates, who published slightly more solo-authored papers than their comparison group (1.73 versus 1.07, Student's t-test, p-value = 0.07). Yet, interestingly, solo-authored papers in early careers turned out to be disproportionally more likely to be prize-winning papers than team-authored ones. Indeed, comparing the fractions of prize-winning papers within solo- and team-authored papers, we find that the former is about twice as high as the latter on average (χ^2 test, p-value $< 10^{-11}$, figure 1*d*).

2.2. Career before the prize

Figure 1 documents the outstanding early performance of future laureates. This is consistent with the innovation literature, which shows that the most important works tend to occur early in the life cycle [3,16,50], speaking to the idea that great, young minds disproportionally break through. Yet, on the other hand, growing evidence shows that ordinary scientific careers are governed by the random impact rule [28], predicting that the highest impact work occurs randomly within the sequence of works. To reconcile these two schools of thought, we focus on the career of laureates before they were awarded the Nobel Prize and measure the positions of the prize-winning work and highest impact work within the sequence of works one produced. Here, the paper impact is measured by rescaled 10 year citation (Methods). We find both types of works tend to occur early within the sequence of papers (figure 2*a*), a result that contradicts the random impact rule governing typical scientific careers [28,46]. Yet, our earlier analysis suggests that a selection effect may offer a potential explanation for this observation [51]—since the Nobel Prize in science has never been awarded posthumously, those who produced groundbreaking works early were more likely to wait long enough to be recognized [20,22]. Indeed, we removed prize-winning papers and calculate among the remaining ones the position of the highest impact papers. We find that the timing of each of the three remaining highest impact works for Nobel laureates all follow clearly uniform patterns [51] (figure 2*b*). This means, apart from the prize-winning works, all other important works in Nobel careers closely follow the random impact rule: they could be, with equal likelihood, the very first work, the last or any one in between. This observation is in line with the recent discovery of hot streaks that occur at random within individual careers [46], and therefore raises an important next question: Are these high-impact works clustered together in time?

To answer this question, we quantify the relative timing between the two most-cited papers (N^* and N^{**}) within each career by calculating the joint probability $P(N^*, N^{**})$ with a null model in which the two papers each follow their independent temporal patterns. We uncovered clear diagonal patterns across all three domains (figure 2*c–e*), showing that high-impact papers are more likely to cluster together than expected

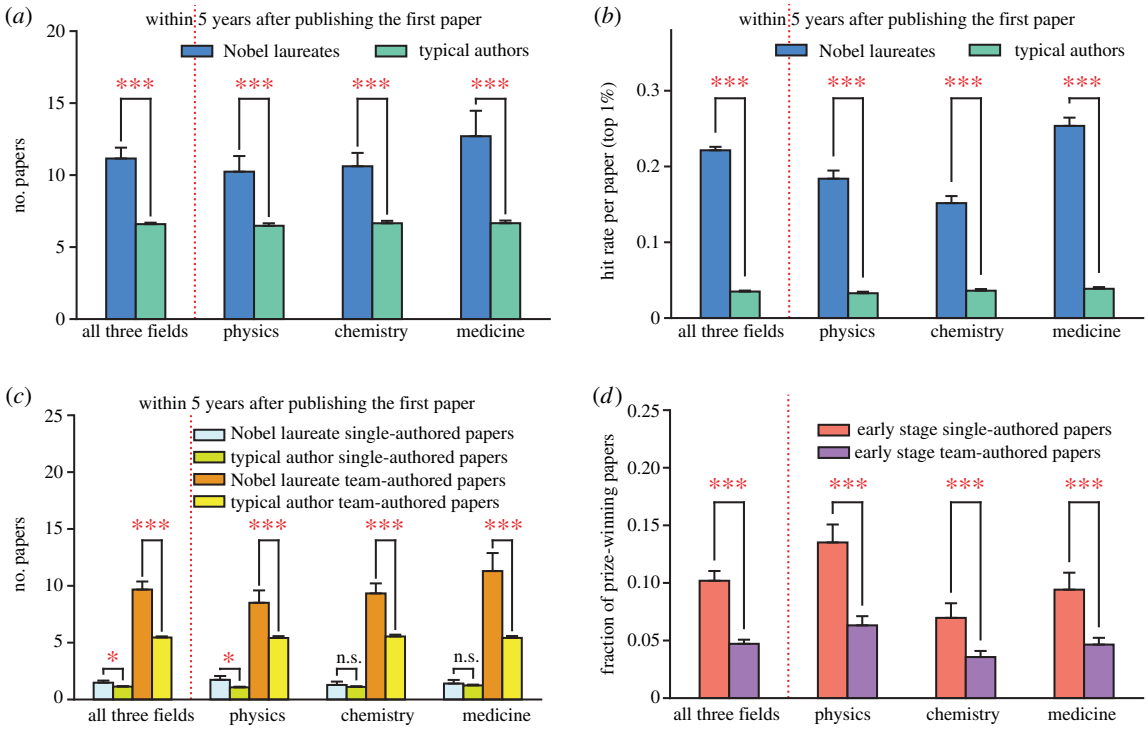

Figure 1. Early career performance. By early career stage, here we mean the first 5 years after publishing the first paper. (*a*,*b*) Early performance of Nobel laureates compared with typical authors in terms of productivity and hit rate per paper (top 1%). We chose typical authors with at least 10 years of career length, and consider Nobel laureates and GS typical scientists with their first paper published after 1960. We randomly selected 20 typical authors with the same first-paper publishing year and research domain, eventually leading to 3540 scientists for 177 Nobel laureates. (*a*) In terms of productivity, the Nobel laureates are indeed more productive (11.15 versus 6.59, Student's *t*-test, *p*-value $< 10^{-7}$). (*b*) When it comes to impact, the two populations are not comparable, and the hit paper rate (probability of publishing papers in the top 1% of rescaled 10 year citations in the same year and field) of the Nobel laureates is 6.33 times higher than for typical authors. (*c*) Much of the difference in early productivity between Nobel laureates and typical scientists resulted from joint papers (9.67 versus 5.46, Student's *t*-test, *p*-value $< 10^{-7}$). In chemistry and medicine, there was no significant difference between the average number of single-authored papers published by laureates in their early stage and the average author. In physics, instead, Nobel laureates publish slightly more single-authored papers than typical authors (Student's *t*-test, *p*-value = 0.07). (*d*) We further compare the fractions of prize-winning papers within all laureates' early stage single-authored papers and team-authored papers published in early stages. The former is 2.16 times as high as the latter on average (χ^2 test, *p*-value $< 10^{-11}$). As for different disciplines, the ratios are 2.13, 1.95 and 2.03 times for physics, chemistry and medicine, respectively. ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$ and n.s. (not significant) for $p > 0.1$. Error bars represent the s.e.m.

by chance. The diagonal pattern disappeared when we shuffle the order of the works, while preserving the random impact rule (figure 2*f–h*). We also measured the distribution of the longest streak within a career L, finding that $P(L)$ follows a broader distribution than that in shuffled careers across all three disciplines (figure 2*i–k*) (electronic supplementary material, S4.3–S4.5). We further find that their hot streaks occur randomly within the sequence of works (figure 2*l*), and are not associated with any detectable change in the overall productivity (figure 2*m*, Kolmogorov–Smirnov test, *p*-value = 0.18). Together, these results demonstrate a remarkable resemblance between the career histories of Nobel laureates and those of ordinary scientists [46].

