Yang and Wang *EPJ Data Science* (2024) 13:15
https://doi.org/10.1140/epjds/s13688-024-00453-6

EPJ Data Science
a SpringerOpen Journal

RESEARCH Open Access

Higher-order structures of local collaboration networks are associated with individual scientific productivity

Wenlong Yang[1] and Yang Wang[1*]

*Correspondence:
yang.wang@xjtu.edu.cn
[1]School of Public Policy and Administration, Xi'an Jiaotong University, Xi'an, China

Abstract

The prevalence of teamwork in contemporary science has raised new questions about collaboration networks and the potential impact on research outcomes. Previous studies primarily focused on pairwise interactions between scientists when constructing collaboration networks, potentially overlooking group interactions among scientists. In this study, we introduce a higher-order network representation using algebraic topology to capture multi-agent interactions, i.e., simplicial complexes. Our main objective is to investigate the influence of higher-order structures in local collaboration networks on the productivity of the focal scientist. Leveraging a dataset comprising more than 3.7 million scientists from the Microsoft Academic Graph, we uncover several intriguing findings. Firstly, we observe an inverted U-shaped relationship between the number of disconnected components in the local collaboration network and scientific productivity. Secondly, there is a positive association between the presence of higher-order loops and individual scientific productivity, indicating the intriguing role of higher-order structures in advancing science. Thirdly, these effects hold across various scientific domains and scientists with different impacts, suggesting strong generalizability of our findings. The findings highlight the role of higher-order loops in shaping the development of individual scientists, thus may have implications for nurturing scientific talent and promoting innovative breakthroughs.

Keywords: Higher-order structures; Local collaboration networks; Disconnected components; Higher-order loops; Productivity

1 Introduction

The advancement of modern science has led to an increase in the complexity of scientific problems, and a rise in the cost of scientific instruments, resulting in the emergence of *big science* [1–5]. This paradigm shift has led to the accumulation of knowledge, making it almost impossible for a single scientist to possess comprehensive expertise required for one scientific project, known as the burden of knowledge [6]. Therefore, scientists have increasingly formed scientific teams to address these challenges [3, 4, 7, 8]. Previous research has demonstrated that teams dominate knowledge creation in contemporary science, operating across institutional and national boundaries [8–10]. Collaboration networks have thus become a powerful tool for studying team structures and scientific collaborations [7].

Past two decades have witnessed numerous studies on the properties of collaboration networks, suggesting that collaboration networks exhibit scale-free, small-world, assortativity and strong community structures [7, 11–13]. Recent studies expanded the scope of collaboration networks from binary to weighted [14, 15], temporal [16–18] and multilayer networks [19]. The availability of large-scale bibliometric datasets as well as quantitative tools enables the study of the relationship between collaboration network structure and scientific performance. From the macroscopic point of view, previous studies showed that macroscopic network properties significantly affect scientists' academic performance, including productivity and citation impact [20–28]. From the individual paper's point of view, empirical studies explored microscopic team formation, examining the association between team diversity, team structures and paper citation, novelty, disruption and multidisciplinarity [9, 28–39]. However, existing studies mainly constructed collaboration networks at a dyadic level, potentially overlooking valuable information, as scientific collaboration now is dominated by group interactions beyond dyadic levels.

In recent years, researchers have made substantial progress in network science and computational topology, leading to the emergence of higher-order representations that capture multi-agent relationships beyond conventional dyadic interactions. Notable examples include simplicial complexes [40, 41] and hypergraphs [42, 43], which have been widely applied in analyzing various types of networks across social systems [44], neuroscience [45, 46], ecology [47, 48], and other biological systems [49, 50]. Despite of similar frameworks in the field of science of science [51–55], to the best of our knowledge, there is limited research exploring the association between higher-order properties and individual scientific productivity. In fact, prior research demonstrated that higher-order holes play necessary roles in biological systems especially the brain functioning [56, 57]. This highlights an encouraging and promising direction in the collaboration system, i.e., investigating how these higher-order characteristics affect scientific outcomes. This calls for a further analysis into translating the original co-authorship data into structures that preserve group interactions. Additionally, existing studies have drawn conclusions from specific scientific domains, raising questions regarding the generalizability of the findings.

In this paper, we fill this gap by leveraging the Microsoft Academic Graph data (MAG), a large-scale scholarly dataset. We utilize a simplicial complex framework to construct local collaboration networks for a cohort of more than 3.7 million scientists. Our primary objective is to investigate the association between higher-order structural properties of local collaboration networks and scientists' productivity. Specifically, we delve into two key higher-order characteristics: the 0th Betti number (β_0), representing the number of disconnected components, and the 1st Betti number (β_1), indicating the presence of higher-order loops. There are three key findings. Firstly, we find that there is an intriguing inverted U-shaped relationship between the number of disconnected components and individual productivity. Secondly, we observe that the presence of higher-order loops within local co-authorship networks is positively associated with scientists' productivity, suggesting interesting underlying forces related to group interactions. Thirdly, the uncovered relationship can be generalizable to major scientific domains, indicating strong generalizability of our results. This study has several contributions. First, we use a simplicial complex approach to depict scientific collaboration networks, which helps to capture group interactions and higher-order structural properties that cannot be obtained in the conventional dyadic view. Second, our work encompasses scientists from diverse scientific disciplines, offering insights that extend beyond specific scientific domains. These results may help us better understand individual careers and have policy implications for nurturing scientists towards high academic performance.

2 Related work

2.1 The impact of macroscopic collaboration structure on scientific output

Recently, there has been significant interest in the science of team science [8, 9, 37, 38, 58, 59]. Previous studies documented several fundamental characteristics of collaboration networks [7, 11–13]. The availability of computational tools also pushes scientists to extend conventional binary collaboration networks to weighted, temporal, multilayer and higher-order networks, enabling a more nuanced analysis of collaboration patterns [14–19, 51–55]. Numerous studies demonstrate the impact of collaboration networks on individual scientist's academic performance. For example, prior studies focus on the association between centrality, tie strength and its configuration, structural hole and scientific productivity and citation impact [20–28, 60]. Recent research has also explored the relationship between collaboration networks and innovative research. Using patent datasets, Wang et al. showed that inventors with a high degree centrality in patent collaboration networks often exhibit low exploratory innovation, whereas inventors spanning structural holes produce more innovative outputs [61]. Using the American Physical Society data, Wang et al. observed that scientists spanning over structural holes in scientific collaboration networks produced more novel and disruptive research and had a higher chance to publish novel/disruptive papers [60].

2.2 The impact of microscopic team structure on scientific output

Recent studies delved into the relationship between microscopic team structures and scientific outputs. For example, Zeng et al. proposed the concept of team freshness, and found that team freshness strongly predicts multidisciplinarity and disruption of individual papers [38]. Liu et al. focused on link freshness and demonstrated an inverted U-shaped relationship between link freshness and citation impact [34]. Xu et al. discovered that author contribution within a team is associated with long-term citations, novelty and disruption [36]. Furthermore, Chen et al. explored new author combinations within scientific teams, revealing that new author combinations positively inspire the emergence of new knowledge units and combinations of knowledge elements [33]. Recent studies also focused on team diversity. Yang et al. demonstrated that gender-diverse teams produce novel and impactful papers [37]. In addition to gender diversity, researchers have examined other dimensions of diversity, including ethnicity, nationality, affiliations, discipline and academic age, finding consistently that diverse teams produce impactful papers [9, 29–32]. Finally, Lin et al. studied the association between collaboration distance and disruption, revealing that remote teams were less likely to produce disruptive research compared with onsite teams [39].

2.3 Higher-order network representations in science of science

Conventional research primarily focused on pairwise interactions in collaboration networks, overlooking higher-order interactions involving three or more researchers [51–54].

To fill this gap, algebraic topologists and network scientists have introduced higher-order network representations such as simplicial complexes [40, 41] and hypergraphs [42, 43]. These advancements have enabled the application of higher-order networks in various fields, including social systems, neuroscience, ecology, and other biological systems [44–50]. In science of science domain, there are a few studies exploring higher-order network representations. For example, Carstens and Horadam were among the first to introduce persistent homology to analyze Betti numbers in weighted collaboration networks, distinguishing them from random networks [51]. Patania et al. studied topological structures by analyzing the distribution of facet size, simplicial degrees, homological hole lengths, and community sizes [54]. Similarly, Salnikov et al. constructed sequential knowledge networks using simplicial complexes, and analyzed the persistence of homological holes [55]. Gebhart and Funk used simplicial complexes to study the evolution of homological holes and their correlations with traditional network properties, as well as their impact on the novelty and impact of papers and patents [52]. Juul et al. investigated the frequency of different hypergraph patterns in random models and empirical data, and explored the relationship between citations and hypergraph patterns [53].

In summary, previous research has explored the relationships between the structural attributes of macroscopic collaboration networks and microscopic team structures and how these factors impact scientists' academic performance. Nonetheless, significant gaps remain within the current body of literature. Firstly, there has been limited emphasis on local collaboration networks, despite their potential role in knowledge spillovers and individual outcomes. Furthermore, while earlier studies have indeed investigated higher-order structural features, the precise influence of these structures on scientists' performance remains an open question. To add to this complexity, the generalizability of these findings across a wide array of scientific domains has yet to be fully addressed. In this paper, we seek to address these gaps by examining the impact of higher-order structural properties within local collaboration networks on the productivity of scientists from diverse academic fields. Our study aims to contribute valuable insights and extend the understanding of these intricate relationships.

