PNAS RESEARCH ARTICLE | APPLIED PHYSICAL SCIENCES SOCIAL SCIENCES

Impactful scientists have higher tendency to involve collaborators in new topics

An Zeng[a], Ying Fan[a], Zengru Di[a], Yougui Wang[a], and Shlomo Havlin[b,1]

Edited by David Weitz, Harvard University, Cambridge, MA; received April 29, 2022; accepted July 4, 2022

In scientific research, collaboration is one of the most effective ways to take advantage of new ideas, skills, and resources and for performing interdisciplinary research. Although collaboration networks have been intensively studied, the question of how individual scientists choose collaborators to study a new research topic remains almost unexplored. Here, we investigate the statistics and mechanisms of collaborations of individual scientists along their careers, revealing that, in general, collaborators are involved in significantly fewer topics than expected from a controlled surrogate. In particular, we find that highly productive scientists tend to have a higher fraction of single-topic collaborators, while highly cited—i.e., impactful—scientists have a higher fraction of multitopic collaborators. We also suggest a plausible mechanism for this distinction. Moreover, we investigate the cases where scientists involve existing collaborators in a new topic. We find that, compared to productive scientists, impactful scientists show strong preference of collaboration with high-impact scientists on a new topic. Finally, we validate our findings by investigating active scientists in different years and across different disciplines.

scientific collaboration | research topics | impactful scientists | controlled surrogate

Significance

Scientific collaboration is an important feature of modern science. The topics involved in scientific collaborations have been associated with creativity and the impact of research. Yet, how scientists involve collaborators in their research topics remains poorly understood. We reveal here the general tendency of collaborators to be involved in a single topic. The tendency is stronger for the collaborators of productive scientists, but weaker for the collaborators of impactful scientists. We further identify the past research productivity and impact of collaborators as key factors affecting their probability to join a new topic of a given scientist. The analysis framework is general and applicable for understanding collaborations in various other systems, such as film-making, patent design, and software development.

Coauthored publications in science have increased significantly during the last century (1, 2). Through collaboration, scientists could bring new ideas and techniques from different fields, which, in many cases, result in high-quality publications. Indeed, it has been found that the number of authors in a paper (3) and the less prior collaboration relations between coauthors (4) are strongly associated with the originality of the paper. Thus, scientific collaboration seems to be an important key to enhance innovation of research teams.

Studies regarding scientific collaborations have a long history and have attracted much attention in recent years (5, 6). Early works on scientific collaboration concentrate on collaboration networks constructed from scientific publication data (6). Numerous topological properties of collaboration networks have been revealed, such as small-world features (7), assortative degree mixing (8), rich motifs (9), and community structure (10). In recent years, attention has been given to further aspects of scientific collaboration. Regarding the collaboration frequency as tie strength, weak, strong, and super-strong ties in scientific careers have been identified, and the super-strong ties have been found to have a positive effect on productivity and citations (11). For coauthored papers, methods have been designed to collectively allocate credits among authors (12). Another trend to understand collaboration relations is from the perspective of scientific teams, with research questions ranging from team-assembly mechanisms (13) to the effect of team characteristics on team performances (14–17). A specific type of collaboration—namely, the mentor–mentee relationship—has been recently shown to influence research performance (18) and academic rewards of scientists (19).

In recent years, numerous works have been devoted to investigate topic switching in individual careers. With the help of the field-classification codes in physics, it has been found that research interest of individual physicists could shift significantly from the beginning to the end of their career (20). The transition map of scientists from field to field has also been extracted from the data (21). By applying the community-detection technique in the cociting networks of individual scientists' papers, scientists have been found to have a narrow distribution of the number of major topics during their lifetime (22). This framework has been later used to understand the careers of Nobel laureates (23) and identify the key mechanisms for hot streaks in scientists' careers (24). However, the characteristics and mechanisms of scientists behind initiating collaboration on a new topic have not been studied. In fact, scientists' choice to collaborate on a new topic is a fundamental process that drives the creativity and impact of the scientific research. The increasingly in-depth development of science requires specialization and accumulated knowledge for researchers to work on a topic (25, 26), suggesting a hypothesis that science

Author affiliations: [a]School of Systems Science, Beijing Normal University, Beijing 100875, China; and [b]Department of Physics, Bar-Ilan University, Ramat-Gan 52900, Israel

Author contributions: A.Z. and S.H. designed research; A.Z. performed research; Y.F., Z.D., and Y.W. contributed new reagents/analytic tools; A.Z. and S.H. analyzed data; and A.Z., Y.F., Z.D., Y.W., and S.H. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: havlin@ophir.ph.biu.ac.il.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2207436119/-/DCSupplemental.

Published August 8, 2022.

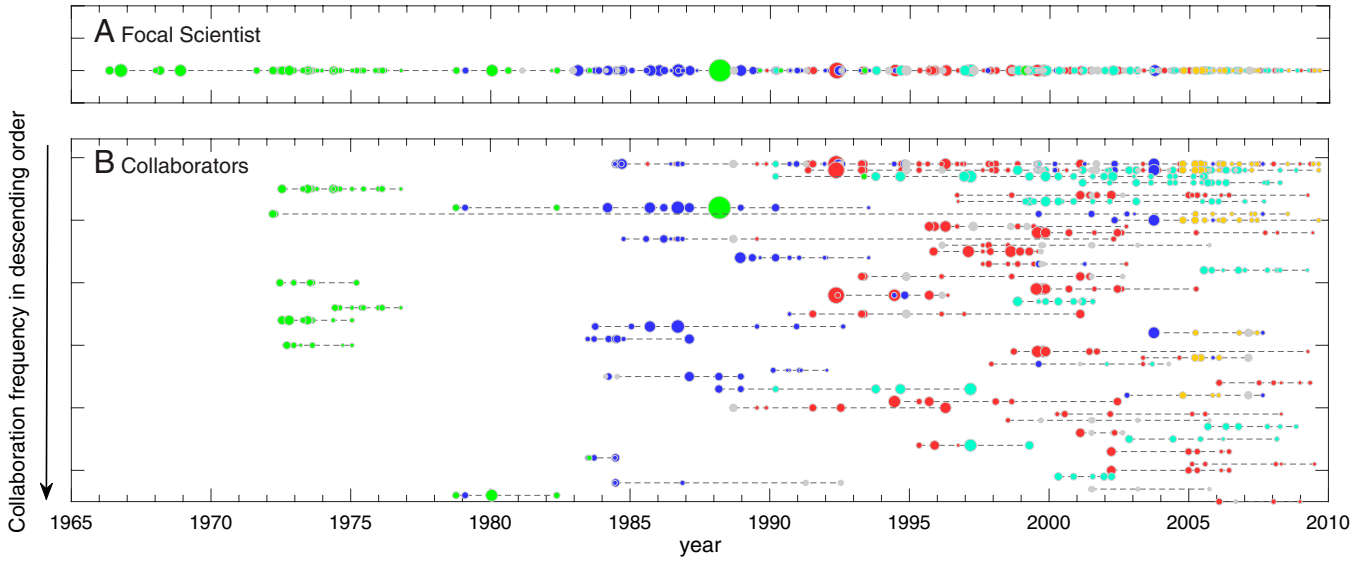

Fig. 1. Illustration of topics in which a typical scientist's collaborators are involved. *A* shows a real typical evolution of research topics during a focal scientist's career. Each node is a paper published by this scientist, and the colors of the nodes represent the research topics of these papers. Node size represents the number of citations of this paper. *B* shows the research topics in which each of the focal scientist's collaborator is involved. The collaborators are sorted in descending order from top to bottom according to the number of coauthored papers with the focal scientist. Each line shows the results of a collaborator, with each node on it representing a coauthored paper with the focal scientist. Thus, the first node and the last node on a line denote, respectively, the starting year and the ending year of the collaboration. We only show the collaborators who published at least five papers with the focal scientist.

might be dominated by single-topic collaboration. On the other hand, interdisciplinarity and atypical combination of knowledge have been shown to promote creativity (27, 28), suggesting another hypothesis that involving collaborators specialized in a topic with another topic may bring fresh ideas and unexpected solutions. Thus, a series of fundamental questions regarding research topics in scientific collaboration naturally arise: On how many different topics do a pair of scientists typically collaborate? How do scientists differ in involving collaborators in their research topics? What factors would affect the probability of a collaborator to join a new topic of a given scientist?