What seems to distinguish the Nobel laureates from ordinary scientists, however, is that they are disproportionately more likely to have more than one hot streak. Indeed, while a hot streak is usually unique for typical scientists [46], Nobel laureates are characterized by 1.93 hot streaks on average (figure 2*n*). Furthermore, their hot streaks also tend to sustain for longer. We measured the duration distribution of hot streaks for Nobel laureates, finding that it peaks around 5.2 years (figure 2*o*), compared with 3.7 years for typical scientists [46]. The longer duration of laureates' hot streaks is also captured by its proportion over career length (figure 2*p*). We also find that prize-winning works are disproportionately more likely to be produced during hot streaks (figure 2*q*). Overall, the vast majority of all Nobel-winning works (88%) occurred within hot streaks.

2.3. Collaboration patterns

One of the most fundamental shifts in science over the past century is the flourishing of large teams across all areas of science [29,39,52,53]. Compared with the overall rate of this shift, Nobel laureates' papers are produced by an even higher proportion of large teams (figure 3*a*). One possible factor that may explain this team-size difference is impact, as larger teams tend to produce papers with higher impacts [37]. To control for this factor, we created a matching sample for each paper published by the laureates by selecting 20 papers from the same field and year but with the most similar number of citations. We find that, after controlling for impact, the Nobel laureates' papers are still more likely to be produced by larger teams in all times across the last century (figure 3*b*).

Figure 3*a*,*b* thus underscores another similarity between Nobel and ordinary careers, highlighting the increasing reliance of team work across all types of scientific careers. Yet, the ubiquitous increase in team size can be in tension with the fact that the Nobel Prize can only be awarded to at most three recipients for each subject every year [1], prompting us to compare the team size of all prize-winning papers with those published immediately before and after them by the

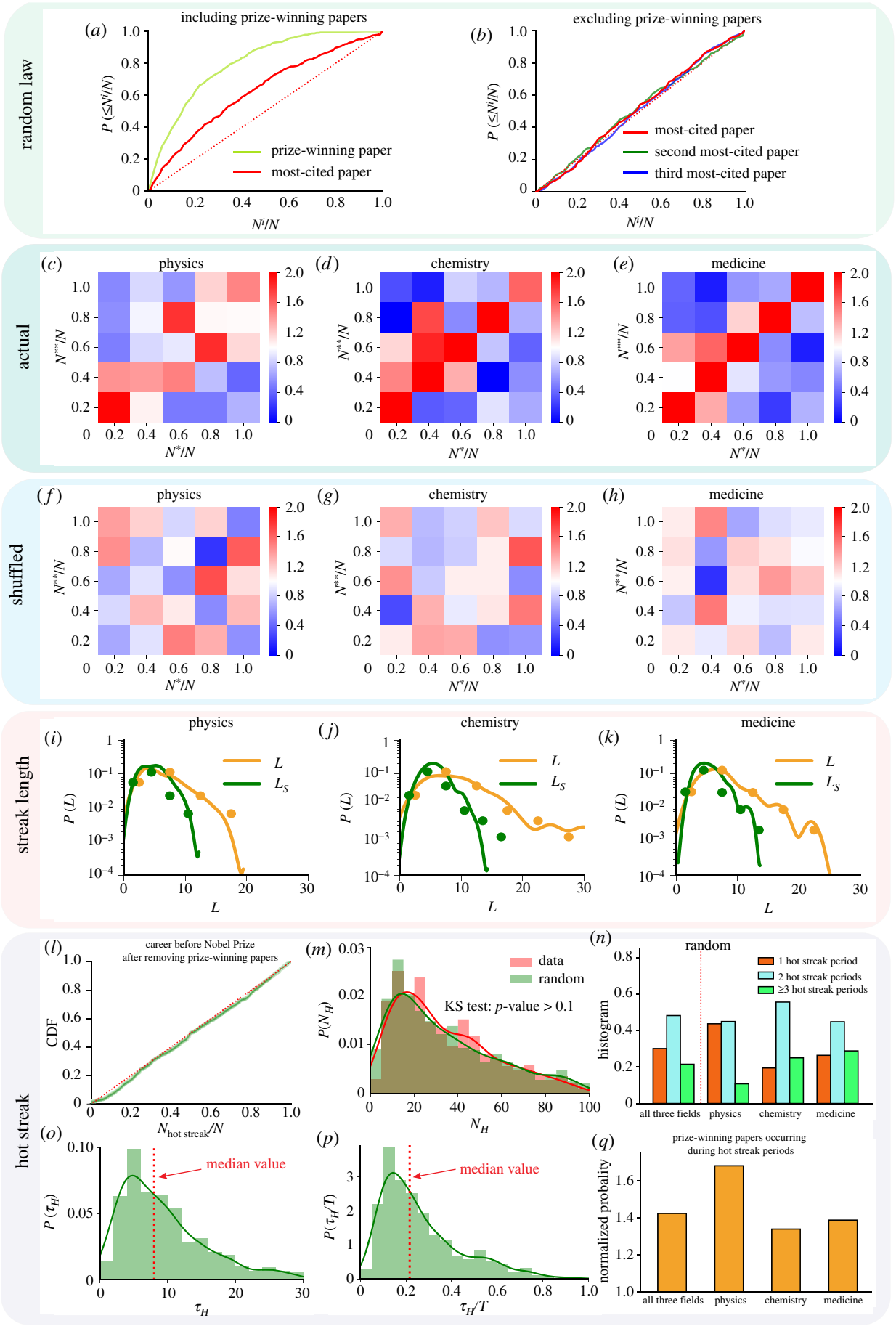

Figure 2. (*Caption opposite.*)

same laureates [51] (electronic supplementary material, S5.1). We find a greater propensity for the prize-winning papers to be written by fewer than three authors [51] (61.43% versus 53.28%, χ^2 test, p-value $< 10^{-4}$, figure 3*c*). We further examine the authorship structure of the prize-winning papers, finding that they are substantially more likely to have the laureates