3 Data

In this paper, we leverage the Microsoft Academic Graph dataset (MAG), which comprises more than 260 million digital publications spanning from 1800 to 2021. MAG offers comprehensive information regarding each publication, including publication year, scientific field(s), and author name(s). It has emerged as a pivotal data source for research on individual careers [62–68]. MAG employs cutting-edge techniques for distinguishing author identities. In addition to machine learning algorithms that leverage publication records for author disambiguation, MAG goes further by harnessing the power of web search engines to access public information such as personal websites and public curricula vitae [69]. Recent studies have established a gold standard dataset for author name disambiguation based on ORCID, finding that MAG author IDs achieve an impressive 81.87% accuracy, 78.13% F1 score, and 98.49% precision, underscoring the reliability of MAG's author identification methods [34, 70].

In this study, we focus on journal articles and conference papers published prior to 2011. Our analysis includes papers with scientific field information as well as venue information, resulting in a dataset of 56,895,201 papers. Furthermore, we focus on scientists who published at least 5 papers and no more than 500 papers during their entire career. This approach helps us mitigate potential errors related to author name disambiguation within the Microsoft Academic Graph (MAG), including instances of author under-conflation, where an author's publication count may be erroneously lower than the actual number, or over-conflation, which involves wrongly assigning additional publication records to an author. This method also allows us to reduce the influence of outliers, which could include authors with very few or exceptionally high numbers of publications. This selection criterion aligns with recent research practices [38, 60]. Moreover, we exclude scientists who have collaborated with more than 36 distinct partners in any given year. The reason for this exclusion is rooted in the considerable computational complexities associated with high-order network analyses. In particular, the computation of homology necessitates enumerating all conceivable combinations of simplices, with computational complexity growing exponentially with the dimension of the simplicial complex [54]. This threshold helps us manage these computational challenges, balancing the need for accuracy with the constraints of available computational resources. Additionally, we focus on scientists who published his/her first paper later than 1960 in order to reduce the noise derived from the relatively small number of publications before.

Our final sample comprises a total of 3,785,807 scientists. For each scientist, we construct his/her yearly local collaboration networks by considering interactions among collaborators (see details in Methodology), resulting in a total of 27,786,774 scientist-year observations till 2011 (see the data frame of "scientist-year observations" in the Appendix, Table A1). Note that scientists with less than a 3-year publication history were excluded to ensure the consistency of the number of samples included into the regression analysis of the panel data.

4 Methodology

4.1 Simplicial complexes

Basic notations and definitions We provide several basic notations and definitions related to simplicial complexes. First, a d-simplex α represents a set of interacting nodes, where d denotes the dimensionality of the simplex. For example, a single node is a 0-simplex, a link between two nodes is a 1-simplex, and a (filled) triangle is a 2-simplex, and so on. Second, a face of a d-simplex α is a lower-dimensional simplex α' formed by a proper subset of nodes of α, i.e., $\alpha' \subset \alpha$. For instance, in the case of a 2-simplex, its faces include three 0-simplices and three 1-simplices. Third, a simplicial complex γ is a collection of simplices that satisfies closure under the inclusion of faces, indicating that for every simplex α belonging to γ, all of its faces α' also belong to γ. For more details, please refer to [71, 72].

Why using simplicial complexes? The use of simplicial complexes can be justified for several reasons. First, it is a natural approach when investigating scientific collaborations, considering that it allows to model multi-agent interactions. Over recent decades, science has witnessed a remarkable increase in complexity and scale, with most knowledge creation by teamwork, or group interactions [8]. When studying collaboration networks through dyadic aspects that originated from scientist-paper bipartite networks, we risk losing crucial information regarding these group interactions. In response to this, recent advancements have been made in higher-order network representations, and such frameworks have found widespread application in the analysis of various network types [44–50].

Second, the use of a simplicial framework is advantageous because it explicitly preserves group interactions that involve more than two scientists. One key benefit of this approach is its ability to encode higher-order "holes" within the collaboration network [54]. To illustrate this, consider two cases: in the first case, three scientists have never co-authored a paper together, but any two of them have collaborated on at least one paper. In the second case, all three scientists have indeed published a paper together previously. When using conventional methods, both situations might be represented as triangles. However, we recognize that only the former case is accurately depicted by an empty triangle, while the latter should be represented by a filled triangle. Similarly, conventional methods cannot distinguish whether quadrilaterals or pentagons are empty or filled. Lastly, the application of higher-order structures empowers us to delve deeper into understanding the functions of these topological features within scientific collaboration networks. Significantly, prior research has illuminated the crucial roles played by higher-order holes in the functioning of the human brain [56, 57]. Nonetheless, it remains unclear how these higher-order holes within collaboration networks are linked to individual scientific careers. This underscores the need to translate original co-authorship data into structures that accurately represent and preserve these group interactions.

4.2 Local collaboration networks

We construct yearly local collaboration networks for each scientist at year t, by extracting his/her collaboration records from preceding year t-5 to t-1 among his/her collaborators. Figure 1 shows an illustrative example of a selected scientist. At year t, the focal scientist collaborated with six scientists (see Fig. 1a). We then identify collaboration relationships among collaborators using publication data between t-5 and t-1 (see Fig. 1b). For example, [1, 5] indicates that scientists 1 and 5 have co-authored a paper during this period, while [1, 2, 6] suggests that scientists 1, 2, and 6 have published a paper together. Using these collaboration records, we obtain the local collaboration network for the selected scientist at year t (see Fig. 1d). It is important to note that we construct this network using higher-order interactions, which differs markedly from the conventional bipartite network projection (see Fig. 1c).

4.3 Betti numbers

In this study, we characterize higher-order structural properties of local collaboration networks using the Betti number, which is a topological measure to quantify the presence of holes in higher-order networks. Each Betti number corresponds to a specific dimension of holes within the network. We provide several related notations below. For details, we refer to these references [73–75].

Boundary operation, d-chain, d-cycles and d-boundary Here, we provide a brief description of key definitions. The boundary of a d-simplex is defined as the sum of its $(d-1)$-dimensional faces, denoted as ∂_d. A d-chain is defined as the sum of d-simplices in a simplicial complex. The group of d-chains is defined as the d-chains with the addition modulo 2, denoted as C_d. A d-cycle is defined as a d-chain with a boundary of zero. The group of d-cycles is defined as the d-cycles with the addition modulo 2, denoted as Z_d. A d-boundary refers to a d-chain that is the boundary of a $(d+1)$-chain. The group of d-boundaries refers to the d-boundaries with the addition modulo 2, denoted as B_d. Note that $B_d \subset Z_d \subset C_d$.

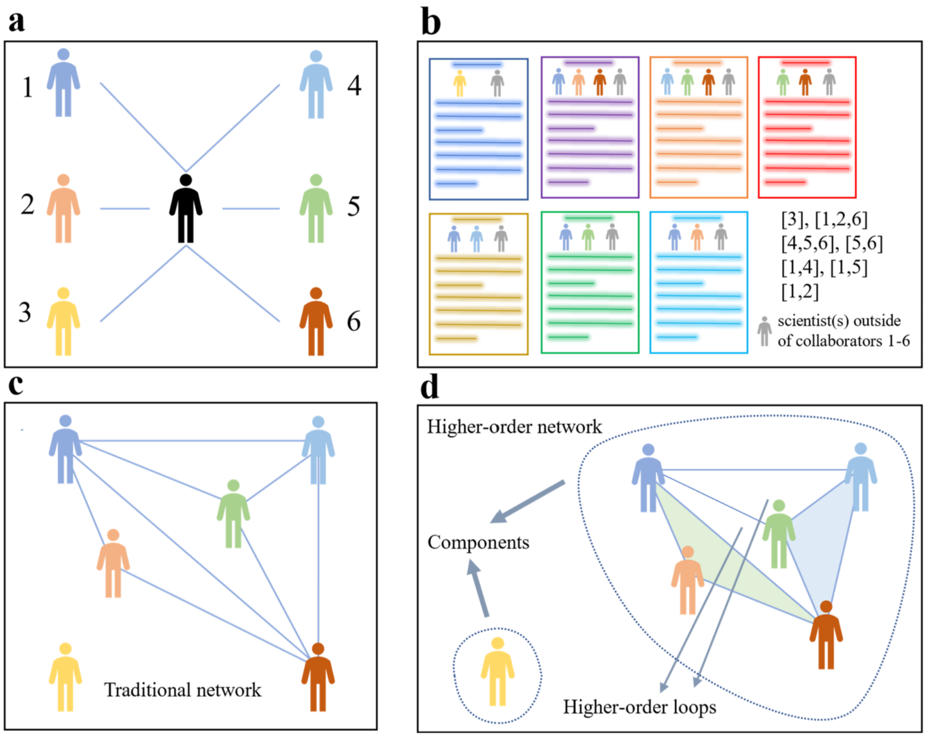

Figure 1 An illustration of constructing higher-order local collaboration networks. (**a**) shows the individual scientist's egocentric network at year *t*. The links indicate that two scientists collaborated at year *t*. (**b**) shows the publication records and collaboration relationships among collaborators between *t*-5 and *t*-1. Note that grey person represents scientists who did not collaborate with the focal scientist at year *t*. (**c**) depicts the individual scientist's local collaboration network based on conventional bipartite network projections. Solid lines suggest that connected two scientists have collaborated at least once between *t*-5 and *t*-1. (**d**) depicts the higher-order local collaboration network with a simplicial description. Solid lines represent that two scientists have collaborated at least once between *t*-5 and *t*-1. Filled triangles indicate that the connected three scientists have at least one joint publication during this period. The empty triangle means the connected three scientists have not collaborated together, whereas any two of them have a pairwise collaboration

Homomorphism, kernel and image If there is a map $f: M \rightarrow S$, which satisfies that $\forall a, b \in M, f(a * b) = f(a) \cdot f(b) \in S$, then f is a homomorphism from M to S. Here M and S are two nonempty sets; $*$ and $\cdot$ are two operations defined on these two sets, respectively. So the boundary operator ∂_d is a homomorphism from C_{d+1} to C_d. The kernel of a homomorphism $f: M \rightarrow S$ is the set of all elements in M that are mapped to zero. Therefore, Z_d is the kernel of ∂_d. The image of a homomorphism f: $M \rightarrow S$ is the set of all elements in S. As a result, B_d is the image of ∂_{d+1}.