In this paper, we address the above questions by systematically investigating the coevolution of topics and collaborators during a scientist's career, aiming to understand how scientists choose to collaborate on a new topic of research. We decompose the publication series of a scientist to partial series that record the coauthored papers with each of his collaborators, allowing us to understand the statistics of the topics in which collaborators are involved. The partial time series also enable us to study the temporal features of the collaboration-topic formation. By comparing the data of highly productive and highly cited scientists, we investigate how successful scientists of these two types differ in involving their collaborators in new topics. We finally compare active scientists along the past 80 y and across different disciplines to understand the evolution and disciplinary differences regarding the topics in scientific collaboration.

Results

We first describe the method (22) to identify the involved topics of each collaborator of a focal scientist. The method begins with constructing a network of the focal scientist's publications, where the links are defined by the cociting relations (*SI Appendix*, Fig. S1). We then detect communities in the cociting network, where each major community represents a different research topic of the focal scientist. In a scientist's publication time series, we mark each paper with a color according to the community to which it belongs (Fig. 1*A*). The colored time series thus exhibits how a scientist switches from one topic to another. To capture the involved topics of the collaborators, we decompose the series of the focal scientist to partial series, each of which consists of all the coauthored papers with a given collaborator. The topics in which a collaborator is involved can be identified by the marked colors of the coauthored papers. In Fig. 1, we illustrate the publication time series of a typical scientist, as well as the decomposed time series of his collaborators. The figure indicates that many collaborators of this scientist are involved in a very small number of his topics.

To statistically test, quantify, and understand the pattern illustrated in Fig. 1, we analyzed the scientific publication data of the American Physical Society (APS) journals, as well as five other datasets from other disciplines (see details in *Materials and Methods*). The present study will mainly focus on the APS data. The results of the other five datasets are similar to those of APS; see Fig. 6 and *SI Appendix*, Figs. S12–S18 for a summary. After name disambiguation (29), the APS data contain 236,884 distinct scientists. We consider as focal scientists, in order to ensure meaningful community-detection results, all scientists that have published at least 50 papers, resulting in 3,420 focal scientists. The rest of the authors are included in the analysis, as they may appear as collaborators of these focal scientists.

The first question we ask is: In how many topics is the collaborator of a focal scientist typically involved? To this end, for each focal scientist, we take all his collaborators who coauthored at least two papers with him and calculate the number of topics in which each collaborator is involved. The distribution of collaborators in a number of topics is computed for each focal scientist. Then, we evaluate over all focal scientists the average fraction of collaborators for a given number of topics, as shown in the probability distribution in Fig. 2*A*. The results indicate that, on average, 63% of collaborators of a scientist are involved in a single topic, and about 25% are involved in two topics, while 12% are in three or more topics. To test whether this phenomenon can be explained by random behavior, we consider a surrogate time-controlled reshuffling of the coauthored papers of the collaborators. In the reshuffling process, each paper coauthored by a collaborator and the focal scientist is exchanged with a randomly selected paper that

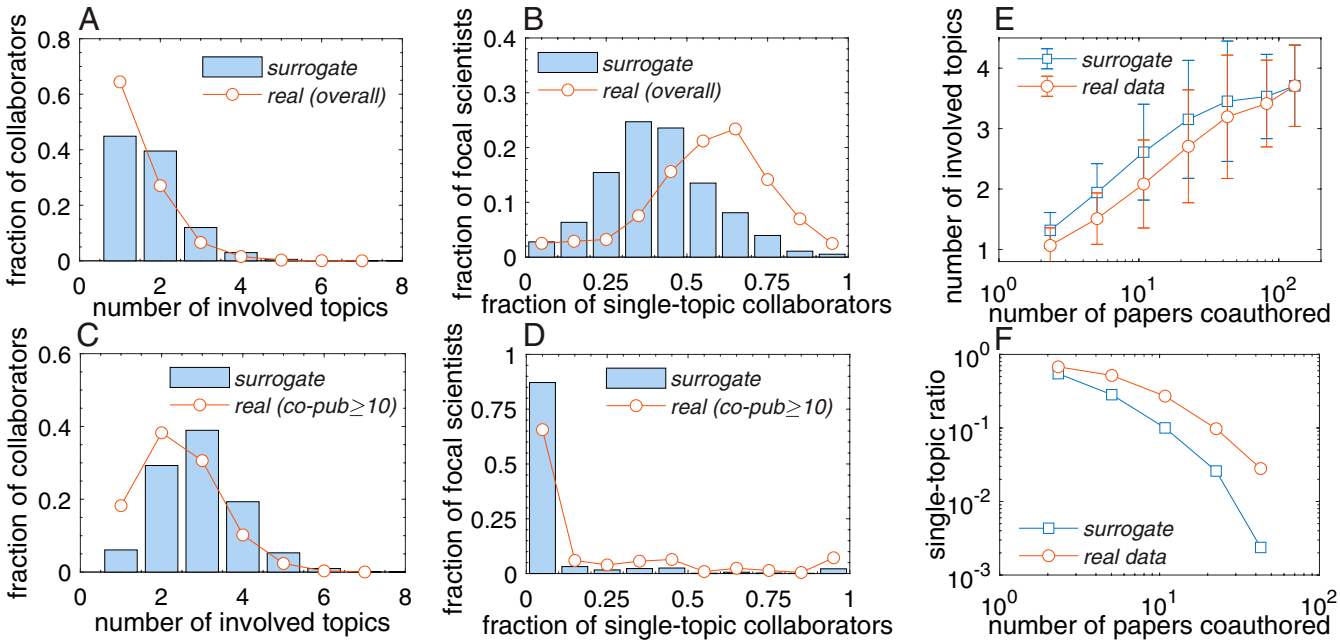

Fig. 2. Number of topics in which collaborators are involved. (*A*) The distribution of the number of topics in which collaborators are involved with a focal scientist. For a scientist, on average, about 63% of his collaborators are involved in only one topic. We also show the results of a controlled surrogate case, where the relations between collaborators and their coauthored papers with the focal scientist are randomly shuffled, which is only about 45%. Note that only the papers published in the same year are allowed to be shuffled in the randomization (*Materials and Methods*). (*B*) For all individual scientists, we calculate the fraction of their collaborators involved in only one topic (denoted as single-topic collaborators). We show in this panel the distribution of the fraction of single-topic collaborators for different scientists. It is clearly seen that the majority of scientists tend to have a high fraction of single-topic collaborators compared to the surrogate control. (*C*) The distribution of the number of involved topics for collaborators who coauthored at least 10 papers with the focal scientists. For a scientist, the fraction of these collaborators involved in one topic is close to 20%, suggesting that the number of involved topics is strongly associated with the number of coauthored papers. In contrast, the controlled surrogate case has about 6% single-topic collaborators. (*D*) The distribution of the single-topic collaborator ratio among those having at least 10 copublications with a focal scientist. Around 65% focal scientists have less than 10% single-topic collaborators. (*E*) The average number of involved topics for collaborators who coauthored different numbers of papers with the focal scientist. The result shows that the number of involved topics increases very slowly—i.e., logarithmically—with the number of coauthored papers. The number of involved topics for the surrogate case is higher than that of the real data, again suggesting the strong tendency of scientists to have single-topic collaborators. (*F*) The fraction of single-topic collaborators for the collaborators who coauthored different numbers of papers with the focal scientist.