Figure 2. (*Opposite*.) Hot streak phenomenon. (*a*) The cumulative distribution function (CDF) $P(\leq N^i/N)$ of relative sequence positions of the prize-winning papers and the most-cited papers (citations are ranked based on 10 year citation counts) during the academic career before the reception of the prize. N^i denotes the order of the hit work within N works in a career. The red dotted line represents the null model, in which the most-cited paper can occur at any position in the sequence of papers. (*b*) To eliminate sample bias brought from prize-winning works (49.7% of the most-cited papers before the prize is given are the prize-winning papers, and the Laureates wait an average of 17.6 years for formal recognition after making prize-winning achievements), prize-winning papers are removed and then we recalculate the top three most-cited papers among the papers published before conferment of the award. (*c–e*) The normalized joint distribution of the relative position of the top two most-cited papers (e.g. N^* and N^{**}) within N works in a career of a Nobel laureate across three domains, compared with a null model in which the two papers each follow their independent timing distributions. Values greater than 1 indicate that two hits are more likely to co-locate than random. (*f–h*) We shuffle the order of each work in a career while keeping their impact intact as a null model for (*c–e*). The longest streak within a career before the Nobel Prize, L, is defined as the maximum number of consecutive works whose impact is above the median impact of the career before the prize. (*i–k*) The distribution $P(L)$ of the longest streak within a career before the Nobel Prize and the corresponding distribution $P(L_s)$ for shuffled careers, for physics, chemistry and medicine, respectively. Orange dots represent empirical observations, whereas green dots correspond to shuffled careers. The orange solid line shows the simulation results produced by a hot streak model (electronic supplementary material, S4.3–S4.5) and the shuffled version is illustrated by the green solid line. (*l*) The hot streak model describes well the laureates' scientific career pattern for different disciplines. $N_{\text{hot streak}}/N$ measures the relative position of the work lying in the middle position of the hot streak period, among works in a career before the Nobel Prize after removing the prize-winning papers. Their cumulative distributions are shown by the green dots. (*m*) The distribution of the number of works produced during hot streaks $P(N_H)$, compared with a null distribution, where we randomly pick one work as the start of the hot streak for Nobel laureates. We use the Kolmogorov–Smirnov (KS) measure to compare $P(N_H)$ of data with the null distribution, finding that we cannot reject the hypothesis that the two distributions are drawn from the same distribution (p-value = 0.18). (*n*) The histogram of the number of hot streak periods. Nobel winners have 1.93 hot streak periods on average; specifically, 1.67 for physics, 2.08 for chemistry and 2.04 for medicine. (*o*) The duration distribution of the hot streak $P(\tau_H)$ for Nobel laureates. The median hot streak duration τ_H is 8 years, which is shown as the red dotted line. (*p*) The relative hot streak duration distribution $P(\tau_H/T)$ for Nobel laureates, where T is the career length of Nobel laureates. The red dotted line shows the median relative duration. (*q*) We show the normalized probability of prize-winning papers occurring during the hot streak periods $P_{\text{winning papers}}/P_{\text{random}}$. We find that the prize-winning papers are about 1.42 times more likely to occur during the hot streak periods than random, especially for physics laureates.

as the first author than other joint papers published by them (45.04% versus 30.64%, χ^2 test, p-value $< 10^{-7}$) (figure 3*d*). We also calculate the probability of being the last author, finding no statistical difference (χ^2 test, p-value = 0.41).

To test if these phenomena are unique to the prize-winning works, we removed the prize-winning papers and repeated the same analysis for the most-cited paper among the remaining papers. We find that there is no statistical difference in their likelihood of being written by small teams [51] (60.18% versus 56.17%, χ^2 test, p-value = 0.1193, figure 3*e*). While the difference in the likelihood of being the first author still exists for chemistry laureates, there is no statistical difference for laureates in physics or medicine (figure 3*f*). Together, these results show that prize-winning papers are more likely to be authored by fewer than three authors, with an intriguing tendency for laureates to claim the first authorship in the prize-winning works. While these observations are consistent with the finding that works produced by small teams tend to disrupt science and technology [37], they are also consistent with Zuckerman's argument that 'the future laureates were especially concerned to have the record clear for their most significant work, and particularly in their prize-winning research papers' [1].

2.4. After the prize

How does winning the Nobel Prize impact one's subsequent career? The Matthew effect [4,54] tells us that winning begets more winnings. Hence, one may expect that works produced after the Nobel Prize garner more impact than those produced before, given their substantially elevated reputation and visibility [15]. Here, we find that, to the contrary, when comparing the average impact of papers (defined in equation (4.4)) published by the laureates in each of the 4 years before and after winning the Nobel Prize, the average impact per paper shows a significant *drop* in the 2 years following the Nobel Prize. The effect is most significant in the year immediately after, where impact dropped by 11.1% on average compared with the year before. Furthermore, the effect is not permanent, with impact quickly bouncing back by year 4 to a similar level to that of the year of the Nobel Prize (figure 4*a*). The 'Nobel dip' is most pronounced for physics laureates, as the impacts of their papers were reduced by 18.1%, compared with 4.8% for chemistry and 13.4% for medicine (electronic supplementary material, S6.2, figure S17). Interestingly, in contrast with the common perception of decreased productivity following the Nobel Prize [1,21], possibly because of 'the disruptive consequences of abrupt upward social mobility' [1], we find that the average number of papers by the laureates shows no significant change (figure 4*b*), indicating that the uncovered Nobel dip mainly pertains to impact rather than productivity. Note that winning the Nobel Prize may introduce citation boosts to prior papers by the laureate [15,26]. To understand if the observed dip in impact may be explained by this factor, we alter the observation window to exclude post-prize citations to pre-prize works, finding that the 'dip and bounce back' pattern remains robust (figure 4*c*; electronic supplementary material, S6.4, figure S20). We also find that the number of solo-authored papers decreased precipitately after the Nobel Prize (Student's t-test, p-value = 0.004, figure 4*d*), whereas the fraction of team-authored papers increased (Student's t-test, p-value = 0.008, figure 4*e*), suggesting that collaboration and teamwork carry an increasing importance for the laureates after winning the Nobel Prize.

The Nobel dip signals that the scientific community's attention is not driven by status but the quality of work. To unearth potential mechanisms underlying the 'dip and bounce back' dynamics, we trace topic changes before and after the Nobel Prize as reflected in their publications. We use an established method [43] that detects research topics based on communities in the co-citing network of papers published by a scientist, offering a discipline-independent method to identify and trace research topics across a career (electronic supplementary material, S6.5). As an illustrative example, figure 4*f* shows the constructed co-citing network and topic communities for the career of Jean-Marie Lehn, who was awarded the 1987 Nobel Prize in Chemistry together with Donald J. Cram and Charles J. Pedersen for the synthesis of cryptands. In his remarkable career, Lehn published more than 700 papers. Figure 4*f* visualizes his publication history by topic, showing that his research agenda

