Physics Reports 714–715 (2017) 1–73

Contents lists available at ScienceDirect

Physics Reports

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

The science of science: From the perspective of complex systems

An Zeng [a], Zhesi Shen [a,c], Jianlin Zhou [a], Jinshan Wu [a], Ying Fan [a], Yougui Wang [a,b,*], H. Eugene Stanley [b,**]

[a] School of Systems Science, Beijing Normal University, Beijing, 100875, PR China
[b] Center for Polymer Studies and Physics Department, Boston University, Boston, MA 02215, USA
[c] National Science Library, Chinese Academy of Sciences, Beijing 100190, PR China

ARTICLE INFO

Article history:
Accepted 25 September 2017
Available online 27 October 2017
Editor: Massimo Vergassola

Keywords:
Science of science
Scholarly data
Complex networks

ABSTRACT

The science of science (SOS) is a rapidly developing field which aims to understand, quantify and predict scientific research and the resulting outcomes. The problem is essentially related to almost all scientific disciplines and thus has attracted attention of scholars from different backgrounds. Progress on SOS will lead to better solutions for many challenging issues, ranging from the selection of candidate faculty members by a university to the development of research fields to which a country should give priority. While different measurements have been designed to evaluate the scientific impact of scholars, journals and academic institutions, the multiplex structure, dynamics and evolution mechanisms of the whole system have been much less studied until recently. In this article, we review the recent advances in SOS, aiming to cover the topics from empirical study, network analysis, mechanistic models, ranking, prediction, and many important related issues. The results summarized in this review significantly deepen our understanding of the underlying mechanisms and statistical rules governing the science system. Finally, we review the forefront of SOS research and point out the specific difficulties as they arise from different contexts, so as to stimulate further efforts in this emerging interdisciplinary field.

Contents

* Corresponding author at: School of Systems Science, Beijing Normal University, Beijing, 100875, PR China.
** Corresponding author at: Center for Polymer Studies and Physics Department, Boston University, Boston, MA 02215, USA.
E-mail addresses: ygwang@bnu.edu.cn (Y. Wang), hes@bu.edu (H. Eugene Stanley).

https://doi.org/10.1016/j.physrep.2017.10.001

1. Introduction

1.1. Science of science

Science is a broad concept that includes a variety of human activities, such as biologists viewing bacteria using microscopes, chemists conducting reaction experiments in a lab, astronomers peering through telescopes to observe galaxies, and physicists solving the equations on a blackboard. Essentially, it is a form of accumulated knowledge through the effort of human in understanding the universe in which we are living. Different from the past, when only a small number of people had the opportunity to work as scientists, modern science is advanced by many researchers from different disciplines. Meanwhile, much effort is being made to accelerate the publication of research findings. As a result, hundreds of new journals have been created in the past decade, and thousands of scientific papers are published every day.

Together with the development of science itself, the science of science (SOS) has become an important research field. It seeks to understand, quantify and predict scientific research and the resulting outcomes. Major research topics

include, for example, measuring the influence of scientific publications, researchers, journals, and universities; modeling scientific collaboration and citation patterns; understanding innovation processes; classifying different scientific domains; and predicting the future evolution of science. The significance of this research direction lies in its wide applications in the faculty hiring process, the job promotion process, and scientific proposal evaluation. Scientists with different backgrounds, including social scientists, information and computer scientists, economists, physicists and mathematicians, are involved in the research on SOS.

The industrialization of science has led to a revolution in the availability of datasets that capture major activities in science. One representative example is the Science Citation Index (SCI) produced by the Institute for Scientific Information (ISI) and created by Eugene Garfield. The SCI includes thousands of notable journals across different disciplines, and is now made available online through different platforms, such as the Web of Science (WoS). The increased access to such large datasets via the Internet has created an unprecedented opportunity to explore the patterns of scientific production and reward using rich mathematical and computational models [1]. The scientific publication data can easily contain millions of authors, papers and their citations. In addition, many large-scale datasets related to scientific prizes, funding, and patents have also been made public. As a result, the related research has been largely boosted, and the outcomes have been very fruitful. The resultant evaluation indicators are currently widely used by scholars, and the empirical trends are commonly referenced by policy makers. For instance, many proposed metrics, such as the well-known h index [2] for individual scientists and the impact factor [3] for scientific journals, are now widely used in practice.

The science of science actually has a broader scope than an already existing discipline called scientometrics which mainly involves measuring scientific impact, understanding scientific citations, mapping scientific fields and developing indicators for decision makers [4]. Specifically, SOS uses models to more deeply probe the mechanisms driving science, from knowledge production to scientific impact, distinguishing predictable patterns from random ones. It has more ambitious and diverse purposes, such as modeling the dynamics of research activities; revealing the rules underlying in scientific discoveries; predicting the development of science; and reformulating policies to stimulate innovations. To this end, one has to systematically investigate the complex structures, dynamics and evolution mechanisms of entire science systems. The emerging complexity science provides effective tools toward achieving the ultimate purposes of SOS.

1.2. Rapid development of complexity science

Complexity science is a science that studies the complex systems, which consist of a large number of components that interact with each other to produce nontrivial phenomena that cannot be explained by analyzing the individual constituent elements. Many real systems have been well accepted as complex systems, such as the human body, with its cells and processes; financial markets; social organizations; traffic; and climate. The key problems of complex systems are the difficulties that arise in their formal modeling and simulation. Because a complex system is usually composed of many components and their interactions, it can be represented by a network in which nodes represent the components and links represent their interactions. Complex networks can largely simplify real systems and preserve the essential information of the interaction structure that leads to emergent complex phenomena. The complex network has thus become an ideal tool for investigating complex systems.

In the early research on network science, the network representation of real systems was usually referred to as graphs. The origin of graph theory dates back to Euler's solution to the puzzle of Konigsberg's bridges in 1736. For a long time, graphs were the subject of many studies in mathematics, sociology, physics and chemistry. Regular lattices and random graphs [5] long served as the most important models of graphs. However, their simple intrinsic structures are unable to match the complexity observed in most real systems. After the small-world [6] and scale-free [7] networks were proposed at the end of the 20th century, a new surge of interest and research in complex networks began. Complex networks are usually irregular and evolve with time. They can have millions of non-trivially connected nodes and dynamical processes can take places on them. Thousands of examples of networks can be cited and envisioned: transportation systems, where nodes are airports and links are airline connections; neural systems, where nodes are neurons and links are synapses; social systems, where nodes are people and links are their social interactions, and power grids, where nodes are power plants and links are power cables. The advantages of modeling real systems by using complex networks are numerous. This type of modeling not only helps us to understand the detailed structures of these systems with high complexity but also allows us to create dynamic models (e.g. spreading and traffic) of the networks to explain and even control the collective patterns observed in these real systems [8–10].

Another important branch of complex system research is human dynamics [11–14]. It is a highly data-driven study of human behavior using methods from statistical physics. The goal is to understand, for example, when to send an email, receive a phone call, and view a web page. While traditional models assume that the timings of human activities are random, with the inter-event time following a Poisson distribution, increasing evidence has shown that they are not. Instead, many human activities exhibit a bursty feature [15–17], with fat tails in the inter-event time distribution. Deeper analyses have shown that the temporal patterns can be captured and predicted by simple models [18]. Human mobility is also one of the focuses of human dynamics research [19–25], which seeks to understand when and where people move in physical space as well as in cyber space. A striking finding is that the predictable and non-predictable co-exist [21]. Most human mobilities are regular and highly predictable, while a small fraction, usually long-distance and non-regular movements are hard to predict. Because many high-quality datasets are becoming available, human dynamics approaches are now being applied to more and more real systems and bring novel solutions to longstanding problems, e.g., spatial epidemic control [26], energy consumption and supply [27] and ride-sharing services [28].

1.3. Science as complex systems

Currently, there is no precise definition of complex systems. A recent article by a philosopher and a mathematician tried to answer the question "what is a complex system?" [29]. After reviewing and analyzing several definitions previously established by scientists working in complexity science, they listed several properties of complex systems: nonlinearity; feedback; spontaneous order; robustness and lack of central control; emergence; hierarchical organization; and numerosity. These properties are not necessary for a system to be qualified as a complex system, nor sufficient, but they are key features that differentiate complex systems from other simple systems. Like other complex systems, the system of science also contains a large number of components (e.g., papers, authors, and research fields) with multiple and evolving interactions (e.g., citations and collaborations). Using approaches from complex networks and statistical physics, researchers have identified many of the abovementioned properties. Examples include the spatial–temporal patterns of researchers' mobility and collaboration [30], the universal distribution of paper citations across different disciplines [31], and the collapse of the citation evolution of different papers [32]. In addition, the evolution of physics over the past century has been studied systematically [33,34]. The main contribution of these complex system approaches is the revelation concerning the hidden rules and patterns in scientific research through the process of building the linkage between the different scales and dimensions of a system.

In this Physics Reports, we will review the recent advances in SOS, seeking to cover topics from empirical study, network analysis, mechanistic models, ranking, prediction, and many important related issues. The results summarized in this review will significantly deepen our understanding of the underlying mechanism and statistical rules of scientific publication systems, as well as the forefront of SOS research. The methodologies will not only be valuable for practical use but will also inspire novel tools for other problems, such as online information filtering, critical part identification, algorithm robustness enhancement, and trend prediction. Despite recent progress, many challenges remain. We will also note the specific difficulties that arise from different contexts to stimulate further efforts in this new interdisciplinary field. The article is organized as follows. Section 2 summarizes the data sources of SOS and how they can be characterized by different types of networks, with their topological features discussed in detail. In Section 3, we focus on the evolution of the SOS networks. The empirical dynamical patterns as well as the models will be reviewed. Section 4 is devoted to the algorithms that quantify the scientific influence of different entities, including papers, authors, journals and institutes. The prediction problems in SOS are discussed in Section 5. In Section 6, we review the empirical works regarding the paths to success. Section 7 discusses the innovation in science and the diffusion of knowledge. The final section contains the summary of the review, along with a discussion about future research directions in this area.

2. Static properties of science systems

2.1. Data sources for SOS research

Scientific publication data. The most common data for SOS research are the scientific publication data. Scientific publications primarily comprise the research articles that report scientific findings made by scholars. Other types of papers (although much fewer than research articles), including reviews, comments, perspectives, and correspondences, are also considered scientific publications. Together with the summary of the major research results, a scientific publication usually consists of additional information such as the journal name, title, authors, affiliations, abstract, keywords, category code (e.g. PACS in physics, MeSH in medical science, JEL in economics), received and publication dates, acknowledgment, and references. This part of the information is extracted as source data for SOS research. In fact, these data can not only be used for the study of papers, but can also be aggregated to analyze scientific journals, institutes and even countries. For instance, the total number of published papers each year, the impact factor of journals, and the citations from one journal to another can be computed from the above mentioned data.

Early study of scientific publication data is based on relatively small datasets, and the data usually contain only one or two types of information. Collaboration, consists of one of these types, in which two scientists have a record if they have coauthored a paper. Newman studied four such databases in 2001 [35]: MEDLINE (biomedical research field), the Los Alamos e-Print Archive (theoretical physics field), SPIRES (high-energy physics field), and NCSTRL (computer science field) [35]. The total numbers of authors in these four databases range from several thousands to more than one million. After this pioneering paper, many follow-up works were done and much collaboration data have been collected and analyzed. Another type is citation data, in which two papers have a record if one cites the other. Unlike the collaboration data, in which the relationship between two scientists is reciprocal, the relationship between two papers in the citation data is directed. In 1998, Redner discussed the distribution of the citations among papers with citation data, e.g., papers published in 1981 in journals that are cataloged by the Institute for Scientific Information and 20 years of publications in Physical Review D [36]. Follow-up works primarily use citation data for studying the evolution of the system. In Table 1, we list numerous representative publicly available collaboration and citation data sources. Many papers on SOS have been done with these datasets.

The rapid development of the research on SOS requires larger and more complete datasets of scientific publications. In addition to citation and collaboration information, other information such as affiliations, received and publication dates, and keywords have attracted researchers' attention. To compare results across different disciplines, the scientific publications data of journals from different research fields are needed. In this context, recent studies on SOS primarily report analyses on large-scale datasets with relatively more complete information. In the following, we summarize several major datasets used for recent SOS research.

Table 1
Some publicly available collaboration and citation data sources. The data without a hyperlink can be obtained upon request to the authors.

Name	N	E	Time	Source	Description
Condensed matter	36458[a]	171735[a]	1995–2005	[35,37] (link)	undirected and weighted collaboration network on the Condensed Matter E-Print Archive
Astrophysics	14845[a]	119652[a]	1995–1999	[35,37] (link)	undirected and weighted collaboration network on the Astrophysics E-Print Archive
High-energy theory	5835[a]	13815[a]	1995–1999	[35,37] (link)	undirected and weighted collaboration network on the High-Energy Theory E-Print Archive
Network science	379[a]	914[a]	∼	[37,38](link)	undirected and weighted collaboration network of scientists working on network theory and experiment
SPIRES	56627	∼4900000	1995–1999	[37,38]	undirected and weighted collaboration network on the SPIRES
NCSTRL	11994	∼21600	1995–1999	[37,38]	undirected and weighted collaboration network on the NCSTRL
MEDLINE	1520251	∼13760000	1995–1999	[35]	undirected and weighted collaboration network on the MEDLINE
cit-HepPh	34546	421578	1993–2003	[39,40] (link)	directed Arxiv HEP-PH (high energy physics phenomenology) paper citation network
cit-HepTh	27770	352807	1993–2003	[39,40] (link)	directed Arxiv HEP-TH (high energy physics theory) paper citation network
APS	203245	1198002	1893–2009	[41] (link)	Weighted network of coauthorships between scientists publishing in APS journals
APS	414977	3992736	1893–2008	[42](link)	directed citation network of papers publishing in APS journals
Econophysics	1992	3485	1995–2010	[43,44]	undirected and weighted co-authorship network in the field of econophysics
Econophysics	2012	10558	1995–2010	[44]	directed citation network in the field of econophysics

[a] In the giant component.

- American Physical Society (APS). This dataset has been made available on request by APS. The metadata consist of all journals published by APS, including Physical Review A, B, C, D, E, I, L, ST and Review of Modern Physics, from 1893 up to now, with regular updates, amounting to over 450,000 publications. For each paper, one can obtain, for example, its title, DOI, author names, affiliations, printed time, received time, references, and PACS codes. See more detailed information in the link: http://journals.aps.org/datasets.
- MEDLINE/PubMed. MEDLINE (Medical Literature Analysis and Retrieval System Online), compiled by United States National Library of Medicine (NLM), is a bibliographic database of life sciences and biomedical information. PubMed (pubmed.gov) is a free resource developed and maintained by the National Center for Biotechnology Information (NCBI) at NLM. PubMed provides access to MEDLINE, links to full-text articles indexed in PubMed Central and supports advanced search.
- Web of Science (WoS). This is an extremely large database with the publications of almost all major scientific journals included. For each paper, all information is formatted and can be acquired. Therefore, this database is ideal for a comparison study across different disciplines. Although most universities and many research institutes have purchased the right to use the database, one can only use its interface to search for individual articles, instead of downloading the whole database for analysis. However, the website provides an Application Programming Interface (API) for batch retrieval. In a recent work [45], a large subset from 1900 to 2011 was analyzed. The extracted database contains approximately 47 million papers, 141 million coauthor entries, and 526 million citations referring to other articles.
- Scopus. Scopus, owned and maintained by Elsevier, is a large abstract and citation database of peer-reviewed documents: scientific journals, books and conference proceedings. It provides a service quite similar to WoS.
- arXiv. arXiv is an e-print service covering the fields of physics, mathematics, computer science, quantitative biology, quantitative finance and statistics. There are other preprint repositories, e.g., bioRxiv for biology and SSRN for social science.
- Google Scholar/Microsoft Academic/Baidu Xueshu. These services are freely accessible web search engines for scholarly literature and patents covering a broad range of disciplines. They index the full text or metadata and provide a scientific personal webpage including published papers together with their citation counts and frequent collaborators.

Funding data. Funding information is another important type of data for SOS research. Scientific funding provides financial support for research, with the amount ranging from several thousand dollars to several million dollars. It is usually used for example to purchase scientific instruments, hire doctoral students or postdocs, and go to conferences. Writing proposals is thus an important task that every researcher must do. Because the success rate of funding applications is closely related to the scientific achievement of researchers, the funding received by a researcher today gradually becomes

Table 2
Some online patent data sources (link).

Name of the databases	Source of the database	Brief description
EPO Worldwide Patent Statistical Database	European Patent Office; World	Bibliographic data, citations and family links of about 70 million patent applications of more than 80 countries.
USPTO Patent data	United States Patent and Trademark Office (USPTO)	Granted United States Patent and Trademark Office (USPTO) patent data, including names of inventors, names of assignees, grant and application dates, technology classes, citations, etc (link).
NBER patent data	National Bureau of Economic Research (NBER)	A subset of USPTO data provided by NBER patent project.
U.S. Patent Inventor Database	Institute of Quantitative Social Science, Harvard University	An update to the original NBER Patent Data, with inventor names disambiguated.
OECD Triadic Patent Families Database	Organization for Economic Co-operation and Development (OECD)	Set of patents filed for at the EPO, the Japan Patent Office (JPO) and granted by the USPTO that share one or more priority applications.
OECD REGPAT	OECD	Patent applications to the EPO and Patent Corporation Treaty (PCT) filings linked to more than 5500 regions using the inventors/applicants addresses.
OECD Citations Database	OECD	Citations from patents published by the EPO, and the World Intellectual Property Organization (WIPO) via Patent Cooperation Treaty (PCT).

an indicator for evaluating scientists. In principle, the useful data of each funding for SOS research contains for example the title, abstract, keywords, references, amount of money, the principal investigator (also called main applicant or coordinator), participants, funder, approval year, and duration. The analysis of these data can help to understand and improve the development policy for science, eventually distributing research resources better.

Although the funding data contain many different types of useful information, it is difficult to obtain all of them. Partial data can be extracted for analysis. For instance, because funding is acknowledged in scientific publications, extracting the funding information from the acknowledgment section is one means of obtaining funding data. By aggregating this data from different papers, one can know all the papers published by a scientific project. Other national funding data of each country can be acquired from their corresponding government and funding agency. Many funders' websites list a history of the basic information of finished projects they support, even including their final reports. For instance, NIH provides a publicly available database comprising the basic grant information and related output supported by that grant. The data can also be obtained by crawling or crowdsourcing. With the collaboration of the funders, some research groups even acquire the rating data in the review process of the funding. Such types of data are commonly used to study the bias of the funding evaluation and the policy of science. Together with the citation data, the funding data can be used to answer a variety of questions, such as how successful is a scientific project? In how many directions is a scientific project working? What is the similarity between different projects?

Patent data. Patent data are a type of SOS data mostly used for studying innovations in technology. The patent data usually contain names of inventors, names of assignees, grant and application dates, technology classes, citations and a key matching patents to firms. One of the most commonly used information items is the forward citation, which can be employed to trace innovation paths, identify the origin of a certain patent and rank the importance of different patents. Additional important information here is the key matching patents to firms, which can reveal the development alliance and the competition between two firms. Many patent data records are made online; a summary of the major data sources is provided in Table 2.

2.2. SOS data in terms of networks

A complex network consists of a large number of nodes (vertices) that are connected by links (edges). Starting at the end of the 20th century [6,7], studies on complex networks have been performed over a decade and continue to develop rapidly, covering issues ranging from the general network theory to modeling empirical systems with complex networks. The fundamental reason for this movement of complex network research is that the structure of real networks is irregular, complex and dynamically evolving in time, completely different from the traditional network models such as lattice and the Erdös–Rényi random network model [46]. Similar to other real systems such as power grids, airlines, and the Internet, the science system is also an ideal system that should be modeled by complex networks [10]. The system consists of thousands or millions of components (e.g., authors and papers) with complex and evolving relationships (e.g., citations and collaborations). In the following, we review some typical network models that have been used to characterize SOS data.

A bipartite graph is a network consisting of two types of nodes, with connections existing only between different node types. In SOS, it is usually used to model the co-authorship data [47]. The scientists form a type of node, and the papers are the other type of node. A connection between a scientist and a paper indicates that the scientist is one of the authors of this paper. The number of links of a scientist represents the number of papers he/she published. The number of links of a paper indicates the number of scientists who coauthored this paper. In addition to the co-authorship data, much other SOS data can

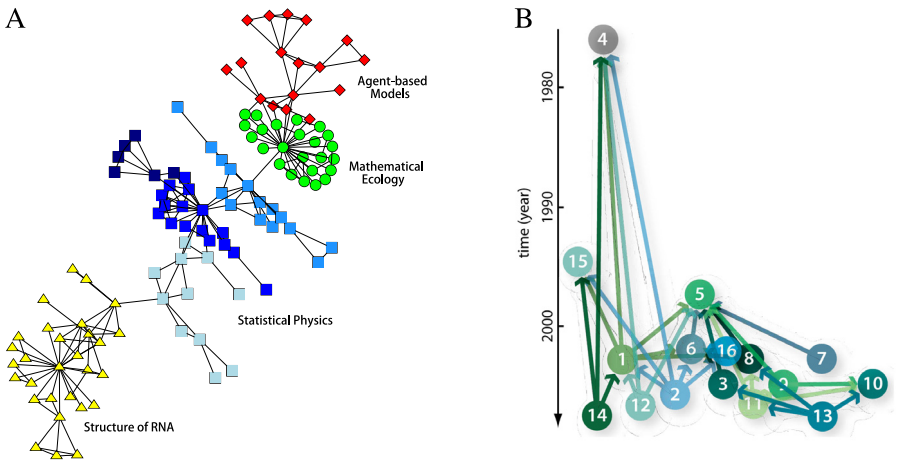

Fig. 1. An example of (A) a collaboration network and (B) a citation network.
Source: The figure is reprinted from Refs. [50] and [51].

be described and analyzed with bipartite networks such as keyword data (scientist-keyword bipartite network), funding data (funding-paper bipartite network), and patent data (inventor-patent bipartite network). Most of abovementioned bipartite networks are unweighted, but a few others, might have weights. For instance, the link weight in the scientist-keyword bipartite network indicates the number of papers authored by a scientist that contains a certain keyword.

Monopartite networks are the major network model for SOS research. Different from the bipartite network, they have only one type of node. A collaboration network is a monopartite network of scientists in which two scientists are considered connected if they have coauthored a paper together [48]. This network is weighted, with the link weights representing the number of papers two scientists have coauthored. A citation network is also a monopartite network, with nodes representing papers and links representing citations from one paper to another [49]. This network is unweighted but directed. Unlike other real directed networks such as the Twitter social network and neural networks, the direction of links in a citation network is strictly constrained with time. The links can only originate from younger nodes and point to older nodes, making the citation network acyclic. A visualization of a collaboration network and a citation network is shown in Fig. 1. Note that the collaboration and citation networks can be constructed at a higher level, such as the collaboration networks between universities and the citation networks between journals. In addition to the collaboration and citation data, many other SOS data have been modeled with a monopartite network. Examples include a co-appearance network of keywords and a co-cited network of papers.

Multilayer networks have recently attracted increasing attention [52]. They are usually networks with multiple types of relationships. Modeling the SOS data with the multilayer network yields valuable insights for understanding the structure of the science system and ranking the scientific impact of papers and scientists. A typical example of a multilayer network in SOS consists of scientist nodes and paper nodes, with the authorship links in one layer and citation links in the other layer [44]. Thus, both collaboration relationships and citation relationships are captured with one network. Authors' contributions to a research article have also been modeled with multilayer networks, with the interactions of authors in each contribution category representing one layer [53]. In fact, because of the high complexity of the science systems, many multilayer networks remain to be explored.

2.3. Name ambiguity as a challenge for network analysis

As mentioned in the previous subsection, many SOS networks involve the data of both scientists and their publications. To associate a scientist with all of his/her publications, it is necessary to infer each author's actual identity. However, some authors have distinctive names, whereas numerous others have similar or even identical names. The scientists with the same name might be considered as one person, which might result in the emergence of some "super productive" scientists. Another risk is that the same author might be referenced in different ways, which might split some true hubs in the network. This *name ambiguity* problem is a long-standing challenge for scientometrics. When the SOS data are modeled with complex networks, name ambiguity also becomes a major barrier to constructing accurate networks for further analysis. From the network point of view, the name ambiguity problem causes many different nodes to be merged as one node with many links. The sensitivity of the collaboration network analysis to name ambiguity was discussed by Klosik et al. [54]. The work seeks to search induced few-node subgraphs (known as motif) in the collaboration network constructed from the APS dataset. Each motif is assigned a significance score based on the citation received by the coauthored papers. An interesting finding is that in a basic method of author disambiguation the box-motif (a ring subgraph with four nodes) score can be suppressed, whereas

a stricter disambiguation scheme yields a high score. This study highlights the fact that name ambiguity is a non-negligible issue for collaboration network analysis.

In the literature, various algorithms for name disambiguation can be found. The most straightforward method is to use surname and the initial of given name to distinguish authors. However, this simple method might lead to significant problems because numerous authors share a surname and the initial of given name. Author names can be further disambiguated by incorporating similarities calculated with additional information including coauthor names, affiliations and research area. For instance, in two recent works on scientists' careers [30,55], author names in APS data are disambiguated with this approach by considering coauthor and affiliation information. After performing the name disambiguation in the APS dataset, the study eventually obtained 236,884 distinct scientists. The disambiguated data are now made available online (see link here). To use the additional information more effectively, a Bayesian model was introduced to develop a probabilistic similarity metric for author name disambiguation in Medline data [56]. Other more complicated techniques such as text mining and natural language processing [57] have also been introduced to distinguish author names, a technique shown to be more effective than the initial-based methods [58].

In collaboration networks, name disambiguation can be improved with topological features and hierarchical characterization [59]. Two approaches were proposed. The first one extends the traditional method based on connectivity with the closest neighbors to the neighbors with longer distance, finding that using the 3rd hierarchy in connections yields the highest accuracy. The second approach uses the average degree and average collaboration strength of immediate neighbors to discriminate different authors. However, these two approaches are validated on an arXiv collaboration network with the ambiguity deliberately introduced. Their effectiveness compared with other simple methods on real data remains to be examined.

A more comprehensive study of name disambiguation was done by Schulz et al. [45]. The work seeks to disambiguate all author names in the WoS database, with special focus on assigning the well-cited publications to the correct authors. With the WoS citation network, the similarity between each pair of publications is evaluated. Papers are assigned to authors based on the assumption that there is higher similarity between two publications written by the same author than between two random papers. Mathematically, the similarity between two papers i and j is calculated as

$$s_{ij} = \alpha_A \left(\frac{|A_i \bigcap A_j|}{\min(|A_i|, |A_j|)} \right) + \alpha_S (|p_i \bigcap R_j| + |p_j \bigcap R_i|) + \alpha_R (|R_i \bigcap R_j|) + \alpha_C \left(\frac{|C_i \bigcap C_j|}{\min(|C_i|, |C_j|)} \right) \tag{1}$$

where for each paper p_i the reference list, the co-author list and the set of citing papers are denoted as R_i, A_i and C_i respectively. As noted in [45], the first term measures the number of co-authors shared by two papers. The second term detects potential self-citations. The third term is the count of common references between the two papers. The fourth term represents the number of papers that cite both publications. Two papers with a similarity score s_{ij} greater than a threshold β_1 are linked together. Each connected component is labeled a cluster. The similarity between cluster γ and κ is computed as

$$S_{\gamma,\kappa} = \sum_{i \in \gamma, j \in \kappa} \frac{s_{ij} \Theta(s_{ij} > \beta_2)}{|\gamma||\kappa|}. \tag{2}$$

Two clusters are connected if the new similarity score $S_{\gamma,\kappa}$ is greater than a threshold β_3. Each connected component is again merged into a single cluster. Remaining individual papers are added to a cluster if they have a similarity score s_{ij} above a threshold β_4 with any paper in that cluster. Each finally obtained cluster contains the published papers of an individual author. The method has eight parameters in total. With the Google Scholar Profile data as the ground truth, these parameters are optimized for the WoS database: $\alpha_A = 0.54, \alpha_S = 0.75, \alpha_R = 0.19, \alpha_C = 1.02, \beta_1 = 1, \beta_2 = 0.19, \beta_3 = 0.011, \beta_4 = 0.49$. Moreover, the method is claimed to be very efficient, with 47 million articles of WoS disambiguated on a single machine in less than a day.

In addition to these algorithms, the Open Researcher and Contributor ID (ORCID) was launched in 2012 to solve the name ambiguity problem in practice [60]. The main purpose is similar to the digital object identifiers (DOIs) for scientific publications. Specifically, ORCID offers an open and independent registry for authorship identification in paper submission, such that each scientist corresponds to an unique ID. By 2017, the number of registered accounts is over 3 million and organizational members include many large publishers such as Elsevier, Springer and Nature Publishing Group.

2.4. Structural properties of the empirical networks

The starting point for understanding science systems is to analyze the structural properties of the corresponding empirical networks. Thus, many endeavors have been made in this direction. In this subsection, we will review the major structural patterns uncovered in collaboration networks, citation networks and several other networks.

2.4.1. Collaboration networks

Collaboration networks are one of the earliest investigated SOS empirical networks. One of the pioneering studies was performed by Newman [35]. By constructing the empirical collaboration networks from MEDLINE (biomedical research), the Los Alamos e-Print Archive (physics) and NCSTRL (computer science), Newman studied a variety of structural properties including the degree distribution, giant component, average shortest paths and clustering coefficient. Both the distribution

of collaborators of scientists and papers they write are well fit by power-law forms with an exponential cutoff. In addition, the network in each discipline exhibits the well-known "small-world" feature, i.e., small average shortest paths (a scientist is only five or six steps away from a randomly chosen scientist) and highly clustered (two scientists tend to have collaborated if they have a third common collaborator). More-detailed analyses of these empirical collaboration networks are reported in two follow-up publications [61,62].

Motifs are the building blocks of complex networks. They are subgraphs in real networks that are significantly higher in number than those in randomized networks. Krumov et al. focused this local property of collaboration networks and investigated its effect on the citations received by scientific publications [63]. Eight undirected motifs consisting of three and four nodes are considered and the success of a motif is defined as the average citation frequency per edge of all involved publications. The empirical analysis indicates that a box motif (four authors forming a closed chain) is most successful, i.e. has the highest average citation frequency per link. However, this effect is found to be much weaker if the name ambiguity problem is solved, as noted in a previous subsection [54].