is published in the same year by another collaborator and the focal scientist (see *SI Appendix*, Fig. S2 for illustration). By comparing the real data and the controlled surrogate in Fig. 2*A*, one can see that the high fraction of scientists involved in a single topic, 0.63, cannot be explained by the controlled surrogate, which is significantly smaller, 0.45, suggesting the significant tendency of focal scientists to involve collaborators in fewer topics than expected by the surrogate (for significance test, see *SI Appendix*, Fig. S3). To further support this, we calculate for all focal scientists the probability density of the fraction of their collaborators involved in only one topic. As seen in Fig. 2*B*, the distribution of the fraction of single-topic collaborators follows a roughly normal distribution, with the most probable value around 0.65, very close to the mean value. The surrogate of reshuffled data also follows a roughly normal distribution, yet with a much smaller most probable value, close to 0.4. We further compute the distribution of the number of involved topics for collaborators who coauthored at least 10 papers with the focal scientists. Collaborators with many joint papers have a higher chance to be involved in more than one topic (Fig. 2*C*). Nevertheless, despite the smaller fraction of single-topic collaborators in real data, 0.2, it is still much higher, over a factor of 3, than that of reshuffled data, 0.06. We also show in Fig. 2*D* the distribution of the fraction of single-topic collaborators among those who coauthored at least 10 papers with a focal scientist. It can be seen that the distribution is no longer normal as in Fig. 2*B*, and the majority of focal scientists in this case have a very low fraction of single-topic collaborators.

The results in Fig. 2 *A*–*D* also indicate that the number of involved topics is strongly associated with the number of coauthored papers. To quantify this effect, we study directly in Fig. 2*E* the relation between the number of involved topics and the number of coauthored papers. The results suggest a positive correlation between these two quantities. Note that the nearly linear relation under the logarithmic x axis indicates that the number of involved topics increases very slowly—i.e., logarithmically—with the number of coauthored papers. Note also, in Fig. 2*E*, that the number of involved topics in real data are consistently smaller than those of reshuffled surrogate data for different number of coauthored papers. This further supports that collaboration with the same collaborator on several topics is limited—i.e., lower than expected in a surrogate control. We also compute, in Fig. 2*F*, the fraction of single-topic collaborators for collaborators with different number of coauthored papers with the focal scientists. One can see that the fraction of single-topic collaborators decreases with the number of coauthored papers. Nevertheless, the fraction of single-topic collaborators in real data is constantly higher than that in the surrogate data, confirming the tendency of collaborators to join efforts in a single topic.

We further ask how successful scientists are associated with their collaborators in different topics. There are many ways to define a successful scientist. In this paper, we consider two widely adopted metrics—namely, the productivity (in terms of total publications) and impact (in terms of the mean citations c_{10} per paper). Here, c_{10} is the number of citations that a paper receives during the 10 years since it was published (29). We show in Fig. 3*A* that these two metrics are almost uncorrelated; thus, selecting the top scientists according to each of the two metrics independently would result in two very different groups of scientists. Indeed, in