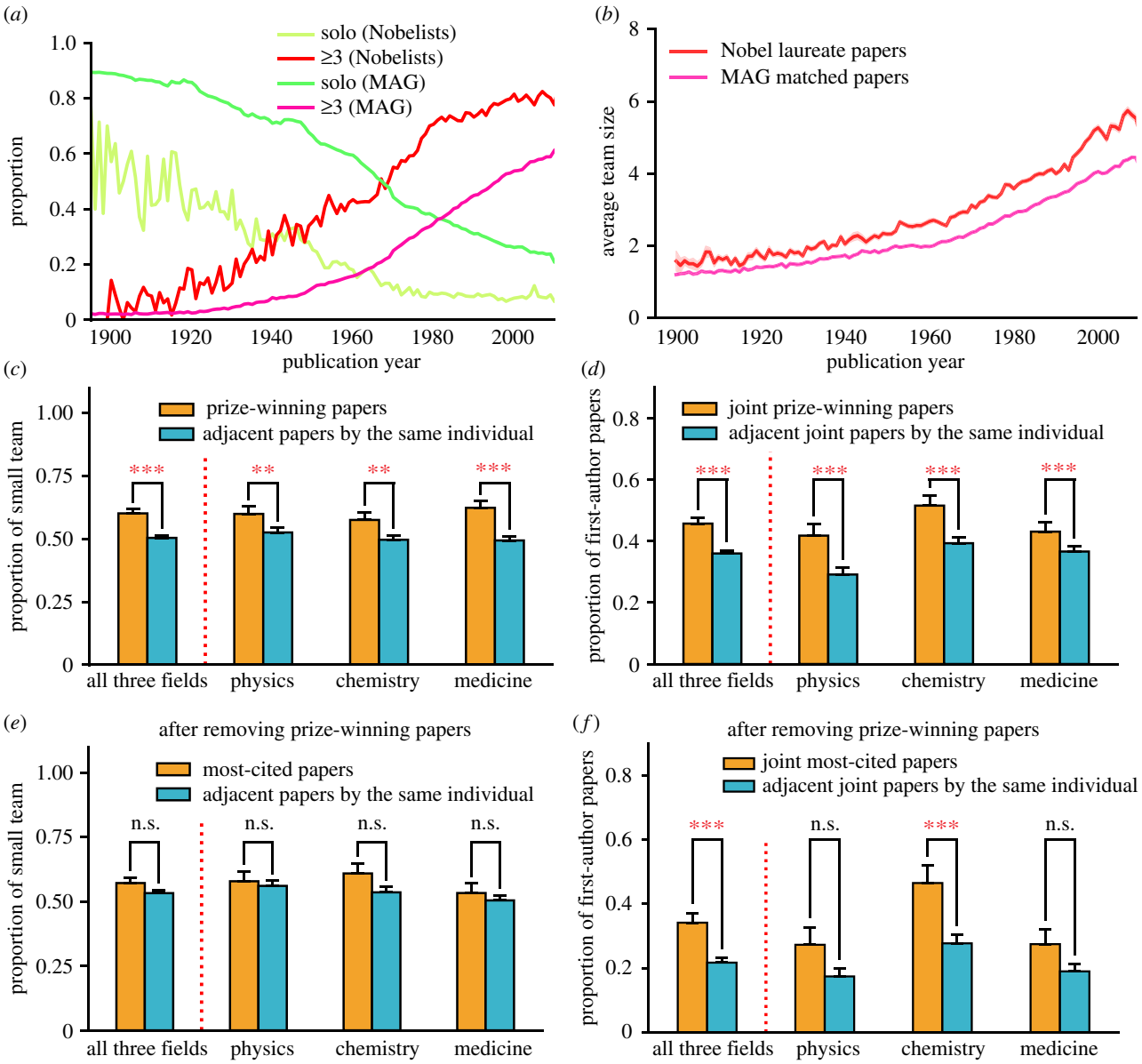

Figure 3. Collaboration patterns. (*a*) A comparison of the proportion of single-authored papers and large-team-authored papers between Nobelists' papers and all MAG papers as a function of the publishing year. The last century saw a decline in the fraction of single-authored papers and an increase in large team collaboration. Moreover, Nobel laureates tend to have a smaller fraction of single-authored papers and a larger fraction of big team papers. (*b*) Team size of Nobel laureates' papers and matched papers as a function of the publication year. For each Nobel laureate's paper, we matched 20 papers with the same publication year, same specific field and the closest number of citations for comparison. It shows that the team size of Nobel laureates' papers is always larger than random. (*c*) The proportion of small team-sized papers (team size ≤ 2) for all the prize-winning papers and the null model. For each prize-winning paper, we chose four non-prize-winning papers with the closest publication time by the same individual as a null model. (*d*) The proportion of first-authored papers of joint prize-winning papers in comparison with a null model. In this case, we only consider joint-authored papers. For each joint-authored prize-winning paper, we chose four joint-authored non-prize-winning papers published by the same Nobel laureate with the closest publication time. (*e*) We also measure the proportion of small team-sized papers for the most-cited papers after removing the prize-winning papers before the prize in comparison with a null model. For each most-cited paper, we chose four papers published by the same Nobel laureate with the closest publication time as the null model. (*f*) We then compare the proportion of first-authored papers among joint most-cited papers after removing the prize-winning papers before the prize with a null model, and we chose four joint-authored papers published by the same Nobel laureate with the closest publication time as the null model. ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$ and n.s. (not significant) for $p > 0.1$. Error bars represent the s.e.m.

was almost exclusively focused on cryptands-related research, until he was awarded the Nobel Prize in 1987. Yet, just as this line of research was officially recognized, we observed a clear shift in the topic right after winning the Nobel Prize (figure 4*g*). In the next 10 years, his research was primarily focused on self-assembly and self-organization. Most interestingly, this is a topic that he had never published on before winning the Nobel Prize.

The intriguing example of Lehn's career prompts us to ask if laureates disproportionately shift research topic after winning the Nobel Prize. We randomly selected two papers, within 4 years before and after the Nobel Prize, respectively, and measured the probability of two papers belonging to the same topic, finding only 36.8% of the two papers cover the same topic before and after winning the prize. We then built a null model by randomly choosing a year as the pretended prize-winning year for comparison, finding that the probability is significantly higher (45.2% versus 36.8%, p-value = 0.004, figure 4*h*), which suggests the laureates have a higher likelihood of shifting research topics after winning the Nobel Prize.