Collaboration networks are also found to have community structure [50]. A community is defined as a subgroup of nodes within which links are dense but between which links are much sparser. A simple method is proposed to identify communities in collaboration networks by repeatedly cutting the links with the highest edge-betweenness (number of shortest paths passing through a link) in the network. The algorithm is applied to a collaboration network of scientists at the Santa Fe Institute, and six communities are found. The communities in general correspond to scientists from different disciplines. A visualization of the detected communities is shown in Fig. 1(A). After this paper, a large number of works focused on developing better algorithms for community detection. For a comprehensive review, see [64].

Another important property of collaboration networks is assortativity mixing [65]. A positive assortativity coefficient indicates a preference for high-degree (low-degree) nodes to attach to other high-degree (low-degree) nodes. In [65], networks including physics coauthorship, biology co-authorship and mathematics co-authorship all have a positive assortativity coefficient. This effect has been analyzed locally. Colizza et al. focused on the highly connected nodes in the collaboration network and revealed the "rich-club" phenomenon [66]. This phenomenon refers to the tendency of hubs (high-degree nodes) to form tightly interconnected communities, which is to some extent consistent with the intuition that the elite formed by more-influential scientists tend to form collaborative groups within specific domains. However, this phenomenon was later challenged by Opsahl et al. [67]. A weighted collaboration was constructed from the arXiv data based on the publications on network science, with the weight on links between two scientists representing the number of papers coauthored by them. When detecting the rich club effect in the weighted network, the researchers compute the total amount of weights among influential scientists instead of the number of links among them. The weighted and topological rich-club effects display strikingly different trends for the scientific collaboration networks. The influential scientists no longer form a large rich-club in which one collaborates with all those in the rich-club. Instead, they maximize their resources per collaboration, tending to intensively collaborate with one another; see Fig. 2.

In fact, weighted collaboration networks preserve more information from the collaboration data than binary collaboration networks do. Many interesting patterns were revealed in the weighted collaboration networks. Ramasco et al. defined social inertia in the weighted collaboration networks as the ratio between node strength and node degree [68]. This ratio measures the tendency of the actors to repeat the same collaborators. Social inertia is found to have a long-tailed distribution and generally grow with a scientist's experience. The distribution of the strong weight connections in network topology was analyzed in [41]. In contrast to other social networks, the weak links in collaboration networks are associated with locally dense network neighborhood, whereas the stronger links largely preserve the overall connectivity of the networks. This feature was later reproduced by a growing network model [37]. By simulating a spreading process on the weighted collaboration network, it was also found that the topological position of the strong links can accelerate spreading dynamics and enhance the flow of information across scientific collaboration networks [41]. Focusing on the longitudinal aspects of collaboration, Petersen quantified the effects of weak, strong, and super ties in scientific careers [69]. The collaboration networks were found dominated by weak ties characterized by high turnover rates. However, some extremely strong collaborations also exist in the networks that have a significantly positive effect on productivity and citations.

Numerous other issues with respect to the collaboration networks have been studied. The friendship paradox (your friends have on average more friends than you do) is an important phenomenon in social science and is revealed to be present in collaboration networks for various characteristics, i.e., your collaborators tend to have more coauthors, higher citations, and more publications [70]. The positioning patterns of authors in the author list are revealed in collaboration networks [71]. Compared with the authors publishing more often as last authors, the authors publishing more first-author articles have fewer publications in the short-term, because the first author usually must devote much time to perform the research, preventing them from attempting as many parallel projects as last authors. The attempts to understand the general structural properties of collaboration networks are followed with a number of case studies. The collaboration networks were constructed and analyzed among the scientists working in econophysics [48], sociology [72], finance [73], scientometrics [74], and other disciplines.

2.4.2. Citation networks

Citation networks are directed networks with the nodes representing papers and links representing the citations from one paper to another. The total number of citations received by a paper is simply the in-degree of the node in the citation network. Conversely, the out-degree of a node describes the number of references a paper has. The in-degree and out-degree

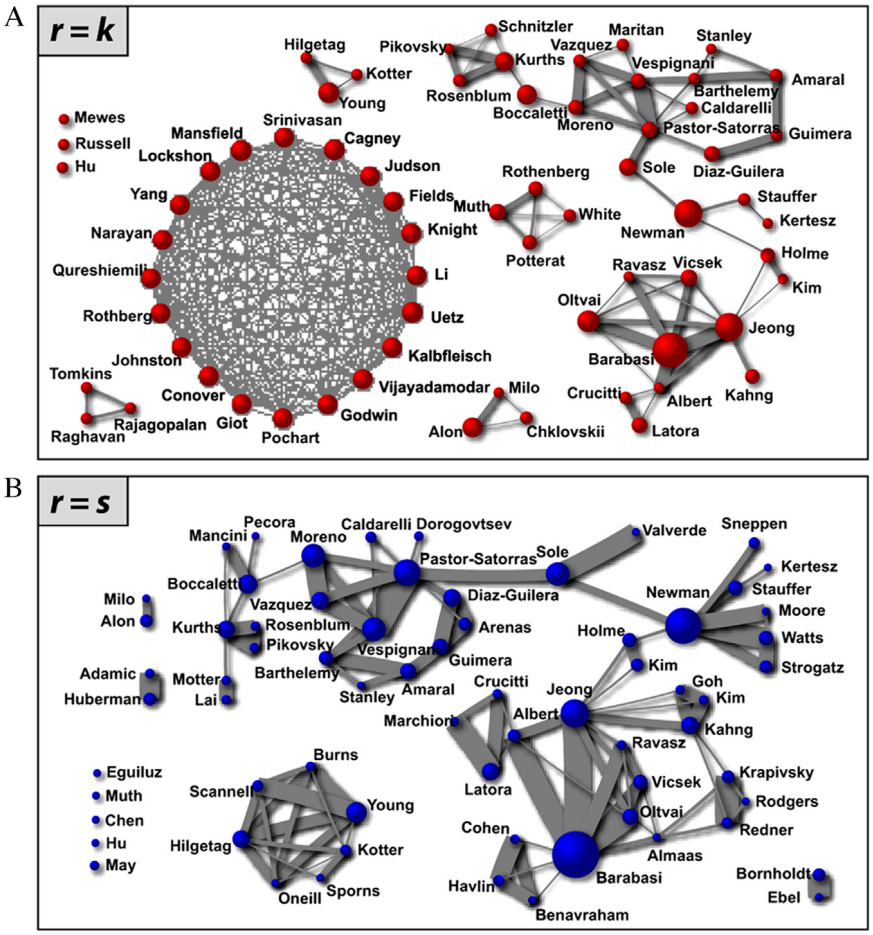

Fig. 2. Subsets of the rich nodes in the network science collaboration network based on degree (A) and strength (B). Only links among the rich nodes are shown. The size of the nodes is proportional to their richness; the width of the links is proportional to their weight.
Source: The figure is reprinted from Ref. [67].

of nodes in citation networks follow a power-law and an exponential distribution, respectively [36]. Specifically, the citation distribution of the WoS dataset and Physical Review D dataset are computed, and a power-law $N(x) \sim x^{-\alpha}$ with $\alpha \approx 3$ is found. Although citation distribution follows a power-law form, the exponent can vary from one dataset to another. For instance, a study of the citation network in high energy physics shows a truncated power law with $\alpha \approx 1.2$ for less highly cited papers and $\alpha \approx 2.3$ for more highly cited papers [49].

In principle, each directed network can be described with the bow-tie structure, with a giant strongly connected component (GSCC) or a weakly connected component (WCC), the in- and out-components (IN and OUT), and tendrils (TEND) [75]. When performing the bow-tie decomposition in citation networks, a significant change in the fraction of different components can be observed when comparing different databases [76]. This work also shows that the citation networks have in general weak disassortativity, relatively high clustering coefficients and small shortest path length, with the detailed values differing from one database to another. The analysis highlights that the coverage of the database and the time span of the literature greatly affect the overall citation topology.

Different from other directed networks, citation networks are acyclic, with links strictly pointing from newer papers to older papers [77]. To identify the significant features of empirical networks, the common approach is to compare them with the counterpart random graphs (usually with the same number of nodes and links, occasionally even with the same degree sequence). Because citation networks are acyclic, their counterpart random graphs must be generated with a more complicated rule. For each node in the random counterpart, the number of links that connects later nodes to earlier nodes must be consistent with real cases [78]. With these acyclic random graphs, one of the most significant motifs in citation networks is the feed-forward loop, i.e., paper a cites b, b cites c and a cites c [51]. Citation networks were also found to have community structure, with each community roughly corresponding to a subfield [79]. A major challenge when analyzing this property is that the original community detection methods generate strange results if directly applied to citation networks. Although many community detection algorithms for directed networks were proposed [80–82], a method

specifically designed for citation networks remains to be developed. The new method is required to estimate the significance of communities by comparing real citation networks with acyclic random graphs. When validating the community detection algorithms with real citation networks, one must be very careful with the ground truth data because the expert-made classifications can include a rather strong bias [83]. A systematic comparison of the performance of numerous clustering methods in citation networks was recently presented by Šubelj et al. [84], showing that map equation methods [80] in general perform best.

The dimensions of citation networks have also been measured. The dimension of a complex network can be defined in various ways. A common characteristic of dimension is that the higher the dimension of a system, the more complex the system is. By simulating the diffusion of a test particle in the citation networks, the spectral dimension of citation networks was found to be 3, which is much less than those of other artificial network models [77]. Considering that citation networks are constrained in time, citation networks are considered embedded in a Minkowski spacetime and their dimensions are measured using Myrheim–Meyer and Midpoint-scaling estimates [85]. Such dimension analysis reveals some properties that cannot be observed by traditional network structural analyses. An interesting finding is that two empirical citation networks in particle physics have similar degree distribution and clustering coefficient, but they differ in the dimension measures, which provides an alternative method of quantitatively characterizing structure by allowing distinct citation networks to be differentiated.

Links in citation networks have diverse properties. Although it is better in general for a paper to receive more citations, some citations might express negative opinions. By automatically extracting negative citations from the in-text references, signed citation networks can be constructed [86]. It was found that both positive citations and negative citations obey heavy-tailed distributions, and the signed citation networks follow weak balance theory [87]. Apart from signs, the heterogeneity of links in citation networks also reflects their significance. The citation count and topological similarity were employed to identify the most important reference of each publication, leading to a treelike backbone of the citation network [88–90]. The backbone of citation networks is analogous to the descendant chart of research papers in which one can identify seminal papers, paper clusters, and in general a synthetic picture of different research fields.

2.4.3. Other networks

A number of other networks can be constructed from scientific publication data. An important one is the citation network between scientists [91]. This network is weighted and directed, with the link weight computed as the number of times that one scientist cites another. Both the in-degree and in-strength of a node in this network follow a broad distribution [91]. Unlike a citation network between papers, the citation network between scientists has cycles. The creation of citations is strongly influenced by the co-authorship relationships between scientists, leading to some delicate structures in this network. Unsurprisingly, direct self-citations (including co-authors of both citing and cited papers) commonly exist and are rather constant in fraction [92]. There are also citation cartels, defined as groups of authors that cite each other disproportionately more than they cite other groups of authors, existing in citation networks [93]. Keyword co-occurrence networks, with links representing a pair of keywords appearing in the same article, are built to understand the organization of scientific knowledge [94] and conflict research [95]. Some keywords are found to have high betweenness, as an important bridge across other keywords. Using similarity in citing patterns, the networks between WoS subject categories can also be built [96]. The network has a clear community structure, with each community corresponding to a discipline. The diversity and coherence in this network are indicators of research interdisciplinarity.

The scientific publication data can be mapped to the citation relationships between entities at higher levels. One of the most commonly analyzed types of networks is citation networks between journals [97]. The centrality of a node in this type of network can measure the influence of a journal [98] (see more in Section 4). Apart from the citation network, similarity networks between journals have been constructed and are considered the backbone of the scientific landscape [99]. Similar to the keyword networks, the betweenness centrality in this network was shown to be an indicator of the interdisciplinarity of journals [100]. Other higher-level networks include universities and countries because the nodes have been built to achieve a more objective ranking of their influence [101]. A more detailed review of this aspect will be provided in Section 4.

The scientific publication data have also been modeled with multilayer networks, primarily by combining the collaboration networks and citation networks. The most straightforward one is a two-layer network with one layer as a weighted collaboration network between authors and the other layer as a weighted citation network between authors [102]. This multilayer network shows a significant overlap of links and a significant correlation between degrees of nodes captured by the Pearson correlation coefficient. Numerous other effects are revealed. Highly cited authors are cited by their co-authors to a much greater extent than poorly cited authors. In the citation network, node strength is found to grow super-linearly as a function of node degree. In other works [103,104], the effect of authors' centrality in the collaboration networks on their received citations was investigated. The degree positively affects citations [103] whereas the betweenness centrality (authors bridging two groups) has instead a negative effect [104]. The study of the co-evolution of collaboration networks and citation networks brings much insight to understanding the success of scientists' career [105,106], which will be discussed later in Section 6.

3. Dynamical properties of science systems

3.1. Dynamics of scientific publications and mechanistic models

Modeling the dynamics of scientific publications is difficult because they are highly complex. The development of each citation is influenced by numerous factors, including topics, authors, journals, publication time, etc. Therefore, numerous

nontrivial empirical phenomena emerge (discussed in detail below). In the past decade, a variety of mechanistic models have been developed to generate citation networks and describe the empirically observed features regarding their structure and evolution.

3.1.1. Basic structure-oriented models

The Barabasi–Albert (BA) model is one of two network models that triggered the study of complex networks. In SOS, this model is often considered to be the earliest and most basic model that describes the dynamics of citation networks. The BA model is based on two fundamental components: growth and preferential attachment [7]. Conceptually, in real networks, nodes with high degrees obtain new links at higher rates than low-degree nodes. Starting with m_0 fully connected nodes, at each time step t, a new node j with $m \leq m_0$ links is added to the network. The probability that a link will connect j to an existing node i is linearly proportional to the present degree of i,

$$P(j \to i) = \frac{k_i}{\sum_l k_l}. \tag{3}$$

Because every new node has m links, the average degree of the final network will be $\langle k \rangle = 2m$. In the limit $t \to \infty$, the model produces a degree distribution $P(k) \sim k^{-\gamma}$, with exponent $\gamma = 3$. Although this exponent differs from certain empirical citation networks, this model explains the underlying mechanism for the power-law citation distribution in SOS data.

Despite its success, the BA model is far different from actual citation networks. Certain differences are straightforward. The BA model is an undirected network, and real citation networks are directed. In citation networks, the in-degree and out-degree refers to the number of citations received by a paper and the number of references in a paper, respectively. A common approach for extending the BA model to become a directed network is to consider the probability that a node acquires a new link according to its in-degree (present citation); the links always point from younger nodes to older nodes. To avoid new nodes (with zero in-degree) that have zero probability to attract links, Eq. (3) can be modified to become a directed case when it is written as follows: $P(j \to i) = (k_i^{\text{in}} + 1)/(\sum_l k_l^{\text{in}} + 1)$ where k_i^{in} denotes the in-degree of node i [107]. This modification ensures that the network includes the acyclic feature of actual citation networks. In addition, this modification will not significantly alter the exponent of the power-law of the original BA model. However, if 1 is replaced by a tunable parameter r, the final exponent γ will depend on r, i.e. $\gamma = 2 + r/m$ [108].

Empirical studies demonstrate that the average out-degree of papers in an APS dataset gradually increases with time; however, the BA model and its variants generally assume that the out-degree of new nodes is a constant. To explain this feature of real networks, a growing network model that is developed by copying mechanism was proposed [109]. The network grows by adding a new node in each step. First, the new node randomly selects an existing node and then connects to it and all its upstream nodes (i.e., the nodes at which the randomly selected node is pointed). The new node copies the linking patterns of the randomly selected node. As the number of nodes in the network grows, the number of links that the new nodes must establish also increases. A primary feature of this model is that the average degree of the network increases logarithmically with the system size, which is consistent with APS data for the past 110 years. The in-degree and out-degree of this model aligns with power-law and Poisson forms, respectively. In addition, the accelerated growth of links can be modeled using a power-law distribution of new nodes' out-degree m where $1 \leq m \leq t$ at each step t [110]. Clustering is a primary characteristic of citation networks. In undirected networks, clustering generally refers to the existence of a subgraph (or motif) with three fully connected nodes [6]. Because citation networks are acyclic, the 3-node motif can only form a feed-forward loop (A cites B, B cites C, and A cites C). Empirical studies reveal that feed-forward loops are abundant in citation networks [111]; however, the BA model and its variants mentioned above cannot capture this feature. In addition, studies have indicated that acyclic random graphs fail to generate sufficient feed-forward loops as real citation networks [51]. Although it is not explicitly addressed in the growing model with copying [109], various studies have noted that feed-forward loops can be significantly increased by using this mechanism [9,112]. Alternatively, the property of rich triangles in citation networks can be modeled using a geometric graph [113] that is similar to the popularity–similarity model [114]. The geometric graph uses angular coordinates of nodes to represent the research content of papers. Citations between papers are drawn according to a geometric rule, which forms many feed-forward loops among similar papers [113,115]. In addition, this model captures another important feature of citation networks, i.e., an exponential increase in the number of scientific papers.

3.1.2. First-mover advantage and the aging effect

The preferential attachment mechanism strongly favors the nodes that appear earlier, which is referred to as the first-mover advantage [117]. Because of this effect, the first papers that are published regarding a new field of study receive significantly more citations than papers that are published later. A theoretical prediction of the average citations that a paper receives as a function of its date of publication is calculated using the preferential-attachment process with a mathematical expression $c(t) = r(t^{-1/(\alpha-1)} - 1)$ in which r represents a parameter in the preferential attachment $P(j \to i) = (k_i^{\text{in}} + r)/(\sum_l k_l^{\text{in}} + r)$ and α represents the exponent of the power-law in-degree distribution [117]. The theoretical prediction fits well with actual citation data, and only a few exceptions have been observed. However, the first-mover advantage is to a certain extent overcome by the aging effect [118]. The citations of a paper within a given year do not refer evenly to papers that were published in prior years, although the number of papers published in prior years is similar, as illustrated

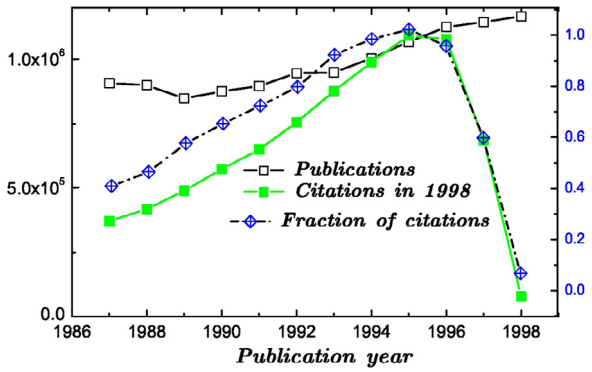

Fig. 3. The aging effect in citation networks. The open squares correspond to the number of papers published between 1987 and 1998. The solid squares correspond to citations from the papers published in 1998 to papers published in a given year. The diamonds stand for the average number of citations received in 1998 by a paper published in a given year. The data in this figure is a subset of WoS database.
Source: The figure is reprinted from Ref. [116].

in Fig. 3 [116]. The time interval between a published and cited paper follows rather universal behavior $T(t) \sim t^{-0.9}$ for $t \leq t_c$ and $T(t) \sim t^{-2}$ for $t > t_c$, where $t_c \sim O(10)$ [119]. Notably, this effect can be represented by an exponential decay factor, $e^{-\beta\tau}$, where τ represents the age of the node. The preferential attachment must be modified as $p_j \sim k_j^{\text{in}} e^{-\beta\tau_j}$, i.e., the probability that new node j will connect to an old node is proportional not only to the in-degree of j but also to the exponential decay of its age. Generally, decay has certain effects on resultant networks because of this growth: (i) a more homogeneous in-degree distribution, (ii) increased clustering, (iii) a hierarchical structure, (iv) a disassortative degree-degree correlation, and (v) a larger average node-node distance [116]. These effects are studied in greater detail in Refs. [120–123].

Later, the aging effect is combined with copying behavior to generate more realistic citation network models and simultaneously fulfill the observed structural and temporal patterns. In contrast to the BA model, the preferential attachment mechanism of this type of model is not directly set but emerges from the local connection rules of nodes. A representative network model was proposed in Ref. [51]. During each stage of network growth, a new node enters the network; its first link connects to an old node according to probability $\Pi_{i\to j}$ proportional to its age $t_j = i - j$ to power α (α should be negative). The remainder of the outgoing links of this new node with probability β attach to the random neighbors of j and $1-\beta$ attaches to an older node with probability $\Pi_{i\to j}$ as the first link. Clearly, α and β are the two primary parameters. The parameter α controls the effect of aging, and β represents the probability of repeating the copying behavior. Three properties of this model are compared with empirical networks: degree distribution, number of triangles, and parameter $\lambda_i = \sum_{j=1}^{i-1} k_j^{\text{in}} - \sum_{j=1}^{i} k_j^{\text{out}}$ that represents the number of links that connect vertices later than i to vertices earlier than i [78]. λ_i is the parameter that captures the temporal patterns of real citation networks. The quantitative analysis indicates that a negative α and a large β (e.g., $\alpha = -1$ and $\beta = 0.99$) can result in a network that most closely represents an actual case. A later study goes one step further by considering the connecting patterns among existing papers when a new node decides which existing nodes to attach to [124]. In this model, the first link is attached to an old node j according to the probability $\Pi_{i\to j} \sim k_j^{\text{in}} t_j^{-\alpha}$, where k_j^{in} represents j's in-degree, $t_j^{\text{in}} = i - j$ represents the age of j and $\alpha > 0$ represents the decaying parameter. With a probability of β, the new node i connects the remainder of the outgoing links to j's clique neighbors, i.e., the nodes in the same clique that j belongs to (note that the neighbors of j that are not involved in j's cliques should not be attached by i). When $1-\beta$ or there are no clique neighbors that i can connect to, the remainder of the outgoing links of j connect to older nodes as the first link. This model can solve the problem that occurs when the copying behavior underestimates the number of triangles (i.e., feed-forward loops) in actual citation networks [124].

The preferential attachment mechanism predicts that the most highly cited papers are always old papers because of the first-mover advantage [117]. However, in the real world, we observe certain recent papers with high citation rates. This is because of the aging of early papers as well as the heterogeneous fitness of papers. The ability of a node to attract new links not only depends on its cumulated links but is also strongly influenced by its fitness, which represents the intrinsic quality of the paper [125]. Mathematically, this mechanism can be expressed as $P_{i\to j} \sim k_j^{\text{in}} f_j$ where f_j represents the fitness that is preassigned to each node prior to the growth of the network. The fitness can be estimated from the real degree time sequence of papers. The in-degree of paper i at time t is denoted as $k_i(t)$, and $C(t, \Delta t)$ represents the new citations that are added to papers during the next Δt days. The expected number of citations received by a paper i according to preferential attachment can be calculated as $\Delta k_i(t, \Delta t)_{\text{PA}} = C(t, \Delta t) k_i(t) / \sum_j k_j(t)$. If i actually receives $\Delta k_i(t, \Delta t)$ citations, the fitness (also referred to as relevance in certain papers) can be calculated as the ratio between the actual number of citations received and the expected number of citations that are received as follows [126]:

$$X_i(t, \Delta t) := \frac{\Delta k_i(t, \Delta t) \sum_j k_j(t)}{C(t, \Delta t) k_i(t)}. \tag{4}$$

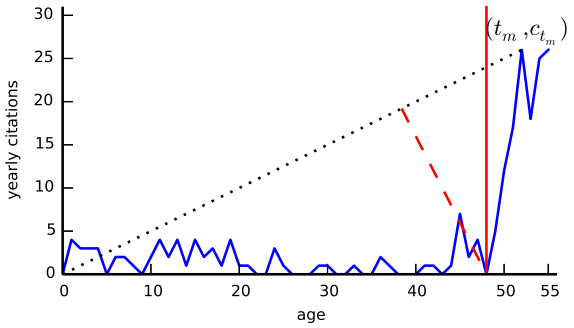

Fig. 4. Illustration of the sleeping beauty phenomenon and the definition of a coefficient quantifying the significance of this effect [129]. The blue curve represents the number of citations c_t received by the paper at year t after its publication. The black dotted line connects the points $(0, c_0)$ and (t_m, c_{t_m}), as a reference line. The awakening time t_a, marked by the red vertical line, is defined as the age that maximizes the distance from (t, c_t) to the reference line, indicated by the red dashed line.
Source: The figure is reprinted from Ref. [129].

This equation can be applied to estimate the fitness of a paper at different times. Interestingly, a decay of the fitness is observed in empirical data, which suggests that fitness is time-dependent [126]. Therefore, a more reasonable modification of the preferential attachment should read as follows:

$$P(i, t) = \frac{k_i(t)R_i(t)}{\Omega(t)}, \tag{5}$$

where $\Omega(t)$ is a normalization factor. Fitness takes the form of $R_i(t) = R_i(0)e^{-\beta(t-t_i)}$, where t_i represents the time when i enters the network and $R_i(0)$ values are drawn from an exponential distribution. The analytical solutions to this growing process indicate that the model can generate a degree of distribution with an exponential, log-normal or power-law decay, depending on the speed of the fitness decay. This model captures two key factors in the citation network growth, i.e., the heterogeneous fitness (relevance) and aging of the nodes (time decay). A mechanistic model that characterizes the citation dynamics of papers was inspired by this network growth model [32], which is discussed in detail in Section 4. The decay of fitness can be modeled in other forms. An alternative method is to assume the probability of receiving new links as $P(i \to j, t) \sim [k_j^{\text{in}} + A_j(t)]$, where $A_j(t)$ represents the fitness of paper j, which will decay with time $A(t) = A_0 e^{-(t-t_0)/\tau}$ [42]. This form can capture the bursty behavior of citation dynamics, i.e., the power-law distribution of $\Delta k/k$.

3.1.3. Sleeping beauty in science and its modeling

Sleeping beauty refers to the phenomenon of delayed recognition of scientific papers. This phenomenon was revealed by Garfield in the 1980s [127], and the term "Sleeping beauty" was used by van Raan to describe this phenomenon in 2004 [128]. Sleeping beauty refers to a type of paper that receives very few citations immediately after its publication but later experiences a sudden increase in the number of citations (referred to as the "awakening time"). A typical example of the sleeping beauty phenomenon is illustrated in Fig. 4 [129]. To quantify this effect, generally, three dimensions are measured, including the length of sleep, the depth of sleep and the intensity of waking [128]. However, one of the most challenging issues is identifying sleeping beauties in citation data. In early empirical studies, sleeping beauties are identified by certain specific criteria [33,128]. Therefore, the identified set of sleeping beauties varies significantly when different rules are adopted.

A metric referred to as a "beauty coefficient" was proposed by Ke et al. to quantify the significance of the sleeping beauty effect in individual papers [129]. The beauty coefficient is based on a reference line that begins with the publication date to the year when the paper receives its maximum number of yearly citations. This coefficient is a line that connects two points $(0, c_0)$ and (t_m, c_{t_m}) in the citation-time plane, as illustrated by Fig. 4. The coefficient is defined as the relative difference between the real citation curve and the reference line from $t = 0$ to $t = t_m$. The mathematical expression is

$$B = \sum_{t=0}^{t=t_m} \frac{\frac{c_{t_m}-c_0}{t_m} + c_0 - c_t}{max\{1, c_t\}}. \tag{6}$$

According to this definition, B increases with the length of the sleeping period and the awakening intensity. A higher value of B indicates that an individual paper is a more significant sleeping beauty. In addition, the framework provides a simple method to identify the time that a sleeping beauty wakens which can be computed as the age that maximizes the distance from the real citation curve to the reference line. The design of the beauty coefficient is followed by a comprehensive empirical analysis on the entire WoS database [129]. Focusing on papers that have beauty coefficients that lie in the top 0.1%, we can compute the percentage of papers regarding a specific topic. Most papers are regarding physics, chemistry and multidisciplinary sciences.

The sleeping beauty effect is not noticeable in the classic preferential attachment because it assumes that the probability that a node will receive new links is proportional to its cumulative citations. Prior to awakening, a sleeping beauty accumulates very few citations. These citations are predicted to be poorly cited by preferential attachments. A detailed empirical analysis regarding the citation history of a large number of physics publications indicates that the citation process cannot be described as a memoryless Markov chain because a substantial correlation exists between the present and recent citation rates of a paper [130]. In addition, a superlinear preferential attachment occurs. Accordingly, Golosovsky et al. proposed a network growth model that combines a paper's accumulated and recent citations [130]. The latent citation rate of a paper i is estimated as

$$\lambda_i = (1-c)A(k_i + k_0)^{\alpha} + c\Delta k_{i,t-1}. \tag{7}$$

where c is a parameter that can be fitted from real data, $A = 3.54/(t+0.3)^2$ and α is the exponent for superlinear preferential attachment. The final citation increase Δk of a paper should be drawn from a Poisson distribution with $\lambda = \lambda_i$ according to the empirical results,

$$P(\Delta k) = e^{-\lambda\Delta t}\frac{(\lambda\Delta t)^{\Delta k}}{(\Delta k)!}. \tag{8}$$

Because the latent citation rate of a paper partially depends on its recent citations, this model captures the sleeping beauty phenomenon better than the pure preferential attachment. Certain papers may initially have a small number of citations but suddenly became popular.