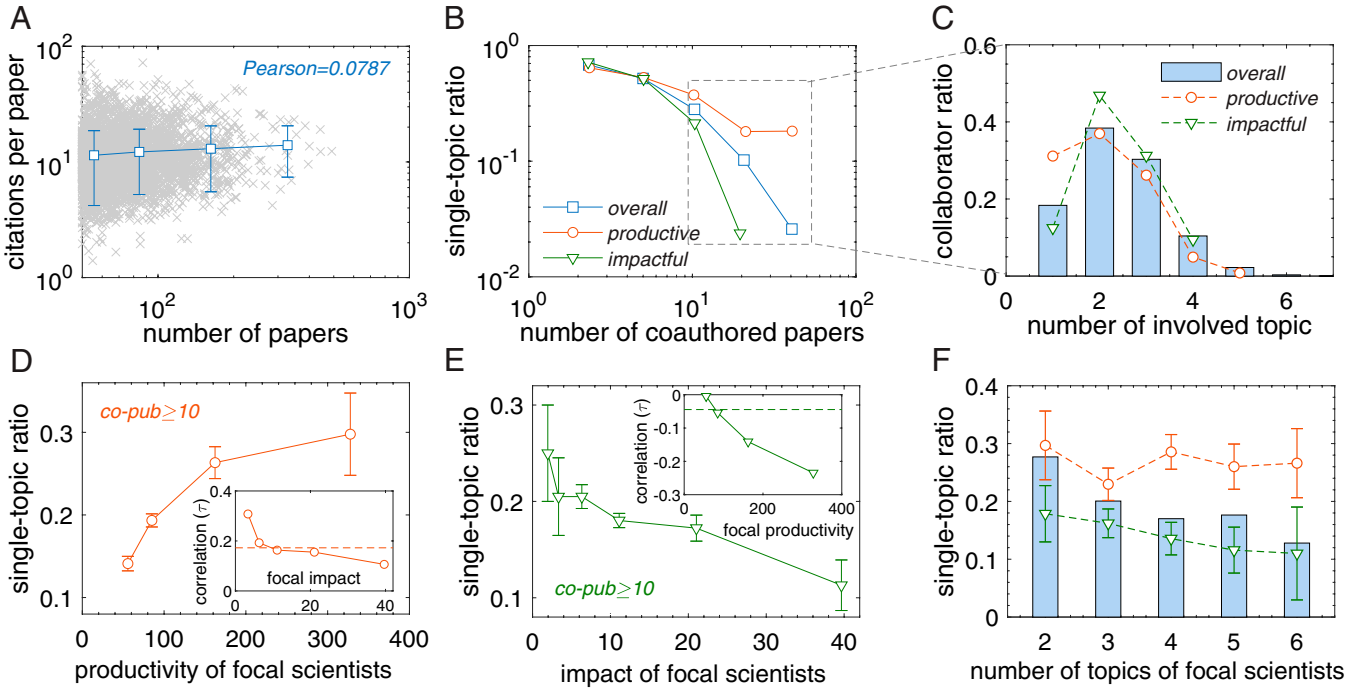

Fig. 3. Productive and impactful scientists associate differently with single-topic collaborators. (*A*) Scatterplot of the productivity (measured by the number of papers) and average impact (measured by the mean citations c_{10} per paper) of scientists, where each dot represents a scientist. c_{10} is the number of citations that a paper receives in the 10 y since it was published. The results show that the correlation between productivity and average impact is very weak, indicated also by the low Pearson correlation of 0.08. Therefore, the scientists with high productivity and the scientists with high impact are two very different groups. (*B*) The fraction of single-topic collaborators for the collaborators who coauthored different numbers of papers with the focal scientist. We compare the 1% most productive scientists (productive in terms of number of published papers) and the 1% most impactful scientists (impactful in terms of mean citations per paper). (*C*) The distribution of the number of topics for the collaborators who coauthored at least 10 papers with the focal scientists. The productive scientists have a significantly higher fraction of single-topic collaborators, while the highly cited scientists have a lower fraction of single-topic collaborators. (*D*) The fraction of single-topic collaborators (among those having at least 10 copublications) for focal scientists with different productivity. More productive scientists have a significantly higher single-topic collaborator ratio. We fix the impact of the focal scientists and show in *D*, *Inset* the Kendall's τ positive correlation between a scientist's single-topic collaborator ratio and his productivity. The dashed line in *D*, *Inset* marks the overall correlation. (*E*) The fraction of single-topic collaborators (among those having at least 10 copublications) for focal scientists with different impact. Higher-impact scientists have a significantly smaller single-topic collaborator ratio. We fix the productivity of the focal scientists and show in *E*, *Inset* the Kendall's τ negative correlation between a scientist's single-topic collaborator ratio and his impact. The dashed line in *E*, *Inset* marks the overall correlation. (*F*) The fraction of single-topic collaborators (among those having at least 10 copublications) for focal scientists with different numbers of topics. The legend in *F* is the same as that in *C*.

SI Appendix, Fig. S5, we show that the mean citations per paper of the scientists with the highest productivity (top 1%) is roughly the same as the mean citations per paper over all scientists. Also, the productivity of scientists with the highest mean citations per paper (top 1%) is almost the same as the average productivity of all scientists. In Fig. 3*B*, we show the relation between the fraction of single-topic collaborators and the number of coauthored papers and compare between the behavior of focal scientists with the 1% highest productivity and the 1% highest impact (the top 5% and top 10% scientists show similar trends; *SI Appendix*, Fig. S6). It is seen that there is no significant difference between these two groups of focal scientists when considering the occasional collaborators (those who coauthored at most five papers). However, for the frequent collaborators who coauthored at least 10 papers with the focal scientists (marked by copub ≥ 10), it is clearly seen that productive and highly cited scientists behave very differently in involving scientists in research topics. Productive scientists have a higher fraction of single-topic collaborators, yet impactful scientists have a lower fraction of single-topic collaborators, which means a higher fraction of multitopic collaborators. This difference is supported by Fig. 3*C*, where we show directly the distributions of the number of involved topics for frequent collaborators.

To support the finding in Fig. 3*B*, we calculate the fraction of single-topic collaborators among frequent collaborators (copub ≥ 10) for scientists with different productivity and impact in Fig. 3 *D* and *E*, respectively. An increasing trend in Fig. 3*D* and a decreasing trend in Fig. 3*E* can be observed. This suggests that the fraction of single-topic collaborators is positively correlated with focal scientists' productivity and negatively correlated with focal scientists' impact. To quantify the correlation, we directly compute the Kendall's tau correlation (τ) between the fraction of single-topic collaborators (copub ≥ 10) of a scientist and the scientist's productivity or impact. Fig. 3 *D*, *Inset* shows the correlation between the fraction of single-topic frequent collaborators and the focal scientists' productivity, given different impact of the focal scientists. The results suggest that the positive correlation exists, even when fixing the impact of the focal scientists, and the correlation is stronger for scientists with smaller impact. Fig. 3 *E*, *Inset* shows the correlation between the fraction of single-topic frequent collaborators and the focal scientists' impact, given different productivity of the focal scientists. The results suggest that the negative correlation exists, even when fixing the productivity of the focal scientists, and the correlation is stronger for scientists with higher productivity. In Fig. 3*F*, we show the fraction of single-topic collaborators (copub ≥ 10) of focal scientists with different numbers of topics. The results indicate that scientists working on more topics tend to have a lower fraction of single-topic collaborators. When fixing the number of topics that a scientist has, the fraction of single-topic collaborators of productive scientists is still higher than average, and the fraction of single-topic collaborators of impactful scientists is consistently lower than average.

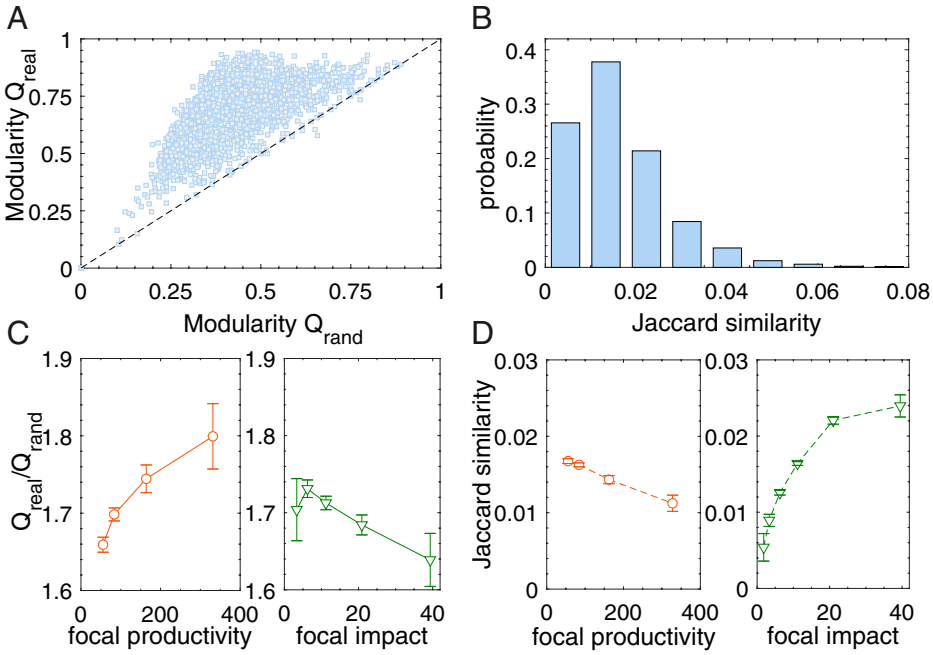

Fig. 4. Features of the collaborators of productive scientists and impactful scientists. (*A*) The relation between the maximized modularity Q_{real} and Q_{rand} for all focal scientists. Each pair of Q_{real} and Q_{rand} are obtained by detecting community structure in the collaboration network among collaborators of a scientist and in the degree-preserved reshuffled counterparts, respectively (*Materials and Methods*). All the points are located above the diagonal line $Q_{real} = Q_{rand}$, indicating that the community structure in real networks is truly significant. (*B*) To quantify the research-interest similarity between a focal scientist and each of his collaborators, we measure the Jaccard similarity of the references given by their papers before collaboration (see *SI Appendix*, Fig. S7 for results of other similarity metrics). We show the distribution of the mean similarity for all focal scientists. (*C*) Q_{real}/Q_{rand} for focal scientists with different productivity or impact. A larger Q_{real}/Q_{rand} indicates a more significant community structure. (*D*) The mean reference similarity for focal scientists with different productivity or impact. Productive scientists and their collaborators have limited research interest in common, while impactful scientists and their collaborators have more common research interest.