Homology group and Betti numbers The dth homology group is defined as the quotient between Z_d and B_d, denoted as

$$H_d(\gamma) = \frac{Z_d}{B_d} = \frac{\ker(\partial_d)}{\operatorname{im}(\partial_{d+1})}.$$

The elements of $H_d(\gamma)$ refers to the d-cycles that are not induced by a d-boundary, namely the d-dimensional holes of our simplicial complex γ. The rank of $H_d(\gamma)$ is defined as the

dth Betti number of γ, denoted as

$$\beta_d = \operatorname{rank}\bigl(H_d(\gamma)\bigr) = \operatorname{rank}(Z_d) - \operatorname{rank}(B_d),$$

which indicates the number of different d-dimensional holes. In this study, we only focus on the effects of β_0 and β_1. β_0 counts the number of disconnected components, and β_1 counts the number of higher-order loops, capturing the presence of circular relationships or cycles within the network.

To illustrate the concept more, let's consider the local collaboration network shown in Fig. 1d. In this network, there are two disconnected components, one consists of node 3, and the other is formed by the rest nodes. Hence, β_0 is 2. Additionally, we observe two empty triangles. One is formed by nodes 1, 5, and 6, while the other is formed by nodes 1, 4, and 5. Therefore, β_1 is also 2. It is worth noting that in the dyadic view, there is no filled triangle within collaboration networks. If the focal scientist has no coauthors at year t, then β_0 and β_1 are set to zero.

4.4 Variables in regression analysis

In this study, we consider scientific productivity, which refers to the total number of papers published at year t as the dependent variable. For independent and control variables, we utilize the 0th Betti number (β_0) and 1st Betti number (β_1) to quantify the higher-order structural properties of local collaboration networks. It is important to note that β_0 is a continuous variable, while β_1 is transformed into a binary variable as the majority of observed values are zero. We consider several explanatory variables that may affect the performance of individual scientists. Specifically, we consider network size, network density, average tie strength and collaborative strength. Network size refers to the number of collaborators at year t. Network density is defined as the fraction of real links with respect to all possible links in conventional collaboration networks [34]. Average tie strength is the average number of papers coauthored between individual scientist and collaborators from t-5 to t-1 [22]. The collaborative strength is the ratio of collaborative papers among all collaborators to the total number of papers published by all collaborators between t-5 and t-1. Prior studies demonstrated that such network properties may be associated with scientists' academic performance [20, 22, 24, 25, 34]. Moreover, we also consider career age at year t [76]. Finally, given that the scientist's academic performance at year t can be affected by previous records [22], we control for the productivity at the last year in which the scientist has publication records. The details of variables are shown in Table 1.

4.5 Regression models

We use Poisson regressions to quantify the relationship between high-order properties and scientific productivity. The application of a Poisson model in our study is grounded in its suitability for regressions where the dependent variable is counted and follows a Poisson distribution. In our context, productivity is denoted by the number of publications, which inherently assumes non-negative integer values. While the distribution of publication counts exhibits characteristics of a fat-tailed distribution [77], it is important to note that prior research has demonstrated the Poisson estimator's reliability in panel data models. This reliability is maintained even when the actual data distribution does not precisely

Table 1 Variables description

Variable	Description
Productivity	Total number of papers published at year *t*.
β_0	The number of disconnected components in the local collaboration network.
β_1 (binary)	The number of higher-order loops in the local collaboration network. $\beta_1 = 1$ if there exists at least one loop, otherwise 0.
Network size	The number of collaborators at year *t*.
Network density	The fraction of real links with respect to all possible links in conventional collaboration networks.
Average tie strength	The average tie strength between the focal scientist and his/her collaborators between year *t*-5 and year *t*-1.
Collaborative strength	The ratio of collaborative papers among collaborators to the total number of papers published by all collaborators between year *t*-5 and year *t*-1.
Career age	Career age of the focal scientist.
$\text{Productivity}_{\overleftarrow{t}}$	Total number of papers published at the last year in which the scientist has publication records.

conform to the Poisson distribution, as long as the mean specification remains accurate [78]. The regression equations are as follows:

$$\begin{aligned}
&\ln(\text{Productivity}_{i,t}) \\
&\quad = a_0 + a_1(\beta_{i,\Delta t}) + a_3(\text{Network density}_{i,\Delta t}) + a_4\big(\log_2(\text{Average tie strength}_{i,\Delta t} + 1)\big) \\
&\qquad + a_5\big(\log_2(\text{Collaborative strength}_{i,\Delta t} + 1)\big) + a_6\big(\log_2(\text{Career age}_{i,t} + 1)\big) \\
&\qquad + a_7\big(\log_2(\text{Productivity}_{i,\overleftarrow{t}})\big) + \sum_j b_j \sigma_{ji,t} + \mu_i + \tau_t + \varepsilon_{i,t},
\end{aligned}$$

where Δt refers to the period from t-5 to t-1, $\overleftarrow{t}$ indicates the last year in which the scientist has publication records. μ_i represents individual fixed effects, which is a vector of unobserved but fixed confounders depending only on individual i [79]. The rationale for adding individual fixed effects is to control for individuals' unobservable characteristics [80]. τ_t represents year fixed effects, and the rationale for adding year fixed effects is to take into account unobserved variables that evolve over time but are constant across entities [80]. σ_{jit} indicates network size fixed effects, and we categorize the network size into six bins: $[0,6]$, $[7,12]$, $[13,18]$, $[19,24]$, $[25,30]$, and $[31,36]$. The reason why we consider fixed effects instead of controlling for its continuous form is that there is collinearity between β_0 and network size, which may influence the precision of estimations [37]. Note that in the regression model we add quadratic terms of β_0 in order to check whether there is an inverted U-shaped relationship, and we also control for β_0 when exploring the effect of β_1. We take logarithms for variables with fat-tail distributions.

We use scientific fields provided by the MAG data to categorize scientists into different scientific domains. This categorization is based on scientific domains to which more than half of a scientist's papers belong to. The Appendix Table A2 shows the number of scientists, as well as scientist-year observations across 19 scientific domains.

5 Results

5.1 Descriptive statistics

Table 2 shows the descriptive statistics of the variables used in our analysis. To assess the presence of multicollinearity, we calculate the variance inflation factor (VIF), and find that

Table 2 Descriptive statistics of different variables

Variable	Mean	Std. dev	Min	Max
Productivity	2.089	1.833	1	312
β_0	3.397	3.136	0	36
β_1 (binary)	0.083	0.438	0	1
Network size	5.983	5.690	0	36
Network density	0.231	0.303	0	1
Average tie strength	0.929	1.486	0	313
Collaborative strength	0.164	0.218	0	1
Career age	8.304	8.371	0	51
Productivity$_{\overleftarrow{t}}$	2.112	1.843	1	312

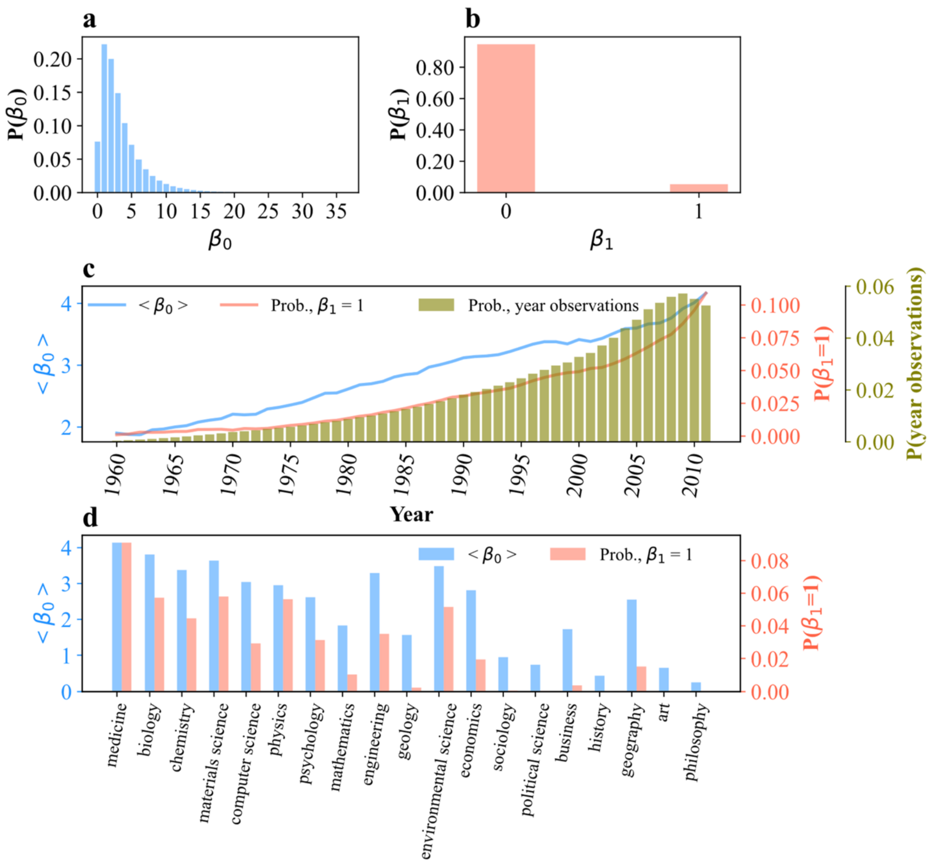

Figure 2 The distribution, evolution and disciplinary variations of β_0 and β_1.(**a-b**) The distribution of β_0 and β_1. (**c**) $\langle\beta_0\rangle$, and $\mathrm{P}(\beta_1 = 1)$ as a function of time. (**d**) $\langle\beta_0\rangle$ and $\mathrm{P}(\beta_1 = 1)$ across different scientific domains

the VIFs for β_0, β_1, network density, average tie strength, collaborative strength, career age are 1.23, 1.05, 1.71, 1.56, 1.42 and 1.04, respectively. These values suggest that there is no strong multicollinearity among these variables.