3.1.4. Multi-mechanism models

Citation dynamics is highly complex and cannot be described by only one mechanism. A model has been developed that combines two natural mechanisms of citation behavior and reveals an interesting "tipping point" property of citation dynamics [131]. The model includes a direct mechanism for when a paper randomly cites an older paper and an indirect mechanism for when a paper cites an older paper by finding it in the reference list of a newer intermediary paper that cites the older paper. The direct mechanism is easy to describe; the probability that an old paper will be cited by a new paper is calculated as $r_{\text{direct}} = 1/N$ where N represents the number of existing papers in the network. The indirect mechanism computes this probability as $r_{\text{indirect}} = k/(nN)$ where n represents the reference number of each new paper (assumed to be a constant), and k represents the number of citations that the old paper has accumulated. These two mechanisms are then combined with a parameter c to obtain the aggregated probability of an old paper being cited,

$$R(k) = \frac{n(1-c)}{N} + \frac{kc}{N}. \tag{9}$$

By solving this equation, the expression for the final citation distribution can be reached,

$$p(k) \approx \frac{(\alpha+1/c)^{\alpha+1/c}}{(\alpha c+1)(\alpha-1)^{\alpha-1}}\left(\frac{\alpha-1+k}{\alpha+1/c+k}\right)^{\alpha+k} \times (\alpha-1+k)^{-1}(\alpha+\frac{1}{c}+k)^{-1/c}, \tag{10}$$

where $\alpha = \frac{n}{c} - n$. If one focuses on the tail of the distribution, i.e., the papers with large k ($k \gg \alpha$), the probability function reduces to

$$p(k) \approx \left[\frac{(\alpha+1/c)^{\alpha+1/c}e^{-(1+1/c)}}{(\alpha c+1)(\alpha-1)^{\alpha-1}}\right]k^{-(1+1/c)}. \tag{11}$$

These two equations indicate that once a paper has a sufficient number of citations, its future citations will rapidly increase because there are numerous ways to find the paper through its citing papers. This type of paper is referred to as a "classic" paper. The critical citation k that determines a classic paper is referred to as a tipping point and is equal to $k = \alpha$. According to Eq. (11), the citation distribution of classic papers becomes a power-law with exponent $\gamma = 1 + 1/c$. Papers that have fewer citations than the tipping point experience much slower accumulations of future citations.

The two-mechanism model was utilized to analyze three actual databases by fitting two key parameters, n and c. The databases include the WoS database in 1981, a database that includes a 2007 list of the living highest h-index chemists, and a database that includes all Physical Review D papers that were published from 1975 to 1994. The tipping points were found to be 25, 37 and 31, which indicate different difficulties in getting into the classic paper club. Furthermore, the models analyze the citation data of authors with different h-indices. The tipping points and power-law exponents are remarkably different, which indicates that the power-law exponent is not a universal feature of all scientific citations [131].

A similar tipping point is detected in another study that investigates the influence of authors' reputations on their papers' popularity [132]. Specifically, a citation crossover $c_x \approx 40$ is observed when one notes the future citation of a publication $\Delta c_p(t+1)$ rather than its cumulated citations $\Delta c_p(t)$. The preferential attachment mechanism breaks down for papers with $c < c_x$. To understand this empirical observation, a model that combines the preferential attachment with aging and authors' reputation was proposed as follows:

$$\Delta c_{i,p}(t+1) = \eta \times \Pi_p(t) \times A_p(\tau) \times R_i(t), \tag{12}$$

where $\Pi_p(t)$ is the preferential attachment term with $\Pi_p(t) = [c_p(t)]^\pi$, $A_p(\tau)$ is the decay term with $A_p(\tau) = e^{-\tau_p/\bar{\tau}}$, and $R_i(t)$ is the author reputation term with $R_i(t) = [C_i(t)]^\rho$. Here, p and i respectively represent a publication p and one of its authors i. An author's citation is obtained by summing all his/her papers' citations, i.e., $C_i(t) = \sum_p c_{i,p}(t)$. τ_p represents the age of paper p. For each author, the model assigns i three parameters: π_i (super-linearity in preferential attachment), $\bar{\tau}_i$ (papers' average life cycle) and $\bar{\tau}_i$ (author's reputation effect). By fitting these three parameters with a multivariate regression using WoS data that includes 450 highly cited scientists, a significant difference can be observed between authors' papers with $c < c_x$ and $c > c_x$. Generally, π_i is much smaller for $c < c_x$ (much smaller than 1), but ρ_i is much larger for $c < c_x$ (approximately 0.2), which indicates that poorly cited papers do not follow preferential attachment, and the author's reputation dominates the annual citation rate for papers with few citations [132]. Clearly, the value of π_i and ρ_i vary for authors.

3.2. Team formation and the evolution of collaboration networks

For scientific endeavors, collaboration is the most effective way to incorporate individuals with different ideas, skills, and resources. Empirical studies have established that the percentage of coauthored papers is growing when compared to single-authored papers, which indicates the increasing importance of scientific collaboration for innovation [133]. Studies regarding scientific collaborations have a long history and have recently attracted even more attention. In the following section, we summarize studies regarding the dynamic properties of collaboration networks and their modeling and mechanisms for scientific team formation.

3.2.1. Modeling scientific collaboration networks

Similar to other networks, scientific collaboration networks follow preferential attachment in growth [134]. The probability of collaboration between two scientists is positively correlated with their number of mutual acquaintances in collaboration networks, which indicates that a new collaboration tends to form triangles in the network. Furthermore, this probability is roughly a linear correlation with the number of prior collaborations and the number of prior collaborators, which supports the preferential attachment assumption for the growth of collaboration networks.

Similar conclusions have been reached in prior studies [135]. By analyzing collaboration networks that are constructed from a database that includes all relevant journals related to mathematics and neuroscience from 1991 to 1998, this study confirms that network evolution is governed by preferential attachment. The results indicate that the average shortest path length and the clustering coefficient decrease with time, but the average degree increases gradually. In contrast to citation networks and the BA network model, where new links are only created between each new node and existing nodes, new links in collaboration networks may also connect nodes that are already present in the network. These two types of links are referred to as external links and internal links. A model was proposed to reproduce the observed features [135]. In this model, the total number of links and the total number of nodes at time t are denoted as $T(t)$ and $N(t)$, respectively. New researchers join the field at a constant rate, which leads to $N(t) = \beta t$. By definition, the average degree of nodes is $\langle k \rangle = T(t)/N(t)$. During the network growth, the probability of two existing nodes i and j connected by an internal link can be expressed as

$$\Pi_{ij} = \frac{k_j k_j}{\sum_{s \neq m} k_s k_m} 2N(t)a, \tag{13}$$

where a is the number of newly created internal links per node in each step. Nodes join the network at a constant rate with a new node connects to an existing node i according to the probability

$$\Pi_i = b\frac{k_i}{\sum_j k_j}, \tag{14}$$

where b is the average number of new links that an incoming node creates. The model can be solved analytically. The degree distribution is a truncated power law, with $P(k) \sim k^{-3/2}$ for $k \ll k_c$ and $P(k) \sim k^{-3}$ for $k \gg k_c$. The crossover point is $k_c = b\sqrt{t(2+\alpha t)^3}$. Monte Carlo simulations of this model indicate that the empirically observed patterns of the average degree, the average shortest path length, and the clustering coefficient are reproduced by this model. In addition, this model is extended to a nonlinear preferential attachment case where external links attach to old nodes with $\Pi_i = bk_i^v/\sum_j k_j^v$ ($v \approx 0.8$ according to empirical results). However, nonlinearity has no distinguishable effect on $P(k)$. One possible explanation is that there are more internal links than external links; therefore, internal attachments dominate the system's behavior. A similar model with nonlinear preferential attachment of external links was analyzed, with a special focus on the effect of the nonlinearity parameter v on the network structure [136]. The results demonstrate that tuning v could lead to four different categories of final degree distribution: exponential, non-power law, semi-power law, and power law.

Scientific collaboration networks are weighted in nature. To simulate the weight evolution during network growth, a weighted network model was developed [137]. The model begins with a fully connected network that includes n_0 nodes. In each step, a new node is added to the network and l number of old nodes are randomly selected. Each of these $1 + l$ nodes (denoted as set n) will establish m links, with the probability for each link connecting to an existing node i as

$$\Pi_{n\to i} = (1-p)\frac{k_i}{\sum_j k_j} + (p-\delta)\frac{w_i}{\sum_j w_j} + \delta\frac{l_{ni}}{\sum_{j\in\partial_n^d} l_{nj}}, \tag{15}$$

where k_i is the degree of node i, w_i is the node strength of i and l_{ni} is the similarity distance from n to i: $l_{ni} = w_{ni}$ if n and i are connected, $l_{ni} = (w_{ns}w_{si}/(w_{ns} + w_{si}))$ if n and i are connected through two links ns and si. ∂_n^d is the set of n's neighbors within distance d. The number of times that two nodes are connected is denoted as T_{ij}, and the weight of the link between two nodes is computed as $w_{ij} = f(T_{ij})$, where several different functions can be applied, such as the tanh function or the linear function. In addition, this model can be solved analytically. The distribution of node strength is expressed as $p(w) \sim (w + 2ml)^{-3}$. The link weight distribution obeys the power-law, which is consistent with actual scientific collaboration networks. Two parameters p and δ control the weight between different growth mechanisms. The factor p adjusts the weight between the degree preferential attachment and the node strength preferential attachment. The term δ in the model uses local information when nodes establish new connections. Adding more weight on this term increases the clustering coefficient but does not have an effect on the distribution of degree and node strength [137].

3.2.2. Evolving bipartite author-paper networks

Collaboration networks between scientists are actually projections from author-paper authorship networks. As bipartite networks, author-paper authorship networks preserve more information regarding scientific collaborations [138]. One example is that a fully connected graph of four scientists can be formed by one paper that is coauthored by all the authors or numerous papers that are coauthored by at least two of the authors. These two examples appear the same in collaboration networks that only include author nodes; however, they are distinguished in bipartite author-paper networks. Therefore, it is meaningful to understand and model the dynamics of scientific collaborations by using author-paper bipartite networks.

Author-paper bipartite networks can be generated by using a preferential edge attachment mechanism that can describe numerous topological characteristics of real monopartite collaboration networks (networks that only have author nodes) [139]. The model is based on a growing process. For each step, a paper enters the network with n authors. n can represent either a constant or random variable that is distributed with a specific distribution $S(n)$. m of these authors are new authors that have not published any prior scientific publications. The remaining $n - m$ authors are selected from the existing author nodes with a probability that is proportional to their degree q in the network. The degree of an author in bipartite network represents the number of papers that an author has published. If the network is sufficiently large, the number of an author's collaborators k can be computed as $k = q(\bar{n} - 1)$, where $\bar{n}$ represents the mean of n. The symbol k represents the degree of the authors if one projects the bipartite network to a monopartite network. The distribution of k is $P(k) \sim [k+(\gamma - 1/2)(\bar{n} - 1)]^{-\gamma}$, where $\gamma = 2+\bar{m}/(\bar{n} - \bar{m})$, following a power-law distribution. The clustering coefficient of nodes with degree k can be solved as $c(k) = (\bar{n} - 2)/(k - 1)$. These two solutions in general overlap with the results in real data. These solutions can be improved by considering the aging effect in the model. This modification generally ensures the survival of each author until a certain age (measured by q) and a subsequent exponential decay. In addition to the degree distribution and clustering coefficient, aging alters the type of degree-degree correlations, i.e., the network is a disassortative mix when aging is not considered but is an assortative mix when aging is considered. According to empirical data, actual collaboration networks are an assortative mix, which indicates the importance of considering aging in the model.

A similar model considered research groups in the generation of author-paper bipartite networks [47]. The bipartite network grows with each new paper that is created in each step. In addition to this new paper, a new author group with N_g all new authors is created with probability α. The authors in this new paper include a first author plus a Poisson-distributed number of additional coauthors that are selected from the new group. If a new group is not created, an existing group will be selected with a probability that is proportional to the number of papers that have been published by the authors in this group, which is referred to as the Yule process. The number of coauthors follows a Poisson distribution. The authors are selected with probability $1 - \beta$ in the existing group that has been selected. The authors in this group are preferentially selected according to the number of papers they have published. With probability β, the coauthors are randomly selected among all the authors in the network, regardless of whether they have authored a paper or not. By tuning the parameters, this model can reproduce the characteristics of six metrics for actual data, i.e., authors per paper distribution, papers per author distribution, coauthorship per author pair distribution, collaborator per author distribution, coauthor clustering coefficient distribution and minimum path between author pairs distribution. Coauthor clustering coefficient distribution and minimum path between author pairs distribution are measured in projected monopartite collaboration networks.

When modeling networks include author and paper nodes, the real challenge is capturing the simultaneous growth of more than one network structure, i.e., coauthorship and citation networks. The evolution of these networks mutually affects each another. Therefore, these two networks should not be modeled separately. A framework model that considers the interactions between topics, aging, and recursive follow-up of links (referred to as the TARL model) was proposed as an attempt to describe the simultaneous growth of coauthor and paper citation networks [140]. The model begins with a set of authors and a set of papers regarding randomly assigned topics. Subsequently, a predefined number of coauthors who share the same topic is randomly selected and assigned to each paper, which ensures that all papers are assigned to authors but certain authors may have no paper. For each step, the author set is updated by adding new authors and subtracting certain existing authors. Then, each author randomly identifies a set of coauthors and produces a specified number of new papers. Each new paper cites a fixed number of existing papers according to a local rule. The references can be selected from a specified number of randomly selected papers P_0 regarding the author's topic and the r-level references of these papers. If $r = 2$, the reference set includes P_0, any paper P_1 that is cited in one of the P_0 papers, and any paper P_2 that is cited in any of the papers in set P_1. The aging effect can be imposed when these papers are selected as references, which compensates the rich-get-richer effect that favors the old and highly cited papers. The model stops when the number of specified time steps

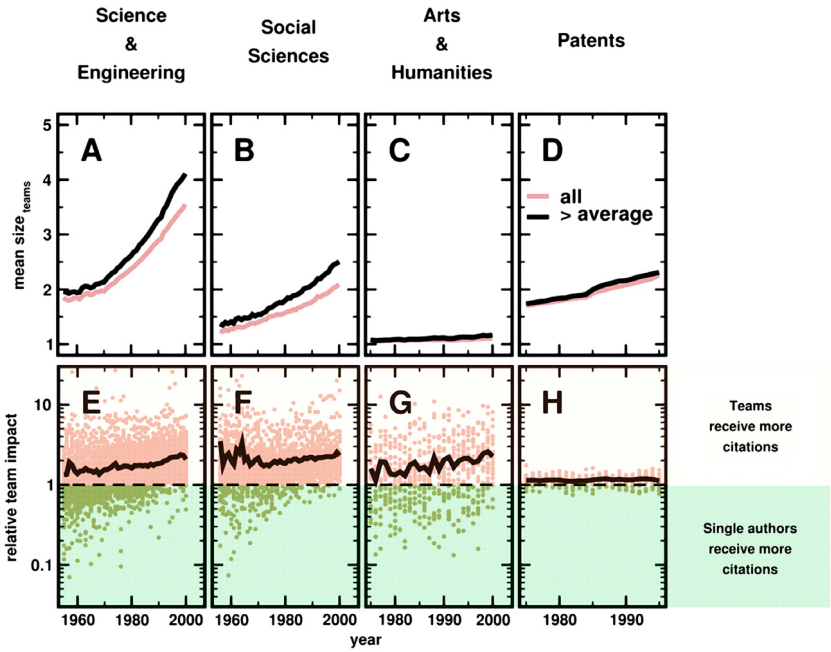

Fig. 5. (A–D) Mean team size comparing all papers and patents with those that received more citations than average in the relevant subfield. (E–H) The RTI, which is the mean number of citations received by team-authored work divided by the mean number of citations received by solo-authored work. A ratio of 1 indicates that team- and solo-authored work have equivalent impact on average.
Source: The figure is reprinted from Ref. [133].

has been reached. The model was calibrated and validated against a 20-year dataset of articles that have been published in PNAS. The model reproduced the small-world properties of coauthorship networks and the power-law citation distribution of citation networks. This study was followed by a number of extensions to better capture the co-evolution of coauthor and paper citation networks [141,142].

3.2.3. Team assembly mechanisms

In contrast to the past, when an individual genius played a significant role in scientific discovery, teamwork is becoming increasingly important for modern science. Investigating collaboration networks is an effective method for understanding this shift, yet a more direct method is to focus on each individual team formation. In regard to scientific publication data, a team is defined as the coauthors of a paper. In regard to patent data, a team can be similarly defined. By studying 19.9 million research articles in the WoS database and 2.1 million patent records, Wuchty et al. determined that the percentage of papers and patents that were written by teams and the mean team size increased with time, which indicates a shift toward teamwork in science and engineering [133]. Specifically, the mean team size increased from 1.9 to 3.5 authors per paper from 1955 to 2000 and increased from 1.7 to 2.3 inventors per patent from 1975 to 2000. In addition, the citations of papers and patents were used to investigate the impact of teamwork. By comparing papers and patents with multiple authors and inventors to papers and patents that have a sole author and inventor, a strong signal favoring teamwork was detected, as illustrated in Fig. 5. These phenomena can be explained by the increasing scale, complexity, and costs of big science [133].

In addition, a team can be defined at higher levels according to the affiliation of the scientific publication. Papers that are submitted by more than one university are referred to as multi-university collaborations [143]. The percentage of multi-university collaborations has considerably increased from 1975 to 2005. The share of single-university collaborations has remained fairly constant over time, and the percentage of sole-author papers continues to decrease. Despite an increase in multi-university collaborations, the geographic distance between collaborators has only slightly increased. More patterns were observed by decomposing universities into schools and categorizing schools according to the total number of citations received by their papers during the corresponding period. If a top-tier school is included, generally, the collaboration results in a high-impact paper. However, stratification in collaboration was also observed. Lower-tier schools can reach across university boundaries, yet they tend to interact within their own tier, which replicates the in-group status-matching behavior of elite schools.

At the country level, the percentage of publications that are a result of multi-national collaborations have steadily increased over time [144]. Examining 14 million publications using WoS data from 2000 to 2009 indicates that the US, the UK, Germany, France, Italy, and Canada produce 82% of the multi-national publications and China, Japan, and South

Korea mostly cooperate. Another study considered the relative percentage of multi-national publications across different disciplines (measured by a well-defined z-score) [145]. A notable result is that although the relative percentages are remarkably different across disciplines, the data indicates an obvious trend regarding the convergence of internationalization across different disciplines. Recently, Hsiehchen et al. examined the influence of international research teams on citation outcomes [146]. They found that additional authors and countries are associated with higher aggregate citation rates. When measuring per capita citation rates, an increasing number of countries and authors lead to contrary effects. An increasing number of countries has a persistent additive citation effect, but an increasing number of authors in large research teams tends to decrease per capita citation rates.

The studies discussed above note that the size of research teams has significantly increased. However, an empirical study regarding the distribution of research team size indicates that a recent distribution of research team size reported a shift toward a larger value for the old distribution of research size and that the shape has fundamentally changed [147]. The size of research teams from 1961 to 1965 follows a simple Poisson distribution; however, a power-law tail was noted for recent team size distributions from 2006 to 2010. This phenomenon can be explained by using a simple model. The model describes how authors write papers over time. Each paper has a lead author who is associated with two types of teams: "core" and "extended". Core teams include the lead author and coauthors and have a Poisson size distribution. However, extended teams evolve constantly. Beginning with the same members as the core team, new members are added to extended teams in proportion to the aggregate productivity of its current members. The rich-get-richer mechanism implies that teams that have initial large core teams or highly productive members generally attract new members. The lead author can either publish with the core team or an extended team at any time, with a fixed probability $p_{\text{ext}} = 0.3$. This model demonstrates that team formation is a multi-modal process. The core team, as a primary mode, leads to team sizes that represent the typical number of researchers needed to produce a research paper. The extended team, as the secondary mode, results in a large team size for research that requires expertise or resources from outside of the core team.

Various models have been developed to better understand team formation in science fields. Guimerá et al. proposed a model that reveals that team assembly mechanisms determine the structure of the collaboration network and team performance [148]. The model begins with an infinite number of newcomers who become incumbents after they are selected for a team for the first time. A new team is assembled at each step t and the collaboration network is updated accordingly. The new team includes $m(t)$ agents and each agent has a probability p of being selected from the incumbents or otherwise, of being selected from newcomers. In the first case, if there another incumbent is already a member of the team, with probability q the new agent is randomly selected among the collaborators of randomly selected existing incumbents on the team; otherwise, the agent is selected at random among all incumbents in the network. Once the team is formed, team members are connected to each other in the collaboration network. This model also considers the aging effect. Agents that remain inactive for longer than τ steps are removed from the network, which ensures that the network is able to maintain a steady state after a transient period. The model was used to study a tradition problem in percolation theory, i.e., the relative size of the giant component S in the network. As noted in Ref. [148], a large connected cluster would be a supporting evidence for the invisible college, a web of social and professional contacts that link scientists across universities. Conversely, a large number of small clusters would indicate that the field includes isolated schools of thought. Notably, a second-order phase transition that exists between S and p was observed in the model, as exhibited in Fig. 6. The phase transition occurs with an increasing number of loops as p increases. In addition, the model helps us understand collaboration networks in different fields and journals by estimating the parameters (e.g., m, p, q, and S) from actual data. Generally, teams that publish in journals that have a high-impact factor typically results in a large giant component, but teams that publish in low-impact journals typically form small isolated clusters.

3.3. Statistical trends encoded in the SOS data

A comprehensive analysis of SOS data revealed various statistical trends. These trends not only enhance our understanding regarding the history of the development of science but also highlight certain policies that lead to a healthier environment for scientific research. The number of citations received by a paper roughly measures its impact. An empirical study that focuses on 1% the most highly cited papers in different journals revealed that their share of total citations increases over time. This trend indicates that although more papers are published, scientists only read and cite a limited number of them [149]. Based on the citations evolution data, another study investigated future citations of papers with certain acquired citations [150,151]. The results indicate that a scaling relationship exists between future citations and currently acquired citations, with a larger exponent indicating a stronger trend in tracing hot topics. The study compares the citation data for papers that are written by authors from several major countries. The phenomenon of tracing hotness occurs more in China than other countries.

The dynamics of scientific publications has been used to quantify the development of physics as a discipline during the past century. A representative study was conducted by Sinatra et al. [34]. In the WoS data spanning over 100 years, the papers published in 242 physics journals are defined as core physics papers. Interdisciplinary physics papers are defined as papers that are connected to the core by references or citations, each of which must fulfill a constraint that the observed percentage of references and citations to core physics papers is significantly larger than the expected random value. The remaining papers are denoted as non-physics papers. This definition leads to several primary results. The percentage of physics core papers has slowly increased over time, but the percentage of interdisciplinary physics papers has remained

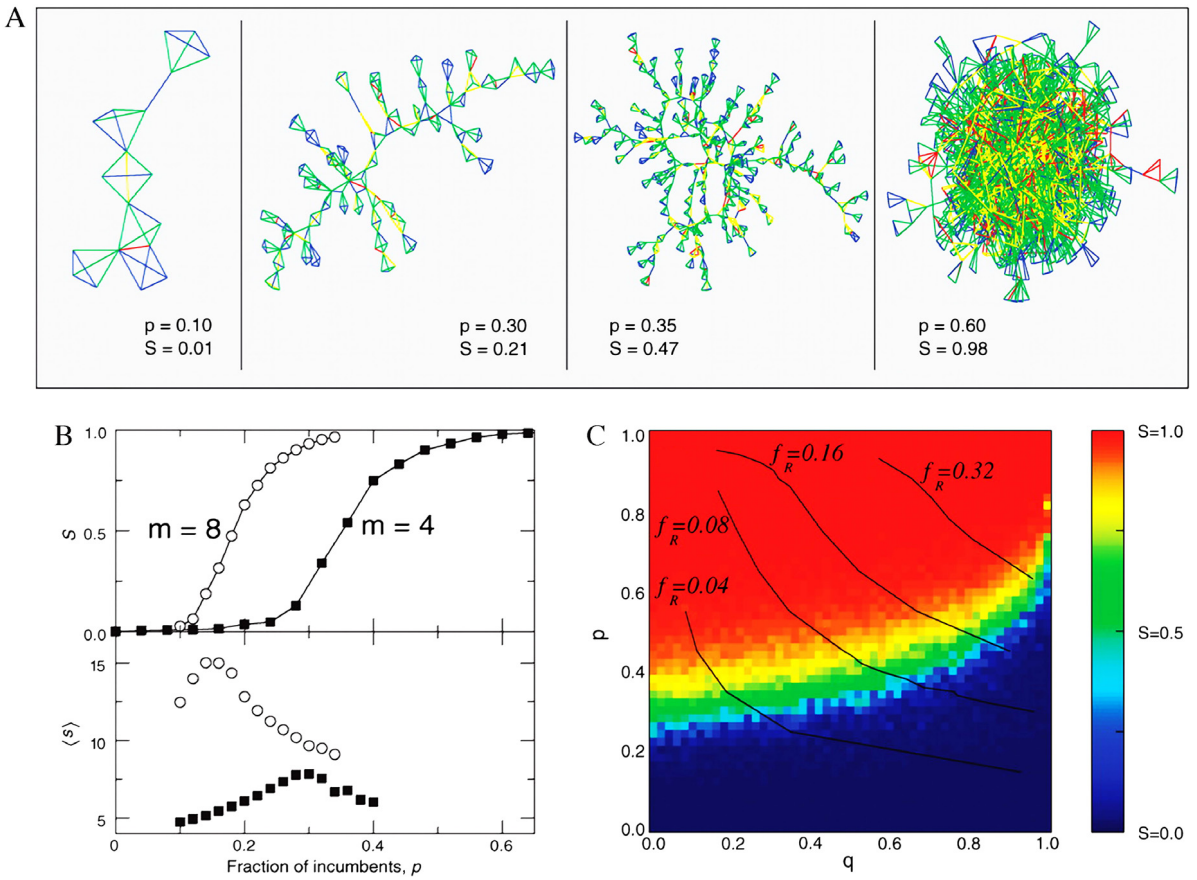

Fig. 6. (A) The structure of the giant component of the collaboration network under different values of parameter p. (B) The dependence of the relative size of the giant component of the collaboration networks S and mean size of the other clusters $\langle S\rangle$ on p. (C) The value of S in the parameter space $[p, q]$. *Source:* The figure is reproduced from Ref. [148].

stable. In contrast to the decade 1950–1960 when core physics paper were heavily self-cited, this self-referential feature of core physics significantly decreased during 1990–2000. Core physics and interdisciplinary physics papers are separated into categories by PACS codes. By constructing a citation network between these categories as illustrated in Fig. 7, the impact and longevity of the corresponding physics subfields are revealed. Certain other studies have also been devoted to understanding the evolution of physics. Based on APS journal data, Pan et al. constructed a network between sub-fields (classified by PACS codes) of Physics, with a link indicating that two PACS codes coexist for papers [152]. The network exhibits a core–periphery organization; the core includes PACS from Condensed Matter and General Physics. In addition, the data indicates a trend of increasing interactions between the different sub-fields over time. Using the publication data for APS journals, Perc considered trends of physics discovery and determined that the upward and downward trends follow a burst behavior, which indicates self-organization [153].

The evolutionary trend of research fields has been analyzed using the community detection technique. Herrera et al. constructed a scientific concept network where nodes represent PACS codes and two nodes are linked if they coexist in APS papers. Then, a community detection algorithm was applied in this network to identify scientific fields [154]. By analyzing the time evolution of these fields using data from 1985–2006, Herrera et al. found that long lived communities tend to have more PACS codes and more papers. Shirazi et al. constructed a shared-word network for medical sciences where two words are linked if they co-appear in the abstract of a paper. Then, the k-clique percolation method was used to identify research communities within the network [155]. The analyses reveal trends of collaboration and splintering among different disciplines in medicine. Sun et al. built a conference network with nodes representing science conferences and links representing that there are same authors publishing papers in two conferences [156]. Different scientific fields are discovered by using a community detection algorithm. The results highlight a trend of interdisciplinary research and indicate the type of authors who are most likely to perform research in this direction. In addition to these studies, the evolution trend of research fields has been studied by using the analogy of biological evolution [157].

Certain trends have been identified at the journal and paper level. The impact factors of scientific journals were analyzed from the dynamic perspective [158]. Scaling laws were indicted in the distribution of impact factors and their growth rates,

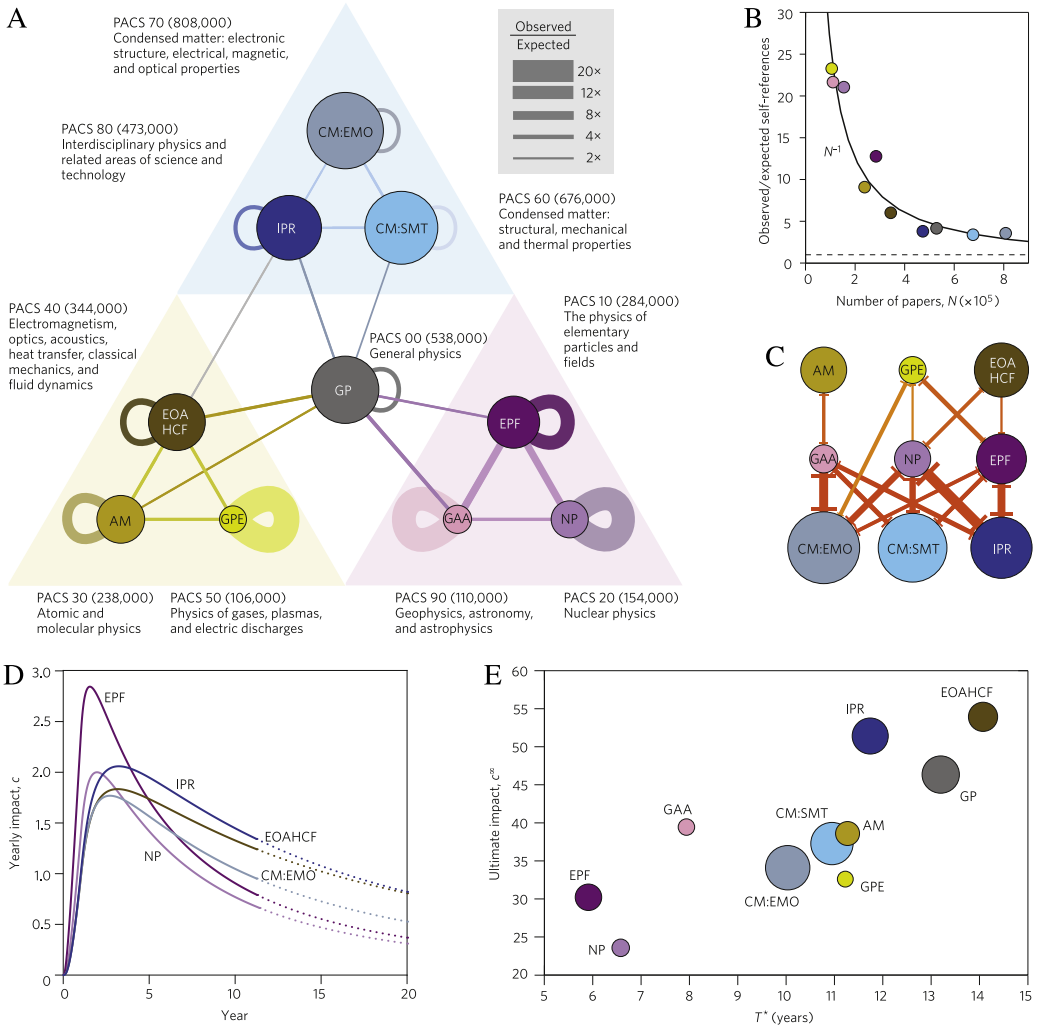

Fig. 7. (A) The citation networks between the different physics subfields. (B) The over/expected self-citation tends to decrease with the size of the subfield. (C) The citation barriers between different physics subfields. A barrier indicates that two areas have a significantly smaller number of citations than expected by chance. (D) The yearly citations of papers published in 2000 in each subfield. (E) The total number of citations a typical paper in each subfield receives versus the typical time over which a paper collects its citations.
Source: The figure is reprinted from Ref. [34].

which are characterized by a model that extends from the simplest model for firm growth [159]. For specific journals, the dynamic trends of papers can be extremely different. A case study regarding the evolution of papers' popularity evolution was conducted for the journal "Europhysics Letters" [160]. Generally, papers are frequently downloaded during the first couple of months after publication and then experience a decay of attention. However, the trajectory of individual papers can be classified in various manners. For example, various effects are detected including the bursty accumulation of downloads, the delayed recognition of papers' value, and the different age of attractiveness of open-access and non-open-access papers. In addition, studies have analyzed coauthorship and citation patterns in Physical Review journals [105]. Certain expected empirical findings are reported such as the exponential growth of papers, high self-citations and citations among coauthors. Interestingly, a strong tendency toward reciprocal citations is noted; however, a weak triadic closure effect is detected for coauthorships, which differs from numerous prior empirical studies.