We further explore the possible reasons leading to the findings in Fig. 3. We first test an interesting hypothesis that our findings are a result of systemic effects that engagement in various fields yields higher impact. If this were the case, the top interdisciplinary scientists would tend to have more citations and higher impact. Their collaborators would, most likely, be interdisciplinary scientists as well—i.e., engaged in multiple topics. However, we find that the mean citations per paper of individual scientists is negatively correlated with their number of research topics (see *SI Appendix*, section 4 for more details). This observed pattern suggests that our findings are not due to systemic effects, but more likely a result of individual behavior of scientists. Specifically, a productive scientist is usually a principal investigator and, thus, has a large research team, in which each topic has a specific group of collaborators working on it. This is supported by the evidence in Fig. 4 *A* and *C* that the collaboration network among collaborators of a productive scientist has more significant community structure, which suggests that collaborators of a productive scientist tend to form clusters (possibly according to topics), and they are more likely to work with each other in the same cluster. As the collaborators of a productive scientist tend to work on the topic in which they are specialized, the fraction of single-topic collaborators would be high. On the other hand, the high fraction of multitopic collaborators of highly cited scientists might be associated with their tendency to work with collaborators who share similar interests. This is indeed supported by the higher fraction of common references between an impactful scientist's papers and his collaborators' papers before their collaboration started, as shown in Fig. 4 *B* and *D*. Therefore, the selected collaborators are not only suitable for the initially collaborated topic, but also are preferred collaborators for further topics, which results in a higher fraction of multitopic collaborators.

The next question we ask is: What are the features of the collaborators involved in single or multiple topics of a focal scientist? We focus on how the collaboration history with the focal scientist is related to the probability of the collaborator to be involved in the next new topic of the focal scientist. The overall probability of an existing collaborator to be involved in the next new topic of a focal scientist is close to 0.11. We show in Fig. 5*A* the probability to be involved in the next topic of a focal scientist as a function of the number of past coauthored papers. The results suggest that collaborators who published more papers with a focal scientist have significantly higher probability to be involved in the next topic. Considering that collaborators with many coauthored papers might have started collaboration with the focal scientist long ago and may no longer be actively collaborating with the scientist, we further show in Fig. 5*A* the probability among recent collaborators who have coauthored papers with the focal scientists within the past 2 y. The average probability of a recent collaborator appearing in the next new topic of a focal scientist is 0.25, much higher than the overall probability, indicating that a scientist is significantly more likely to involve recent collaborators in a new topic. When considering recent collaborators, the probability to be involved in the next topic still significantly increases with the number of past coauthored papers. The increasing relations can be further quantified by the Kendalls' τ correlations, given in the legend of Fig. 5*A*. In Fig. 5*C*, we also show the correlation for focal scientists with different productivity or impact. One can see that the correlation becomes weaker for productive scientists, yet it becomes stronger for impactful scientists. Fig. 5*B* depicts the relation between the probability to be involved in the next topic and the mean citations of past coauthored papers. Like in Fig. 5*A*, we compute here also the probability among recent collaborators to become a collaborator of a new topic. Interestingly, in both

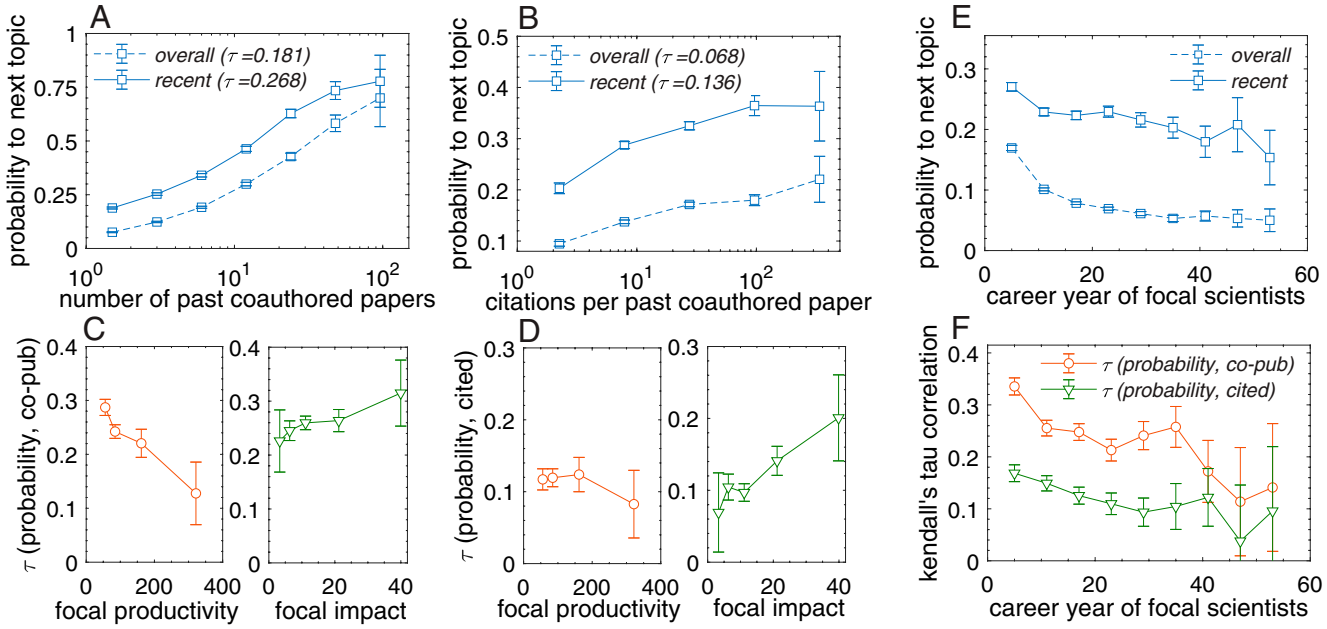

Fig. 5. Factors associated with the probability of an existing collaborator to join a new topic of a focal scientist. (*A*) For a focal scientist that starts to work on a new topic, we calculate the probability of his existing collaborators to join him in the new topic. The overall probability of an existing collaborator to join the new topic of a focal scientist is close to 0.11. We compute also the probability among recent collaborators who have coauthored papers with the focal scientists within the past 2 y and find it to be much higher (i.e., the mean is close to 0.25). Both probabilities show an increasing trend with the number of past coauthored papers, indicating that more intensive past collaboration increases the probability of a collaborator to join a new topic of a focal scientist. (*B*) The probability of a collaborator to join the next topic of the focal scientist versus the mean citations of their past coauthored papers. Both the overall probability and the probability among recent collaborators (which is much higher) show an increasing trend, suggesting that the collaborators who published highly cited papers with the focal scientist have higher probability to join the next topic of the focal scientist. (*C*) The Kendall's τ correlation between the probability to join the next topic and the number of past coauthored papers of a collaborator, for focal scientists with different productivity or impact. (*D*) The Kendall's τ correlation between the probability to join the next topic and the citations per past coauthored paper of a collaborator, for focal scientists with different productivity or impact. (*E*) The probability of a collaborator to join the next topic in different career stages of the focal scientists. Both the overall probability and the probability among recent collaborators show a decreasing trend, suggesting that the focal scientists tend to have higher fraction of single-topic collaborators in their later careers. (*F*) The Kendall's τ correlation of the collaborators' performance and the probability to join the next topic in different career stages of the focal scientists. The correlations tend to be weaker in the later career stage of the focal scientists.

cases, positive correlations are again observed, suggesting that collaborators having published higher-impact papers with a focal scientist have significantly higher probability to become involved in the next topic of the scientist. In Fig. 5*D*, we show that correlation becomes even stronger for impactful scientists. We also investigate the features of selected collaborators for their first topic with a focal scientist, finding similarly that initial collaborators of impactful scientists tend to have much higher mean citations per past paper, up to a factor of 4 compared to low-impact scientists (*SI Appendix*, section 2 and Figs. S8 and S9). These results imply that a pair of high-impact scientists have significantly higher probability to initiate collaboration on a new topic, compared to a pair of low- and high-impact or a pair of low-impact scientists.