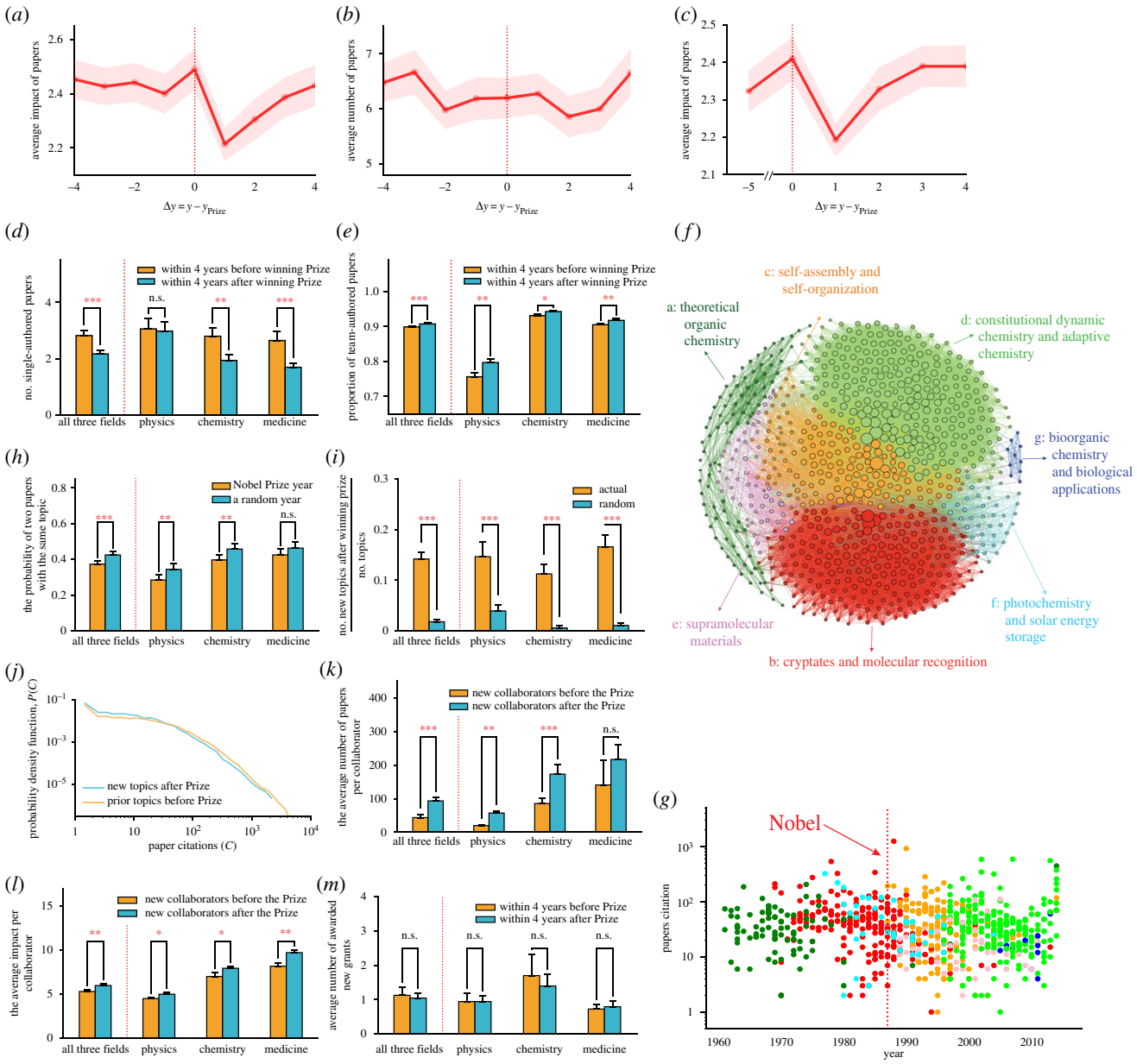

Figure 4. After the Nobel Prize: the temporary dip and bounce back. (*a*) Average impact per paper (defined in equation (4.4)) as a function of time. The year when the Nobel Prize is given is marked as 0. For each laureate, we calculate the average impact of papers the laureate published in each of the 4 years before and after the Nobel Prize, as well as the prize-winning year. The solid line indicates the average across all laureates in our sample, with the shaded area denoting the standard error of the mean. (*b*) The average number of papers before and after winning the Nobel Prize. The solid line indicates the average across all laureates in our sample, with the shaded area denoting the standard error of the mean. The change following the Nobel Prize mainly pertains to impact rather than productivity. In contrast with the common belief of decreased productivity following the Nobel Prize, the average number of publications by the laureates shows no significant changes. (*c*) Average impact per paper as a function of time. We set the observation window as 5 years and calculate the average impact of papers based on the 5 year citation counts. For each laureate, we compare the average impact of papers published in the 5th year before and each of the 4 years after winning the Nobel Prize. (*d*) Comparison of the number of individual papers within 4 years before and after receipt of the Nobel Prize. It shows a significant decrease in individual work after the Nobel Prize. The change is significant for chemistry and medicine, while there is no significant difference for physics. (*e*) Comparison of the proportion of joint papers within 4 years before and after the prize. It shows an increase of joint works after the Nobel Prize. (*f*) Communities of topics for Nobel laureate Jean-Marie Lehn. Each paper is represented by a node, and two papers are connected if they share at least one reference; thus, constructing a co-citing network. Here, communities are detected using the fast unfolding algorithm [43,55], and each community represents a research topic. (*g*) The time series of all the papers published by Nobel laureate Jean-Marie Lehn; the *Y*-axis shows the citation for each paper. Each paper is represented by a point and the colour corresponds to the topic community in the co-citing network. It shows a clear topic switching from 'cryptates and molecular recognition' to 'self-assembly and self-organization' for Nobel laureate Jean-Marie Lehn immediately after winning the Nobel Prize. (*h*) Comparison of the probability of two papers belonging to the same topic within 4 years before and after the reception of the prize and a random year. The probability is significantly lower after winning the prize, suggesting that Nobel laureates tend to shift research topic after winning the Nobel Prize. (*i*) We measure the chance of Nobel laureates shifting to a new topic after winning the Nobel Prize, by no. new topics after winning the prize/no. topics. We also shuffled the topic of the works and repeated the measurement as a null model, finding that the laureates are much more likely to shift to a new topic after winning the Nobel Prize than random (14.2% versus 1.8%, p-value $< 10^{-14}$). The change is significant for physics and chemistry, while there is no significant difference for medicine. (*j*) The distribution of paper citations for prior topics before the Nobel Prize and new topics after the Nobel Prize. We use the Kolmogorov–Smirnov measure to compare the two distributions, finding that papers of the prior topics receive higher citations than those of the new topics after the prize (Kolmogorov–Smirnov test, p-value $< 10^{-71}$). (*k*) Comparison of the average number of papers of new collaborators before and after the Nobel Prize, showing post-prize collaborators tend to be more productive (Student's t-test, p-value $< 10^{-3}$). (*l*) Comparison of the average impact of new collaborators before and after the Nobel Prize, showing post-prize collaborators tend to have a higher impact (Student's t-test, p-value $= 0.02$). (*m*) Comparison of the average number of awarded new grants within 4 years before and after winning the Nobel. There is no significant change in funding before and after the prize (Student's t-test, p-value $= 0.77$). ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$ and n.s. (not significant) for $p > 0.1$. Error bars represent the s.e.m.