At a more microscopic level, scientific memes were defined and extracted in scientific literature [161]. As defined in Ref. [161], a scientific meme is a word or a phrase that appears in a paper and is replicated in its citing publications. This definition compares scientific memes to inheritance mechanisms, which supports the ability to quantify the significance of scientific memes by using a diffusion process. Accordingly, a propagation score p_m is developed to evaluate whether a certain text in the title and abstract can be regarded as a significant scientific meme. Generally, p_m is high if meme m frequently appears in publications that cite publications containing the meme, but rarely appears in publications that do not cite a publication without the meme. This mathematical expression rules out certain common terms such as "method". An

interesting temporal pattern is that top memes exhibit bursty dynamics in their propagation scores, which is interpreted as a rapid rise and fall in their popularity because of the emergence of new scientific paradigms.

4. Quantification of scientific significance

4.1. Quantification of the influence of scientific publications

As the number of scientific publications exponentially increases over time, it is almost impossible for individuals to review all of the published papers in order to find the important ones in certain fields. Therefore, designing methods to automatically quantify the impact of scientific publications based on citation data is an important research topic from both theoretical and practical viewpoints. However, this is a challenging problem as various mechanisms co-determine the dynamics of papers' citations and their positions in citation networks. To date, many metrics have been introduced to address this problem. In this subsection, we review the key methods used to quantify the influence of scientific publications, particularly focusing on those methods developed from the perspective of complexity science.

4.1.1. Citation and its variants

The most straightforward way to compare paper impact is to use the citation count [36]. This metric is commonly adopted by scientists. In a citation network, the citation of a paper i is simply the in-degree of the corresponding node $c_i = \sum_j A_{ij}$ where A_{ij} is an element in the adjacency matrix of the citation network, representing a link pointing from node j to node i.

Despite its simplicity and wide usage, the drawback of using the citation count is obvious. The citation distributions can differ widely among disciplines. Papers in some fields (e.g., physics) are typically cited much more often than papers in other fields (e.g., mathematics). Therefore, it is unfair to use citation counts to directly compare papers from different disciplines. To address this problem, Radicchi et al. proposed a relative index $c_f = c/c_0$ and found that the distribution of c_f of papers in various disciplines in the same year overlap in a single curve that resembles a log-normal distribution [31]. Here, c_0 is the mean number of citations received by the papers within a discipline published in the same year. Additionally, similar rescaling can be observed for papers in the same discipline but in different years. These features indicate that c_f can be used as an unbiased metric to compare the scientific impact of papers across fields and years. This work is followed by numerous extensions. c_f was later used to rescale the citation distribution of papers from different academic institutions, journals and even time [162]. In these cases, c_0 is adopted as the mean citation number of papers from an institution, a journal or a time period. Similar universality is observed for the rescaled citation distribution. The c_f indicator was also applied to the publications in APS journals which cover all the subfields of physics. The PACS code was used to identify papers that belong to different fields [163]. Again, the significant difference in the citation distribution across different fields can be avoided when the distributions of the relative index c_f in these fields are compared. In addition, the minimum value of the rescaled citation count $c_f(q)$ for a paper to be in the top q% most-cited papers has been derived theoretically [164].

Although the aforementioned rescaled method c_f can effectively remove the bias of comparing papers from different disciplines, the problem remains when comparing papers from fields that are not well defined. In this case, the c_0 of a field is difficult to estimate. This problem was addressed with an index called the relative citation ratio (RCR) in which a paper's co-citation network was introduced to field-normalize the number of citations it has received [165]. The co-citation network of a paper i is defined as the network that consists of papers that have been cited by the papers that also cite i. To calculate the relative citation ratio of a paper i, one has to compute the article citation rate (ACR) and field citation rate (FCR) of i. ACR is simply the total number of citations of i divided by the number of years since publication. In principle, FCR can be obtained by averaging the ACR of papers in i's co-citation networks. However, this value can be highly vulnerable to finite-size effects. Therefore, FCR is computed by averaging another index called the journal citation rate (JCR) which averages the citation rates of the journals represented by the collection of articles in i's co-citation network. For normalization, FCR is rescaled to an expected citation rate (ECR). Finally, RCR can be obtained by dividing ACR by ECR. The relative citation ratio is field independent and can be used to compare any two papers in science. The method is eventually validated with database recording experts' rating scores on papers.

As discussed in the previous section, the citation networks exhibit first-mover advantage. The earliest papers in a field tend to have more citations than papers that are published later. To remove the temporal effect when comparing the citations of papers published at different times, Newman proposed ranking papers based on a z-score [117]. First, the mean number of citations and the standard deviation are computed for papers published in the same time window. The z-score is then obtained by calculating the number of standard deviations by which that paper's citation count differs from the mean. Recent yet high quality papers usually have only been cited a few times, they may still have a high z-score because they accumulate more citations than the average among their peers. In this sense, this z-score is also effective for predicting papers that will be highly cited in the future [166], which we will discuss again in the next section. Essentially, the z-score shares similar idea to c_f, with only a small difference in their mathematical expressions.

Though citation count and its variants are widely adopted, they have been pointed out to be poor proxies for a paper's quality because a citation might result from different reasons [167]. In some cases, a paper could be cited not because it is seminal and relevant, but because it has some error (e.g. cited by comment papers) [168] or an author insists on its citation (e.g. self-citations) [169]. Therefore, it is important to distinguish the context of each citation. There are already some efforts in the literature devoted to address this issue. These works consider that not all cited papers are equal to the citing

paper [88,90,170–172]. Methods are proposed to filter out less relevant citations, which results in a significant reduction of the citation network. Gualdi et al. computed the structural similarity between the citing and cited papers and obtained a tree-like backbone by keeping a most similar cited paper for each paper [88]. Clough et al. took into account the constraints of causality and revealed the fundamental causal skeleton of the citation network by removing the unnecessary links for the flow of information between papers [90]. To mine the key references in a paper, Zhu et al. conducted automatic feature selection with a supervised machine learning and found the best feature as the number of times a reference is mentioned in the body of a citing paper [170]. By giving these key citations more weight, they further designed an influence-primed h-index for evaluating scientists. Valenzuela et al. modeled the task of identifying important citations in scholarly literature also by a supervised classification problem [171]. They eventually classified citations into four types including related works, comparison, using the work, extending the work. A scientific search engine, Semantic Scholar, was established based on the algorithm to automatically identify the subset of a paper's citations in which the paper had a strong impact on the citing work [172].

4.1.2. PageRank and its variants

PageRank. The impact of a paper depends not only on how many citations it receives, but also which papers cite it. Implementing the well-known PageRank algorithm in citation networks is a commonly used approach to capture this structure factor in ranking scientific publications. PageRank is a famous ranking algorithm designed by computer scientists and forms the basis of the Google search engine, ranking the importance of webpages [173]. Running in a citation network, this algorithm assigns a score s to each node which is updated in each iteration step by sending its own score evenly to downstream neighbors and meanwhile receiving scores from its upstream neighbors. A paper obtains a higher score if many other important papers cite it. The final stable scores are used to indicate the significance of a paper. From a physical perspective, PageRank describes a random walk process on a directed network, where the score is proportional to the frequency of visits to a particular node by a random walker. Unlike the traditional random walk process, PageRank uses a parameter c ($0 < c < 1$) called the return probability or damping factor. With probability c, a random walker will jump to a random node; otherwise the random walker continues walking through the directed links. The PageRank process can be expressed by a single iterative equation:

$$s_i(n) = c + (1-c)\sum_{j=1}^{N}[\frac{A_{ij}}{k_j^{out}}(1-\delta_{k_j^{out},0}) + \frac{1}{N}\delta_{k_j^{out},0}]s_j(n-1), \tag{16}$$

where N is the network size, $\delta_{a,b} = 1$ when $a = b$, and $\delta_{a,b} = 0$ otherwise. Initially, a random walker is assigned to each node, namely $s_i(0) = 1$ for $i = 1, 2, \ldots, N$. The typical value of the return probability is approximately 0.15 in computer science [173] and 0.5 in the science of science [174]. The final score of each node is defined as the steady value after the convergence of $s_i(n)$. In the literature, Chen et al. directly applied the PageRank algorithm to assess the relative importance of all publications in the Physical Review family of journals from 1893 to 2003 [174]. Ma et al. also directly used PageRank to measure the importance of scientific papers in Biochemistry and Molecular Biology [175]. Moreover, numerous studies have developed modified PageRank algorithms that are more suitable for ranking papers in citation networks. In the following, we review some representative variants. For more complete generalizations and applications of PageRank, readers are encouraged to refer to two recent review articles [176,177].

CiteRank. To have a high PageRank score, a paper has to be cited by many highly cited papers, which generally favors older papers. To account for the strong aging characteristics of citation networks, Walker et al. proposed the CiteRank algorithm to improve the rank of recent papers [178]. Similar to PageRank, CiteRank is also based on a propagation process in a citation network. First, a transfer matrix is defined in CiteRank, with each component as $W_{ij} = 1/k_j^{\text{out}}$ if j cites i and $W_{ij} = 0$ otherwise. The initial distribution of the score on nodes is recorded in a vector $\vec{\rho}$ with each component $\rho_i = e^{-age_i/\tau_{dir}}$ where age_i is the number of years since the publication of i and τ_{dir} is a parameter that determines the characteristic decay time. The CiteRank score of nodes can be obtained by iteratively computing

$$s_i(n) = \alpha\sum_{k_j^{out}>0} A_{ij}\frac{s_j(n-1)}{k_j^{out}} + \alpha\sum_{k_j^{out}=0}\frac{s_j(n-1)}{N} + (1-\alpha)\frac{e^{-(t-t_i)/\tau_{dir}}}{\sum_j e^{-(t-t_j)/\tau_{dir}}}, \tag{17}$$

where α is similar to the return probability c in PageRank. Owing to the decay effect, old papers have less initial score; thus, those papers cited by many old papers tend to have a small final $\vec{T}$ score. By applying CiteRank to a dataset from the "high energy physics theory" archive (hep-th) and APS data, studies have shown that this method is able to give more recent papers a higher rank than PageRank. The Spearman rank correlation coefficient between recent citations accrued (Δk_{in}) and CiteRank score $\vec{T}$ is computed in the parameter space of (τ_{dir}, α). The optimal parameters were found to be $\alpha = 0.31$, $\tau_{dir} = 1.6$ years for the hep-th dataset and $\alpha = 0.5$, $\tau_{dir} = 8$ years for the APS dataset.

Rescaled PageRank. Mariani et al. also tackled the problem of temporal bias in PageRank and offered an alternative solution [179]. The idea is similar to the rescaled citation method presented by Newman [117]. To calculate the rescaled PageRank score, one has to first compute the original PageRank score s_i of each paper i, as well as the mean $\mu_i(s)$ and standard deviation $\sigma_i(s)$ of the PageRank score for papers published in a similar time as i. The time is simply defined as the order of

the papers entering the network sorted to represent decreasing age of papers. The rescaled PageRank score of paper i is then defined as

$$R_i(s) = \frac{s_i - \mu_i(s)}{\sigma_i(s)}. \tag{18}$$

Old papers will have peers that have similar high PageRank score, and their scores will be largely suppressed after rescaling. In this way, papers of different ages can be compared fairly. Mariani et al. compared five quantitative metrics (i.e. citation, PageRank, CiteRank, rescaled citation and rescaled PageRank) with respect to their ability to identify the Milestone Letters selected by the Physical Review Letters editors. They found that the rescaled PageRank can rank these Milestone papers higher than the other methods.

DivRank. Mei et al. proposed an algorithm called DivRank to improve diversity in ranking, i.e., to make the highly ranked nodes come from different clusters of the network [180]. The DivRank is simply a reinforced PageRank process in networks. From the random walk point of view, it means that the probability that the walk stays at the current node is reinforced by the number of previous visits at the current node. To obtain DivRank scores, a transition probability matrix needs to be constructed, with each component as

$$p_t(u, v) = (1 - \lambda) \cdot p^*(v) + \lambda \cdot \frac{p_0(u, v)N_t(v)}{D_t(u)}, \tag{19}$$

where $D_t(u) = \sum_{v \in V} p_0(u, v)N_t(v)$ as a normalization factor. $p^*(v)$ represents the prior preference of visiting node v. If $p^*(v)$ is set to be uniform, then the first term is similar to the random jumping process in PageRank. The key for DivRank is that $N_t(v)$ as the number of times the walk has visited v up to time t introduces the reinforcing process in the random walk. The final DivRank score can be obtained after a stationary distribution of score $\pi(v)$ is obtained with sufficient iterations based on the following equation:

$$\pi(v) = \sum_{v \in V} p_t(u, v)\pi(u). \tag{20}$$

$\pi(v)$ is then used to rank nodes in the citation networks. The diversity in the ranking results from one node absorbing the scores from other nodes in the same cluster. As such, a node with a large score emerges in each cluster. Therefore, the top rank would contain nodes from different clusters.

PrestigeRank. Su et al. focused on the effect of missing data on the results of PageRank and proposed an improved algorithm called PrestigeRank [181]. This algorithm adds a virtual node to the network with N nodes. The virtual node is supposed to represent the references not included in the collection and receives all the citations that come from papers in the collection. The transition matrix W thus has $(N + 1) \times (N + 1)$ dimensions. The components for $i < N$ and $j < N$ are $W_{ij} = 1/k_j^{out}$ if j cites i, and $W_{ij} = 0$ otherwise. The component corresponding to the virtual node is defined differently: $W_{i(N+1)} = 1 - \sum_{i=1}^{N} W_{ij}$ and $W_{(N+1)j} = k_j^{in} / \sum_{i=1}^{N+1} k_i^{in}$. The iterative formula for the PrestigeRank algorithm is written as

$$\pi(n)^T = \pi(n - 1)^T(\alpha W + (\alpha a + (1 - \alpha)e)\frac{1}{N}e^T), \tag{21}$$

where α is the return probability set as 0.5, a is the binary dangling node (papers without any out-link) vector, e^T is a row vector of all ones. PrestigeRank has been applied to all physics papers in the Chinese Scientific and Technology Papers and Citation Database (CSTPCD) published between 2004 and 2006. The results show that PrestigeRank is significantly correlated with PageRank and citation counts. However, the analysis provides insufficient evidence for claiming that PrestigeRank is better than PageRank or citation counts.

NonlinearRank. Instead of missing data, NonlinearRank seeks to overcome the effect of malicious data in citation networks [182]. To improve the robustness of the PageRank to malicious citations, Yao et al. introduced nonlinearity into PageRank. The essential difference between NonlinearRank and PageRank is the way the target nodes aggregate the score from downstream neighboring nodes. Mathematically, the iterative equation is expressed as

$$s_i(n) = c + (1 - c)\left[\sum_{j=1}^{n} \frac{1}{N}\delta_{k_j^{out},0}s_j(n - 1) + \sqrt[\theta+1]{\sum_{j=1}^{n} A_{ij}(1 - \delta_{k_j^{out},0})(\frac{s_j(n - 1)}{k_j^{out}})^{\theta+1}}\right] \tag{22}$$

where θ is a tunable parameter. The score aggregated from the citing papers is controlled by adjusting the key parameter θ, so that only papers cited by papers with high scores receives higher scores and so that papers that are cited by many low-score papers cannot obtain a final high score. NonlinearRank reduces to PageRank when $\theta = 0$. The effectiveness of this algorithm has been validated in various ways including using Nobel prize-winning papers, simulation of epidemic spreading models, predictions and a robustness test in APS data. The robustness test is carried out by mimicking a common case in which some articles deliberately cite a target paper to enhance its ranking. First, a paper with in-degree $k^{in} = 0$ is randomly picked. Then, n new papers with m links each are added to the citation network. All of these new papers cite the target paper, and the rest of their links are randomly connected to other nodes. The ranking of the citation count and PageRank are largely distorted by this manipulation. However, the ranking of NonlinearRank is remarkably tolerant of these malicious papers, as their contribution to the scores of other papers is largely suppressed by the nonlinear parameter θ.

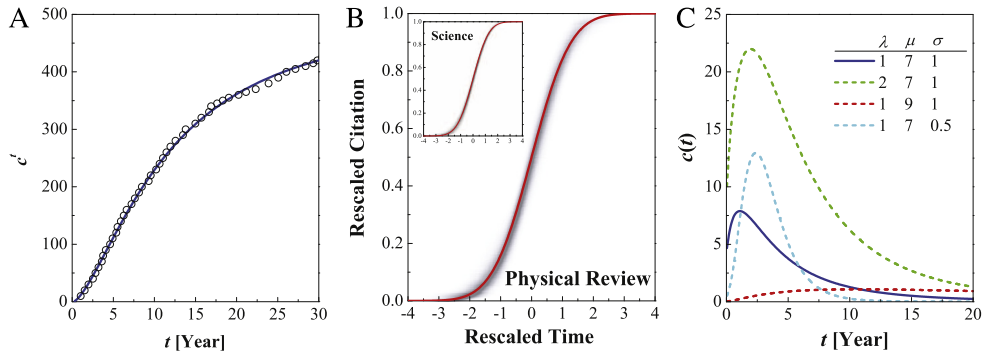

Fig. 8. Results for the mechanistic model developed in Ref. [32]. (A) The cumulative citation c^t for a research paper in APS dataset (circles) and the best fit of the model (solid line). (B) Rescaling the citation evolution of 7775 APS papers into one single curve by the model. Inset is the data collapse for papers published in *Science*. (C) The evolution of yearly citations $c(t)$ when the parameters λ, μ and σ in the model are set as different values.
Source: The figure is reprinted from Ref. [32].

SPRank. Also targeting malicious data, SPRank seeks to enhance the robustness of ranking against unreliable citations [183]. Zhou et al. introduced a preferential mechanism to the PageRank algorithm when aggregating resources from different nodes, with the scores from similar citing nodes being promoted. The logic is that if a paper is cited by many dissimilar papers, it could be the case that the authors select references carelessly or maliciously aim to push up a low-quality paper. As such, the citation from similar citing papers should be placed with more weight in ranking. The formula for SPRank is

$$s_i(n) = c + (1-c)\sum_{j=1}^{N}[\frac{f_{ij}^{\theta}A_{ij}}{k_j^{out}}(1-\delta_{k_j^{out},0}) + \frac{1}{N}\delta_{k_j^{out},0}]s_j(n-1), \tag{23}$$

where f_{ij} is the similarity between node i and j ($f_{ij} \in [0, 1]$) and θ is a tunable parameter ($\theta \geq 0$). When $\theta = 0$, SPRank reduces to the classic PageRank. As θ becomes larger, the score sent by dissimilar citing nodes is suppressed more severely. In this way, even if a node is cited by many papers, its final score cannot be high if the papers are too dissimilar. The similarity between two papers is measured topologically. The cosine metric [184] is used to measure the similarity of nodes based on their outgoing links:

$$f_{ij} = \frac{|\tau(i) \cap \tau(j)|}{\sqrt{k_i^{out}k_j^{out}}}, \tag{24}$$

where $\tau(i)$ and $\tau(j)$ are respectively the set of the upstream neighbors of node i and j. Similar to NonlinearRank, SPRank is also examined with Nobel prize-winning papers, predictions and a robustness test in APS data. Specifically, the mean ranking of Nobel prize-winning papers is improved the most when the parameter θ is approximately 0.1. In addition, the future citation of papers is more highly correlated with SPRank scores than with PageRank scores. The rankings of the papers that are supposed to be pushed up by spamming papers are effectively suppressed by SPRank.

4.1.3. Model-based approaches

In addition to the propagation methods, growing models also provide a way to measure the impact of scientific publications. Empirical evidence suggests that the growth of citation networks is mainly driven by preferential attachment [185], with also the influence of nodes' fitness [125]. In mathematical form, it is expressed as $p_i \sim \eta_i k_i^{in}$ where p_i is the probability of paper i attracting new citations and η_i is the fitness of i. The parameter η_i can actually be estimated from real data. The ratio between the number of citations a paper actually receives and the expected number of received citations by preferential attachment defines the paper's fitness η_i. For a more detailed discussion and formula to compute η_i, readers can refer to Section 3. In some literature, the obtained η_i is interpreted as a paper's intrinsic quality [125]. Therefore, η_i is also used as a metric to rank scientific publications.

Wang et al. moved further and derived a mechanistic model for the citation dynamics of individual papers [32]. The model combines three fundamental mechanisms, i.e., preferential attachment, fitness and aging, and eventually derives a formula for the probability that paper i is cited at time t after publication as

$$\Pi_i(t) \sim \eta_i c_i^t P_i(t), \tag{25}$$

c_i^t is the cumulative citations of paper i at time t, $P_i(t)$ describes the aging effect with a log-normal survival probability,

$$P_i(t) = \frac{1}{\sqrt{2\pi}\sigma_i t}\exp[-\frac{(\ln t - \mu_i)^2}{2\sigma_i^2}]. \tag{26}$$

The cumulative degree of paper i can be obtained by solving the associated master equation of $\Pi_i(t)$,

$$c_i^t = m[e^{\frac{\beta\eta_i}{A}\Phi(\frac{lnt-\mu_i}{\sigma_i})} - 1], \tag{27}$$

where

$$\Phi(x) = (2\phi)^{-1/2}\int_{-\infty}^{x} e^{-y^2/2}dy \tag{28}$$

is the cumulative normal distribution. Some newly introduced parameters are m (the average number of references each new paper contains), β (the growth rate of the total number of publications), and A (a normalization constant). These three parameters are identical for each paper and can be obtained with easy calculation in real data. The rest of the parameters for paper i, η_i, μ_i and σ_i need to be fitted using i's citation evolution data. One remarkable finding with this mechanistic model is that the evolution of all papers' citations can be rescaled to one curve. With the scaled variables $\tilde{t} = (lnt - \mu_i)/\sigma_i$ and $\tilde{c} = ln(1 + c_i^t/m)/\lambda_i$ with $\lambda_i = \eta_i\beta/A$, the collapsed curve is expressed as $\tilde{c} = \Phi(\tilde{t})$, see Fig. 8. This universal curve indicates the validity of the model capturing the real mechanisms of papers' citation dynamics. In addition, the model shows that the final cumulative citation of a paper is determined by its fitness, i.e., $c_i^\infty = m(e_i^\lambda - 1)$. This result supports the assertion that the fitness of papers is a fair metric for quantifying the impact of papers published at different times. However, estimating η_i by fitting is not easy and could be influenced by many factors such as data fluctuations and the fitting technique [186]. Therefore, for the purpose of ranking papers, a simpler method for roughly estimating η_i is to use the formula in Eq. (4). This model also serves as an effective tool for prediction, which we will discuss in detail in Section 5.

4.2. Evaluating scientists

Evaluating scientists is an important issue for all disciplines. The ongoing rapid development of information technology has greatly accelerated the publication of scientific findings, resulting in a large number of scientific papers. The quantitative studies of these papers have thus become an important method for evaluating the scientific influence of researchers. As a result, many metrics for ranking scientists have been developed, and these methods serve as important references when evaluating grant proposals and promotion applications. In this subsection, we review some representative metrics, emphasizing on the metrics with the concept of complex systems.

4.2.1. Traditional methods

Several traditional metrics are defined straightforwardly based on the citation data of papers. The widely used ones are the number of publications; the total number of citations (the total citations received by all papers published by a scientist); the average number of citations per publication; the number of highly cited publications (this index requires a threshold to define highly cited papers); and the proportion of highly cited publications. The first four metrics can be easily computed by scientists themselves while the last one needs global information which can only be provided by organizations with access to the complete scientific publication database. As an extension, the threshold for determining highly cited papers may vary from one discipline to another, which can lead to a fairer comparison of scientists from different fields. In practice, Essential Science Indicators (ESI) provides each year highly cited papers in different fields for reference. For a comprehensive review of the literature on citation impact indicators, see Ref. [187].

An important index based on citation count is the well-known ***h*-index**. This index was first proposed by Hirsch to quantify an individual's research performance [2]. A scientist's h-index equals h if he/she has h papers with at least h citations. This metric is a combined measure of the quantity and quality of a scientist's publications. The h-index will not be high if a scientist publishes either many lowly cited papers or one very highly cited paper. Owing to its simplicity, the h-index has been widely adopted by the scientific community. Many large websites such as Google Scholar and Web of Science provide users with their h-index values. The h-index has some obvious advantages over the single-value quantitative indicators mentioned above. One of the admirable properties is its robustness against manipulations. Simply increasing the number of lowly cited publications does not improve the h-index. However, the h-index has some nonnegligible disadvantages. First, it cannot be used to compare scientists from different disciplines because of the large variance in productivity and citation patterns across fields. Second, the h-index strongly depends on the number of researchers' publications, so it is not fair for evaluating the scientific performance of researchers with a few highly cited scientific papers or to compare scientists at different stages of their career. In addition, it has been shown that the h-index is simply the one-half of the square-root of the number of citations for the large majority of condensed-matter physicists, indicating that the h-index does not add anything new beyond the citation count [188]. To correct these shortcomings, dozens of improved metrics have been proposed so far, including h-dependent variants, h-independent variants, h-adjusted variants to field, and h-adjusted variants for co-authorship, and so on. Recently, it has also been shown that the well-known k-shell centrality will be acquired if one iteratively apply h-index in networks [189]. We list some representative modifications of the h-index below. For more detailed discussion of advantages and disadvantages of the h-index and its other variants, readers can refer to three previously published review papers [187,190,191].

- **g- and hg-index.** Because the h-index treats all h papers equally, this index neglects the importance of highly cited papers when evaluating a researcher's scientific impact. Egghe proposed the g-index, which is defined as the largest number g of individual publications that together have at least g^2 citations [192]. Though the g-index has the advantage of taking into account the citations of highly cited papers, the index has its own limitations. A researcher's g-index usually increases dramatically when he/she has a paper with a lot of citations. Alonso et al. proposed a hg-index that combines the h-index and g-index [193]. The hg-index is simply defined as $hg = \sqrt{h \times g}$. The new metric breaks the degeneracy in the h-index by taking into account the influence of successful papers, and softens the strong sensitivity of the g-index to some very highly cited papers. However, the hg-index depends on other indices (the h- and g-indices). The equivalence classes of hg are questionable and the substitution rate between h and g may arbitrarily change depending on the specific h and g values.
- **A- and R-index.** Similar to the g-index, the A-index considers the exact number of citations of articles and highlights the importance of high-impact papers. This index is defined as the average number of citations received by articles in the Hirsch core (i.e., papers with more than h citations). Mathematically, it can be computed as $A = \frac{1}{h}\sum_{j=1}^{h} c_j$ where h is the h-index and c_j is the citation count of the jth most cited paper. The disadvantage of the A-index is that it may overly "punish" scientists with high h-index. In some cases, we can observe that the A-index of scientist i is lower than that of scientist j despite scientist i's h-index being higher than that of scientist j. The principal reason for this is that the A-index involves a division by h. Jin et al. proposed a new index called the R-index by eliminating the division by h and adding a square root [194]. The R-index is defined as $R = \sqrt{\sum_{j=1}^{h} c_j}$, which makes $R = \sqrt{h \cdot A}$. However, R-index can be very sensitive to just a few papers receiving extremely high citation counts.
- **o-index.** Dorogovtsev and Mendes studied a sample of researchers from physics and complex systems and found that the h-index tends to favor scientists with many papers yet unfairly punishes researchers with a high mean number of citations per paper [195]. They proposed a new indicator called the o-index, which focuses on a researcher's most cited paper. The o index can be defined as $o = \sqrt{mh}$, where m is the number of citations for a researcher's most cited paper. The o index is more difficult to be manipulated compared with h-index and can distinguish some successful researchers who are hidden in the h-based ranking. The disadvantage of o-index is that it grows a lot when the author has one paper with extremely high citation count.
- **AR-index.** Jin et al. suggested that the age of the publication is also necessary to be used to evaluate a researcher's performance in addition to the number of citations [194]. Accordingly, they defined an age-dependent index called the AR-index, which is calculated as $AR = \sqrt{\sum_{j=1}^{h} c_j/a_j}$, where a_j is the age of the jth most cited paper. Unlike the h-index, which always increases throughout a scientist's career, the AR-index may decrease over time. Due to this feature, the AR-index can evaluate authors' performance changes. However, the AR-index overlooks the fact that the aging speed of papers from different disciplines can be vastly different.
- **h_w-index.** Focusing on the disadvantage of the h-index being insensitive to the varying performance of scientists, Egghe and Rousseau proposed a citation-weighted h-index called the h_w-index [196]. The h_w-index is defined as $h_w = \sqrt{\sum_{i=1}^{r_0} c_i}$ where r_0 is the largest row index j ensuring $\sum_{i=1}^{j} c_i/j \leq c_j$. The h_w-index can also be defined as a continuous version with properties similar to those of the discrete version. The h_w-index is sensitive to performance changes but it requires multiple and advanced calculations.
- **h_I-index.** Empirical studies have shown that the number of citations that a paper receives can be influenced by the number of authors [197]. Batista et al. considered the co-authorship effect and proposed the h_I-index as $h_I = h/\langle N_h \rangle$ where $\langle N_h \rangle$ is the average number of authors for h papers. [198]. The advantages of the h_I-index are that it diminishes the degeneracy of h and is less sensitive to different research fields (as papers from different fields can have different average numbers of authors). However, the h_I-index is not suitable for the publications with large-scale collaborations.