Figures 2a and 2b display the distribution of β_0 and β_1, respectively. We find that over 90% of local collaboration networks exhibit less than eight disconnected components. Moreover, the occurrence of higher-order loops in these networks is relatively rare. Specifically, local collaboration networks that contain at least one higher-order loop account for around 5% of the total networks. Figure 2c illustrates the temporal evolution of β_0 and β_1. We observe that the average number of components in local collaboration networks steadily increased. Additionally, there is a significant rise in the proportion of local collaboration networks that exhibit at least one higher-order loop. Notably, approximately 11% of the local collaboration networks display the presence of higher-order loop structures at year 2011, highlighting the growing prevalence of higher-order structures within local collaboration networks. Figure 2d illustrates the average value of β_0 and probability of $\beta_1 = 1$ across different scientific domains, revealing distinct disciplinary variations. Generally, scientists in medicine, biology, material science and environmental science are more likely to have local collaboration networks with disconnected components and higher-order loops. Besides, additional descriptive analyses show that scientists with higher-order loops are typically more senior, with higher productivity and citation impact than those without higher-order loops.

5.2 Scientific productivity

Figures 3a and 3b show the relationship between β_0, β_1, and the number of papers published at year t, respectively. We find several noteworthy patterns. First, scientific productivity shows an initial increase with each additional component until β_0 reaches 22, beyond which it starts to decline, suggesting that having a moderate number of disconnected components in the collaboration network is associated with high productivity. Second, scientists whose local collaboration networks contain at least one higher-order loop tend to publish more papers compared to those without loops, indicating the positive impact of higher-order loops on scientific productivity (2.00 versus 3.65, Two-sided Welch's t-test, p-value < 0.001).

To eliminate the effects of potential explanatory factors, we perform fixed effects Poisson regressions (see Table 3). The results confirm an inverted U-shaped relationship between β_0 and scientific productivity, with a turning point estimated at 15 (Table 3 model 5). Figure 3c visualizes the estimated scientific productivity as a function of β_0 based on the regression, holding other variables at the sample means. And it demonstrates that the productivity increases by 645.0% when β_0 rises from 0 to 15, but decreases by 94.7% when β_0 increases from 15 to 36. We find that β_1 is positively associated with scientific productivity (Table 4 model 5). Adjusting for all factors, having at least a higher-order loop

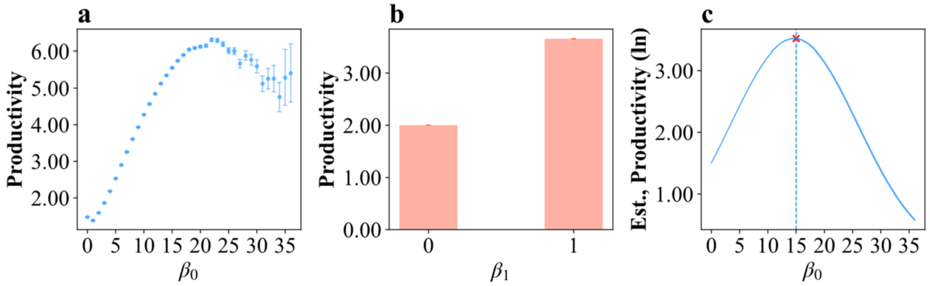

Figure 3 The relationship between β_0, β_1 and scientific productivity. (**a**) The scatter plot between β_0 and scientific productivity. The point represents mean value and the error bar represents standard error of the mean. (**b**) The bar chart between β_1 and scientific productivity. The bar represents mean value and the error bar represents standard error of the mean. (**c**) The estimated association between β_0 and scientific productivity based on Table 3 model (5) using the "margins" function of STATA. The red cross mark represents the turning point. The error bar represents the 95% confidence interval

Table 3 Fixed-effects Poisson regressions regarding the association between β_0 and scientific productivity

Productivity(ln)	(1)	(2)	(3)	(4)	(5)
β_0	0.094***	0.164***	0.157***	0.139***	0.116***
	(0.00006)	(0.00013)	(0.00014)	(0.00015)	(0.00017)
β_0^2		−0.004***	−0.004***	−0.003***	−0.004***
		(0.00001)	(0.00001)	(0.00001)	(0.00001)
Network density			−0.255***	−0.198***	−0.026***
			(0.00059)	(0.00059)	(0.00060)
Average tie strength (log)			0.262***	0.101***	0.071***
			(0.00023)	(0.00033)	(0.00025)
Collaborative strength				0.128***	0.240***
				(0.00068)	(0.00074)
Career age (log)				−0.003***	−0.026***
				(0.00014)	(0.00031)
$\text{Productivity}_{\overleftarrow{t}}$(log)				0.232***	0.020***
				(0.00035)	(0.00018)
Year fixed effects	N	N	N	N	Y
Individual fixed effects	N	N	N	N	Y
Network size effects	N	N	N	N	Y
Log pseudolikelihood	−44,919,517	−44,482,073	−43,456,407	−37,342,694	−34,025,920
R^2	0.0848	0.0937	0.1146	0.1428	0.2189
Observations	27,786,774	27,786,774	27,786,774	24,000,967	24,000,967
Individuals	3,785,807	3,785,807	3,785,807	3,785,807	3,785,807

Notes: Robust standard errors in parentheses; *** p-value < 0.001, ** p-value < 0.01, * p-value < 0.05; 3,785,807 observations dropped when we control for $\text{Productivity}_{\overleftarrow{t}}$.

Table 4 Fixed-effects Poisson regressions regarding the association between β_1 and scientific productivity

Productivity(ln)	(1)	(2)	(3)	(4)	(5)
β_1	0.600***	0.360***	0.308***	0.238***	0.111***
	(0.00064)	(0.00062)	(0.00055)	(0.00052)	(0.00046)
β_0		0.090***	0.088***	0.077***	0.044***
		(0.00006)	(0.00006)	(0.00006)	(0.00007)
Network density			−0.401***	−0.332***	−0.168***
			(0.00059)	(0.00058)	(0.00055)
Average tie strength (log)			0.262***	0.103***	0.065***
			(0.00023)	(0.00033)	(0.00025)
Collaborative strength			0.166***	0.207***	0.284***
			(0.00068)	(0.00068)	(0.00074)
Career age (log)				−0.002***	−0.023***
				(0.00014)	(0.00031)
$\text{Productivity}_{\overleftarrow{t}}$(log)				0.233***	0.022***
				(0.00034)	(0.00019)
Year fixed effects	N	N	N	N	Y
Individual fixed effects	N	N	N	N	Y
Network size effects	N	N	N	N	Y
Log pseudolikelihood	−48,324,649	−44,638,277	−43,575,288	−37,461,268	−34,164,091
R^2	0.0154	0.0905	0.1122	0.1401	0.2158
Observations	27,786,774	27,786,774	27,786,774	24,000,967	24,000,967
Individuals	3,785,807	3,785,807	3,785,807	3,785,807	3,785,807

Notes: Robust standard errors in parentheses; *** p-value < 0.001, ** p-value < 0.01, * p-value < 0.05; 3,785,807 observations dropped when we control for $\text{Productivity}_{\overleftarrow{t}}$.

in local collaboration networks is associated with an increase of 11.7%, on average, more publications for individual scientists. Overall, these observations highlight the critical role of higher-order structures of local collaboration networks.