We further measured the likelihood of laureates studying a new topic after winning the prize, and compare it with a null model where we shuffled the topic of the works, finding that the laureates are much more likely to study a new topic after winning the prize than expected (14.2% versus 1.8%, p-value $< 10^{-14}$, figure 4*i*). To ensure that these results are not affected by specific community detection methods used to detect topics, we repeated our analyses with another well-known algorithm (Infomap [56]), obtaining the same conclusions (electronic supplementary material, S6.6).

To understand potential forces behind the uncovered change in research agenda following the Nobel Prize, we examined several different factors, including the popularity of research topics before and after the prize (electronic supplementary material, S6.7), changes in collaborators (electronic supplementary material, S6.8) and funding opportunities (electronic supplementary material, S6.9). We find that the topic studied after the Nobel Prize tends to be less popular at the time. The number of new collaborators does not increase after the Nobel Prize, but these collaborators tend to be more established in terms of productivity and impact. And somewhat surprisingly, the overall funding to each laureate remains mostly constant around the time of the award. Although none of these factors can directly explain the observed topic change and the associated citation dip (figure 4*j–m*; electronic supplementary material, S6.7–S6.9, figures S24–S26), they appear consistent with an endogenous shift in the laureate's interest to explore new directions. Note that, although the uncovered dip–bounce-back dynamics and topic shifting behaviour both occur around the same time (when awarded the Nobel Prize), it does not imply that the two are causally related. On the other hand, while one may be better at anticipating which work will be recognized by the Nobel Prize eventually (https://en.wikipedia.org/wiki/Clarivate_Citation_Laureates), it remains difficult to precisely predict the year of winning, indicating that the award year can be viewed as a largely exogenous variation in a career [57], which then coincides with topic-shifting behaviour that is largely endogenous to the individual. Regardless, these results highlight the unwavering scientific efforts by the laureates, actively pursuing new lines of enquiry while undeterred by the extra burdens imposed by growing duties and responsibilities [1].

3. Discussion

In summary, building on Zuckerman's canonical work on scientific elites [1], here we present a systematic empirical investigation of the careers of Nobel laureates by studying patterns of productivity, collaboration, authorship and impact. This analysis is now possible owing to a novel dataset we curated—both algorithmically and manually—which links several disparate biographical and bibliographical data sources, offering a unique opportunity to quantitatively study the scientific contributions and recognitions of scientific elites. Despite the clear difference between the Nobel laureates and 'ordinary' scientists, we find universal career patterns that are applicable to both ordinary and elite scientists. Indeed, we find the careers of the laureates before winning the prize are governed by remarkably similar patterns to those of ordinary scientists, characterized by hot streaks and increasing reliance on team work. Hence, these results help advance the canonical innovation literature by offering new empirical evidence from large-scale datasets. At the same time, we also uncovered notable but previously unknown variations along their careers associated with the Nobel Prize, including shifting coauthorship structure in the prize-winning works, and a temporary but significant dip in the impact of works they produce after winning the Nobel Prize. Overall, these results represent new empirical patterns that further enrich our understanding of careers of the scientific elite.

This paper takes an initial step probing our quantitative understanding of career patterns of the scientific elite, which not only offers an empirical basis for future studies of individual careers and creativity in broader domains [16,50], but also deepens our quantitative understanding of patterns governing exceptional careers in science.

4. Methods

4.1. Rescaled number of citations

To approximate the scientific impact of each paper, we calculate the number of citations the paper received after 10 years, C_{10}, and use it as a proxy for the paper's impact. Previous studies [29,37,44] have shown that the average number of citations per paper changes over time. To be able to compare the impact of papers published at different times and to adjust for temporal effects, the rescaled number of citations a paper receives after 10 years, $\hat{C}_{10,i}$, is suggested as a good proxy for publication impact. According to Fortunato *et al.* [29], given a paper i, $\hat{C}_{10,i}$, is defined as follows:

$$\hat{C}_{10,i} = 10 \times \frac{C_{10,i}}{\langle C_{10} \rangle}, \tag{4.1}$$

where $C_{10,i}$ is the raw number of 10 year citations for paper i, and $\langle C_{10} \rangle$ is the average C_{10} calculated over all publications published in the same year and field.

4.2. Definition of hit paper rate

In figure 1*b*, we compare the 'hit' paper rate—defined as the probability of publishing papers in the top 1% of rescaled 10 year citations in the same year and field—for Nobel laureates and typical authors. Our collected Nobel laureate dataset is based on information provided by the MAG, which assigns the field of subject for each paper. It is worth noting that the field of subject is a hierarchal structure with six levels. The first level contains 19 main fields, such as 'physics', 'chemistry', 'medicine' and 'biology.' The second level contains 295 subfields, such as 'astrophysics', 'biophysics' and 'geophysics'. In this paper, we choose the second-level fields in calculating the hit paper rate for Nobel laureates. The GS typical scientist dataset is based on information from the WOS, and it is almost impossible to precisely match the career histories of 3540 GS scientists from the WOS to the MAG. Thus, the hit rate analysis of the GS scientists is based on the WOS database itself. Papers in the WOS are also assigned to one of 234 specific field categories, such as 'astronomy & astrophysics', 'biophysics' and 'geochemistry & geophysics'. The hit paper rate for typical scientists is calculated using these 234 specific fields from the WOS.

4.3. Selecting matching papers

In figure 3*b*, we created a matching sample for each paper published by Nobel laureates. The procedure for selecting matching papers is introduced here in detail. For each Nobel prize-winner's work, we first determine its year of publication, total citation number and subject categories based on the MAG dataset. Next, all the MAG papers with the same publishing year and specific field are obtained and sorted according to their number of citations. It is worth noting that, when a

laureate's paper spans multiple subjects, we deem MAG papers appropriate matches if they share at least one common subject with the laureate's. We then select the 20 papers with citation counts that are most similar to the laureate's paper and use these as matching papers.

4.4. Quantifying impact

In figure 4*a*, we compare the average impact of papers published by the laureates in each of the 4 years before and after winning the Nobel Prize. We propose a measure to quantify the average impact of papers: we first calculate the average impact within all papers in specific years, and then we take the individual heterogeneity of Nobel laureates into consideration when quantifying the average impact of papers.