4.2.2. Network-aware methods

Scientists are connected through multiple relations, which form several networks, such as collaboration and citation networks, among them. The impact of scientists can be measured by analyzing their positions in networks. Some network centrality measures have therefore been introduced to evaluate scientists. Based on collaboration networks, node centrality indices, including degree, closeness (i.e., mean average shortest paths to other nodes), betweenness (i.e., number of shortest paths passing through a node) and PageRank, have been used to estimate the importance of authors in the network [199]. These four indices are strongly correlated with author citations but also provide additional information. They are therefore suggested as alternative indicators for author impact analysis. Similar analyses have also been performed in other studies, considering some additional network measures such as link density and clique [200], a new centrality index called the Q-measure being proposed [201], and the betweenness being used as a driving force for the evolution of collaboration networks [202]. A survey about applying social network analysis in scientist evaluation can be found in a previously published review article [203].

Other majority of studies that seek to incorporate network structure into author ranking are mainly based on PageRank and its variants. They can be roughly classified into two categories. The first type is to design iterative algorithms on collaboration or citation networks among authors. The obtained final score is a direct measurement of an author's impact.

The second type is to first estimate the impact of papers using PageRank or other algorithms, and then aggregate the scores of an author's papers to rank this author.

Ranking algorithms on author networks. Liu et al. built a weighted and directed network co-authorship network and proposed a modification of PageRank called AuthorRank, which considers link weight [204]. The definition of link weights is essential for this work. For each paper α with m authors, a quantity $g_{i,j,\alpha} = 1/(m-1)$ representing the degree to which author i and j have an exclusive co-authorship relation is computed. Then a co-authorship frequency is computed by summing $g_{i,j,\alpha}$ over all papers that i and j have co-authored, as $c_{ij} = \sum_\alpha g_{i,j,\alpha}$. The weight from i to j is simply a normalized c_{ij} over the summation of c_{ij} over all i's collaborators, $w_{ij} = c_{ij}/\sum_l c_{il}$. As the denominator in w_{ji} is different from that in w_{ij}, the weight matrix is asymmetric and the network is directed. AuthorRank is actually a straightforward extension of PageRank to weighted networks:

$$s_i(t) = (1-c) + c\frac{w_{ij}}{\sum_l w_{lj}} s_j(t-1). \tag{29}$$

They applied AuthorRank in the co-authorship network to rank authors' prestige. They compared the AuthorRank with the degree, betweenness, closeness and original PageRank in this co-authorship network. Using the dataset of Joint Conference on Digital Libraries (JCDL) program committee members, they found that betweenness, PageRank, and AuthorRank all show good results, i.e., these metrics gave high ranks to the committee members. PageRank and AuthorRank are highly correlated, but no conclusive evidence indicates which one performs better.

Several other studies adopt the framework of personalized PageRank to design better ranking algorithms for authors. In personalized PageRank algorithms, the damping factor is multiplied by a variable depending on some characteristic of nodes. Ding et al. proposed a personalized PageRank algorithm and applied it to various networks among authors [205–207]. Different from Eq. (29), the personalized PageRank is defined as

$$s_i(t) = c\frac{a_i}{\sum_l a_l} + (1-c)\frac{w_{ij}}{\sum_l w_{lj}} s_j(t-1); \tag{30}$$

where a_i is the initial assigned weight to node i, which can be used as any preassigned vectors. For instance, a_i can be the number of publications of an author or the number of citations of an author. w_{ij} is the link weight matrix, depending on which weighted network is considered. Yan and Ding constructed an undirected and weighted co-authorship network with the link weights representing the co-author frequency among authors [205]. They first applied PageRank with different damping factors c in the co-authorship network and found that damping factors do not have much influence on the PageRank score in the co-authorship network. Then, the personalized PageRank algorithm was applied to the co-authorship network, with the node weight a_i set as the number of citations pointing to each author i. Comparing with Prize Award winners, the personalized PageRank outperformed PageRank and h-index in identifying these winners. Ding et al. constructed an author co-citation network in which edges represent the co-cited relations among authors and the weights of the edges represent the author co-citation frequencies [206]. As a result, the network is weighted and undirected. The personalized PageRank algorithm was also applied to this network, with a_i set as the number of publications or the number of citations of author i. They examined the correlations among PageRank, personalized PageRank, centrality measures, h-index and citation rank in the author co-citation network. They found that citation rank is highly correlated with PageRank under different damping factors and personalized PageRank but is not strongly correlated with centrality measures and h-index. The above personalized PageRank is also applied to author citation networks that are weighted and directed with the link weights w_{ij} representing the number of times that one author cites the other [207]. The node weights a_i are again given with two different vectors: the number of citations or the number of publications of the authors. They found that the personalized PageRank is highly correlated with the author's total number of citations and the number of times the author is cited by highly cited papers [208]. To rank the prize winners within the Information retrieval field, the prestige rank outperformed the other two measures.

As a group of physicists, Radicchi et al. constructed a new weighted author citation network in which the nodes are authors and the weights of the edges are calculated using normalized citation counts [91]. They proposed a diffusion-based method called SARA (Science Author Rank Algorithm), which mimics the spread of scientific credits in the author citation network to rank scientists. They assumed that each author initially has a unit of credit that is then distributed to its neighbors proportionally to the weight of the directed link and the node i receives the scientific credits from its neighbors following

$$P_i = (1-q)\sum_j \frac{P_j}{s_j^{out}} w_{ij} + qz_i + (1-q)z_i \sum_j P_j \delta(s_j^{out}), \tag{31}$$

where q is the damping factor, P_i is the SARA score of node i, w_{ij} is the weight of the directed link from node j to node i in author citation network. $s_j^{out} = \sum_k w_{jk}$, and z_i is related to the productivity of author i with the normalization $\sum_i z_i = 1$. The function $\delta(x) = 1$ if $x = 0$ and $\delta(x) = 0$ otherwise. The SARA algorithm can be regarded as a personalized PageRank algorithm as it is based on a mixed process that contains a biased random walk and a personalized redistribution of the credits among the nodes. The biased random walk ensures that citations from highly ranked authors are more important than citations from authors with a low rank. The method was validated in APS data with physics Nobel prize winners. Compared with the citation count and balanced citation count (normalizing the citation weight by the total number of authors of the cited paper), the SARA algorithm can rank these important scientists more highly.

In addition to the personalized PageRank modifications, Fiala et al. assumed that a citation obtained from a frequent co-author is less valuable than that from a foreign researcher and proposed a family of variations of PageRank that combines co-authorship and citation information [209]. The score of an author in the iteration can be expressed as

$$R(u) = \frac{1-d}{|A|} + d \sum_{(v,u)\in E} R(v) \frac{\sigma_{v,u}}{\sum_{(v,k)\in E} \sigma_{v,k}}, \tag{32}$$

where

$$\sigma_{v,k} = \frac{w_{v,k}}{\frac{c_{v,k}+1}{b_{v,k}+1} \sum_{(v,j)\in E} w_{v,j}}, \tag{33}$$

and d is the damping factor set to 0.9. This variant is again a biased random walk with damping. The factor $\sigma_{v,u}$ determines the degree to which the random walk is biased to link from v to u. In the definition of $\sigma_{v,u}$, $w_{v,u}$ is the number of citations from author v to u, $c_{v,u}$ is the common publication between author v and u, $b_{v,u}$ is a normalization factor that can be defined according to different similarity indices (such as Jaccard, Cosine). The factor $\sigma_{u,v}$ suppresses the effect of citations from an authors' frequent collaborators, i.e., these citations have small $\sigma_{u,v}$ so less score will be passed through these links. This algorithm is applied to the author citation network and found to rank awarded authors higher than that by the standard PageRank. After this work, Fiala took into account the time of publications and citations to propose a time-aware version of this PageRank variant, which further improves its performance [210].

Ranking algorithms by aggregating impacts from paper networks. Nykl et al. applied the PageRank algorithm to author and publication citation networks to evaluate scientists [211]. The main question they raised is whether better evaluation results were based directly on an author network or on a publication network. Several variants of author or publication citation networks with various types of self-citations are investigated. As discussed above, author ranking can be directly obtained by applying PageRank to author citation networks. In publication citation networks, the PageRank values of papers need to be distributed to authors in order to then rank authors. Two approaches are compared: assigning sums of the publication values to all of their authors or assigning them proportionally according to the number of authors. Some prestigious awards such as the ACM A. M. Turing Award, ACM SIGMOD Edgar F. Codd Innovations Award, ACM Fellows and ISI highly cited researchers are used to compare different cases (variance in algorithms and networks). By checking the mean ranking of these award winners, the optimal combination of aggregation approach and network can be identified. Their main conclusion is that the best ranking of authors is obtained by using a publication citation network without self-citations and by distributing the same proportional parts of the publications' values to their authors. As an extension, Nykl et al. considered the importance of journals for ranking authors and proposed several personalized PageRank algorithms to incorporate this information for ranking scientists [212].

Wang et al. compared the author ranking methods that are directly based on author networks and the author ranking methods by aggregating impacts from paper networks [213]. For author citation networks, the RLPR (see Eq. (30)) and SARA (see Eq. (31)) algorithms are considered. For paper citation networks, the original PageRank algorithm (see Eq. (16)) is employed. They interpreted the PageRank algorithm and its variants as graph ranking methods that mimic the spreading of scientific credit among researchers or their publications. They found that paper level credit diffusion is equivalent to authorship level credit diffusion and that these two type of methods are actually two aggregated versions of a fine-grained authorship level credit diffusion. With APS data, the paper level method is shown to be more reliable and robust compared with the authorship level method as this method may cause misallocation of credit among researchers and their publications. When aggregating impact scores from papers, a fair assignment of credits to individual authors in a paper is essential. This is a very important research topic in SOS, which we will discuss in detail below.

4.2.3. Dynamics-aware methods

The impact of authors can be extracted from the dynamics of their publication data. Sinatra et al. addressed this problem with an extremely simple model [55]. Using the cumulative citations received by paper i 10 years after it was published c_{10}^i, the impact of scientific papers published at different times can be quantified and compared, as shown in Fig. 9A. The evolution of c_{10}^i is examined throughout thousands of scientific careers. The paper with highest c_{10} (i.e. c_{10}^*) for each scientist is identified. Denoting N and N^* as the number of publications of an author and the order of the c_{10}^* paper of this author in his/her publication sequence, respectively, the distribution of N^*/N of scientists is shown to follow a uniform distribution, indicating that the highest-impact work in a scientist's career is randomly distributed within the sequence of papers published by this scientist (see Fig. 9B,C). This result inspires a simple model for describing the time sequence of the impact of individual author publications:

$$c_{10,i\alpha} = Q_i p_\alpha, \tag{34}$$

where Q_i is a unique value for scientist i describing his/her ability in scientific research, p_α is a completely random variable represents the potential impact of paper α. As a scientist-independent variable, p_α can also be interpreted as luck in scientific research. This model is shown to well reproduce the dynamic patterns in the impact of authors' papers observed in real data. Apparently, to measure the ability of a scientist i, one just has to compute Q_i, which can be written as

$$Q_i = e^{\langle \log c_{10,i}\rangle - \mu_p}, \tag{35}$$

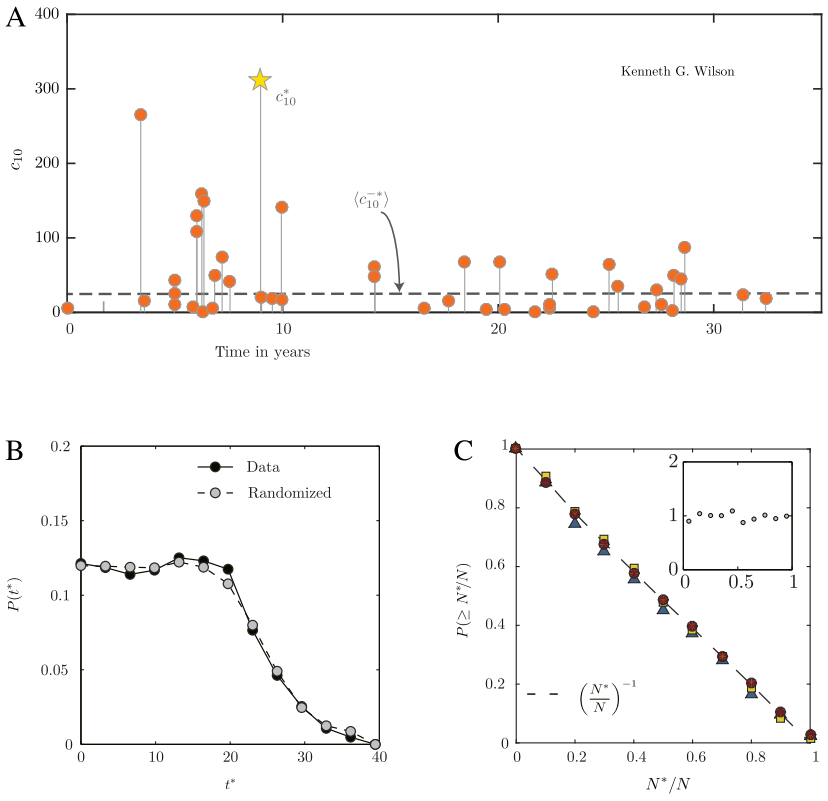

Fig. 9. (A) Illustration of c_{10}^* with publication history of Kenneth G. Wilson (Nobel Prize in Physics, 1982). Each vertical line represents a research paper. The height of each line corresponds to c_{10}. The highest line is denoted as c_{10}^* as the scientist's most influential work. (B) Distribution of the publication time t^* of the highest-impact paper for different scientists in real data and random counterparts in which a scientist's sequence of publication is randomized. The highest-impact work is more likely to appear in a scientist's early career. (C) Cumulative distribution $P(\geq N^*/N)$ for scientists. The inset is the probability density distribution $P(N^*/N)$. The results indicate that N^* has the same probability to occur anywhere in the sequence of papers published by a scientist. *Source:* The figure is reprinted from Ref. [55].

where μ_p is the mean of p_α. Here, $\langle logc_{10,i} \rangle$ is easy to calculate from i's citation data. μ_p, however, needs to be fitted with a maximum-likelihood approach through all authors' career data. In the APS data considered in the paper, μ_p is estimated as 0.92. In addition, it is argued that the Q value of the considered scientists is relatively stable over time, making it reliable to use Q for ranking scientists. However, one drawback is that this model relies on c_{10} (i.e., the citation of papers after publishing for 10 years), so it is difficult to apply this model to younger scientists with career of less than 10 years. It remains still unclear whether the effectiveness of this model will be affected if c_{10} is reduced to c_5 or even smaller values. Apart from that, this model is shown to be very effective for predicting the future career of scientists (more precisely, the future h-index). We discuss this part of the work in the next section.

4.2.4. Credit allocation for individual papers

It has been revealed that co-authored publications in science have increased significantly. Through collaboration, scientists combine ideas and techniques from different fields, which can result in higher quality publications. In a multi-author paper, the authors often make different contributions to the paper. How to distribute credit among a paper's authors is not a trivial problem. In physics, the authors who contribute most to a paper are usually listed as the first author or the corresponding authors. However, it is not this case for many other fields such as mathematics in which the authors are usually listed in alphabetic order regardless of who contributes more. A fair credit allocation to authors will lead to a more objective evaluation of a scientist's impact by aggregating his/her contributions in each of his/her papers. Thus, this problem has very practical meaning.

In numerous studies, an author's contribution to a paper is assigned using some simple methods. The simplest method is to follow the traditional practice of assigning full authorship credit or an equal fraction of credit to the co-authors of a paper. This method is referred to as the fractional counting method in some papers [214]. Denoting N as the number of authors in a paper, each author will receive $1/N$ credit. This method favors those authors who participate in many studies but contribute little to each work. Owing to its simplicity, this method is used as a benchmark for comparing newly proposed methods. Other methods consider the order of authors in a paper. There are four main methods: geometric [215], arithmetic [216],

harmonic [217] and network-based [218]. Denoting R as the order of an author for a paper, his/her credit can be computed as

- $s(R) = 2^{N-R}/2^N - 1$ in geometric counting;
- $s(R) = 2(N - R + 1)/N(N + 1)$ in arithmetic counting;
- $s(R) = 1/R/\sum_{i=1}^{N} 1/i$ in harmonic counting;
- $s(R) = 1/N + 1/Nd\sum_{i=1}^{N-R} 1/(N-i)$ for $R = 1, s(R) = 1/N(1-d) + 1/Nd\sum_{i=1}^{N-R} 1/(N-i)$ for $1 < R < N, s(R) = 1/N(1-d)$ for $R = N$, in the network-based method. Here, d is a free parameter between 0 and 1.

The network-based method simplifies to fractional counting when $d = 0$. Similar to the network-based method, the first three methods can also be extended to a hybrid method combining fractional counting $1/N$ and the method based on author orders $s(R)$ with a tuning parameter. These four methods are compared with empirical datasets from economics, marketing, psychology, chemistry and medicine, with a list of credit allocation based on expert surveys [219]. The results show that harmonic scheme performs the best overall.

An alternative framework for assigning author credit according to author order has been proposed [220]. This method assumes that n authors in a paper can be divided into $m < n$ groups, and c_i is denoted as the number of authors in the ith group. The authors in each group are assumed to provide equal contributions to the paper. Then, a credit vector is defined as $\vec{x} = \{x_1, x_2, \ldots, x_m\}$, where x_i is the credit assigned to each of the c_i members in group i. The expectation of $\vec{x}$ is derived as

$$E(x_i) = \frac{1}{m}\sum_{j=i}^{m} \frac{1}{\sum_{k=1}^{j} c_k}. \tag{36}$$

If no authors claim equal contribution, then $c_i = 1$, $m = n$, and

$$E(x_i) = \frac{1}{n}\sum_{j=i}^{n} \frac{1}{j}. \tag{37}$$

This method is called the A-index in the paper. The corresponding author is considered to be as important as the first author in the A-index calculation, so both are assigned to the first group. The A-index is a single-paper metric and has an upper bound of 1. It can be further extended to the C-index and P-index for evaluating researchers, with $C = \sum_{k=1}^{K} A_k$ and $P = \sum_{k=1}^{K} A_k \mathrm{JIF}_k$. Here, A_k is the credit assigned to an author in his/her paper k by A-index, and JIF_k is the Journal Impact Factor of the journal that paper k is published in. According to the definitions, the C-index is a weighted count of an author's publications, while the P-index is a weighted sum of the expected citation counts of an author's papers. These two indices are compared with the well-known h-index in the bibliometric data of 186 biomedical engineering faculty members. It is found that neglecting co-author information (h-index) will inhibit the ability to distinguish a researcher's achievements. The P-index is very useful for young researchers, as they usually publish few papers in their early career but are often the main contributors for these papers.

Based on a local diffusion process in citation networks, Shen et al. proposed a collective credit allocation method for authors in a paper [221]. The basic idea of the method is closely connected to the collaborative filtering (CF) approaches used for recommendation in computer science [222]. In CF, the product a user may be interested in is similar to the products chosen by this user before. In collective credit allocation, an author who designs more credit in a paper is the author who has previously published many similar papers. The similarity is measured with topology in the citation networks, i.e. two papers are considered similar if they are co-cited by many papers. Unlike the previously mentioned methods, this method is independent of the ordering of author list and just let the credit allocation emerges from the diffusion process, as shown in Fig. 10. The method requires computing a credit allocation matrix A and a co-citation strength vector s. Denoting a_i as an author of the target paper whose credit we want to allocate and p_j as one of the papers written by a_i and co-cited by the papers which cite the target paper, the component A_{ij} in the credit allocation matrix, A_{ij} depicts the amount of credit that author a_i gets from the co-cited paper p_j. p_j distributes its credit evenly to its authors, so a_i will receive $A_{ij} = 1/n$, where n is the number of authors in p_j. The component s_j records how many times a paper p_j is co-cited by the papers citing the target paper. The final credit of a_i in the target paper can be computed by

$$c_i = \sum_j A_{ij} s_j. \tag{38}$$

c_i is simply the weighted sum of its local credit obtained from all co-cited papers. To validate this method, the data regarding the Nobel prize winners in physics, chemistry and medicine is examined. Many Nobel laureates are found to be neither the first author nor the most highly cited author in the Nobel prize-winning papers, yet they are assigned a higher credit than other authors by the collective credit allocation method. This is because a Nobel prize winner usually publishes more important papers in the related field than his/her co-authors, so he/she should receive more credit in the Nobel prize-winning paper. This method was also used to investigate the evolution of the credit share by authors in the Nobel prize-winning papers. Based on a similar idea, a local credit diffusion algorithm was proposed by Niu et al. to identify the representative publications for individual scientists [223]. The basic idea is that a representative work of a scientist is an important paper

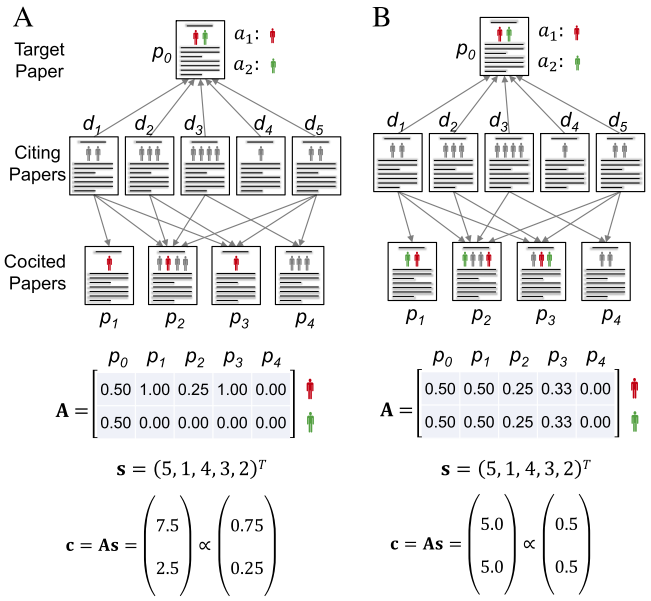

Fig. 10. The illustration of the collective credit allocation method proposed in Ref. [221]. The local structure of the target paper p_0 in the citation network is shown. d_1 to d_5 are the papers citing p_0. p_1 to p_4, denoted as co-cited papers, are the papers cited by d_1 to d_4. A is the credit allocation matrix; Vector s represents the number of times that each co-cited paper is cited by d_1 to d_4; and c is the final credit share for the authors of the target paper p_0. (A) and (B) describe two different examples.
Source: This figure is reprinted from Ref. [221].

in his/her area of expertise. The measure of importance and relevance of a paper is naturally combined in the local diffusion algorithm. The method is found to be more effective than the citation count and PageRank in identifying the Nobel prize-winning paper among the publications of each Nobel laureate.

Some journals currently require authors to declare their contributions in their papers. Contribution categories may contain, for example, who designed the project, who conducted the experiments, who performed the analyses, and who wrote the paper. Such data can be regarded as a multi-layer network. By using tensor analysis, the centrality of nodes can be computed, which can be used as an alternative method for determining the general contributions of the authors [53].

4.3. Measuring the impact of scientific journals

In science evaluation, quantifying the impact of scientific journals is a very practical issue, influencing authors where to submit their research papers and where to search for the latest notable progress in their research fields [3]. Because accurately measuring the influence of a paper requires many years of citation data, most methods fail to evaluate newly published papers. In such cases, one can obtain a quick and preliminary estimation of a new paper's influence using the characteristics of the journal it is published in. Therefore, an objective ranking of the impact of scientific journals serves as a useful indicator for providing early information for paper and author evaluation. In this subsection, we review some mainstream impact metrics for scientific journals. For more information on evaluating the citation impact of journals, readers can refer to another review article in Scientometrics [187].

4.3.1. Metrics based on citation count

The most used impact metric for scientific journals is the impact factor (IF) proposed by Garfield et al. [3]. It is the yearly average number of citations to recent articles published in a journal. In practice, the impact factor of a year considers the papers published by a journal over the past two years and the citations they received in the current year. Impact factors are calculated annually for all the journals listed in the Journal Citation Reports. Despite its wide usage, whether the impact factor is a fair metric for journals is still under debate [224,225]. The main criticisms of the impact factor may be summarized as follows: it is inappropriate to calculate IF as the arithmetic mean of a highly skewed distribution of citations; IF ignores where the citations are from; IF is based on a narrow two-year time window that is inappropriate for many disciplines; IF is not reproducible as the data for computing it are not publicly available; and IF may be manipulated by the policy that editors adopt. Therefore, numerous alternative metrics based on journal citation count have been developed, with some of them being direct variants of IF. The well-known ones include the following [187,226]:

- 5-Year Impact Factor (IF5): the average IF over the last five years
- Immediacy Index (IM): the average number of times an article is cited in the year it is published

- Citation Half-Life (HL): the median age of articles cited by the journal in the Journal Citation Reports year
- Source-Normalized Impact Per Paper (SNIP): the ratio of a journal's citation count per paper and the citation potential in its subject field [227]
- Journal Relative Impact (JRI): ratio of the actual citation of the papers published in that journal to the expected citations of papers published in all journals that are grouped by the WoS in a certain subject category [228]
- Impact Per Publication (IPP): the ratio of citations in a year to papers published in the three previous years, divided by the number of papers published in those same years [227]
- Google 5-Year h-index (h5): the largest number h such that h articles published in the past five years have at least h citations each [2]
- Citation success index: the probability that a random paper in one journal has more citations than a random paper in another journal [229]
- Integrated Impact Indicator (I3): the weighted summation of the number of publications in each percentile rank class according to paper citations, i.e., $I3 = \sum_i x_i f(x_i)$, where x_i is the weight of each percentile class and $f(x_i)$ is the number of papers in class i [230]

The SNIP and IPP are mainly used by Elsevier journals. Among these indices, IF5 is highly correlated with IF but with a smaller fluctuation over time. However, which metrics produce the fairest journal ranking remains to be examined.

4.3.2. Metrics incorporating network effects

More complicated metrics for journals consider the global citation networks among journals. The basic ideas for these metrics is that a journal mainly cited by prestigious journals is more likely to have high impact, which is consistent with the fundamental assumption of the well-known PageRank algorithm. Accordingly, Bollen et al. employed a weighted PageRank algorithm to citation networks among journals and obtained a metric that reflects prestige [231]. The weighted PageRank value $PR_w(v_i)$ after convergence can be used as a measure of journal v_i's prestige. $PR_w(v_i)$ is then a hybrid with the journal impact factor to define a Y-factor, $Y(v_i) = IF(v_i) \times PR_w(v_i)$. Journals that have high $Y(v_i)$ scores must be ranked highly by either or both the IF and weighted PageRank. The Y-factor was applied to a journal citation network constructed on the basis of the WoS dataset in year 2003. It was found that IF rankings highly promote review journals and those that frequently publish background materials that is likely to be cited, while the Weighted PageRank focuses more on a set of journals typically appreciated by domain experts. The Y-factor is a reasonable balance between them.

Based on the same principle, Bergstrom et al. proposed an Eigenfactor metric that can be regarded as a personalized PageRank algorithm for journal citation networks [232]. The Eigenfactor considers not only how many times a journal is cited but also which journals have contributed these citations; thus, highly cited journals will influence the network more than lesser cited journals. Based on the Eigenfactor, a normalized score called the Article Influence Score (AIS) is devised. AIS is calculated by dividing a journal's Eigenfactor by the number of articles in the journal, measuring the average influence per article in a journal. As such, the average AIS of all papers in the database is 1.

To remove the size effect from the journal evaluation, a size-independent metric called the SCImago Journal Rank (SJR) was proposed by González-Pereira [98]. The SJR metric also relies on an iterative process, where the journal score in each step is

$$\mathrm{PSJR}_i = \frac{1-d-e}{N} + e \cdot \frac{Art_i}{\sum_{j=1}^{N} Art_j} + d\left[\sum_{j=1}^{N} C_{ji} \cdot \frac{\mathrm{PSJR}_j}{C_j} \cdot CF + \frac{Art_i}{\sum_{j=1}^{N} Art_j} \cdot \sum_{k \in DN} \mathrm{PSJR}_k\right], \tag{39}$$

where C_{ji} is the number of citations from journal j to journal i, C_j is the number of references of journal j, N is number of journals in the database, Art_j is the number of papers in journal j, $d = 0.9$ and $e = 0.0999$. This metric can be regarded as a modification of the weighted PageRank, with the terms containing Art_j aiming to remove the effect of the different numbers of papers published by journals. The SJR indicator and the IF are found to be strongly correlated, with some level of difference mainly in the top-ranking section. In addition, the SJR indicator tends to be concentrated in fewer journals than the citation quantity measured by IF.

4.4. Evaluations at institution, country and research field levels

Instead of papers and authors, numerous studies focus on science rankings at higher levels. Related works seek to provide various tools to evaluate the scientific achievements of institutions, countries and research fields. This task is driven by the need to justify funding and to assess the effects of policy changes. As this task can be regarded as evaluating a large group of scientific publications to some degree, the simplest method is to aggregate the citations of these papers as a size-dependent approach, or to compare the average citations of these papers as a size-independent approach [175]. The metrics for journal evaluations discussed above can also be naturally extended to institutions, countries and research fields. In addition to those methods, numerous more complicated methods have been developed.