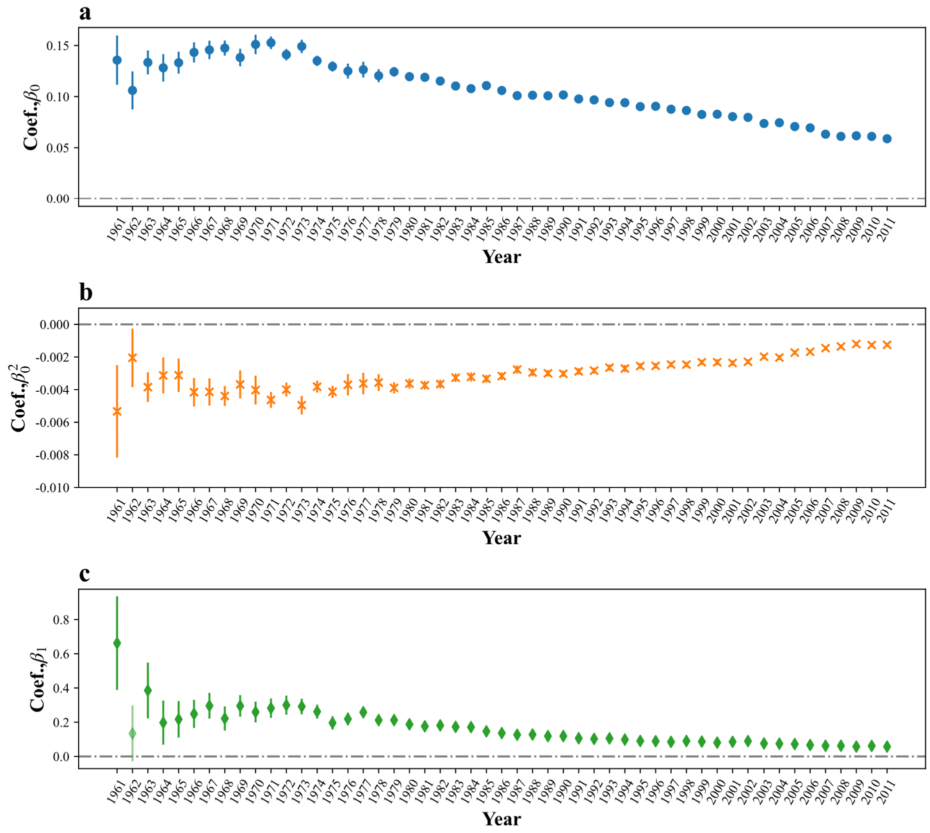

Figure 4 The regression coefficients of β_0, β_0^2 and β_1 across years. The point represents the coefficient and the error bar represents the 95% confidence interval. Darker coloring represents significant coefficients (p-value < 0.05), whereas lighter coloring represents insignificant coefficients (p-value > 0.05)

Moreover, we run the same fixed-effects Poisson regression separately for each scientific field. Table 5 indicates that the findings are strongly generalizable across various scientific domains. The 19 scientific domains are sorted according to the number of scientists in descending order. Specifically, we find that all scientific domains have significantly positive coefficients of β_0 and significantly negative coefficients of β_0^2, indicating that there is an inverted U-shaped relationship between β_0 and scientific productivity. Moreover, we observe that β_1 is significantly and positively associated with scientific productivity for scientists in 18 out of 19 fields (except for art). For example, forming at least one higher-order loop is associated with an increase of 8.9% more papers in medicine, 9.1% in biology, and 12.2% in chemistry.

5.3 Robustness checks

We conduct a series of robustness tests to strengthen the validity of our findings. Initially, we run Poisson regressions separately for each year. Since each scientist occurs exactly once in a given year, we thus eliminate the effect of duplicated scientist has in the aggregated regression. In this analysis, we consider the same control variables as the main regression, while we do not add individual and year fixed effects, as each scientist only has one row in the dataset. We observe that the inverted U-shaped with β_0 and the positive effects of β_1 on productivity remain statistically significant across years (see Fig. 4).

Table 5 Fixed-effects Poisson regressions on scientific productivity across fields

Field	Medicine	Biology	Chemistry	Materials science	Computer science	Physics	Psychology	Mathematics	Engineering	Geology
β_0	0.078***	0.108***	0.115***	0.097***	0.141***	0.150***	0.160***	0.223***	0.146***	0.131***
	(0.00032)	(0.00040)	(0.00054)	(0.00051)	(0.00081)	(0.00108)	(0.00105)	(0.00166)	(0.00161)	(0.00109)
β_0^2	−0.003***	−0.004***	−0.004***	−0.003***	−0.005***	−0.006***	−0.006***	−0.008***	−0.005***	−0.005***
	(0.00002)	(0.00002)	(0.00003)	(0.00003)	(0.00005)	(0.00008)	(0.00008)	(0.00016)	(0.00010)	(0.00007)
β_1	0.085***	0.087***	0.115***	0.097***	0.148***	0.127***	0.169***	0.279***	0.177***	0.129***
	(0.00078)	(0.00100)	(0.00119)	(0.00150)	(0.00266)	(0.00260)	(0.00333)	(0.00602)	(0.00603)	(0.00423)

Field	Environmental science	Economics	Sociology	Political science	Business	History	Geography	Art	Philosophy
β_0	0.108***	0.244***	0.187***	0.165***	0.237***	0.136***	0.131***	0.132***	0.162***
	(0.00264)	(0.00209)	(0.00227)	(0.00296)	(0.00397)	(0.00358)	(0.00297)	(0.00612)	(0.00525)
β_0^2	−0.004***	−0.009***	−0.009***	−0.005***	−0.009***	−0.004***	−0.005***	−0.004***	−0.006***
	(0.00017)	(0.00025)	(0.00030)	(0.00039)	(0.00045)	(0.00033)	(0.00020)	(0.00046)	(0.00049)
β_1	0.081***	0.384***	0.529***	0.410***	0.295***	0.610***	0.195***	−0.296	0.900***
	(0.00553)	(0.02750)	(0.07723)	(0.08394)	(0.02797)	(0.12555)	(0.01857)	(0.31150)	(0.14972)

Notes: The coefficients are derived from the same fixed-effects Poisson regressions shown in Table (3-4) model (5); robust standard errors in parentheses; *** p-value < 0.001, ** p-value < 0.01, * p-value < 0.05.

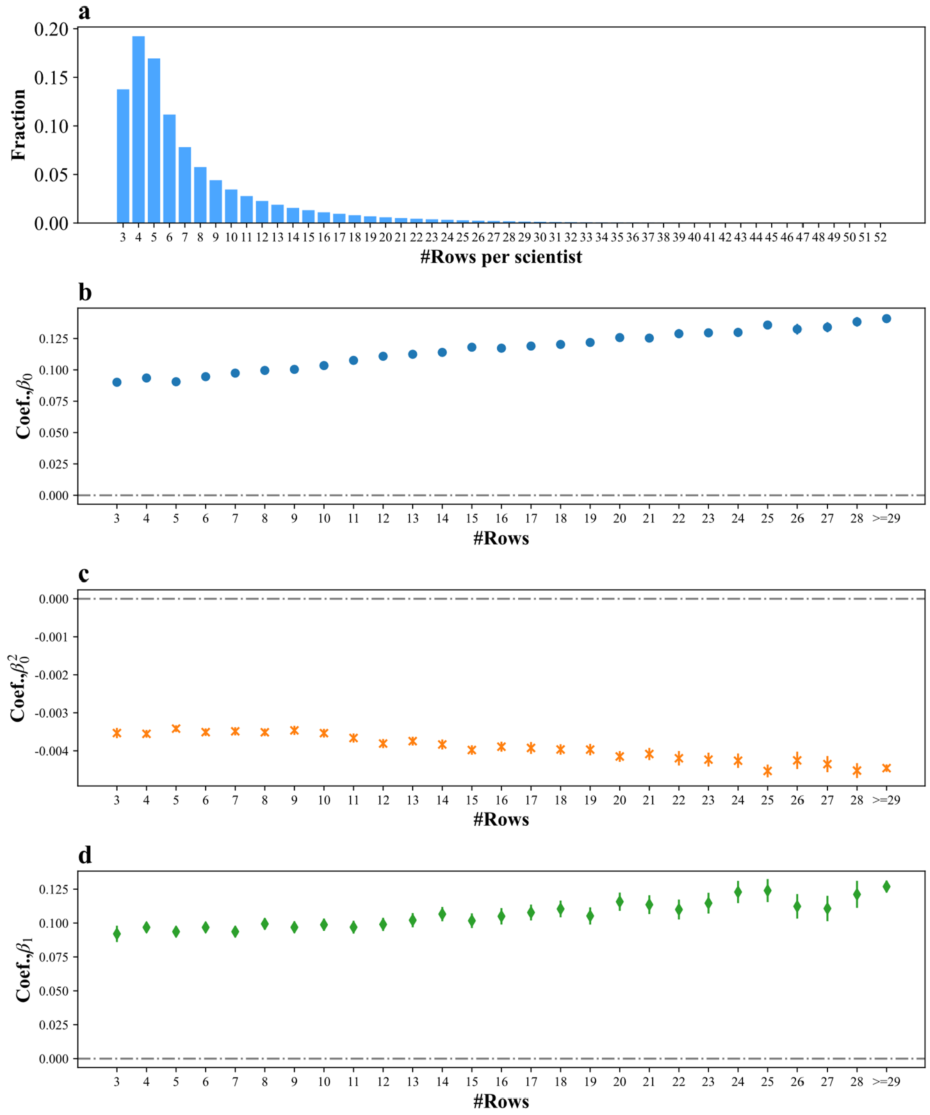

Figure 5 (**a**) The distribution of the number of rows per scientist. (**b-d)** The coefficients β_0, β_0^2 and β_1 in Poisson regression models across rows. The point represents the coefficient and the error bar represents the 95% confidence interval. The grey line indicates y = 0. Darker coloring represents significant coefficients (p-value < 0.05), whereas lighter coloring represents insignificant coefficients (p-value > 0.05)

Besides, we separate scientists into subgroups according to their number of “rows” in the data, and run the Poisson regressions separately for each group. The distribution of the number of rows is depicted in Fig. 5a, and we find that most scientists show less than 10 years of observations. Figure 5b-d depicts the coefficients of β_0, β_0^2 and β_1 for different subgroups. We observe that the inverted U-shaped associations induced by β_0 and the positive effects of β_1 on productivity remain statistically significant for every subgroup, indicating that our results are not affected by high-prolific scientists.