The impact of paper i is quantified by $\Gamma_i = \log(\hat{C}_{10,i} + 1)$, where $\hat{C}_{10,i}$ measures the rescaled number of citations within 10 years of publication. We denote $\Delta y = y_i - y_{\mathrm{Prize}}$ as a laureate's relative publishing time after winning the Nobel Prize, where y_i is the publication year of paper i. Assuming there are $N_{\Delta y}$ papers publishing in the Δy year after winning the prize, we define the average impact of papers as follows:

$$\langle \Gamma_P \rangle_{\Delta y} = \frac{\sum_{i=1}^{N_{\Delta y}} \Gamma_i}{N_{\Delta y}}. \tag{4.2}$$

However, the above measure did not consider the individual heterogeneity of Nobel laureates. For example, average impact may be driven by those laureates with high productivity as well as high paper quality. Thus, we first measure the average impact of papers for each laureate and then calculate the average for all Nobel laureates. For laureate j, the average impact of papers published in the Δy year after winning the prize is defined as:

$$\langle \Gamma_P \rangle_{\Delta y,j} = \frac{\sum_{i=1}^{N_{\Delta y,j}} \Gamma_i}{N_{\Delta y,j}}, \tag{4.3}$$

where $N_{\Delta y,j}$ is the number of papers published in the Δy year of laureate j. Factoring in individual heterogeneity, the average impact of papers is defined as follows:

$$\langle \Gamma_N \rangle_{\Delta y} = \frac{\sum_{j=1}^{M_{\Delta y}} \langle \Gamma_P \rangle_{\Delta y,j}}{M_{\Delta y}}, \tag{4.4}$$

where $M_{\Delta y}$ denotes the number of laureates who still publish papers in the Δy year after winning the Nobel Prize. In the main text (figure 4*a*), we use $\langle \Gamma_N \rangle$ to measure the average impact of papers.

4.5. Topic changing after winning the Nobel Prize

To quantify the topic of a paper, we adopt a recent method based on community structure of the co-citing network of a scientist's papers [50]. To ensure meaningful community detection results, we consider all Nobel laureates who have published at least 50 papers. We also excluded Nobel laureates who published fewer than five papers after winning the prize. Finally, we selected 283 Nobel laureates (74 for physics, 96 for chemistry, 113 for medicine) who satisfied these requirements.

In figure 4*g*, we measure the probability of two papers belonging to the same topic within 4 years before and after the reception of the prize and a random year. To measure the probability of changing topics of Nobel laureates after winning the prize, we randomly selected two papers, within 4 years before and after the Nobel, respectively, and measured the probability of those two papers belonging to the same topic. We then built a null model by randomly choosing a year as the pretended prize-winning year for comparison.

To test if Nobel laureates tend to study a new topic after winning the prize, we measure the chance of Nobel laureates shifting to a new topic after winning the Nobel Prize, by no. new topics after winning Prize/no. topics. We also shuffled the topic of the works and repeated the measurement as a null model for comparison.

Data accessibility. The main data that support the findings of this study are freely available. They are deposited in public repositories with detailed descriptions in Harvard Dataverse (https://doi.org/10.7910/DVN/6NJ5RN).

Authors' contributions. D.W. and S.F. conceived the project, D.W. designed the experiments; J.L. and Y.Y. collected data and performed empirical analyses with help from S.F. and D.W.; all authors discussed and interpreted results; D.W., J.L. and Y.Y. wrote the manuscript; all authors edited the manuscript.

Competing interests. We declare we have no competing interests.

Funding. This work is supported by the Air Force Office of Scientific Research under award nos. FA9550-15-1-0162, FA9550-17-1-0089 and FA9550-19-1-0354, National Science Foundation grant no. SBE 1829344 and Northwestern University's Data Science Initiative.

Acknowledgements. The authors thank L. Liu, Y. Wang, Y. Ma and all members of the Northwestern Institute on Complex Systems (NICO) for invaluable comments. Funding data sourced from Dimensions, an inter-linked research information system provided by Digital Science (https://www.dimensions.ai).

References

1. Zuckerman H. 1977 *Scientific elite: Nobel laureates in the United States*. New York, NY: Free Press.
2. Barabasi AL, Song CM, Wang DS. 2012 Handful of papers dominates citation. *Nature* **491**, 40. (doi:10.1038/491040a)
3. Jones BF, Weinberg BA. 2011 Age dynamics in scientific creativity. *Proc. Natl Acad. Sci. USA* **108**, 18 910–18 914. (doi:10.1073/pnas.1102895108)
4. Merton RK. 1968 The Matthew effect in science. *Science* **159**, 56–63. (doi:10.1126/science.159.3810.56)
5. Newman MEJ. 2001 The structure of scientific collaboration networks. *Proc. Natl Acad. Sci. USA* **98**, 404–409. (doi:10.1073/pnas.021544898)
6. Price DJD. 1976 General theory of bibliometric and other cumulative advantage processes. *J. Am. Soc. Inform. Sci.* **27**, 292–306. (doi:10.1002/asi.4630270505)
7. Petersen AM, Jung WS, Yang JS, Stanley HE. 2011 Quantitative and empirical demonstration of the Matthew effect in a study of career longevity. *Proc. Natl Acad. Sci. USA* **108**, 18–23. (doi:10.1073/pnas.1016733108)
8. Zuckerman H. 1967 The sociology of the Nobel Prizes. *Sci. Am.* **217**, 25. (doi:10.1038/scientificamerican1167-25)
9. Zuckerman H. 1972 Interviewing an ultra-elite. *Public Opin. Quart.* **36**, 159. (doi:10.1086/267989)
10. Moulin L. 1961 Sociology of Nobel-prize winners for science, 1901–1960, with special-reference to nationality. *Cah. Int. Sociol.* **31**, 145–163.
11. Zuckerman H. 1967 Nobel laureates in science: patterns of productivity, collaboration, and authorship. *Am. Sociol. Rev.* **32**, 391–403. (doi:10.2307/2091086)
12. Hansson N, Halling T, Fangerau H. 2018 Nobel nomination letters point to a winning formula. *Nature* **555**, 311. (doi:10.1038/d41586-018-03057-z)
13. Garfield E. 1986 Do Nobel-prize winners write citation-classics. *Curr. Contents* **23**, 3–8.
14. Garfield E, Welljams-Dorof A. 1992 Of Nobel class: a citation perspective on high impact research authors. *Theor. Med.* **13**, 117–135. (doi:10.1007/BF02163625)
15. Mazloumian A, Eom YH, Helbing D, Lozano S, Fortunato S. 2011 How citation boosts promote scientific paradigm shifts and Nobel Prizes. *PLoS ONE* **6**, e18975. (doi:10.1371/journal.pone.0018975)
16. Simonton DK. 1984 *Genius, creativity, and leadership*. Cambridge, MA: Harvard University Press.