The observation in Fig. 5 *C* and *D* might be well explainable by the results in Fig. 4. A focal scientist with high productivity usually has a large research team, in which each topic has a specific group of collaborators working on it. Therefore, the collaborators are very different in their specialization. When a focal scientist selects collaborators for a new topic, he has to take into account both their past performance and their suitability for this topic. Therefore, the productive focal scientists exhibit a lower correlation between the collaborators' past performance and their probability to join the next topic. The impactful focal scientists, on the other hand, tend to work with collaborators who share similar interests to them. The collaborators are generally more likely to be suitable candidates for the new topic of the focal scientists. Taking out the factor of suitability, the past performance of the collaborators thus plays a more important role in affecting the probability to join the next topic. Therefore, for impactful focal scientists, one can observe a higher correlation between the collaborators' past performance and their probability to join the next topic.

Another factor that may affect the probability of collaborators to become involved in a new topic of a focal scientist is the career stage of the focal scientist. We thus show in Fig. 5*E* the probability of a collaborator to join the next topic in different career stages of the focal scientist. In addition to the overall probability, we also provide the probability of recent collaborators. One can see that the probabilities decrease with the career years of focal scientists, suggesting that scientists in later career stages tend to have a lower fraction of multitopic collaborators (i.e., a higher fraction of single-topic collaborators); see *SI Appendix*, Fig. S11 for further support. A possible reason for this could be that senior scientists may have research groups, each of which consists of specialized collaborators. In Fig. 5*F*, we compute the Kendall's τ correlation of the collaborators' past performance (coauthored papers or citations per coauthored paper) and the probability to join the next topic in different career stages of the focal scientists. The results show that both correlations decrease with the career years of the focal scientists, suggesting that senior scientists are less sensitive to the past collaboration performance when involving existing collaborators in a new topic.

We further study how the single-topic-collaboration phenomenon evolved in the past decades. To this end, we consider focal scientists who started their career in different years and calculate the fraction of their single-topic collaborators. We consider only scientists in their first 30 career years, making scientists who start their career in different years comparable. In Fig. 6*A*, we observe a decreasing fraction of collaborators

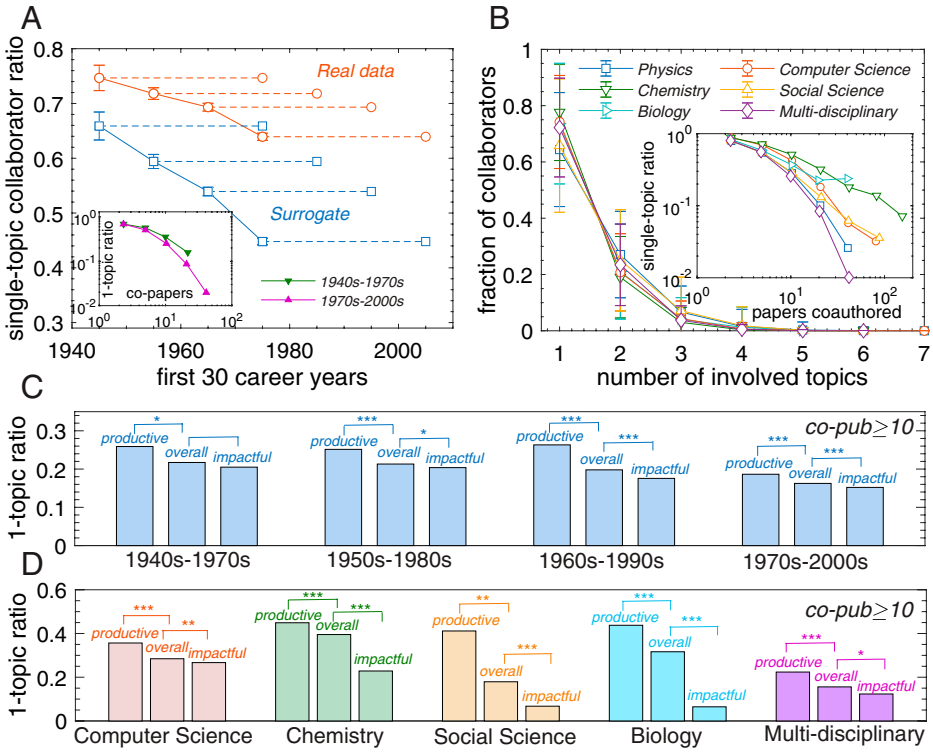

Fig. 6. Evolution in the last century and discipline comparison. (*A*) The fraction of single-topic collaborators of scientists who started their career in different years. We consider only scientists' first 30 career years, making scientists that started their careers in different years comparable. The career starting years are marked by symbols, and the first 30 career years are denoted by the dashed lines. One can see that in more recent years, scientists have a lower fraction of single-topic collaborators, yet the fraction is always significantly higher than surrogate control. The observed trend is supported by *A*, *Inset*, where we show the fraction of single-topic collaborators for the collaborators who coauthored different numbers of papers with the focal scientist. We compare two groups of scientists whose first 30 career years are, respectively, from the 1940s to the 1970s and from the 1970s to the 2000s. (*B*) The distribution of the number of topics in which collaborators are being involved. We compare data from different disciplines, including physics, chemistry, biology, computer science, social science, and multidisciplinary science. (*B*, *Inset*) The fraction of single-topic collaborators for the collaborators who coauthored different numbers of papers with the focal scientist. We observe a strong tendency of single-topic collaboration, with the fraction of single-topic collaborators being particularly high in biology and chemistry. (*C*) The average fraction of single-topic collaborators for the top 10% productive and the top 10% impactful scientists whose first 30-y careers are in different periods. (*D*) The average fraction of single-topic collaborators for productive scientists and impactful scientists in different disciplines. In *C* and *D*, the fraction of single-topic collaborators is calculated among collaborators who coauthored at least 10 papers with the focal scientists. Asterisks between two adjacent bars indicate the *P* values from the Kolmogorov–Smirnov test of the corresponding distributions. * $P \leq 0.1$; ** $P \leq 0.01$; *** $P \leq 0.001$. Almost all pairs of distributions significantly differ from one another. The large *P* values for the 1940*s* through the 1970*s* in *C* are due to the small sample sizes.

involved in a single topic, indicating that in the last century, as time evolved, more collaborators of scientists tended to work in multiple topics. Nevertheless, the fraction of scientists involved in a single topic is still significantly higher than surrogate control, and the difference seems to have become more prominent as time evolved, supporting the significant tendency of single-topic collaborations. We further compare in Fig. 6 *A*, *Inset* two groups of scientists whose first 30 career years are, respectively, from the 1940s to the 1970s and from the 1970s to the 2000s. The results show that recent scientists (career from the 1970s to the 2000s) indeed have a lower fraction of single-topic collaborators for a given number of coauthored papers. In Fig. 6*C*, we show the average fraction of single-topic collaborators for top-10% productive and top-10% impactful scientists whose first 30-y careers are in different periods. One can see that, in each time period, the impactful scientists have a lower fraction of single-topic collaborators compared to overall, while the fraction is higher for productive scientists. In *SI Appendix*, Fig. S10, we additionally examine the correlation between the probability to join the next topic and the past collaboration performance of a collaborator for focal scientists who started their career in different years. The results show that scientists from different years exhibit similar trends as Fig. 5 *C* and *D*.