In addition, we separate scientists according to their citation impact (i.e., average citations within 10 years after publication, i.e., c_{10}), i.e., less-impact scientists whose average c_{10} are in the bottom 25% (949,048 scientists and 5,791,795 observations), median-impact scientists whose average c_{10} are between 37.5% and 62.5% (948,860 scientists and 7,547,067

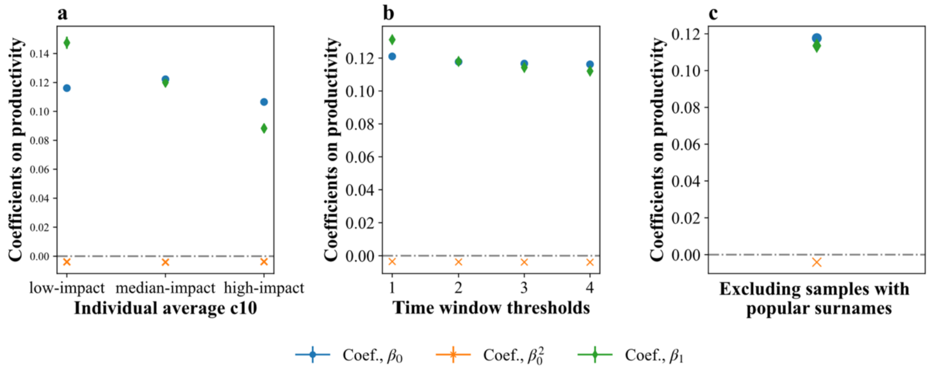

Figure 6 (**a**) The coefficients β_0, β_0^2 and β_1 in Poisson regression models across low-impact, median, and high-impact scientists. (**b**) The coefficients of β_0, β_0^2 and β_1 in Poisson regression models when adopting different time window thresholds. (**c**) The coefficients of β_0, β_0^2 and β_1 in Poisson regression models when excluding samples with popular surnames. The point represents the coefficient and the error bar represents the 95% confidence interval. The grey line indicates y = 0. Darker coloring represents significant coefficients (p-value < 0.05), whereas lighter coloring represents insignificant coefficients (p-value > 0.05)

observations), as well as high-impact scientists whose average c_{10} are in the top 25 percent (946,512 scientists and 7,165,398 observations). We run Poisson regressions for each group separately. Figure 6a depicts the coefficients of β_0, β_0^2 and β_1 in Poisson regression models for each group. We again observe that the inverted U-shaped associations induced by β_0 and the positive effects of β_1 on productivity remain statistically significant. This finding suggests that the main results hold for scientists with different citation impact.

Furthermore, we employ various thresholds to construct local collaboration networks, from 1 to 4 years. Through these iterations, we perform the same regression analyses as in our primary investigation. Notably, the inverted U-shaped associations influenced by β_0 and the positive effects of β_1 persisted as statistically significant (see Fig. 6b).

To address concerns related to the accuracy of disambiguation methods for common names, we compile a list of the 1000 most popular surnames worldwide, which encompass commonly occurring surnames from both Asian and Western regions [accessed from https://forebears.io/earth/surnames]. We repeat the analyses and find the primary findings in our study still hold (see Fig. 6c). Moreover, we repeat our analysis by employing the conventional Ordinary Least Squares (OLS) regression model. In this model, the dependent variable is the logarithm of productivity. It is noteworthy that the outcomes of these analyses aligned with the results of our primary Poisson regression approach, providing further evidence of the robustness of our findings.

6 Conclusions

In an era where scientific knowledge creation is dominated by collaborative teams, it is of paramount importance to delve into the higher-order structures inherent in scientific collaboration networks. The conventional approach, which primarily adopts a dyadic perspective to construct local collaboration networks, may inadvertently overlook invaluable information for group interactions. Leveraging a vast dataset encompassing over 56 million research articles from 1960 to 2011 from the Microsoft Academic Graph, our objective is to explore the intricate link between the higher-order structural features characterizing local collaboration networks and their impact on scientific productivity. Furthermore, we endeavor to ascertain the generalizability of these findings across a diverse set of scientific domains. Throughout our analysis, a noteworthy trend becomes apparent – both the number of disconnected components and the prevalence of higher-order holes exhibit a consistent upward trajectory over time. The fraction of local networks featuring higher-order holes reached 11% in 2011. This surge may be attributed to the remarkable expansion of the scientific community during this period. While higher-order holes are indeed evident in various domains, with domains such as medicine and biology sharing common features, the dominance of triatic closure remains a prevailing characteristic within scientific collaboration networks.

Furthermore, our investigation reveals an intriguing inverted U-shaped association between the number of disconnected components in local collaboration networks and scientific productivity. These results partly speak to the strength of weak tie theory [81], which suggests that individuals spanning over structural holes in social networks can gain significant advantages in accessing new opportunities, fostering innovation [82], and enhancing their overall performance [83]. Previous research, largely rooted in macroscopic collaboration networks, has consistently demonstrated the advantages reaped by scientists who span structural holes. These benefits include paper publication, citation counts, and a higher likelihood of contributing novel research [20, 25, 60]. However, such studies have rarely ventured into the intricate realm of scientists' local networks. Structural holes [84, 85], which foster diversity within local collaboration networks, are primed to play a pivotal role [86]. One would expect significant advantages upon scientists in the realms of productivity. It is plausible that structural diversity acts as a catalyst for resource-sharing and the seamless transmission of knowledge, empowering scientists to harness a spectrum of expertise, diverse ideas, and even the valuable lessons extracted from failure across a heterogeneous pool of collaborators [87–91]. These diverse local collaboration structures equip scientists to acquire a wide array of skills. Ultimately, this dynamic bolsters their productivity. This interpretation aligns with prior findings that suggest novel and multidisciplinary research flourishes within newly-formed teams [38]. This research reinforces this perspective by illuminating a positive correlation between the number of disconnected components within local collaboration networks and scientific productivity – up to a certain threshold. These empirical results effectively substantiate the tenets of structural holes and the significance of weak ties.

This study reveals that as the number of disconnected components reaches a certain threshold, a negative correlation emerges with regard to productivity. This intriguing discovery propels us to explore the potential underlying forces at play. In the realm of scientific collaborations, where the advantages of structural holes and disconnected team members are evident, effective communication and coordination between individuals remain critical [92, 93]. A key facilitator in this regard is familiarity, which results in positive outcomes. Earlier research spotlighted the benefits of strong ties between scientists, often referred to as "super-ties," underscoring their substantial contributions to productivity and citations [94]. Furthermore, the diverse structures present within local collaboration networks can have the unintended consequence of slowing down the assimilation of ideas, leading to lower consensus and, in some cases, potential conflicts [32, 95, 96]. For example, international collaborations tend to produce less novel papers [32], and remote collaborations show a negative association with disruptive research [39]. Similarly, Liu et al. found an inverted-U shaped relationship between team freshness and citations using paper-level data [34].

This study makes a pivotal observation: the presence of higher-order loops within local collaboration networks is positively correlated with productivity in scientific careers. These higher-order loops shed light on the dynamic interplay among multiple agents that goes beyond the typical dyadic interactions. For instance, the phenomena of complex contagion, where an influence requires the involvement of more than two individuals, may exhibit unique characteristics. As highlighted by Iacopini et al. [97], "*the simplicial model of contagion is able to capture the basic mechanisms and effects of higher-order interactions in social contagion processes*." In scientific collaboration, researchers engage in discussions, knowledge diffusion, and the adoption of innovative ideas. Describing these intricate interactions through the lens of higher-order networks provides invaluable insights. This leads to intriguing questions about how resources and knowledge are transmitted within these higher-order loops, as well as the underlying forces driving the positive correlation between higher-order loops and scientific performance. As we conclude, these findings not only provide answers but also raise stimulating questions, paving the way for promising directions in future research within this domain.

In conclusion, these results remain consistent across a spectrum of scientific domains, highlighting its generalizability. This work contributes significantly to the understanding of higher-order collaboration networks by delving into the roles of higher-order holes. Furthermore, it advances our comprehension of how network structures can influence the scientific performance of researchers. Of paramount significance is our discovery of an intriguing inverted U-shaped relationship driven by the number of disconnected components within local collaboration networks. This insight offers a nuanced understanding of the interplay between structural complexity and scientific output. Additionally, our work transcends disciplinary boundaries by encompassing scientists from diverse fields. The insights gleaned from this study hold the potential to benefit a wide array of research areas, extending beyond specific scientific domains. Our findings have important policy implications for nurturing scientific personnel and accelerating innovative breakthroughs. Scientists need to carefully consider the structure of his/her collaboration network. It is crucial for scientists to strive for a well-balanced and properly disconnected or loosely connected local co-authorship network, which is crucial to high productivity.

This study contains several limitations. First, we use publication data to describe collaboration patterns, while collaborative work does not always result in written outputs, and the presence of ghost authors, where individuals contribute to research but are not acknowledged as authors, cannot be ruled out [34, 98, 99]. This may introduce possible biases in our findings and limit the generalizability of our results to all forms of scientific collaboration. Secondly, we gauge scientific productivity using the number of publications. However, the number of publications alone may not be a perfect indicator that captures scientists' scientific performance [100]. Prior research proposed various indicators to measure the quality of academic outputs, such as citations [101], novelty indicators [102, 103] aligning with Schumpeter's innovation economics that "innovation combines components in a new way" [104], disruption index [59, 105], as well as other metrics capturing the interdisciplinarity [106]. It is thus interesting to understand the effect of higher-order structures on scientists' academic performance taking into account the quality of works. Thirdly, it is worth noting that despite we control for possible confounding variables, our study is still of a correlational nature and does not establish causal relationships. Despite these limitations, our study offers valuable insights into the relationship between higher-order structural properties and scientific outcomes, contributing to a growing body of literature in the field of science of science and data science.