Country level. King proposed ranking nations based on their share of top 1% of the most highly cited publications in the WoS database [233]. Fairclough and Thelwall introduced two new methods based on linear modeling for normalized data

and the geometric mean for comparing the national differences in average scientific impact [234]. The first one establishes a linear regression model for each country:

$$log(1 + citations) = a + \sum_c \beta_c p_c, \tag{40}$$

where p_c is the proportion of authors from country c, a is a constant and β_c is a parameter representing the individual contribution rate of country c. With ordinary least squares method, the value of a and β_c can be fitted and later are used to estimate the national bias:

$$b = (e^{a+\beta_c} - 1)/(e^{\frac{1}{n}\sum log(1+citations)} - 1). \tag{41}$$

The national bias is interpreted as follows: a typical paper in a typical subject from country c receives b times more citations than the world average for that subject. The second method directly computes the geometric mean of the article citation from each country:

$$\mu_{gc} = exp(\frac{\sum_c \log(1 + citations)p_c}{\sum_c p_c}) - 1. \tag{42}$$

This value needs to be divided by μ_g to obtain the national bias, $b = \mu_{gc}/\mu_g$. The national bias computed in this way has a similar interpretation as the first one. They vary slightly when analyzing real data but are significantly differently from other methods, such as the share of the most cited X% articles [233].

Mazloumian et al. introduced a network-based index quantifying knowledge flow from one country to all other countries [235]. The basic idea is to construct a scientific food web and to consider a country as particularly successful (“fit”) if its knowledge is consumed (cited) more than expected. The network flow index is defined as

$$F_{ji} = (\frac{C_{ij}}{R_i - C_{ii}} - \frac{P_j}{P - P_i})\frac{P_i}{P}, \tag{43}$$

where C_{ij} is the number of citations received by papers of country j from papers of country i, $R_i = \sum_j C_{ij}$ is the total number of references listed in the papers of country i, P_i denotes the number of papers produced by country i, $P = \sum_i P_i$ is the total number of papers generated in the considered period. The network flow F_{ji} index is the actual citations per reference (excluding self-citations) from country j to country i abstracting its expected value, which is called the excess citations per reference or the excess knowledge flow from j to i. Finally, the fitness of a country i is defined as the summation of the excess knowledge flows from i to all other countries $K_i = \sum_j F_{ij}$. Although defined with countries, this index can also be used to analyze cities. A map of K_i for major cities around the world is shown in Fig. 11. This figure confirms our general understanding of knowledge-producing areas in the world. By looking at the historical evolution, the productivity of Asian countries has increased yet its dependence on knowledge produced in North America and Europe has also increased over time. Similar analyses have been performed by defining a knowledge diffusion proxy on city-to-city citation networks constructed from APS data, which reveal the spatio-temporal dynamics of physics knowledge worldwide [236].

Cimini et al. introduced a ranking method originally designed for measuring economic complexity to evaluate the scientific competitiveness of nations [101]. The method considers a binary bipartite network in which nations and scientific domains are two distinct types of nodes and a link represents a citation relation from a domain to a nation. The links can be defined in either an extensive or an intensive way. In the former case, a link indicates that a nation ranks in the top-T for the total number of citations received from a domain. In the latter case, a link indicates that a nation ranks in the top-T for the number of citations per paper received from a domain. Based on the Scopus database, $N = 238$ nations and $d = 296$ domains are identified. The method assigns each nation node i a fitness score f_i measuring how competitive a nation is in scientific research, and each domain α a complexity score c_α measuring not only the technical requirements but also the complex social and economic substrate that allows making research on them. These two types of scores depend on each other and are iteratively updated with a nonlinear diffusive process until they converge to stationary values. The motivation of the method is an empirical observation of the triangular structure in the nation-domain adjacency matrix, indicating that a scientifically developed nation diversifies its research in almost all domains whereas a less developed nation can work in only a few given domains that require a low level of sophistication. The iterative equations are accordingly designed as

$$\tilde{f}_i^{(t)} = \sum_\alpha a_{i\alpha} c_\alpha^{(t-1)}, \tag{44}$$

$$\tilde{c}_\alpha^{(t)} = [\sum_i a_{i\alpha}/f_i^{(t-1)}]^{-1}, \tag{45}$$

where $a_{i\alpha}$ is a component in the binary nation-domain adjacency matrix. To avoid divergence, the sum of both $\tilde{f}_i^{(t)}$ and $\tilde{c}_\alpha^{(t)}$ needs to normalize to 1 after each iteration. The nonlinearity is considered in the equation for $\tilde{c}_\alpha^{(t)}$. The complexity of a field depends on the lowest fitness nations that can conduct research on it. Applying this method to the extensive matrix, the highly ranked nations are the G8 countries, and the top domains belong to life sciences and earth sciences. However, the intensive matrix provides completely different results. The top ranking list is occupied by nations like Switzerland, Israel,

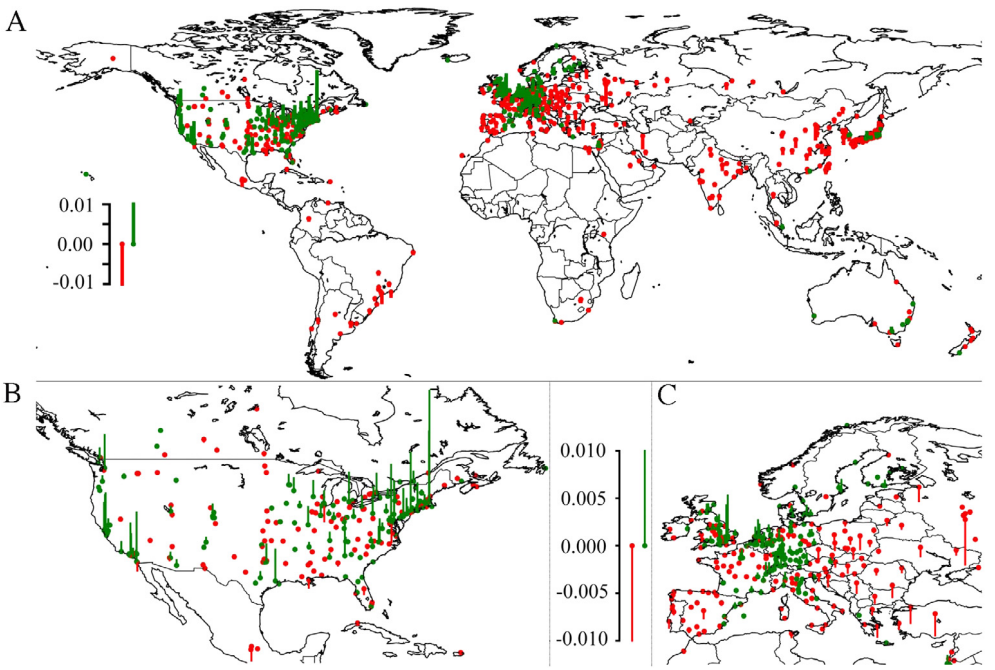

Fig. 11. World map of the major knowledge sources (green) and sinks (red), indicated by the signs of the scientific fitness index.
Source: This figure is reproduced from Ref. [235].

Australia and New Zealand, which are generally considered to be "efficient". In the domain ranking, the social sciences and humanities as well as some medical sciences are ranked at the top.

Institution level. Molinari and Molinari extended the h-index to rank institutions [237]. The original h-index is size dependent, which means that h increases with the number of papers N. In this work, a scaling relation between h and N is discovered, $h = h_m N^{\beta}$, where h_m is a size-corrected impact index and N^{β} is a factor depending on the population size. The exponent β is found to be approximately 0.4, which is universal for different universities. Therefore, it is suggested to rank universities with the size-independent metric h_m. This approach was later re-examined by Kinney with the aim of setting the highest standard of comparison for US federal investment in science [238]. As biology is generally more highly cited than other disciplines, Kinney limits the data of the research fields excluding biomedical research. The universal growth rate $h = h_m N^{\beta}$ with $\beta = 0.4$ is confirmed across different institutions and research fields. Comparing h_m between private and public institutions in the US, it is found that many national science facilities are ranked higher than the leading universities.

Instead of citation data, Clauset et al. ranked universities based on the hiring relations between them [239]. They selected 19,000 university tenure-track or tenured faculty and 461 North American departmental or school-level academic units in the disciplines of computer science, business, and history. The institutions where these scientists obtained their doctorate are recorded. The hiring networks are then constructed, with the nodes representing the institutions and the directed and weighted links indicating the number of doctoral graduates that one institution hires from another. The hiring network exhibits a heterogeneous out-degree distribution, indicating a strong inequality in faculty production. Furthermore, clear prestige hierarchies can be observed in the hiring networks, i.e., institutions tend to hire doctoral graduates from other institutions with higher prestige. The prestige hierarchies can be obtained by altering the ranking positions of institutions to minimize the total weight of the "upward" links. The hierarchy ranking is a measure of the prestige of the institutions, which well overlaps with the authoritative rankings from the U.S. News & World Report and the National Research Council (NRC). Finally, a core–periphery pattern is uncovered in the hiring network. The standard measures of network centrality correlate strongly with prestige rank, which implies that ideas originating in the high-prestige core spread more easily throughout the discipline with the mobility of scientists.

In another study, Crespo et al. proposed a new method to assess the scientific merit of a group of papers in a specific research field [240]. The metric calculates the probability that a randomly set of n articles drawn from a given pool of articles in that field has a lower citation impact (measured by h-index) than the target group of papers. By regarding the publications of different institutions as this group of papers, the metric can compare the impact of institutions in specific fields. The Nature publishing group also proposes its own metric called the *Nature Index* to evaluate scientific achievement at the institutional, national and regional levels. It is simply a database of author affiliation information collated from research articles published in a group of 68 high-quality science journals that are selected independently. This index is publicly accessible via http://www.natureindex.com/ and is frequently updated.

Research field level. Ranking research fields is critical for understanding the development of science. Measuring the relative importance of all scientific subfields and their interrelations is crucial for science funding agencies and policymakers to decide which science or technology fields to prioritize and allocate resources efficiently. The fitness-complexity iterative model mentioned above can estimate the complexity of different research domains [101]. Furthermore, Shen et al. [241] developed a specific method to evaluate the significance of different research domains under the framework of input–output analysis. Specifically, they modified the open-system Leontief input–output analysis in economics into a closed-system analysis focusing on eigenvalues and eigenvectors and reveal the importance and interrelations via the hypothetical extraction method. The output X^i of a target subfield i is represented as the total input x^i_j to other subfields including itself (i.e., the total citations projected from the scientific publications). X^i can be expressed as

$$X^i = \sum_j x^i_j = \sum_j b^i_j X^j \rightarrow X = BX \tag{46}$$

where $b^i_j = x^i_j / X^j$ and B is regarded as the direct input–output coefficient matrix having one as the largest eigenvalue. Using the hypothetical extraction method, the relative importance of subfield i is defined as the change in the largest eigenvalue $\lambda^{(-i)}$ of $B^{(-i)}$, and the interrelations among subfields are calculated as the relative change of the corresponding eigenvector $|\lambda^{(-i)}\rangle$ as following

$$\Delta^i_k = \frac{X^k - \lambda^{(-i)}(\sum_{l\neq i} X^l)\langle k|\lambda^{(-i)}\rangle}{X^k}. \tag{47}$$

After applying this method to the subfields of physics classified by the PACS codes, the authors identified some important subfields and their influences on each other. For example, *statistical physics* shows more influence than that measured directly from citation; in addition, the closely relation between subfield of *quantum mechanics* and subfield of *mechanical control*, which is hidden in the direct citation measurement, is uncovered. This method is potentially applicable to other systems having input–output relations.

4.5. Ranking in science with multilayer networks

The idea of describing real complex systems with multiplayer networks began around 2010. Various issues are studied in those pioneer works including the explosive cascading caused by the interdependency between networks, community detection with multi-slide networks, spreading on multilayer networks and so on [52]. The scientific publication data consists of various types of information such as references, keywords, author names, affiliations, and countries. A more complete model of such data is to describe it as multilayer networks. Naturally, evaluating the scientific impact of papers, authors or entities at higher levels would benefit from considering the multilayer structure. Some ranking methods and even some validation models have been developed by considering the science system as a multilayer network; however, many more studies are needed on this topic, and we believe that the outcome of these studies will be fruitful.

With the citation data and authorship data, a multilayer network consisting of both author and paper nodes can be constructed with two types of links between these nodes. Based on the publication data collected in the econophysics field, Zhou et al. constructed a bipartite multilayer network based on this idea [44]. The citation links point from author nodes to paper nodes, while the authorship links start from paper nodes and end at author nodes. They also proposed an iterative algorithm called author-paper rank (AP rank) in this network to quantify scientist prestige Q_{s_i} and publication quality Q_{p_α}. They argue that a paper is expected to be of high quality if it is cited by prestigious scientists and that high-quality papers will raise the prestige of their authors. Mathematically, this method can be expressed as

$$Q_{s_i} = \sum_{\alpha\in P} \frac{Q_{p_\alpha}}{k^{\text{out}}_{p_\alpha}} \cdot b_{\alpha i}, \tag{48}$$

$$Q_{p_\alpha} = 1 + \sum_{i\in S} Q_{s_i} \cdot a_{i\alpha}, \tag{49}$$

where $k^{\text{out}}_{p_\alpha}$ is the number of authors a paper p_α cites, $b_{\alpha i}$ is a component of the citation matrix (if $b_{\alpha i} = 1$, paper α cites author i), $a_{i\alpha}$ is a component of the authorship matrix (if $a_{i\alpha} = 1$, author i is an author of paper α). Q_{s_i} and Q_{p_α} are iteratively updated until they are converged. The stationary values are respectively the measure of scientist prestige and publication quality. Considering the aging effect in citation, a time-dependent AP rank is proposed as an extension. The mean value of Q_{p_α} for papers published in a journal is computed to evaluate the impact of this journal. The results of the AP rank algorithm are not strictly validated but are shown to be consistent with general understanding.

The citation data and authorship data are also used to create another form of multilayer network, with citation links between papers and authorship links between authors and papers. Based on this network, Sayyadi et al. designed an iterative method called FutureRank, which ranks papers by estimating the expected future PageRank score [242]. This multilayer network is characterized by the two adjacency matrices $M^C_{i,j}$ and $M^A_{i,j}$. $M^C_{i,j} = 1$ if paper p_i cites p_j, and 0 otherwise. $M^A_{i,j} = 1$ if researcher a_i is an author of paper p_j, and 0 otherwise. By assuming that important articles are written by researchers with

high-reputation and are cited by many important articles, the FutureRank is simply an iterative algorithm running one step of PageRank on the citation layer and one step of HITS on the authorship layer and then combining their results. The paper scores and author scores are denoted by R^P and R^A, respectively, and the iterative process can be written as

$$R^A = M^A * R^P, \tag{50}$$

$$R^P = \alpha * M^C * R^C + \beta * M^{A^T} * R^A + \gamma * R^{Time} + (1 - \alpha - \beta - \gamma) * [1/n], \tag{51}$$

where α, β and γ are the parameters that tune the weight of the PageRank score in the citation network, the authorship score in the authorship network and a pre-computed"personalized" PageRank decaying with time $R_i^{Time} = e^{-\rho(T_{current} - T_i)}$, respectively. Similar to other iterative methods, FutureRank also needs to be iterated until reaching convergence. FutureRank is validated on an arXiv dataset by predicting the future citation numbers for papers. A higher prediction accuracy is obtained with FutureRank than with the standard PageRank method.

Zhou et al. considered an additional relation between authors and constructed a multilayer network consisting of a social network that links authors, a citation network that connects publications, and an authorship network that ties the previous two together [243]. A co-ranking method is devised by coupling the PageRank processes in the social network layer and the citation network via the authorship relations. The top-20 authors identified with the co-ranking method are merged and reshuffled with the top-20 authors ranked by four other traditional methods based on publication count and citation count, and then submitted for expert judgment. The assessment scores indicate that the new co-ranking method outperforms the other four ranking methods.

Numerous multiplayer networks are composed of layers that go beyond the author and paper nodes. Based on a scientific forum, Liao et al. extracted the interaction data including paper submission, download, abstract view between users and papers, and authorship between papers and authors [244]. They constructed a multilayer network with user nodes, paper nodes and author nodes, which are assigned a user reputation score, a paper quality score and an author credit score, respectively. These scores are coupled and are iteratively updated following a HITS paradigm. One advantage of this method is that its rankings are relatively robust with respect to active authors with low fitness content. Yan et al. introduced journal nodes into the multilayer structure and constructed a tripartite network [245]. In addition to the authorship links and citation links, a link between a paper and a journal represents that the paper is published in this journal. Several principles are assumed in this network: authors and journals have higher impact if their papers are cited by important papers, and papers are more influential if they are cited by important papers, prestigious authors or high-impact journals. Following these principles, a P-Rank algorithm is proposed. Denoting the A_{journal} and A_{author} as the paper-journal publishing adjacency matrix and the paper-author authorship adjacency matrix, respectively, the P-Rank scores of authors and journals can be written as

$$x(v)_{\text{author}} = A^T_{\text{author}} \times x(v)_{\text{article}}, \tag{52}$$

$$x(v)_{\text{journal}} = A^T_{\text{journal}} \times x(v)_{\text{article}}, \tag{53}$$

where $x(v)_{\text{article}}$ is the P-Rank score of articles which should be obtained in another equation coupled with $x(v)_{\text{author}}$ and $x(v)_{\text{journal}}$. Denoting $\bar{M}$ as the fractioned citation matrix with $\bar{M}_{ij} = 1/k_i^{out}$, $x(v)_{\text{article}}$ can be updated via

$$x(v)_{\text{article}} = (1 - d)(I - d\bar{M})^{-1} v, \tag{54}$$

where d is the damping factor and v is a personalized factor combining the effect from author and journal nodes,

$$v = (\alpha((x(v)_{\text{author}}/np_A)^T \times A^T_{\text{author}}) + \beta((x(v)_{\text{journal}}/np_J)^T \times A^T_{\text{journal}}))^T; \alpha + \beta = 1, \tag{55}$$

where np_A and np_J are vectors with the number of publications for each author and journal, respectively. α and β tune the mutual dependencies of papers, authors and journals. This method is not directly validated with prize-winning data but is examined based on a correlation analysis with other major indices. P-Rank is shown to be a size-dependent metric revealing prestige, but the results also show popularity and prestige are highly correlated.

Based on the multiplayer networks of papers, authors and journals, Jiang et al. proposed a framework named MutualRank to simultaneously rank papers, researchers, and venues [246]. The focus of this work is alleviating ranking bias by leveraging heterogeneous network structures. Yu et al. constructed a multilayer network with four intra-networks and three inter-networks of papers, authors and journals [247]. The intra-networks contain a paper citation network, author citation network, author co-authorship network, and journal citation network. The inter-networks include an author-paper network, paper-journal network and author-journal network. A method based on the PageRank [173] and the Hyperlink-Induced Topics Search (HITS) [248] algorithms is then proposed to rank journals.

The multilayer networks from scholarly data are also adopted by researchers as real examples to validate new definitions of node centrality in multiplex networks. Bianconi et al. performed multiplex PageRank analysis of the Physical Review E Citation–Collaboration Network and defined it as a new centrality for multiplex networks [249,250]. In the same dataset, Iacovacci et al. defined another centrality called Functional PageRank to the citation/collaboration network of PRE authors, which shows different success patterns for scientists [251].

Validating the proposed metrics is a critical issue. Many of the proposed indices are not strictly validated but are only compared with the existing ones using a correlation analysis. A number of methods are validated with real data by either predicting future citations or the mean ranks of the awarded papers/authors. In a recent paper, Medo et al. developed a model mimicking the dynamics of authorship and citation relations in a multilayer network consisting of authors and papers [142]. This work provides a unique benchmark for evaluating different ranking metrics with predefined ground truths. The model starts with A fixed number of authors in the network with preassigned ability a_i and productivity k_i. The model evolves with a growing process; in each step, papers are gradually introduced into the network. The number of authors for each paper $d_\alpha = |A_\alpha|$ is drawn from an author-number distribution identical to that of the Microsoft Academic Search (MAS) data, where the author names have already been disambiguated. d_α different authors are randomly selected from the authors with probability proportional to the remaining productivity (i.e., k_i subtract the number of papers i has authored). Each paper has a fitness value f_α, which is set as proportional to the highest ability value among its authors with a level of noise η, i.e., $f_\alpha = \eta \max(max_{i\in A_\alpha}(a_i))$, where η is generated from a uniform distribution in $[1-\eta^*, 1+\eta^*]$. Every new paper cites q existing papers. The citation from a newly entered paper α to an existing paper β is established according to the probability

$$P_{\alpha\to\beta}(t) = \frac{[c_\beta(t)+1]f_\beta D(t-\tau_\beta)}{\sum_\gamma [c_\gamma(t)+1]f_\gamma D(t-\tau_\gamma)}, \tag{56}$$

where $c_\beta(t)$ is the cumulative citations of paper β, and τ_β is its appearance time in the network, $D(\cdot)$ is the aging factor following an exponential decay. The model can generate a consistent distribution for a variety of quantities to real data, including the number of authored papers, the number of co-authors and the number of citations. More importantly, this model provides the quantities a_i and f_α, reflecting the intrinsic ability of researchers and the quality of papers. Therefore, the previously designed ranking metrics can be validated with this model, and their accuracies can be systematically compared. Medo et al. consider several traditional metrics and find that the average citation number efficiently measures author ability, while the total citations and h-index measure the joint outcome of the ability and productivity of researchers. We encourage readers to examine other metrics with this model.

Remarks for this section. The increasingly available scholarly big data provide an unprecedented opportunity to explore the networks that represent research activity and to thus better evaluate researcher performance. Despite considerable research efforts, scientist productivity remains difficult to measure. Therefore, in the future, we expect to see continued effort in the developing more objective evaluation metrics for science. Combined with the previous studies on this topic, we would like to make several remarks at the end of this section: (i) Given that there are already numerous metrics in the literature, newly proposed methods need to have a clear motivation, which is overlooked by the previous methods. Otherwise, the new method will remain unnoticed. (ii) The proposed methods need to be validated in order to clearly show their advantages and disadvantages. Several validation methods can be considered including the expert judgment, future prediction, academia prizes, ground truth from mechanistic model and so on. (iii) Although evaluation metrics would directly benefit the academic community and funding agencies by lowering the necessary workload of the experts who are instrumental in research evaluation, they cannot completely replace expert evaluation. If some biased metrics are used, funding decisions might be misdirected, which directly result in ineffective allocation of research funds and promising talent leaving academia because of the lack of support. Therefore, we emphasize that the ranking metrics based on scholarly data are more appropriate to be treated as a reference for evaluation in science, complementary to expert assessment.

5. Microscopic and macroscopic prediction in science

Prediction of a system is made possible by the accumulation of our understandings of it. In science, prediction is everywhere, ranging from an individual scientist predicting which research topics will more likely result in a high-quality paper to a funding agency predicting which proposals will have a higher chance of producing more fruitful outcomes. Although there are strong uncertainties in scientific discoveries, the progress of science remains to some degree predictable in the sense that various mechanisms can be extracted from scholarly data and be extended to forecast the future evolution of the system. The research of prediction in science has largely advanced in the past decade with big data and the rapid development of the Internet, which accordingly stimulate a variety of prediction tools [252]. In this secton, we will review the prediction methods devised to tackle the microscopic and the macroscopic prediction issues in science.

5.1. Link prediction in citation and collaboration networks

Link prediction is a microscopic prediction seeking to retrieve missing links and forecast future links in a network [253]. By estimating the likelihood that two nodes will interact with each other with the observed network structure, prediction of links is realized. Such a research topic is strongly connected to many other fields such as online product recommendation [222], biological network reconstruction [254,255] and community detection [64]. The existing methods can be roughly divided into three categories: node-based similarity algorithms [184,256–258], path-based similarity algorithms [258–260] and Bayesian estimation algorithms [261]. In all of these methods, a score s_{ij} is computed for each node pair ij as a measure of the likelihood that the nodes will be connected. The not-yet-connected node pairs with a high s_{ij} are the predicted future links. Some of these methods have been applied to identify spurious interactions in networks by assuming that the connected node pairs with low s_{ij} are more likely to be fake. To date, the validation of these link-prediction algorithms is usually done

within the framework of a training set (considered the observed network) and a probe set (considered the future network) data division [262]. In practice, 10% of the links in real network data are usually randomly selected and placed in the probe set E^P and the remaining links form the training set E^T. Link-prediction algorithms are run on the training set, whereas the testing set is used to measure the accuracy of the prediction. The metric called *area under the receiver operating characteristic curve* (AUC) is usually employed to quantify the accuracy of the prediction. In practice, the metric represents the probability that a true link has a higher link-prediction score s_{ij} than a nonexistent link. Out of n times of comparisons, there are n_1 times the probe set links having a higher s_{ij} and n_2 times the probe set links having the same rank as the nonexisting links. The mathematical form of the AUC metric is $AUC = (n_1 + 0.5n_2)/n$. A higher AUC indicates a more accurate prediction of links. Although this validation framework is widely adopted in the literature, researchers have recently started to consider a more realistic validation manner in networks with temporal information. The training set and probe set are no longer divided randomly but strictly according to time. Preliminary analyses showed that the accuracy of many link-prediction methods is seriously lowered in the temporal validation, indicating the difficulty in realistic link prediction [263,264]. The question of how to improve the prediction accuracy for these cases, however, remains for further investigation.

Here, we will focus on the link prediction in networks constructed from scholarly data, primarily including the citation and collaboration networks. Prediction of links in these networks is particularly meaningful because it allows the inference of the missing citations in the incomplete citation map and forecasting of future collaborations between scientists. As extensions, link-prediction algorithms can also be used as tools for removing unreliable citations from the citation data and recommending possible collaborators for scientists. The key factor in link prediction is to design an algorithm that can calculate an accurate estimation of s_{ij}. Numerous topological similarity metrics are introduced for this purpose. The basic motivation is that similar nodes are more likely to be connected in the future, which is also supported by empirical studies. Some representative similarity measures are listed below.

- Common neighbors (CN): $S_{ij}^{\mathrm{CN}} = \|\Gamma_i \cap \Gamma_j\|$, in which Γ_i is the set of neighboring nodes of node i and $\|...\|$ indicates the number of nodes in a set.
- Jaccard (JA) [256]: $S_{ij}^{\mathrm{JA}} = \|\Gamma_i \cap \Gamma_j\|/\|\Gamma_i \cup \Gamma_j\|$, in which the denominator seeks to remove the bias of the *CN* metric toward large-degree nodes. Similar indices with the same purpose but with different forms are many, such as Cosine index [184], Sørensen index [257], Hub promoted index [265], and Leicht–Holme–Newman index [258], to name a few.
- Resource Allocation (RA) [266]: $S_{ij}^{\mathrm{RA}} = \sum_{k\in\Gamma_i\cap\Gamma_j}\frac{1}{\|\Gamma_k\|}$. This metric suppresses the similarity between two nodes by having many large-degree common nodes. Adamic-Adar index is a similar metric.
- Katze (KA) [259]: $S_{ij}^{\mathrm{KA}} = \sum_{l=1}^{\infty}[(\beta A)^l]_{ij}$, in which β is a free parameter that must be lower than the reciprocal of the largest eigenvalue of A. This metric evaluates the similarity between nodes via paths of all lengths.
- Local Path (LP) [260,266]: $S_{ij}^{\mathrm{LP}} = (A^2)_{ij} + \epsilon(A^3)_{ij}$, in which A is the network's adjacency matrix and $\epsilon < 1$ is a free parameter. This metric computes node similarity with path length up to 3, which can effectively break the degeneracy of node similarity in sparse networks but with low computational complexity.

For a more comprehensive review of link-prediction algorithms, readers can refer to [267]. Citation and collaboration networks have their own features apart from other real systems, which requires specially designed link-prediction algorithms. Thus, development of various link-prediction algorithms targeted for these science-related networks appears in the literature.

Citation networks. In this type of network, new links (citations) are created with new nodes (publications) entering the networks. Because link-prediction algorithms can only predict links between existing nodes, in citation networks, related algorithms are not targeting the future links but rather the missing links. Because citation networks are directed, the methods originally designed for undirected networks might perform poorly in these networks. Ciotti et al. thus proposed a novel similarity metric for directed networks, based on the Statistically Validated Network (SVN) approach [268]. The basic idea is that the similarity between two papers i and j is measured by the number of co-references, with the statistical significance determined by an associate p-value. They also showed that the probability of a paper citing another increases with the similarity between them, which indicates that the similarity can be applied to identifying the missing citations. The method can also be treated as a recommendation procedure that proposes a list of papers that the authors might find relevant to their works. The proportion of missing citations is then used to quantify the dissemination of knowledge in different journals and research fields.

Another method proposed by Zhang et al. predicted missing links in citation networks via directed motifs [269]. For a specific type of motif, the number of such motif will be formed by adding a link from i to j defined as the "potential" from i to j. The node pair with higher potential is assumed connected with a higher probability. The accuracy of 12 motifs in link prediction is compared. It is found that the Bi-Fan motif (consisting of four nodes a, b, c, d, with a and b citing c and d, respectively) has the highest AUC among these 12 motifs. Shibata et al. designed a supervised machine-learning model to predict links in citation networks [270]. Eleven features are used as the input for the machine-learning model and three of them, i.e., link-based Jaccard coefficient, difference in betweenness centrality, and cosine similarity of term frequency-inverse document frequency vectors, are found to be key factors that largely affect the predictions. By examining the resultant weights of each feature, they concluded that the best link-prediction method varies in citation networks of different research areas.

Collaboration networks. A collaboration network is an instance of social networks; therefore, most of the link-prediction methods developed for social networks can be naturally applied to collaboration networks. In an early work, Liben-Nowell and Kleinberg [271] proposed to measure the "proximity" of nodes in co-author networks in physics, based on which missing links are predicted in these networks. They explored various indices including common neighbor, Jaccard and Adamic-Adar coefficients, preferential attachment, Katz, hitting time, rooted PageRank (at each time step random walkers return to the root with some probability), SimRank, Low-rank approximation, unseen bigrams and clustering. After comparing all of these approaches on the dataset from five sections of ArXiv, they concluded that the Admic-Adar metric performs consistently well across all considered sections. The same approach can show distinct performance on the five sections, indicating that the collaboration patterns in subfields of physics are formed differently.