Finally, we compare data from different disciplines, including physics, chemistry, biology, computer science, social science, and multidisciplinary science. In Fig. 6*B*, we find a similar form of the distribution of collaborators' topic numbers in different fields. We further find in *SI Appendix*, Fig. S12 that the fraction of single-topic collaborators in these disciplines is higher than that of the corresponding surrogate control. Fig. 6 *B*, *Inset* shows that the fraction of single-topic collaborators is particularly high in biology and chemistry. The reason for this is probably since these two disciplines have many experimentalists whose research fields require expensive equipment and long-term accumulation of knowledge and mastery of techniques, which makes them focus on fewer topics (*SI Appendix*, section 3 and Fig. S13). We additionally show in Fig. 6*D* that in all considered disciplines, impactful scientists have a lower fraction of single topic than overall, while productive scientists have higher fraction of single topic than overall (see *SI Appendix*, Figs. S14–S18 for more details).

Discussion

Scientific research increasingly depends on teamwork. It is thus critical to understand the collaboration behavior of scientists. Despite much effort that has been made to investigate the structure and the strength of collaboration networks, how scientists involve collaborators in their research topics remains poorly understood. In this paper, we find that the actual number of topics in which the same collaborator is involved is significantly smaller than

expected from surrogate time-controlled reshuffling, suggesting the preference of recruiting collaborators for a single topic. We interestingly find that productive scientists have a higher fraction of single-topic collaborators, yet highly cited scientists have a higher fraction of multitopic collaborators. Our analysis suggest that the observed difference is associated with their tendency in selecting collaborators. The impactful scientists tend to have collaborators sharing similar research interests, while productive scientists tend to have collaborators specialized in a topic. We further study for a focal scientist: What are the features of his existing collaborators when starting a new topic? We find a stronger tendency of highly cited scientists to involve collaborators with many publications and high citations per paper, yet, in contrast, highly productive scientists have a much weaker such tendency. By comparing active scientists in different years, we observe a rising probability, but still significantly smaller than the controlled surrogate, of involving collaborators in multiple topics. We finally validate our findings across different disciplines, finding that in all considered disciplines, impactful scientists have a higher tendency to involve collaborators in new topics.

Our findings can be useful for improving the organization of science. First, our analysis shows that the productivity of a scientist and the average impact per paper of the scientist are almost uncorrelated. Productive scientists usually derive their productiveness from large teams, but our results suggest that these teams do not produce works with above-average impact. Therefore, policy makers could consider balancing resources between large and small teams. Secondly, despite much literature having pointed out that the challenges of the modern world are increasingly interdisciplinary (7), our work shows that science is still dominated by single-topic collaborations. As multitopic collaborations are associated with higher impact, proper reorganization of science in terms of encouraging multitopic collaboration might be helpful for advancing science. Finally, we find that impactful scientists tend to choose impactful scientists as collaborators for a new topic. It implies that successfully breaking new ground is still a task that is hard to do alone. Thus, there are probably still obstacles to performing interdisciplinary science that need to be removed.

This work may provide a perspective for understanding individual scientists' careers. In recent years, numerous patterns in individual scientists' careers, such as the random-impact rule (29) and the hot streak (30), have been revealed. However, related analyses inevitably take into account coauthored papers in scientists' careers, causing the risk of regarding collaborators' behavior as the focal scientist's behavior. It is thus still unclear how to separate the true behavior of a scientist from the publication records. The method of decomposing publication time series developed in this paper may shed light on this challenging issue. In addition, the framework proposed in this paper can be easily extended to other systems with collaboration, such as film actors, patent design, and software development. Finally, we note that our research has limitations. Despite revealing the distribution of topics in scientific collaboration, our work cannot distinguish who is the one initiating their collaboration on these topics. Is it dominated by the focal scientists or by their collaborators? Future investigation on this issue could deepen our understanding on the origin of the observed phenomena in this paper.

Materials and Methods

Data. We study in this paper six large-scale datasets, including the disciplines of physics, chemistry, biology, computer science, social science, and multidisciplinary science. The physics dataset consists of the scientific publication data of the APS journals (29). The computer science data are the AMiner dataset, obtained by extracting scientists' profiles from online web databases (31). The chemistry data contain the publication data of the American Chemical Society journals. The biology data contain the publication data of Cell Press journals. The social science data contain the publication data of SAGE publishing group journals. The multidisciplinary science data contain all papers in five representative multidisciplinary journals, including *Nature*, *Science*, PNAS, *Nature Communications*, and *Science Advances*. The data of chemistry, biology, social science, and multidisciplinary science are extracted according to the DOIs of papers from a large publication dataset freely downloaded from Microsoft Academic Graph (32). More detailed data description is given in *SI Appendix*, section 1.

Decomposing Time Series. We first construct for each scientist a cociting network, in which each node is a paper authored by this scientist, and two papers have a link if they share at least one reference (*SI Appendix*, Fig. S1). The communities in the cociting networks are detected via the fast-unfolding algorithm (33), with each significant community (more than 5% of papers) representing a major topic of the scientist. As the cociting network needs to be large enough to ensure meaningful community-detection results, we consider only the focal scientists with at least 50 papers. For each focal scientist, we generate the time series presented in Fig. 1*A* describing the growth history of the network. In the time series, each point is a paper, and different colors represent different communities in the cociting network. Since many of the publications of a focal scientist have resulted from teamwork, the time series is actually aggregated from coauthored papers with different collaborators. We then decompose the publication time series of a scientist to various time series, each of which records the coauthored papers with a specific collaborator, as shown in Fig. 1*B*. The time series of a collaborator clearly exhibits the key information of the collaboration, including the number of involved topics, the starting year of the collaboration, the collaboration length, and so on. For better illustration, we show in Fig. 1*B* the time series of the collaborators with at least five coauthored paper with the focal scientist. The illustration of the time series of all collaborators is given in *SI Appendix*, Fig. S1. We show also in *SI Appendix*, Fig. S4 the statistics of collaboration years and the number of coauthored papers on a topic.

Surrogate Time-Controlled Reshuffling. To examine the significance of an observed pattern in real data, one has to compare it to the result of randomized cases. In this paper, we consider a surrogate time-controlled reshuffling procedure in which the relations between a scientist's collaborators and his papers are iteratively randomized. Specifically, a paper coauthored by a collaborator and the focal scientist is exchanged with a randomly selected paper coauthored by another collaborator and the focal scientist. There is time constraint in the procedure that these two papers must be published in the same year, avoiding the case where a collaborator is assigned to a paper that was published even before they started collaboration. In this way, the timing of the collaboration is preserved for each collaborator, yet their involved topics are randomized. The illustration of the surrogate time-controlled reshuffling procedure is presented in *SI Appendix*, Fig. S2.