Further research is needed to conduct systematic investigations to unravel the underlying mechanisms driving these associations between higher-order properties and productivity. What are the factors that prompt scientists with higher-order structures to publish significantly more papers than their counterparts without higher-order structures? In an era of big science, there are a tremendous number of publications and citations each year, future work could examine the evolution of the effect of high-order structures on scientific achievements, which may untangle the effect of the growth of science and higher-order structures. Finally, future research could go beyond scientific productivity and explore how higher-order structures affect knowledge recombination, originality and interdisciplinarity.

Appendix

Table A1 The data frame of "scientist-year observations"

Author	Year	Structural properties of local collaboration network			Academic performance	Other control variables	
		β_0	β_1 (binary)	…	Productivity	Network size	…
1	t_{11}	a_{11}	b_{11}	…	i_{11}	x_{11}	…
1	t_{12}	a_{12}	b_{12}	…	i_{12}	x_{12}	…
…	…	…	…	…	…	…	…
1	t_{1n_1}	a_{1n_1}	b_{1n_1}	…	i_{1n_1}	x_{1n_1}	…
…	…	…	…	…	…	…	…
N	t_{N1}	a_{N1}	b_{N1}	…	i_{N1}	x_{N1}	…
N	t_{N2}	a_{N2}	b_{N2}	…	i_{N2}	x_{N2}	…
…	…	…	…	…	…	…	…
N	t_{Nn_n}	a_{Nn_n}	b_{Nn_n}	…	i_{Nn_n}	x_{Nn_n}	…

Table A2 Number of scientists across 19 fields

Level-0 fields	Individuals	Scientist-year observations	Level-0 fields	Individuals	Scientist-year observations
Medicine	925,630	6,311,153	Environmental science	33,163	230,966
Biology	617,975	4,628,649	Economics	29,908	265,206
Chemistry	545,615	4,107,767	Sociology	23,342	169,258
Materials science	317,503	2,180,448	Political science	16,914	130,481
Computer science	180,155	1,227,045	Business	14,379	95,820
Physics	132,634	1,010,505	History	13,759	104,225
Psychology	111,025	901,590	Geography	9404	59,661
Mathematics	86,601	767,830	Art	9012	59,621
Engineering	65,940	410,746	Philosophy	8580	66,183
Geology	51,930	445,478	Total	3,193,469	23,172,632

Acknowledgements

We would like to thank the referees and the associate editor for their insightful comments and suggestions.

Funding

This work was supported by the National Natural Science Foundation of China under Grant nos. 72004177, L2324122, L1924078, China Association of Higher Education under Grant no. 23YZF0204, and the Fundamental Research Funds for the Central Universities.

Abbreviations

MAG, Microsoft Academic Graph; VIF, Variance Inflation Factor; OLS, Ordinary least squares; ORCID, Open Researcher and Contributor ID.

Data availability

The Microsoft Academic Graph data can be downloaded via https://zenodo.org/record/2628216#.Y-7RR_5Bz-g.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author contributions

YW conceived and designed the study. WY wrote the code for data analysis and visualization, and conducted the experiments. Both authors analyzed and discussed the data and wrote the manuscript. Both authors read and approved the final manuscript.

Received: 29 May 2023 Accepted: 15 February 2024 Published online: 28 February 2024

References

1. Fortunato S, Bergstrom C, Borner K, Evans J, Helbing D, Milojevic S, Petersen A, Radicchi F, Sinatra R, Uzzi B, Vespignani A, Waltman L, Wang D, Barabasi A (2018) Science of science. Science 359(1):6379
2. Zeng A, Shen Z, Zhou J, Wu J, Fan Y, Wang Y, Stanley H (2017) The science of science: from the perspective of complex systems. Phys Rep 714–715:1–73
3. Shrum W, Genuth J, Chompalov I (2007) Structures of scientific collaboration. MIT Press, Cambridge
4. Katz J, Martin B (1997) What is research collaboration? Res Policy 26(1):1–18
5. de Solla Price D (1963) Little science, big science. Columbia University Press, New York
6. Jones B (2009) The burden of knowledge and the "death of the renaissance man": is innovation getting harder? Rev Econ Stud 76(1):283–317
7. Newman M (2001) The structure of scientific collaboration networks. Proc Natl Acad Sci USA 98(2):404–409
8. Wuchty S, Jones B, Uzzi B (2007) The increasing dominance of teams in production of knowledge. Science 316(5827):1036–1039
9. Jones B, Wuchty S, Uzzi B (2008) Multi-university research teams: shifting impact, geography, and stratification in science. Science 322(5905):1259–1262
10. Adams J (2013) The fourth age of research. Nature 497(7451):557–560
11. Newman M (2001) Scientific collaboration networks. I. Network construction and fundamental results. Phys Rev E 64(1):016131
12. Newman M (2001) Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Phys Rev E 64(1):016132
13. Newman M (2002) Assortative mixing in networks. Phys Rev Lett 89(20):208701
14. Ke Q, Ahn Y (2014) Tie strength distribution in scientific collaboration networks. Phys Rev E 90(3):032804
15. Pan R, Saramaki J (2012) The strength of strong ties in scientific collaboration networks. Europhys Lett 97(1):18007
16. Martin T, Ball B, Karrer B, Newman M (2013) Coauthorship and citation patterns in the physical review. Phys Rev E 88(1):012814
17. Ding Y (2011) Scientific collaboration and endorsement: network analysis of coauthorship and citation networks. J Informetr 5(1):187–203
18. Abbasi A, Hossain L, Uddin S, Rasmussen K (2011) Evolutionary dynamics of scientific collaboration networks: multi-levels and cross-time analysis. Scientometrics 89(2):687–710
19. Menichetti G, Remondini D, Panzarasa P, Mondragon R, Bianconi G (2014) Weighted multiplex networks. PLoS ONE 9(6):e97857
20. Tahmooresnejad L, Beaudry C, Mirnezami S (2021) The study of network effects on research impact in Africa. Sci Public Policy 48(4):462–473
21. Tahmooresnejad L, Beaudry C (2018) The importance of collaborative networks in Canadian scientific research. Ind Innov 25(10):990–1029
22. Wang J (2016) Knowledge creation in collaboration networks: effects of tie configuration. Res Policy 45(1):68–80
23. Gonzalez-Brambila C, Veloso F, Krackhardt D (2013) The impact of network embeddedness on research output. Res Policy 42(9):1555–1567
24. Li E, Liao C, Yen H (2013) Co-authorship networks and research impact: a social capital perspective. Res Policy 42(9):1515–1530
25. Abbasi A, Altmann J, Hossain L (2011) Identifying the effects of co-authorship networks on the performance of scholars: a correlation and regression analysis of performance measures and social network analysis measures. J Informetr 5(4):594–607