Numerous other methods are devised. Sun et al. [272] developed an agent-based model for predicting scientific collaborations. The model assumes that agents search for their collaborators via the random walk in the collaboration networks and that the similarity between authors' interests is reflected by their publications' titles. For the authors belonging to disconnected groups, a random attachment mechanism is introduced for establishing their collaboration links. The model is calibrated with three empirical collaboration networks. The prediction accuracy is comparable to the Katz method; however, the same predicted links are quite few. Backstorm and Leskovec [273] proposed an algorithm based on supervised random walks and validate it on an ArXiv dataset. They formulated the problem as a supervised learning task in which the goal is to learn a function of node and edge attributes that assigns bias strengths to edges such that the random walker is more likely to visit the nodes that will form new links to the root node. This approach is shown to outperform the Admic-Adar metric and other unsupervised approaches. Zhang and Yu [274] exploited the semantic content of an author's research profile together with co-authorship network connectivity to predict biomedical research collaborations. They found that similarity of out-going citations and similarity of abstracts are the most informative semantic features.

Recently, works based on graph embedding show great improvement in the accuracy of link prediction. The basic procedure of graph embedding is to initially assign a vector to each node and then embed it into Euclidean or hyperbolic space [114,275–279]. The node pair with the closer distance in the space has a higher probability of forming a connection. One representative work is the DeepWalk [275], which is an extension of the word2vec [280,281] algorithm in Natural language processing (NLP). Leskovec [282] proposed a Node2Vec algorithm and finds that the bias random walk, e.g., depth first search and breadth first search, can reveal the structural similarity and structural equivalence of nodes. After graph embedding of the network, one can easily predict the missing links and recommend future connections based on the distance between nodes. The validation on the arXiv dataset shows that these link-prediction methods based on graph embedding outperform all of the other existing approaches that are based on structural similarity. Kleineberg et al. mapped a multiplex network, within which each connection represents a collaboration between authors and each layer represents an ArXiv category, into a hyperbolic space [277]. They found that these layers were correlated in a hidden geometry and further enabled accurate trans-layer link prediction.

5.2. Predicting future impact and performance

Knowing which paper will have a high impact in the future and which scientist will perform well in research is a very practical issue. On the one hand, it can help scientists to select which research directions to follow. On the other hand, it can guide policymakers in deciding who they should hire and to whom the research fund should bestow a grant. However, predicting future impact and performance in science is not an easy task because of the high complexity of the dynamics. After a persistent effort has been made in this direction, the literature provides us several methods to predict some of the impact metrics for papers and scientists.

Paper impact prediction. Citation is the simplest measure of the impact of a paper; thus, most of the prediction methods for papers seek to capture this metric in the future. Because the preferential attachment is the major driving force of the growth of the citation networks, linear extrapolation is a handy means of predicting the citation of a paper in the short term [283]. However, the substantial heterogeneity in paper quality and the strong aging of papers will make the prediction accuracy of this method decay quickly over time. Therefore, numerous works instead focus on predicting the long-term impact of papers. A method with concrete theoretical foundation is the mechanistic model developed by Wang et al. [32]; a detailed description of the method appears in Section 4. This model can be used not only to quantify the intrinsic quality of papers but also to predict the future citation of papers. With the equation for c_i^t (see Eq. (28)), one can obtain the predicted citation of paper i at any time t. As preparation for the prediction, one must estimate the key parameters η_i, μ_i and σ_i for the model. The prediction accuracy is examined with the APS dataset. The prediction of the total citation of individual papers with a 5-year history and a 10-year history is shown in Fig. 12. Statistically, with a 5-year history as the training data, only 6.5% papers' 30-year cumulative citations fall outside of one standard deviation of the prediction. The prediction will be more accurate if more historical data are available. This method was also shown to outperform some classic models for prediction including the logistic [284], Bass [285], and Gompertz [284,286] models. However, it was later noted that the prediction accuracy of this model heavily depends upon the fitting quality of the key parameters [186].

With the historical citation data, Cao et al. also devised a completely data-driven method for predicting papers' future impact [287]. The method seeks to predict the future yearly citation $x_{N+1}, x_{N+2}, \ldots, x_{N+M}$ of a paper given that its $x_N, x_{N-1}, \ldots, x_1$ are known. It calculates for each past paper in the database the matching error of the first N years' citation dynamics between this paper and the target paper as $e_x(y) = \sum_{n=1}^{N}(y_n - x_n)^2$. The matching error can be viewed as the

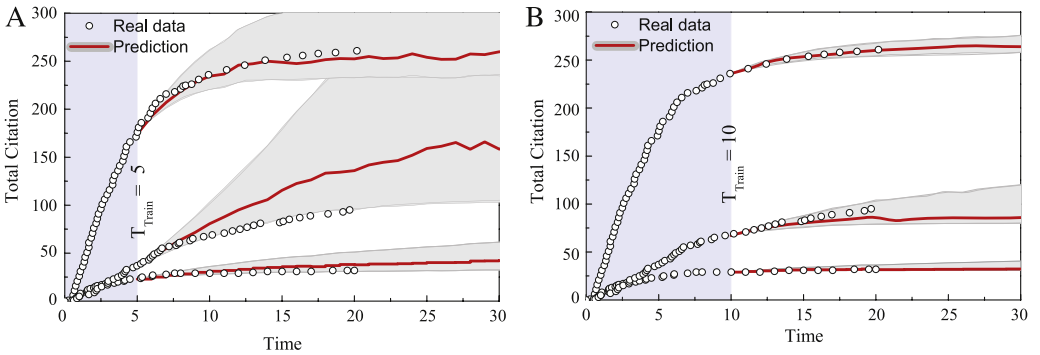

Fig. 12. The prediction of the total citation of individual papers using the mechanistic model in Ref. [32] with (A) 5-year history and (B) 10-year history.[1] Each envelope illustrates the range for which the standard deviation $z < 1$
Source: The figure is reprinted from Ref. [32].

temporal similarity between these two papers. The basic idea is that the citation dynamics of a similar paper can be used to predict the future citation of the target paper. In practice, two methods are considered. If L similar papers are found, the first method, which is called average method, simply predicts the future citations of x as $\hat{x}_p = (1/L)\sum_{l=1}^{L} y_p^{(l)}$ where $p = N+1, \ldots, N+M$. The second method is called a Gaussian mixture model method and predicts several possible trends of the target paper's future citations. L matched papers are grouped into K clusters by fitting a Gaussian mixture model with K Gaussian components. The mean citation of the papers in each cluster is calculated and considered a possible trend for the target paper's future citation. The weights of each Gaussian component can be considered the probability of that corresponding trend. In general, the second method performs better than the first one. Both methods are shown to achieve greater prediction accuracy than is achieved by the method in Ref. [32]. In addition, their accuracy can be maintained as satisfactory if only short-term historical data are available (e.g., 3 years).

Another type of methods seek to incorporate additional information [288–290] apart from citation history to predict the future impact of a paper. Sarigöl et al. examined the relationship between the centrality of authors in the collaboration network and the citation received by their publications [291]. If an author is within the top 10% k-core centrality in the collaboration networks, there is a positive conditional probability of his/her papers to be among the most successful. This conditional probability differs from one centrality measure to another, and it becomes even higher if the author is required to be within the top 10% on all four considered centrality metrics (i.e., degree, k-core, betweenness, and eigenvector). Inspired by this empirical observation, a machine learning classifier using the multidimensional feature vector of collaboration-network centrality metrics is advanced to predict the future citation success of a paper. The precision of identifying a highly cited paper is improved by a factor of six compared with a random guess. Similarly, Brizan et al. also predicted an influential set of papers with a machine learning method that considers 48 different features including author, community, longevity, loyalty, publication, title and abstract categories as input [292]. The key features for prediction are identified and discussed. Yu et al. constructed a feature space that contains 24 features of four types (e.g., external features of a paper, features of authors, published journals and citations) to describe scientific papers [293]. They performed a stepwise multiple regression analysis between paper citation impact and these features and obtain an optimal regression model with only 6 key features involved. This model was shown to be relatively effective and applicable to predict paper citation impact 5 years after publication. McKeown et al. decomposed the scientific publications in scientific concepts (i.e., technical terms), and used machine learning with full-text features to predict the impact of these concepts quantified as relative growth of term appearance in unique documents [294].

Author performance prediction. The h-index is widely adopted as a measure of the scientific success of researchers. Therefore, numerous methods seek to predict scientists' future h-index as a prediction of their future performance in scientific research. In an early work, Hirsch examined the correlation between authors' current performance metrics and authors' future h-index and find that the h-index itself is a better predictor of future h-index [295]. In addition, a modified index combining an author's h-index and total citation counts yields an even better predictor of an author's future performance. This work, however, has been recently challenged by Schreiber, who shows that the increase of the h-index with time often depends for a long time upon citations to rather old publications, making it an inappropriate metric for predicting future scientific output [296].

In the past decade, considerable efforts have been made by researchers to predict h-index. Acuna et al. obtained the formulae for h-index prediction with machine learning [297]. Data in the field of neuroscience are used for the training, and the approximate equations for predicting the future h-index for neuroscientists are

$$h_{+1} = 0.76 + 0.37\sqrt{n} + 0.97h - 0.07y + 0.02j + 0.03q, \tag{57}$$

$$h_{+5} = 4 + 1.58\sqrt{n} + 0.86h - 0.35y + 0.06j + 0.2q, \tag{58}$$

$$h_{+10} = 8.73 + 1.33\sqrt{n} + 0.48h - 0.41y + 0.52j + 0.82q, \tag{59}$$

where h_{+1}, h_{+5} and h_{+10} are, respectively, the prediction of h-index for the future year, 5 years and 10 years. In these equations, n is the number of articles written by the author; h is the current h-index of the author; y is years since the author published the first article; j is the number of distinct journals in which the author has published; and q is the number of the author's top journal articles in *Nature, Science, Nature Neuroscience, PNAS and Neuron*. Several important messages can be identified by comparing these three equations. The length of the career so far is the only factor that contributes negatively to an author's future h-index. In the long term, the diversity of the journals an author published in and the number of top journal papers play an increasingly significant role. Conversely, the current h-index, becomes increasingly less important. Although this method is claimed to have high accuracy, its validity was questioned by a following work testing the regression equations in sample data of Spanish psychologists [298]. It was found that the equations generally overestimate the h indices of authors, and the error grows over years. It was also argued that the error does not show a different pattern in the scientists working within or outside neuroscience. McCarty et al. considered the effect of coauthor networks on an author's future h-index and accordingly propose a regression model for future h-index prediction [299]. The characteristics are used as input, but only the number of co-authors and highly productive co-authors are significant variables. The network structural metrics, however, do not have much predictive power. The results also support the critical role of collaboration for an author's career development.

The predictability of these regression models has been examined more systematically by Penner et al. [300]. They adopted the same factors as Ref. [297] and considered a more general form,

$$h(t+\Delta t) = \beta_0(t,\Delta t) + \beta_h(t,\Delta t)h(t) + \beta_{\sqrt{n_p}}(t,\Delta t)\sqrt{n_p(t)} + \beta_j(t,\Delta t)j(t) + \beta_q(t,\Delta t)q(t), \tag{60}$$

where $\beta_0(t,\Delta t)$, $\beta_h(t,\Delta t$, $\beta_{\sqrt{n_p}}(t,\Delta t)$, $\beta_j(t,\Delta t)$ and $\beta_q(t,\Delta t)$ are parameters that must be fitted in the regression, and the notations of rest parameters are the same in Eqs. (57)–(59). Different from the original model, the β parameter depends not only upon Δt but also upon t, indicating that the parameters relate to not only the future length but also the career age of the author. The model is thus tested on authors with different career ages, and its predictive power for early career researchers is found to be substantially lower than the original model in which all career ages were aggregated. In addition, the authors derive the expected correlation between $h(t+\Delta t)$ and $h(t)$ as $Cor[h(t+\Delta t), h(t)] = \sqrt{t/(t+\Delta t)}$. Apparently, $h(t+\Delta t)$ is highly correlated with $h(t)$ when t is large, suggesting that the observed predictive power of the regression models is largely due to the general properties of the evolution of cumulative measures. In addition, the overestimation of h-index by the regression model is confirmed. By establishing a regression equation of the incremental h-index,

$$\Delta h(t+\Delta t) = \alpha_0(t,\Delta t) + \alpha_h(t,\Delta t)h(t) + \alpha_{\sqrt{n_p}}(t,\Delta t)\sqrt{n_p(t)} + \alpha_j(t,\Delta t)j(t) + \alpha_q(t,\Delta t)q(t), \tag{61}$$

one can immediately see that the predictive power becomes much lower. This work is followed by a commentary by the same authors highlighting the caution for using the regression models in predicting scientists' future impact [301].

Apart from the regression models, Sinatra et al. predicted scientists' h-index via a mechanistic model described in Section 4 [55]. In the model, the 10-year impact of a paper published by an author $c_{10,i\alpha}$ is simply the product of a random variable as luck for the paper p_α and the ability parameter of the author Q_i. To predict the future h-index of a scientist, one should initially estimate the ability parameter via $Q_i = e^{\langle \log c_{10,i}\rangle - \mu_p}$. With Q, one can simulate the impact of future publication of an author and thus predict his/her h-index. Note that because the model relies on the number of the author's publications N to denote time, the prediction in this work is the h-index of the author when he/she publishes his/her Nth paper. The prediction accuracy, as with other mechanistic models, would depend upon the fitting quality of the parameters (μ_p in this case).

As an alternative metric, authors' scientific achievement can also be measured with the total citation of their publications. Thus, if one can predict an author's future total citation, the future performance of this researcher can be captured. Mazloumian considered several predictors, including the average number of citations per paper, h-index, and annual citations, and found that annual citations is the best predictor of future citations [302]. However, it was shown that the future citations can only be accurately predicted in the short-term with this predictor; the long-term accuracy decays very fast over time. In another work, Dorta-González proposed to measure the citation potential of individual authors by computing a ratio between the production (journal papers) and impact (journal citations) dimensions [303]. They found that the resulting citation potential index is a discipline-independent measure, which can be used as a predictor for scientists' future performance and a reference in the selection and promotion process.

5.3. Early identification of the potential publications and scientists

Because the preferential attachment is the major driving force for the growth of citation networks and collaboration networks, predicting the future citations of old papers and future performance of authors with long careers is much easier than the task of focusing on new papers and young scholars. From a practical point of view, early identification of these potential publications and scientists is remarkably more important. Forecasting that a highly cited paper will be more highly cited and that a productive scientist will publish many papers provides little additional information. Identifying the future most highly cited paper from thousands of new papers published each month and the highest-ability young scholars from hundreds of job applicants would provide very valuable references for researchers and policymakers. However, this problem is not an easy one. Although numerous prediction works are published in the literature, relatively few of them are devoted to the early prediction of scientific impact of papers and authors.

This dominant role of preferential attachment in citation-network growth results in a strong effect called first-mover advantage, meaning that early nodes in the networks tend to have more links [117]. Most of the prestigious algorithms (e.g., citation and PageRank) fail to correct this bias, resulting in an extremely skewed prediction of papers citations. In other words, the future most highly cited papers predicted by these methods are simply the most highly cited papers previously. Newman thus designed a predictor by calculating the z-score as the number of standard deviations above the mean for citations in papers published around the same date [117]. This method can filter out preferential attachment and identify promising papers even when they have not yet received many citations. In a paper published in 2013, Newman revisited the promising papers in network science identified by the z-score method in 2008 and finds that these papers indeed receive much more citations than their peers published in similar periods [166]. A similar approach can be designed with the PageRank algorithm. Mariani et al. proposed a rescaled PageRank algorithm by defining a z-score as the number of standard deviations above the mean PageRank score for the peer papers [179]. Although the method is not yet applied for prediction, it is expectable that the method will perform similarly well in predicting the promising papers.

Incorporating additional information with a citation is a natural means of improving prediction accuracy when the amount of historical citation data is limited. Stegehuis et al. proposed a regression model for predicting a recent paper's future citation by considering the impact factor of the journal in which the paper has been published and the number of citations this paper received in the first year after publication [304]. The corresponding equation is

$$\ln(q(p|IF, c_1)) = \gamma_p \ln(c_1 + k_0) + \beta_p \ln(IF) + C_p, \tag{62}$$

where c_1 is the number of citations in the first year, IF is the impact factor of the journal, k_0 is a constant chosen as 0.5. $q(p|IF, c_1)$ is the pth quantile of the long-term citation distribution of a publication, conditioned on the impact factor and the number of citations in the first year. This model is fitted using quantile regression to estimate the values for γ_p, β_p and C_p. Both the coefficient β_p and γ_p are decreasing with p, while C_p increases with p. It can be interpreted as indicating less influence of the IF and c_1 on the long-term impact of highly cited publications but more influence on the long-term impact of average cited publications. Note that this regression model seeks to predict the probability distribution for the future number of citations of a publication instead of the future number of citations itself. Because the distribution can be wide, given that the quantile is accurately predicted, the real number of future citations of a publication can still vary substantially.

Recently, Medo et al. identified a group of users in the user–item online commercial bipartite networks capable of being constantly among the firsts in selecting future popular items when they have only a small number of links [305]. These users are denoted as discoverers and an index called surprisal is derived to measure the significance of a scientist being a discoverer. Identification of these discoverers allows an early prediction of item popularity by determining whether there is any discoverer selecting the items in their early stage. Similarly, a scientist-paper bipartite network with a link between them representing the citation from a scientist to a paper can be constructed. The approach can reveal whether a group of scientists exist who are very sensitive to the latest high-quality publications and cite them even when they have very few citations. The prediction of promising papers can be done accordingly.

The future performance of young scientists is even more difficult to predict because there are high levels of uncertainty in their subsequent career [306]. As discussed above, the scientist-oriented prediction methods in general perform poorly when focusing on young scientists with short-term historical data available. However, there remain several works pushing this research direction forward. Qi et al. provided empirical evidence with American Physical Society (APS) data showing that collaboration with outstanding scientists (measured by their total citation) will significantly improve young researchers' careers [307]. Amjad et al. confirmed this phenomenon and additionally showed that working with leading experts is not the only path to a successful career [308]. These findings indicate the possibility of foreseeing a young scientist's future performance from his/her early collaboration relationships; however, a concrete prediction method is not proposed. Laurance et al. made use of the scholarly data of biologists to evaluate the effect of various factors including gender, native language, prestige of the institution at which they received their PhD, the date of their first publication, and their pre-PhD publication record on long-term publication success of young scientists [309]. The future success is measured with scientific productivity in this work, i.e., the future publication number. With a generalized linear-model format with a gamma error distribution and log-link function, the predictive power of individual factors and their combinations is assessed with the Bayesian information criterion weights and the structural goodness of fit. Four conclusions were reached in this study. First, long-term publication success is largely determined by early publication success, which is suggested by the strong correlation between Pre-PhD publication number and future publication number. Second, publishing early in one's career has a clear advantage for one's future career. Third, being a native English speaker and a male researcher indeed contributes to future success. Finally, a prestigious university does not promote one's future success significantly. A more detailed discussion of the key factors for a successful scientific career will be provided in Section 6.

5.4. Prediction of collective trends in science

Predicting the collective trends of science – the birth, growth and decline of research fields or disciplines – is meaningful for scientists to make research plans and for policymakers to allocate resources. A direct means of quantifying the evolution of research disciplines is to use the collective trends of scientific concepts. Although the precise scientific concepts and their relationships are hard to define, co-word analysis, a content technique commonly used in natural language processing, has been acknowledged an effective means of capturing the intuitive and cognitive picture of knowledge embedded in

publications and of modeling the dynamics of knowledge structures. It is essentially a type of analysis based on the co-occurrence of words in documents. One of the improved versions of co-word analysis with a diffusion process has been used to reveal the scientific memes in physics [161]. Here, the constructed co-word networks are employed to discern relationships among various scientific concepts and furthermore to reveal the evolution of research topics and knowledge. Choudhury and Uddin attempted to predict future co-occurrence trends among scientific concepts via combining network topological similarities and time-series prediction methods, which are verified on three different research domains [310]. Specifically, they extracted the keywords selected by authors representing the authors' understanding of their work, then purified them to a standard form and construct a keyword co-occurrence network. Supervised learning models are applied to predict future associations among keywords, given a set of commonly used topological metrics and their temporal evolution information. In addition to predicting the interactions, they also predicted the future values of specific structural features, e.g., the hotness of keywords reflected by their node degree in the co-word networks, by using the time-series analysis methods. Despite having developed the methods for predicting the interactions and popularity of existing concepts, predicting the emergence of new concepts (new nodes) remains a challenge. Future effort toward addressing this problem is expected.

The evolution of disciplines and research topics is largely driven by the social dynamics of scholar communities; therefore, the future trends in science can be captured by predicting the collective behavior of scientists. With this idea, Sun et al. offered an agent-based model to describe various dynamics of discipline evolution, in which agents represent scholars and their formed groups by collaboration represent disciplines [311]. The key idea of their model is that new scientific fields emerge from the splitting and merging of these groups. The model can be described briefly as follows. Each scholar is characterized by a list of disciplines representing the scientific fields they have been working in, each discipline contains a list of papers, and each paper is assigned a major discipline and several additional disciplines. The model starts with one scholar writing a paper in a discipline. The network then evolves as new scholars join, new papers are written, and new disciplines emerge over time. At each time step, a new paper is added to the network, with the first author randomly selected from the network and co-authors selected via a biased random walk. Starting from the first author, the random walker moves to an adjacent node, with probability $1-p_w$ favoring collaborated authors, and stops on the current node with probability p_w. Every visited node becomes an additional co-author. The major discipline of this new paper is selected as a discipline that is shared by the majority of the authors. Each coauthor will add this major discipline to his/her own list. A new scholar will also be added to the network with a certain probability at each step together with a new paper authored by this new scholar. Like other new papers, the first author will be randomly selected from the collaboration network, and the remaining coauthors are selected via biased random walk. To model the evolution of the disciplines, the collaboration network is split or merged based on the change in the modularity of the new networks. Specifically, for a split event, a collaborator network corresponding to a random discipline is extracted, and this collaborator network will split if the modularity of the partition is higher than the original one. For a merging event, two randomly selected disciplines with at least one coauthor are extracted, and two clusters will be merged if the modularity obtained by merging these two single groups increases. The simulation of this model reproduces the stylized facts about the relationships between scholars, papers and disciplines reflected from real data. The model can also predict the collective research trend of any interested discipline by running the model with the settings of the current condition.

In addition to the splitting and merging of disciplines, predicting the decline of specific disciplines also has practically value. Such a macro phenomenon can be decomposed into micro levels as researchers move from an old topic to new topics on which they have not have published. Such a change is always triggered by some pioneering scientists from whom the new scientific ideas are introduced to the old community and rapidly spread through multiple influence channels, e.g., mutual connectivity through friendship, collaborations and directed links representing the leader and follower effect. Hu et al. formulated such a phenomenon as a problem of influence spreading and establish a percolation model on a multiplex-correlated graph with hidden "influence links" representing the idea flows among scientists [312]. The network under which the viral spreading is occurring has two types of links: undirected connectivity links corresponding to the co-authorship relationships and influence links quantified by the citing behavior. Some pioneers' leaving will make their collaborators disconnect from the giant component (the research community) and result in their leaving too. This process is called the "connectivity-percolation" process. The "influence-induced" process follows thereafter. The researchers will leave the community following the influence links, and extra researchers will leave the community if they are no longer connected to the giant connected component anymore. These two processes will induce a cascading effect, as illustrated in Fig. 13. By combining generating function and percolation theory, Hu et al. analytically predicted the percolation phase transition, which shows good agreement with the empirical study on the APS data. Given that the threshold of the phase transition is identified, one can easily predict the change of collective trends. If the fraction of departing pioneers has exceeded the critical value, the remaining researchers will soon leave the old research topic, and the community will soon break down. If it is far from the critical value, the community for this particular research topic will remain active for a relatively long time.

6. Paths to success in science

Success is desired by people working in different jobs, yet many critical issues related to success still remain unclear, and the key factors of success are also under debate. The high-quality data provided by scientific publications provide a unique opportunity to study the paths to success quantitatively. With the scholarly data, one can comprehensively compare

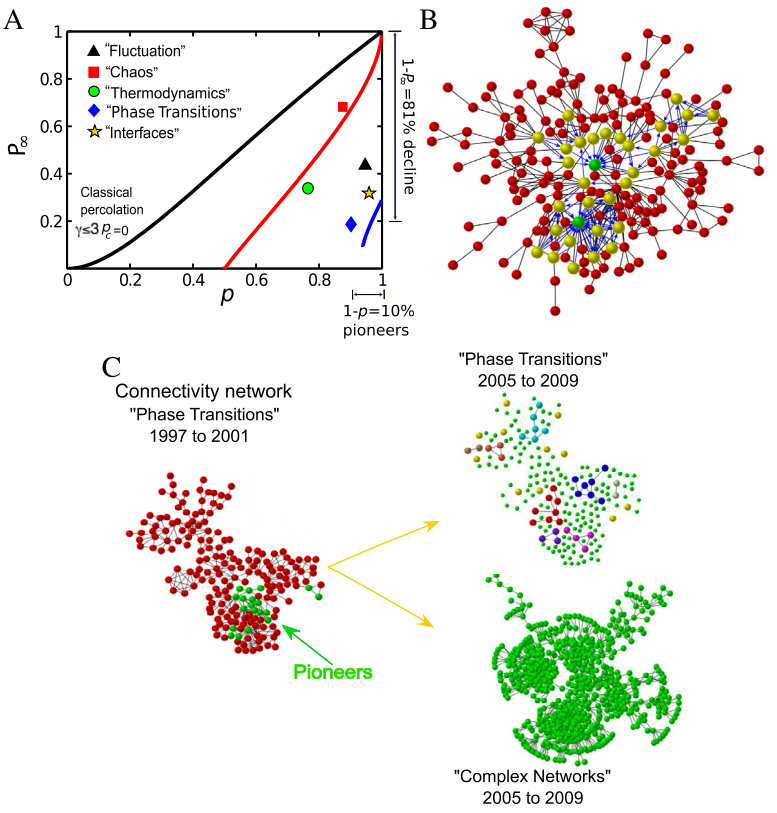

Fig. 13. The illustration of the cascades of followers in APS after some pioneers leave the field. (A) The relative size of the giant component of the collaborating scientists $P_\infty(p)$ versus the fraction of $1-p$ pioneers departing from the field. The black, red and blue curves are respectively the results predicted by classical percolation theory, influence-induced correlated percolation with two different parameter settings. The red and blue curves are bounds to the real data. (B) The influence network of collaborating scientists in the field of "Phase Transitions", with the blue and black links respectively as the influence and connectivity links. Green nodes are a sample of pioneers of the field of "Complex Networks", and the yellow nodes are their closest followers that departed afterward. (C) An empirical example of the cascade process. The giant component of the collaboration networks of "Phase Transitions" up to 2001 and its reduced state from 2005 to 2009 with the concomitant creation of "Complex Networks".
Source: The figure is reprinted from Ref. [312].

overlooked publications with highly cited ones, the dropout researchers with well performing ones, barely read journals with high-impact ones, and failed funding proposals with successful ones to understand and model the dominant factors of success. In the science of science, the topic of paths to success has been intensively studied. In this section, we will review the major research achievements that have been made in this direction.

6.1. Characteristics of a high-impact research work

Scientific publications show significant heterogeneity with respect to citations. Regarding a published paper, its success is usually judged based on the number of citations and whether these citing papers are important ones. In the literature, there exist several studies seeking to understand what are the key factors influencing the impact of a publication. The related works are summarized in the following subsections.

6.1.1. Success with interdisciplinary research

Interdisciplinary research (IDR) is a type of study that connects two or more disciplines to gain a more insightful perspective regarding a longstanding challenge or to discover new research topics. There are also many terms, such as multidisciplinary and transdisciplinary, that are used to describe interdisciplinary research. In the past decade, scientific communities have witnessed an increase in interdisciplinary research [313]. Even though most scientific institutes and universities encourage interdisciplinary research, numerous related issues are still under discussion. Two major issues are how to quantify the interdisciplinarity of a scientific work [314] and what is the relation between interdisciplinarity and the impact of research output [315].

Because multiple disciplines are involved in interdisciplinary research, knowledge integration is commonly considered as the key aspect of interdisciplinary research. It is more difficult to measure the process of integration than the results of

integration; thus, measuring the interdisciplinarity of research is usually done by considering its outcomes, i.e., scientific publications. One major approach is from the perspective of disciplinary diversity [96,316,317]. Stirling proposed a framework for measuring the disciplinary diversity, which contains three basic properties: variety, balance, and disparity [314]. Specifically, variety is the number of disciplinary categories, balance is related to the evenness of the distribution of disciplinary categories, and disparity describes the degree to which these categories are similar or different. For samples of journal articles, the interdisciplinary metric (also called the integration index and the Stirling index) can be expressed as

$$I = \sum_{i,j} (1 - s_{ij})^{\alpha} (p_i p_j)^{\beta}, \tag{63}$$

where p_i is the proportion of references citing subject categories (SC) i in a given paper, s_{ij} is the cosine measure of similarity between SCs i and j, and α and β are tunable parameters (both are set as 1). Porter and Rafols computed the integration index of papers from six research domains according to the Web of Science Subject Categories [316]. They found that the six research domains are becoming more interdisciplinary, as suggested by the integration index. Rafols and Meyer proposed a conceptual framework that includes diversity and network coherence to measure interdisciplinary [96]. Diversity, reflecting the disciplinary heterogeneity, was measured based on indicators such as the number of distinctive categories, Shannon entropy, Simpson index, Stirling index and its extension. Network coherence, capturing the intensity of similarity relations within a bibliometric set, was measured based on the mean linkage strength and mean path length in bibliographic coupling networks. In addition, there are several studies that directly apply network analysis to measure interdisciplinarity. Leydesdorff [100] measured the interdisciplinarity of journals using the betweenness centrality in a local citation network, which contains all journals that cite or are cited by a specific journal above a given threshold [318]. Based on a relatedness index between journals, Lee constructed a journal citation network for the fields related to technology management (TM). The analysis of the relation between TM and other disciplines reveals the brokerage roles of TM specialty journals, indicating the multidisciplinary characteristics of TM [319]. Leydesdorff and Rafols measured the interdisciplinarity of journals by combining the vector-based indicators (e.g., Shannon entropy, Gini index), network indicators (e.g., betweenness) and Stirling index, assuming that these indicators may capture different aspects of interdisciplinarity [320].