17. Simonton DK. 1997 Creative productivity: a predictive and explanatory model of career trajectories and landmarks. *Psychol. Rev.* **104**, 66–89. (doi:10.1037/0033-295x.104.1.66)
18. Moreira JAG, Zeng XHT, Amaral LAN. 2015 The distribution of the asymptotic number of citations to sets of publications by a researcher or from an academic department are consistent with a discrete lognormal model. *PLoS ONE* **10**, e0143108. (doi:10.1371/journal.pone.0143108)
19. Shen HW, Barabasi AL. 2014 Collective credit allocation in science. *Proc. Natl Acad. Sci. USA* **111**, 12 325–12 330. (doi:10.1073/pnas.1401992111)
20. Fortunato S. 2014 Growing time lag threatens Nobels. *Nature* **508**, 186. (doi:10.1038/508186a)
21. 2017 Nobel reactions. *Nat. Phys.* **13**, 921. (doi:10.1038/nphys4296)
22. Chan HF, Torgler B. 2013 Time-lapsed awards for excellence. *Nature* **500**, 29. (doi:10.1038/500029c)
23. Seeman JI. 2017 Synthesis and the Nobel Prize in chemistry. *Nat. Chem.* **9**, 925–929. (doi:10.1038/nchem.2864)
24. Fleming L, Mingo S, Chen D. 2007 Collaborative brokerage, generative creativity, and creative success. *Admin. Sci. Quart.* **52**, 443–475. (doi:10.2189/asqu.52.3.443)
25. Singh J, Fleming L. 2010 Lone inventors as sources of breakthroughs: myth or reality? *Manage Sci.* **56**, 41–56. (doi:10.1287/mnsc.1090.1072)
26. Azoulay P, Stuart T, Wang YB. 2014 Matthew: effect or fable? *Manage Sci.* **60**, 92–109. (doi:10.1287/mnsc.2013.1755)
27. Weinberg BA, Galenson DW. 2019 Creative careers: the life cycles of Nobel laureates in economics. *De Economist* **167**, 221–239. (doi:10.1007/s10645-019-09339-9)
28. Sinatra R, Wang D, Deville P, Song CM, Barabasi AL. 2016 Quantifying the evolution of individual scientific impact. *Science* **354**, aaf5239. (doi:10.1126/science.aaf5239)
29. Fortunato S *et al.* 2018 Science of science. *Science* **359**, eaao0185. (doi:10.1126/science.aao0185)
30. Azoulay P, Jones BF, Kim JD, Miranda J. 2020 Age and high-growth entrepreneurship. *Am. Econ. Rev. Insights* **2**, 65–82.
31. Runco MA. 2014 *Creativity: theories and themes: research, development, and practice*. Amsterdam, The Netherlands: Elsevier.
32. Taleb NN. 2007 *The black swan: the impact of the highly improbable*, vol. 2. New York, NY: Random House.
33. Latour B, Woolgar S. 2013 *Laboratory life: the construction of scientific facts*. Princeton, NJ: Princeton University Press.
34. Taubes G. 1988 *Nobel dreams: power, deceit and the ultimate experiment*. Redmond, WA: Microsoft Press.
35. Cole S, Cole JR. 1967 Scientific output and recognition: a study in the operation of the reward system in science. *Am. Sociol. Rev* **32**, 377–390. (doi:10.2307/2091085)
36. Zeng A, Shen Z, Zhou J, Wu J, Fan Y, Wang Y, Stanley HE. 2017 The science of science: from the perspective of complex systems. *Phys. Rep.* **714**, 1–73. (doi:10.1016/j.physrep.2017.10.001)
37. Wu L, Wang D, Evans JA. 2019 Large teams develop and small teams disrupt science and technology. *Nature* **566**, 378–382. (doi:10.1038/s41586-019-0941-9)
38. Clauset A, Arbesman S, Larremore DB. 2015 Systematic inequality and hierarchy in faculty hiring networks. *Sci. Adv.* **1**, e1400005. (doi:10.1126/sciadv.1400005)
39. Milojevic S. 2014 Principles of scientific research team formation and evolution. *Proc. Natl Acad. Sci. USA* **111**, 3984–3989. (doi:10.1073/pnas.1309723111)
40. Petersen AM. 2018 Multiscale impact of researcher mobility. *J. R. Soc. Interface* **15**, 20180580. (doi:10.1098/rsif.2018.0580)
41. Pfeiffer T, Tran L, Krumme C, Rand DG. 2012 The value of reputation. *J. R. Soc. Interface* **9**, 2791–2797. (doi:10.1098/rsif.2012.0332)
42. Youn H, Strumsky D, Bettencourt LM, Lobo J. 2015 Invention as a combinatorial process: evidence from US patents. *J. R. Soc. Interface* **12**, 20150272. (doi:10.1098/rsif.2015.0272)
43. Zeng A, Shen Z, Zhou J, Fan Y, Di Z, Wang Y, Stanley HE, Havlin S. 2019 Increasing trend of scientists to switch between topics. *Nat. Commun.* **10**, 3439. (doi:10.1038/s41467-019-11401-8)
44. Ioannidis JP, Baas J, Klavans R, Boyack KW. 2019 A standardized citation metrics author database annotated for scientific field. *PLoS Biol.* **17**, e3000384. (doi:10.1371/journal.pbio.3000384)
45. Vuong QH, La VP, Vuong TT, Ho MT, Nguyen HK, Nguyen VH, Pham HH, Ho MT. 2018 An open database of productivity in Vietnam's social sciences and humanities for public use. *Sci. Data* **5**, 180188. (doi:10.1038/sdata.2018.188)
46. Liu L *et al.* 2018 Hot streaks in artistic, cultural, and scientific careers. *Nature* **559**, 396–399. (doi:10.1038/s41586-018-0315-8)
47. Li J, Yin Y, Fortunato S, Wang D. 2019 A dataset of publication records for Nobel laureates. *Sci. Data* **6**, 33. (doi:10.1038/s41597-019-0033-6)
48. Li J, Yin Y, Fortunato S, Wang D. 2018 A dataset of publication records for Nobel laureates. *Harvard Dataverse*. (doi:10.7910/DVN/6NJ5RN)
49. Azoulay P *et al.* 2018 Toward a more scientific science. *Science* **361**, 1194–1197. (doi:10.1126/science.aav2484)
50. Simonton DK. 1988 *Scientific genius: a psychology of science*. Cambridge, UK: Cambridge University Press.
51. Li J, Yin Y, Fortunato S, Wang D. 2019 Nobel laureates are almost the same as us. *Nat. Rev. Phys.* **1**, 301–303. (doi:10.1038/s42254-019-0057-z)
52. Wuchty S, Jones BF, Uzzi B. 2007 The increasing dominance of teams in production of knowledge. *Science* **316**, 1036–1039. (doi:10.1126/science.1136099)
53. Petersen AM. 2015 Quantifying the impact of weak, strong, and super ties in scientific careers. *Proc. Natl Acad. Sci. USA* **112**, E4671–E4680. (doi:10.1073/pnas.1501444112)
54. Perc M. 2014 The Matthew effect in empirical data. *J. R. Soc. Interface* **11**, 20140378. (doi:10.1098/rsif.2014.0378)
55. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. 2008 Fast unfolding of communities in large networks. *J. Stat. Mech: Theory Exp.* **2008**, P10008. (doi:10.1088/1742-5468/2008/10/P10008)
56. Rosvall M, Bergstrom CT. 2008 Maps of random walks on complex networks reveal community structure. *Proc. Natl Acad. Sci. USA* **105**, 1118–1123. (doi:10.1073/pnas.0706851105)
57. Adams D. 1992 *Mostly harmless*. London, UK: William Heinemann.