Computing the Probability to Join the Next Topic. A scientist may work on multiple topics during his career. When a scientist starts to work on a new topic, we calculate the fraction of his existing collaborators that will coauthor at least one paper with the scientist in the new topic. The overall probability is obtained by averaging the fraction over all topics, except the first topic (as the scientist has no existing collaborators when starting the first topic). One possible concern is that the probability might be underestimated, as some collaborators may have already stopped working with the focal scientist long before the focal scientist starts a new topic. We thus further examine the case where all inactive collaborators are removed. Specifically, we calculate the probability to join the next topic only among the collaborators who have coauthored at least one paper with the focal scientist in the testing year or 1 y before.

Detecting Communities in the Collaboration Network among Collaborators of a Scientist. In Fig. 4, we construct a collaboration network for each focal scientist in which nodes are collaborators of this scientist and links are their coauthorship relations in the scientist's papers. We detect community structure in each of these collaboration networks with the fast-unfolding algorithm

(33). We calculate the maximized modularity Q_{real} of the real networks and the maximized modularity, Q_{rand}, in their degree-preserved reshuffled counterparts. The modularity function (34) is defined as

$$Q = \frac{1}{2m}\sum_{i,j}[A_{ij} - \frac{k_i k_j}{2m}]\delta(c_i, c_j), \quad \textbf{[1]}$$

where A is the adjacency matrix of the network, k_i is the degree of node i, m is the total number of links in the network, c_i is the community to which node i is assigned, and the δ function $\delta(c_i, c_j)$ is one for $c_i = c_j$, and zero otherwise. The communities are obtained when the function Q is maximized.

Data Availability. Previously published data were used for this work. The APS data are available upon request submitted to https://journals.aps.org/datasets (35), the AMiner data can be freely downloaded via https://www.aminer.cn/aminernetwork (36), and the Microsoft Academic Graph data can be accessed in Zenodo (37).

ACKNOWLEDGMENTS. We thank the anonymous referees for their excellent suggestions that helped us to improve our manuscript. This work is supported by the National Natural Science Foundation of China under Grant 71731002. S.H. thanks the Israel Science Foundation and the NSF-US–Israel Binational Science Foundation for financial support.

1. S. Wuchty, B. F. Jones, B. Uzzi, The increasing dominance of teams in production of knowledge. *Science* **316**, 1036–1039 (2007).
2. S. Milojević, Principles of scientific research team formation and evolution. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 3984–3989 (2014).
3. L. Wu, D. Wang, J. A. Evans, Large teams develop and small teams disrupt science and technology. *Nature* **566**, 378–382 (2019).
4. A. Zeng, Y. Fan, Z. Di, Y. Wang, S. Havlin, Fresh teams are associated with original and multidisciplinary research. *Nat. Hum. Behav.* **5**, 1314–1322 (2021).
5. S. Fortunato *et al.*, Science of science. *Science* **359**, eaao0185 (2018).
6. A. Zeng *et al.*, The science of science: From the perspective of complex systems. *Phys. Rep.* **714-715**, 1–73 (2017).
7. M. E. J. Newman, The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 404–409 (2001).
8. M. E. Newman, Assortative mixing in networks. *Phys. Rev. Lett.* **89**, 208701 (2002).
9. L. Krumov *et al.*, Motifs in co-authorship networks and their relation to the impact of scientific publications. *Eur. Phys. J. B* **84**, 535–540 (2011).
10. M. Girvan, M. E. J. Newman, Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 7821–7826 (2002).
11. A. M. Petersen, Quantifying the impact of weak, strong, and super ties in scientific careers. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E4671–E4680 (2015).
12. H.-W. Shen, A.-L. Barabási, Collective credit allocation in science. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 12325–12330 (2014).
13. R. Guimerà, B. Uzzi, J. Spiro, L. A. Amaral, Team assembly mechanisms determine collaboration network structure and team performance. *Science* **308**, 697–702 (2005).
14. M. Klug, J. P. Bagrow, Understanding the group dynamics and success of teams. *R. Soc. Open Sci.* **3**, 160007 (2016).
15. D. Hsiehchen, M. Espinoza, A. Hsieh, Multinational teams and diseconomies of scale in collaborative research. *Sci. Adv.* **1**, e1500211 (2015).
16. M. Coccia, L. Wang, Evolution and convergence of the patterns of international scientific collaboration. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 2057–2061 (2016).
17. B. F. Jones, S. Wuchty, B. Uzzi, Multi-university research teams: Shifting impact, geography, and stratification in science. *Science* **322**, 1259–1262 (2008).
18. R. D. Malmgren, J. M. Ottino, L. A. Nunes Amaral, The role of mentorship in protégé performance. *Nature* **465**, 622–626 (2010).
19. C. Jin, Y. Ma, B. Uzzi, Scientific prizes and the extraordinary growth of scientific topics. *Nat. Commun.* **12**, 5619 (2021).
20. T. Jia, D. Wang, B. K. Szymanski, Quantifying patterns of research-interest evolution. *Nat. Hum. Behav.* **1**, 0078 (2017).
21. F. Battiston *et al.*, Taking census of physics. *Nat. Rev. Phys.* **1**, 89–97 (2019).
22. A. Zeng *et al.*, Increasing trend of scientists to switch between topics. *Nat. Commun.* **10**, 3439 (2019).
23. J. Li, Y. Yin, S. Fortunato, D. Wang, Scientific elite revisited: Patterns of productivity, collaboration, authorship and impact. *J. R. Soc. Interface* **17**, 20200135 (2020).
24. L. Liu, N. Dehmamy, J. Chown, C. L. Giles, D. Wang, Understanding the onset of hot streaks across artistic, cultural, and scientific careers. *Nat. Commun.* **12**, 5392 (2021).
25. B. F. Jones, The burden of knowledge and the death of the Renaissance man: Is innovation getting harder? *Rev. Econ. Stud.* **76**, 283–317 (2009).
26. B. F. Jones, B. A. Weinberg, Age dynamics in scientific creativity. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 18910–18914 (2011).
27. J. G. Foster, A. Rzhetsky, J. A. Evans, Tradition and innovation in scientists' research strategies. *Am. Sociol. Rev.* **80**, 875–908 (2015).
28. B. Uzzi, S. Mukherjee, M. Stringer, B. Jones, Atypical combinations and scientific impact. *Science* **342**, 468–472 (2013).
29. R. Sinatra, D. Wang, P. Deville, C. Song, A. L. Barabási, Quantifying the evolution of individual scientific impact. *Science* **354**, aaf5239 (2016).
30. L. Liu *et al.*, Hot streaks in artistic, cultural, and scientific careers. *Nature* **559**, 396–399 (2018).
31. J. Tang *et al.*, "ArnetMiner: Extraction and mining of academic social networks" in *Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008)* (Association for Computing Machinery, New York, 2008), pp. 990–998.
32. A. Sinha *et al.*, "An overview of Microsoft Academic Service (MAS) and applications" in *Proceedings of the 24th International Conference on World Wide Web (WWW 15 Companion)* (Association for Computing Machinery, New York, 2015), pp. 243–246.
33. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, P10008 (2008).
34. M. E. J. Newman, Fast algorithm for detecting community structure in networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **69**, 066133 (2004).
35. Physical Review Journals, APS Data Sets for Research. https://journals.aps.org/datasets. Accessed 26 July 2022.
36. AMiner, Extraction and Mining of Academic Social Networks. https://www.aminer.cn/aminernetwork. Accessed 26 July 2022.
37. Microsoft Academic, Microsoft Academic Graph (22 March 2019). Zenodo. https://doi.org/10.5281/zenodo.2628216. Accessed 5 June 2020.