26. Guan J, Pang L (2018) Bidirectional relationship between network position and knowledge creation in scientometrics. Scientometrics 115(1):201–222
27. Guan J, Zhang J, Yan Y (2015) The impact of multilevel networks on innovation. Res Policy 44(3):545–559
28. Fronczak A, Mrowinski M, Fronczak P (2022) Scientific success from the perspective of the strength of weak ties. Sci Rep 12(1):5074
29. AlShebli B, Rahwan T, Woon W (2018) The preeminence of ethnic diversity in scientific collaboration. Nat Commun 9(1):5163
30. Dong Y, Ma H, Tang J, Wang K (2018) Collaboration diversity and scientific impact. Preprint. arXiv:1806.03694
31. Freeman R, Huang W (2014) Strength in diversity. Nature 513(7518):305
32. Wagner C, Whetsell T, Mukherjee S (2019) International research collaboration: novelty, conventionality, and atypicality in knowledge recombination. Res Policy 48(5):1260–1270
33. Chen W, Yan Y (2023) New components and combinations: the perspective of the internal collaboration networks of scientific teams. J Informetr 17(2):101407
34. Liu M, Jaiswal A, Bu Y, Min C, Yang S, Liu Z, Acuna D, Ding Y (2022) Team formation and team impact: the balance between team freshness and repeat collaboration. J Informetr 16(4):101337
35. Petersen A (2015) Quantifying the impact of weak, strong, and super ties in scientific careers. Proc Natl Acad Sci USA 112(34):E4671–E4680
36. Xu F, Wu L, Evans J (2022) Flat teams drive scientific innovation. Proc Natl Acad Sci USA 119(23):e2200927119
37. Yang Y, Tian T, Woodruff T, Jones B, Uzzi B (2022) Gender-diverse teams produce more novel and higher-impact scientific ideas. Proc Natl Acad Sci USA 119(36):e2200841119
38. Zeng A, Fan Y, Di Z, Wang Y, Havlin S (2021) Fresh teams are associated with original and multidisciplinary research. Nat Hum Behav 5(10):1314–1322
39. Lin Y, Frey CB, Wu L (2023) Remote collaboration fuses fewer breakthrough ideas. Nature 623(7989):987–991
40. Horak D, Jost J (2013) Spectra of combinatorial Laplace operators on simplicial complexes. Adv Math 244(2):303–336
41. Jiang B, Omer I (2007) Spatial topology and its structural analysis based on the concept of simplicial complex. Trans GIS 11(6):943–960
42. Cooper J, Dutle A (2012) Spectra of uniform hypergraphs. Linear Algebra Appl 436(9):3268–3292
43. Ghoshal G, Zlatic V, Caldarelli G, Newman M (2009) Random hypergraphs and their applications. Phys Rev E 79(6):066118
44. Gao T, Li F (2018) Studying the utility preservation in social network anonymization via persistent homology. Comput Secur 77:49–64
45. Saggar M, Sporns O, Gonzalez-Castillo J, Bandettini P, Carlsson G, Glover G, Reiss A (2018) Towards a new approach to reveal dynamical organization of the brain using topological data analysis. Nat Commun 9(1):1399
46. Santos F, Raposo E, Coutinho M, Copelli M, Stam C, Douw L (2019) Topological phase transitions in functional brain networks. Phys Rev E 100(3):032414
47. Mariani M, Ren Z, Bascompte J, Tessone C (2019) Nestedness in complex networks: observation, emergence, and implications. Phys Rep 813:1–90
48. Valverde S, Vidiella B, Montanez R, Fraile A, Sacristan S, Garcia-Arenal F (2020) Coexistence of nestedness and modularity in host-pathogen infection networks. Nat Ecol Evol 4(4):568–577
49. Sanchez A (2019) Defining higher-order interactions in synthetic ecology: lessons from physics and quantitative genetics. Cell Syst 9(6):519–520
50. Guerrero R, Scarpino S, Rodrigues J, Hartl D, Ogbunugafor C (2019) Proteostasis environment shapes higher-order epistasis operating on antibiotic resistance. Genetics 212(2):565–575
51. Carstens C, Horadam K (2013) Persistent homology of collaboration networks. Math Probl Eng 2013(1):815035
52. Gebhart T, Funk R (2020) The emergence of higher-order structure in scientific and technological knowledge networks. Preprint. arXiv:2009.13620
53. Juul J, Benson A, Kleinberg J (2022) Hypergraph patterns and collaboration structure. Preprint. arXiv:2210.02163
54. Patania A, Petri G, Vaccarino F (2017) The shape of collaborations. EPJ Data Sci 6:18
55. Salnikov V, Cassese D, Lambiotte R (2018) Co-occurrence simplicial complexes in mathematics: identifying the holes of knowledge. Appl Netw Sci 31(1):37
56. Reimann M, Nolte M, Scolamiero M, Turner K, Perin R, Chindemi G, Dlotko P, Levi R, Hess K, Markram H (2017) Cliques of neurons bound into cavities provide a missing link between structure and function. Front Comput Neurosci 11:48
57. Sizemore A, Giusti C, Kahn A, Vettel J, Betzel R, Bassett D (2018) Cliques and cavities in the human connectome. J Comput Neurosci 44(1):115–145
58. Milojevic S (2014) Principles of scientific research team formation and evolution. Proc Natl Acad Sci USA 111(11):3984–3989
59. Wu L, Wang D, Evans J (2019) Large teams develop and small teams disrupt science and technology. Nature 566(7744):378
60. Wang Y, Li N, Zhang B, Huang Q, Wu J, Wang Y (2023) The effect of structural holes on producing novel and disruptive research in physics. Scientometrics 128(3):1801–1823
61. Wang C, Rodan S, Fruin M, Xu XY (2014) Knowledge networks, collaboration networks, and exploratory innovation. Acad Manag J 57(2):484–514
62. Liu F, Holme P, Chiesa M, AlShebli B, Rahwan T (2023) Gender inequality and self-publication are common among academic editors. Nat Hum Behav 7(3):353–364
63. Liu F, Rahwan T, AlShebli B (2023) Non-white scientists appear on fewer editorial boards, spend more time under review, and receive fewer citations. Proc Natl Acad Sci USA 120(13):e2215324120
64. AlShebli B, Makovi K, Rahwan T (2020) The association between early career informal mentorship in academic collaborations and junior author performance. Nat Commun 11(1):6446
65. Sun Y, Livan G, Ma A, Latora V (2021) Interdisciplinary researchers attain better long-term funding performance. Commun Phys 4(1):263
66. Xie Y, Lin XH, Li J, He Q, Huang JM (2023) Caught in the crossfire: fears of Chinese-American scientists. Proc Natl Acad Sci USA 120(27):e2216248120

67. Huang J, Gates A, Sinatra R, Barabasi A (2020) Historical comparison of gender inequality in scientific careers across countries and disciplines. Proc Natl Acad Sci USA 117(9):4609–4616
68. Zeng A, Fan Y, Di ZG, Wang YG, Havlin S (2022) Impactful scientists have higher tendency to involve collaborators in new topics. Proc Natl Acad Sci USA 119(33):e2207436119
69. Wang K, Shen Z, Huang C, Wu C-H, Dong Y, Kanakia A (2020) Microsoft academic graph: when experts are not enough. Quant Sci Stud 1(1):396–413
70. Zhang L, Lu W, Yang J (2021) LAGOS-AND: a large gold standard dataset for scholarly author name disambiguation. Preprint. arXiv:2104.01821
71. Battiston F, Cencetti G, Iacopini I, Latora V, Lucas M, Patania A, Young JG, Petri G (2020) Networks beyond pairwise interactions: structure and dynamics. Phys Rep 874(1):1–92
72. Bianconi G (2021) Higher-order networks: an introduction to simplicial complexes. Cambridge University Press, Cambridge
73. Carlsson G (2009) Topology and data. Bull Am Math Soc 46(2):255–308
74. Horak D, Maletic S, Rajkovic M (2009) Persistent homology of complex networks. J Stat Mech Theory Exp 2009(3):P03034
75. Otter N, Porter M, Tillmann U, Grindrod P, Harrington H (2017) A roadmap for the computation of persistent homology. EPJ Data Sci 6:17
76. Blau DM, Weinberg BA (2017) Why the US science and engineering workforce is aging rapidly. Proc Natl Acad Sci USA 114(15):3879–3884
77. Fronczak P, Fronczak A, Holyst JA (2007) Analysis of scientific productivity using maximum entropy principle and fluctuation-dissipation theorem. Phys Rev E 75(2):026103
78. Gourieroux C, Monfort A, Trognon A (1984) Pseudo maximum-likelihood methods – applications to Poisson models. Econometrica 52(3):701–720
79. Angrist J, Pischke J (2009) Mostly harmless econometrics: an empiricist's companion. Princeton University Press, Princeton
80. Dehaan E (2021) Using and interpreting fixed effects models. Working paper, University of Washington. www.ssrn.com/abstract_id=3699777
81. Granovetter M (1973) The strength of weak ties. Am J Sociol 78(6):1360–1380
82. Rodan S, Galunic C (2004) More than network structure: how knowledge heterogeneity influences managerial performance and innovativeness. Strateg Manag J 25(6):541–562
83. Eagle N, Macy M, Claxton R (2010) Network diversity and economic development. Science 328(5981):1029–1031
84. Burt R (2004) Structural holes and good ideas. Am J Sociol 110(2):349–399
85. Hargadon A, Sutton R (1997) Technology brokering and innovation in a product development firm. Adm Sci Q 42(4):716–749
86. Ugander J, Backstrom L, Marlow C, Kleinberg J (2012) Structural diversity in social contagion. Proc Natl Acad Sci USA 109(16):5962–5966
87. Arora A, Gambardella A (1990) Complementarity and external linkages: the strategies of the large firms in biotechnology. J Ind Econ 38(4):361–379
88. Berg S, Duncan J, Friedman P (1982) Joint venture strategies and corporate innovation. Oelgeschlager, Gunn & Hain. xvi, 192 pages: illustrations
89. Richardson G (1972) The organisation of industry. Econ J 82(327):883–896
90. Ahuja G (2000) Collaboration networks, structural holes, and innovation: a longitudinal study. Adm Sci Q 45(3):425–455
91. Jaffe A, Trajtenberg M, Henderson R (1993) Geographic localization of knowledge spillovers as evidenced by patent citations. Q J Econ 108(3):577–598
92. Bikard M, Murray F, Gans J (2015) Exploring trade-offs in the organization of scientific work: collaboration and scientific reward. Manag Sci 61(7):1473–1495
93. Leahey E (2016) From sole investigator to team scientist: trends in the practice and study of research collaboration. Annu Rev Sociol 42(1):81–100
94. Petersen AM (2015) Quantifying the impact of weak, strong, and super ties in scientific careers. Proc Natl Acad Sci 112(34):E4671–E4680
95. Lorenz J, Rauhut H, Schweitzer F, Helbing D (2011) How social influence can undermine the wisdom of crowd effect. Proc Natl Acad Sci USA 108(22):9020–9025
96. Amason A, Sapienza H (1997) The effects of top management team size and interaction norms on cognitive and affective conflict. J Manag 23(4):495–516
97. Iacopini I, Petri G, Barrat A, Latora V (2019) Simplicial models of social contagion. Nat Commun 10:2485
98. Jabbehdari S, Walsh J (2017) Authorship norms and project structures in science. Sci Technol Human Values 42(5):872–900
99. Shapin S (1989) The invisible technician. Am Sci 77(6):554–563
100. Conroy G (2023) Surge in number of 'extremely productive' authors concerns scientists. Nature. https://doi.org/10.1038/d41586-023-03865-y
101. Aksnes DW, Langfeldt L, Wouters P (2019) Citations, citation indicators, and research quality: an overview of basic concepts and theories. SAGE Open 9(1):1–17
102. Wang J, Veugelers R, Stephan P (2017) Bias against novelty in science: a cautionary tale for users of bibliometric indicators. Res Policy 46(8):1416–1436
103. Uzzi B, Mukherjee S, Stringer M, Jones B (2013) Atypical combinations and scientific impact. Science 342(6157):468–472
104. Schumpeter J (1934) The theory of economic development. Harvard University Press, Cambridge
105. Funk R, Owen-Smith J (2017) A dynamic network measure of echnological change. Manag Sci 63(3):791–817
106. Stirling A (2007) A general framework for analysing diversity in science, technology and society. J R Soc Interface 4(15):707–719

Publisher's Note