The purpose of interdisciplinary research is to exchange ideas and techniques across disciplines to inspire innovations. A natural question to ask is whether interdisciplinarity can enhance the impact of a paper. In the literature, there are considerable efforts devoted to answering this question. However, distinct conclusions are reached in different works. We will briefly review the major related works below. For clarity, these existing works are summarized in Table 3, which is built primarily based on a similar table made by Yegros-Yegros et al. [328].

By estimating the degree of interdisciplinary of a paper's references using Brillouin's diversity index, Steele and Stier found a positive influence of IDR on the citation impact [321]. Focusing on physics research programs, Rinia et al. measured the degree of interdisciplinarity as the proportion of papers published by physicists in other disciplines; however, no significant correlation between interdisciplinarity and citation impact was found [322]. Adams et al. quantified the interdisciplinarity of a paper via the Shannon entropy of disciplinary categories in the references of papers and explored its relation to the number of citations received by papers; however, they found no evidence showing any systematic correlation between them [323]. They argued that the relation seems to be an inverted U-shape, indicating that papers with intermediate levels of interdisciplinarity are more likely to be cited. This claim was later supported by Lariviére and Gingras who also obtained an inverted U-shape relationship when analyzing all articles included in the WoS in 2000 [325]. Uzzi et al. divided research papers into conventional reference combinations and atypical ones based on their references [326]. A conventional reference combination represents a paper that cites journals that are proximate in cognitive space, while atypical combination represents a paper that cites journals that are usually not co-cited. It was revealed that highly cited papers are those that have mostly conventional reference combinations but also a small proportion of atypical combinations. This work, to some degree, could also be regarded as in support of the U-shape relationship theory. Focusing also on the reference combinations, Lariviére et al. classified references into intra- (same discipline) or interdisciplinary co-citations [327]. They found that most interdisciplinary combinations have a positive effect on citation impact. Wang et al. used factor analysis to uncover three distinct dimensions of interdisciplinarity: variety, balance, and disparity [317]. The long-term citations of a paper exhibit accelerating increase with variety, decelerating increase with disparity, and decrease with balance. Regarding short-term citations, variety and disparity have a negative effect, while balance no longer has a significant negative effect. Yegros-Yegros et al. also considered these three dimensions and used a Tobit regression model to quantify their effects on the citation impacts of papers [328]. They found that variety has a positive effect on impact, whereas balance and disparity have a negative effect. In addition, they noted that all three dimensions of interdisciplinarity display an inverted U-shape relationship with citation impact. At the journal level, Levitt and Thelwal found that multidisciplinary journals are 50% less cited than monodisciplinary journals on average [324]. However, such effect disappears if a similar analysis is applied to the social sciences instead of the natural sciences.

6.1.2. Factors of scientific publications

In the final stage of a research work, the findings are summarized in one or several papers, which are submitted for publication. The impact of a work is strongly connected to the writing of a paper. Some factors have been revealed in the empirical studies.

Table 3
A summary of the studies investigating the relation between interdisciplinarity and impact of research works.

	Sample	Database	Unit of analysis	IDR indicator	Aspect of diversity	Measure of citation impact	Correlation IDR vs impact	Regression controls
Steele & Stier (2000) [321]	750 articles in forestry (1985–1994)	Journal Forest Science	Article	Brillouin's diversity index	Variety, balance	Average annual citation rate	Positive	Yes
Rinia et al. (2001) [322]	All academic groups in physics the Netherlands	WoS	Journal	% papers not published in physics	Balance	Normalized indicators	No effect	No
Adams et al. (2007) [323]	Articlesfrom two UK universities	WoS	Article	Shannon diversity& % cited refs. to other SC	Variety, balance	Normalized indicators	Visual evidence of inverted U	No
Levitt & Thelwall (2008) [324]	All science and social science articles	WoS, Scopus	Journal	Number of disciplines assigned to journals	Variety	Normalized indicators	Negativeeffect in some disciplines	No
Larivière &Gingras (2010) [325]	All papers published in WoS in 2000	WoS	Article	% cited refs. to other SC	Balance	Normalized indicators	Inverted U shape	No
Uzzi et al. (2013) [326]	All papers in WoS (1990–2000)	WoS	Article	Median disparity, 10%percentile disparity	Disparity	Not normalized	Low median disparity, with high 10%disparity	Yes
Larivière et al. (2015) [327]	All papers in WoS (2000–2012)	WoS	Co-citation	Dichotomous:Intra. vs. Inter subsdiscipline	Disparity	Normalized indicators	Mainly positive	No
Wang et al. (2015) [317]	All articles published in WoS in 2001	WoS	Article	Number of referenced SCs, % cited refs. to other SC, 1-Gini, Simpson index, Shannon entropy, Average dissimilarity, Rao–Stirling diversity	Variety, balance, disparity	Not normalized	Variety and disparity have a positive effect on long-term impact, balance produce negative effect on long-term impact	Yes
Yegros-Yegros et al. (2015) [328]	62408 papers published in 2005 in the four fields	WoS	Article	Number of referenced SCs, Shannon diversity, Average dissimilarity, Rao–Stirling diversity	Variety, balance, disparity	Normalized indicators	Inverted U shape	Yes

Title and abstract. By exploiting a large amount data of 140,000 papers from Scopus, Letchford et al. investigated whether a paper's title length is at all related to the number of citations it receives [329]. In this work, the length of a title was defined as the number of characters it contains, including spaces and punctuation. Analysis of the 20 000 most highly cited papers suggests that papers with shorter titles do receive a larger number of citations and that such relation becomes weaker if the effect of a journal's reputation is removed. In addition, their results reveal that journals that publish papers with shorter titles tend to receive more citations per paper. The analysis is later extended to abstracts of papers by the same group of researchers, and a similar trend is identified [330]. Specifically, journals that publish papers with shorter abstracts and more frequently used words receive slightly more citations per paper. One possible explanation is that high-impact journals might restrict the lengths of the titles and abstracts of their papers and require writing suitable for a wider audience. For instance, abstracts in *Science* are restricted to 125 words. Another explanation is that papers reporting greater scientific advances tend to be written with shorter abstracts and contain less technical language. A third potential explanation is that shorter titles and abstracts with more commonly used words may be easier to read and hence attract more citations.

Equations. The relation between the number of equations a paper has and the impact of the paper is still under debate. Fawcett and Higginson first noticed that many empirical studies are based largely on other empirical studies, with few direct references to relevant theories [331]. They claimed that this observation implies a breakdown of communication that may impede scientific progress. Furthermore, they systematically investigated how the use of mathematical equations affects the scientific impact (measured based on citation count) of studies in ecology and evolution based on 649 publications from the top three specialized journals in 1998. It was observed that the equation density of an article has a substantial negative impact on citation rates. However, this work was followed by much discussion. One major point that needs reconsideration is whether it is appropriate to represent the communication level between researchers by using the citation rate [332]. Physics is a discipline that depends heavily on mathematical formulations, and thus physicists are generally required to be well trained in mathematics. Kollmer, Pöschel and Gallas later reported a similar study based on the publications in physics [333]. By performing a detailed analysis of papers published in *Physical Review Letters*, they found that the distribution of citations as a function of the equation density is similar to that for ecology and evolution. However, the distribution shows a plateau with huge standard deviations when using a smaller binning size for equation density. Kollmer et al. thus argued that the data do not display a reliable dependence between equation density and received citations. We do not want to give a conclusion regarding the effect of equation density in this review. However, we do agree with the sentiment of Fawcett and Higginson that more effort should be made to make the theory of a work more accessible [334]. Mathematical theory plays a critical role in almost all disciplines, being rooted in numerous essential parts of a research work. A clearer expression and better explanation of the mathematical equations will help researchers understand the basic assumptions, their probable implications and their possible extensions.

Editorial delay. The peer review process seeks to evaluate the quality of a research work submitted for publication. Intuitively, the value of high quality papers can be identified more easily, thus resulting in a fast decision by the referees. However, very innovative articles often encounter serious obstacles in the publication process. Siler et al. even found that at three elite medical journals, the 14 most highly cited submissions were rejected and were published elsewhere [335]. Among these 14 articles, 12 were desk-rejected. Regarding the peer review process, more works have been devoted to investigating the effect of editorial delay. In fact, various factors can affect the editorial duration, e.g., the prestige of authors, the processing speed of an editor and the responsibility of a reviewer. For ecology journals, Pautasso and Schafer found a statistically significant negative correlation between average log-transformed editorial delay for journals (in days) and log-transformed impact factors [336]. Shen et al. selected the most well-known multidisciplinary journals *Nature*, *Science* and *Cell*, and conducted a direct empirical statistical analysis of the correlation between the editorial delay and received citations for all academic papers published from 2005–2009. They found that there is a weak correlation between them [337]. Lin et al. also focused on the papers published in top journals including Nature, Science and PRL [338]. They suspected that the existence of noisy signals originating from low quality papers, which might distort the overall correlation between editorial delay and received citations. Therefore, they investigated the correlation between editorial delay and the ratio of highly cited papers, which are defined as those receiving more citations than the median citation number of all collected papers. They found a strong negative correlation between the two, indicating that papers with shorter editorial delays do have higher probabilities of being highly cited. To some degree, the review process indeed provides a reasonable judgment regarding the quality of papers, at least among those that are published and highly cited.

Resubmission flow. The editorial process reflects the peer evaluation of the quality of a work. The submission and resubmission processes reflect the self-assessment of the quality of a work to some extent. Calcagno et al. conducted a large survey of the submission process of papers published in 923 scientific journals from the biological sciences for the period from 2006 to 2008 [339]. Based on the final retrieved submission history of 80748 articles (37% of the total inquiries), they construct a network of manuscript flows among scientific journals, with directed links representing the resubmission process, e.g., an arrow from journal A to journal B means that an article was submitted to and published by journal B after being submitted to and rejected by journal A. The resubmission network is densely connected with clusters representing the research subject categories and high-impact journals positioned in the center. The out-degree of a journal represents the number of times that a journal was reported as an earlier choice for submission. The out-degree is found to be highly correlated with a journal's impact factor, showing that high IF journals are more often earlier choices and that the resubmission flow is downstream oriented. Another way to measure the centrality of the journals is to consider the fraction of published articles that are first-intent submissions. Overall, a high percentage, 75%, is reported, indicating that authors

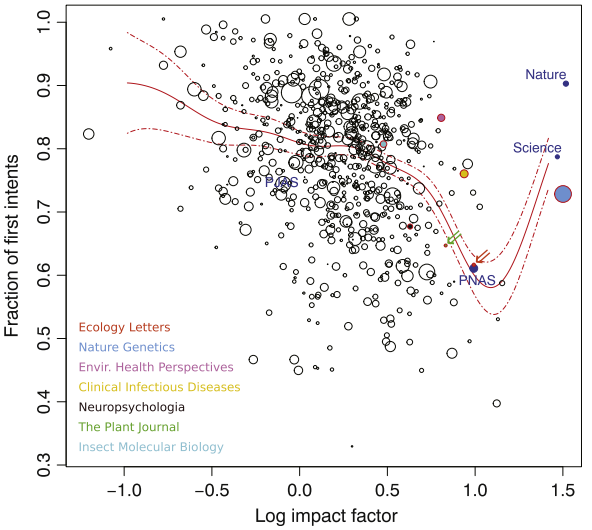

Fig. 14. The scatter plot of the fraction of first-intent submissions to a journal versus the impact factor of the journal. Each circle represents a journal, with the size proportional to the number of articles of each journal. Multidisciplinary journals are highlighted in blue.
Source: The figure is reprinted from Ref. [339].

are overall efficient at targeting their research and limiting the risk of rejection. The scatter between fraction of first-intent submissions and impact factor shows an interesting U-shape curve, i.e., the fraction of first intent submissions to a journal decreases with the journal's impact factor but increases again to a high value for the three most prestigious journals, as shown in Fig. 14. One possible explanation is that high-impact journals have many competitors and high rejection rates, with the manuscripts circulating among them before acceptance occurs. Low-impact journals are more specialized and have lower rejection rates; thus, a large number of manuscripts are accepted the first time they are submitted. Submission history is also connected to the impact of articles after publication. Resubmissions from other journals are usually more highly cited than first-intent submissions, though the number of citations is much lower if the resubmission originated from a journal from a different field.

Apart from the factors mentioned above, there are several other factors influencing the impact of a paper, such as the impact factor of the journal, the number of references and their average citations [340]. Finally, we remark that despite the observed statistical correlations between the abovementioned factors and success (impact), whether they are independent is still not clear. Some of them may be the consequence of others. Therefore, it is important to dig deeper to determine the original mechanisms for these phenomena and uncover the possible hidden causalities from the correlations.

6.1.3. Negative views play a positive role

Some journals publish regular articles as well as comments that usually correct or criticize regular articles previously published in the same journal. The criticism can be either in regard to a deficiency in the techniques used or an incorrect conclusion. If an article has a corresponding comment paper, the natural assumption is that some content in the original article is questionable; thus, this article is less likely to be successful. However, Radicchi revealed that this is not the case in reality [168]. Papers from 13 major journals, including Nature, Science, Phys. Rev. Lett. et al., that have received comments are identified, and their citations are compared with those of the papers that have not received comments. Surprisingly, the papers in a journal that have received comments are found to be more highly cited than papers that have not received comments on average, and they are more likely to appear in the most highly cited list of a journal. A possible interpretation of the results is given. Scientists can recognize the potential scientific value of newly published papers, and those with high potential have a higher chance of being carefully reexamined and initializing a debate. In some sense, papers that received comments are likely those selected by scientists to be sufficiently important such that they require reverification, which leads to an increase attention and citations from the scientific community. Smaldino et al. focused on the false-positive findings and confirmed via a case study that they persist in scientific publications [341]. In addition, an evolutionary model was set up to explain this phenomenon. The model consists of a number of competing laboratories that investigate novel or previously published hypotheses using culturally transmitted research methods. The model shows that pursuing high productivity leads to poorer methods and increasingly high false discovery rates. It is suggested that institutional incentives for publication quality instead of quantity could eventually solve this problem.

Citations from regular papers can be heterogeneous, in the sense that they can be roughly classified as either positive, negative or neutral. Positive citations indicate endorsement, while negative citations usually represent disagreement and criticism. Catalini et al. used bibliometric data and natural language processing (NLP) to extract negative citations from the

references of manuscripts and revealed a number of empirical patterns of those negative citations [342]. The real data are full-text articles from the Journal of Immunology for the period from 1998 to 2007. Some negative citations are first identified by experts and then used as a training set for NLP to classify the remaining citations. The non-negative citations are denoted as "objective" citations in the paper. The analysis shows that the likelihood of an article receiving its first negative citation is higher in the first few years after publication. In addition, the majority of negative citations appear in the "Results and Discussion" section of a paper, while only less than half of objective citations appear in this section. Negative citations are more likely to originate from researchers who work in the same discipline and are topologically closer to the authors of the cited paper in the co-authorship network. With respect to the influence on a paper's impact, the long-term citations of a paper are found to be slightly suppressed by the negative citations. The extracted negative citations, together with the remaining citations, can form a signed citation network between papers [86]. Kumar studied the structural properties of these networks specifically those comprising articles from the field of computational linguistics [87]. The "negative degree" (i.e., the number of negative citations received by a node) follows a broad distribution. The signed citation networks obey a weak balance theory [343]. A signed author citation network is also constructed. Kumar finds that authors do not reciprocate the sentiment they receive from other authors.

Some papers containing significant flaws are retracted from journals to prevent the further propagation of misinformation. Based on three early works [344–346], in a correspondence paper, Campanario argued that some retracted articles are still being positively cited and proposed some possible solutions for making researchers more aware of retracted papers [347]. Fang et al. studied a relatively large number of retracted papers in PubMed, finding that the number of retracted papers increases over time [348]. A more detailed analysis indicates that more than half of these retractions are attributable to misconduct. Among the different reasons for retraction, most articles retracted because of fraud were published in high-impact journals and originated from countries commonly recognized as strong in terms of research (USA, Germany, and Japan). Retractions due to duplicate publications were mainly published in low-impact journals and authored by researchers working in countries with a shorter history of research (China). Similar to previous works, many retracted papers are identified as being highly cited. Lu et al. focused on the penalty caused by retracting papers [349]. They first provided some empirical evidence showing that biology & medicine and multidisciplinary sciences have considerably higher retraction rates than other disciplines and that most retractions are not due to self-reported errors. More importantly, the retraction of a paper on average causes the number of citations to decrease by an average of 6.9% per year for each prior publication, affecting the works of authors published up to a decade earlier and papers four steps away from the retracted paper in the citation networks. However, such penalty on prior works is absent if the retraction is due to a self-reported error.

6.2. Determinants of a successful scientific career

Universities and research institutes are currently training more and more PhDs; however, many of them will eventually leave academia because of strong competition. For those who stay in science, a successful career is not easy to obtain and is strongly associated with whether or not they can acquired future resources, such as receiving tenure or grants. Therefore, much research effort has been devoted to understanding the determinants leading to a successful scientific career.

Scientific creativity. The success of a scientific career primarily depends on the scientific creativity of the researcher. In general, scientific creativity is the ability of a researcher in possessing a body of systematic knowledge, motivating and formulating novel research problems, and eventually conducting efficient search in methodological knowledge to find solutions [350]. In the classic literature, scientific creativity was discussed mainly from the perspective of psychology [351]. Issues include, for instance, connection of scientific creativity to intelligence and personality characteristics [352]; models of the creative process [353]; factors for creative situation [354]; and measuring scientific creativity of a researcher [355].

Regarding the relation to intelligence, it was found that the intelligence quotient (IQ) is not linearly correlated with creativity, but above a certain level of IQ is required for mastery of a field [352]. Six personalities including autonomy, personal flexibility, openness to experience, aesthetic sensitivity, commitment to work and need for professional recognition have consistently been shown to correlate with success in science [352]. These personalities play roles at different stages of a research project. In the model of Busse and Mansfield [353], a scientific creative process contains with two stages, selection of the problem and extended effort to solve the problem. The first four personalities relate to the first stage and the last two personalities relate to the second stage.

With respect to creative situation, a large array of psychological, social, cultural, political, and historical factors have been examined [351]. The psychological factors mainly refer to the atmosphere of working place. Factors favoring scientific creativity include, for instance, the absence of serious threat, the readiness to take risks, openness to ideas of others, confidence in one's perceptions of reality or ideas. Cultural factors include speculations that certain world views such as materialism, individualism, conceptualism, and skepticism favor scientific growth. Political factors such as war and civil disturbances could decrease scientific creativity. As for the historical factors, it was pointed out that at specific points in history, certain discoveries or the development of certain theories were inevitable.

As scientific creativity is hard to quantify directly, a common approach is to design indicators to measure scientific performance as proxy [355]. The simplest indicators count the number of publications of a researcher and the number of citations to him/her in the literature. These two indicators were further improved to incorporate other factors such as author contribution and decay of papers' relevance [191]. Yet, the quantitative measure of scientific creativity is still a challenge.

With the accelerating access to big data of scientific publications, numerous models have been developed in recent years to model the process of scientific discoveries [55,132]. These models usually contain the scientific ability of a researcher as a parameter, and it can be estimated by fitting the model with real citation data. Comparing data in different periods, the evolution of researchers' scientific ability can be investigated. However, the citation-based measures have some drawbacks as scientific creativity is not always expressed in publications. For instance, contributions to science also include patents, talk in conferences, reviewing manuscripts, etc.

Personal characteristics. Some of the personal characteristics of researchers, including gender and age, have been investigated. Larivière et al. focused on the gender-related differences in research performance and conducted a case study of professors in Québec universities [356]. The first observation is that there are more male than female researchers. After passing the age of 38, an obvious disadvantage for women emerges with regard to receiving research funds, producing more publications and having scientific impact (measured based on the number of citations). Such a phenomenon is attributed to several possible reasons, such as the impact of family, rank within the hierarchy of the scientific community and access to resources, and level of specialization. Similarly, Pohlhaus et al. investigated the gender-related differences that exist in science, with a special focus on research grants [357]. With data from the National Institutes of Health (NIH) extramural grants, it was found that women receive larger awards than men on average, while men receive more awards than women at all points in their careers. Because age is usually considered as an important factor in research, Jones et al. quantified the influence of age on scientific creativity [358]. The analysis is based on the data of Nobel laureates, specifically regarding the ages of Nobel laureates at the time their Nobel winning works were completed. The means of such ages are calculated for the whole sample period and separately for an early period (prize-winning work before 1905) and a late period (prize-winning work after 1985). The values were also compared across different disciplines, including physics, chemistry and medicine. A clear pattern is identified. Compared with the early period, the mean ages with respect to prize-winning work all increase in the late period. Such shift in age can be interpreted as that more knowledge currently needs to be accumulated so that great discoveries can be made. The mean ages across fields are significantly different but unstable over periods. To better quantify the effect of age on creativity, a regression model is set up as follows

$$Pr[AchievementAge_i < 30|Year_i, x_i] = (1 + exp\{x_i\beta_0 + Year_i^{P_1}\beta_1 + Year_i^{P_2}\beta_2\})^{-1}, \tag{64}$$

where $AchievementAge_i$ and $Year_i$ denote the age at which and year in which laureate i made his/her great achievement; P_j is searched for within the set $\{-2, -1, -1/2, 0, 1/2, 1, 2\}$ for $j = 1, 2$. The aim of the regression model is to smooth the real data so that a clearer trend can be revealed. The relation between the frequency of great achievement by age 30 (or 40) and year of great achievement is illustrated using a regression model and compared across disciplines. In all three fields, the frequency of great achievement before age 30 is low and becomes almost 0 by the end of the 20th century. In physics, the share of physics Nobel laureates who completed their prize-winning work at young ages exhibits a peak before the middle of the 20th century. This phenomenon is much weaker in chemistry and almost absent in medicine. Finally, the age at which a laureate accomplishes a great achievement is found to be related to a laureate's age upon obtaining his/her highest degree and whether the great achievement had a theoretical component. A predictor of age at which a great achievement is accomplished is obtained via regression

$$Age_i = 31.927 + 0.304PhDAge_i - 4.434Theoretical_i + \epsilon_i, \tag{65}$$

where $Theoretical_i = 1$ is for the theorists, and 0 is for empiricists. The equation indicates that theorists on average accomplish their great achievements 4.434 years earlier than empiricists. In addition, a one year delay in obtaining a PhD is associated with a 0.304 year delay in the average age at which a Nobel winning work is completed.

Cumulative advantage. Petersen et al. revealed the persistence and uncertainty in an academic career by analyzing the career profiles of 300 physicists [306]. The data of these physicists are divided into three groups according to their h-index. Their cumulative production (cumulative number of published papers, $n_i(t)$) exhibits persistent and accelerating growth with time t, as shown in Fig. 15. The exponents are larger than 1 yet different for these three groups. The uncertainty of a career can be observed when the annual production change is computed as $r_i(t) = n_i(t) - n_i(t - \Delta t)$, where $\Delta t = 1$ and $r_i(t)$ can be positive or negative. The distributions of $r_i(t)$ in all three groups are leptokurtic and remarkably symmetric. The deviations in the tails are likely signatures of the exogenous career shocks, indicating the uncertainty in an academic career. The production fluctuation scale $\sigma_i(r)$ as the standard deviation of r has a scaling relation with the median of coauthors per year, $\sigma_i(r) \sim S_i^{\gamma}$, with $\gamma < 1$. Eventually, a proportional growth model is put forward to understand the effects of long-term and shot-term contracts on scientist performance. The model basically presents agents that attract opportunities (i.e., new scientific publications) according to the appraisal based on their history of production, i.e.,

$$P_i(t) = \frac{w_i(t)^{\pi}}{\sum_{i=1}^{l} w_i(t)^{\pi}}, \tag{66}$$

where $P_i(t)$ is the rate of attracting new opportunities, π is a parameter controlling the effect of $w_i(t)$, which is the appraisal of researcher i, defined as

$$w_i(t) = \sum_{\Delta t=1}^{t-1} n_i(t - \Delta t)e^{-c\Delta t}, \tag{67}$$

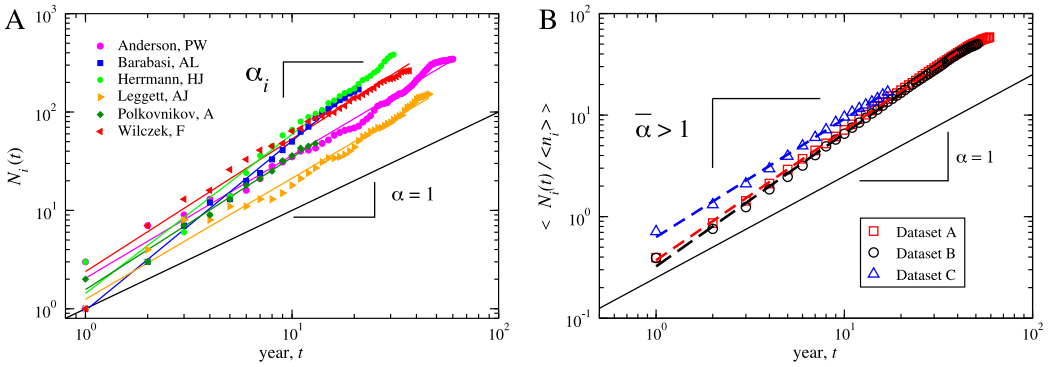

Fig. 15. The accelerating growth of cumulative publication numbers $N_i(t)$ with career year t. (A) The career trajectories of six well-known network scientists are shown. (B) The average $N_i(t)$ of three individual groups with different h-index. A robust accelerating growth is observed in each group.
Source: The figure is reprinted from Ref. [306].

where c is the parameter determining whether it is a long-term appraisal/tenure system ($c = 0$) or a short-term appraisal system ($c \gg 1$). The Monte Carlo simulation of this model shows that when $c \to 0$ and $\pi < 1$, the long-term appraisal time scale averages out early career fluctuations, leading to sustainable production throughout a career for researchers. When $c > 1$, the labor system is driven by fluctuations that can cause sudden career termination. The accelerating growth of $n_i(t)$ with time was later analyzed by Petersen et al. from the perspective of reputation in academic careers [132]. For publications with a cumulative citation smaller than a threshold, the author's reputation (measured based on his/her total citations) is found to dominate the annual citation rate. A model has also been proposed that reveals the impact of a scientist's reputation on the success of his/her publications; see Eq. (12) in Section 3.

The cumulative advantage of scientists in their careers can also be associated with the Matthew effect: "For to all those who have, more will be given". Petersen et al. quantitatively described this effect on academic careers using a simple model based on the stochastic Poisson process [359]. In the model, a career evolves with a position-dependent progress rate $g(x)$ and stagnancy rate $1 - g(x)$. Therefore, the probability of moving from a career position from x to $x + 1$ is

$$g(x) = 1 - exp[-(x/x_c)^{\alpha}], \tag{68}$$

where both x_c and α can be inferred from empirical data. The model is exactly solvable, with career longevity being

$$P(x) = \frac{g(x)^{x-1}}{x_c[\frac{1}{x_c} + g(x)]^x} \approx \frac{1}{g(x)x_c} e^{-\frac{x}{g(x)x_c}}, \tag{69}$$

indicating that $P(x)$ follows a truncated power law for the case of concave $\alpha < 1$, $P(x) \approx x^{-\alpha}$ for $x < x_c$, and $P(x) \approx e^{-(x/x_c)}$ for $x > x_c$. The model shows a universal statistical law including both short and long careers. The predicted career longevity distribution $P(x)$ of this model is validated using data of the career profiles of scientists from six high-impact journals, with x representing the duration between an author's first and last paper in a particular top-ranked journal. In another work by Petersen et al., the top journals data were used to investigate the inequality among science careers [360]. The analysis provides strong evidence that scientists are remarkably heterogeneous in productivity and impact measures in high-impact journals, indicated by the high Gini coefficients. Some other patterns are uncovered. Owing to the cumulative advantage, the average time interval between two successive publications of an author in top journals tends to decrease with each subsequent publication. However, as an author publishes more papers in top journals, the relative citations received by each subsequent publication are prone to decrease.

An academic prize is found to be another important factor that increases one's cumulative advantage. By studying the career trajectories of Nobel laureates, Mazloumian et al. found that achievement of a Nobel prize can boost the citation rates of the previous publications apart from the landmark papers [361]. A "boost factor" $R_w(t)$ is proposed to quantify such effect of a paper, which is defined as a ratio $R_w(t) = R_{>t,w}/R_{<t,w}$, with $R_{>t,w}$ representing the average number of citations received per paper per year in the period from $t - w + 1$ to t. According to the definition, $R_w(t)$ detects the events that substantially increase a scientist's citation rates. The factor thus can be used to identify the boost effect of the landmark works of general scientists. These landmark papers trigger the discovery of the value of older papers that have not yet been recognized. Though not precisely discussed in the paper, the landmark works might be one of the reasons that "sleeping beauties" in science are "woken up" [129].

Collaborators. As most scientific research includes team work, the positions of scientists in the collaboration networks are naturally an unneglectable factor of their success. This assumption has been confirmed in numerous empirical studies. An analysis of the CVs of 73 researchers affiliated with an academic research network in Canada shows that the high betweenness centrality of a researcher in a collaboration network considerably promotes his/her scientific performance,

measured using the h-index and customized productivity index (i.e., a measure combining one's grants, peer reviewed publications, students, and other publications) [362]. Larger-scale empirical analyses were also conducted using data related to physics, computer science and other disciplines. Consistent with the previous case study, it was found for 250,000 papers from *ArXiv* that brokerage positions that connect otherwise unconnected parts of a network enhance researchers' chances of publishing articles in high-impact journals [363]. In addition, it was also observed that maintaining a moderate number of persistent ties is crucial for scientific success. The career data of computer scientists was also used to examine the relation between the volume of collaborations and collaborators of one author and his/her research performance over time [364]. In this work, success is considered based on the volume of publications and citations to these publications. The correlations confirmed that the centrality of computer scientists (degree, betweenness, closeness, and eigenvector) in a co-authorship network is related to the patterns that their publications/citations exhibit in the following years. The phenomenon is attributed to the role of collaboration in improving productivity and the visibility of one's research.

Considering the critical role of collaboration networks in researchers' success, Ebadi et al. sought to quantify the factors that can influence the network positions of researchers [365]. A regression model is set up to estimate the network centrality of researchers based on several independent variables. By fitting the parameters, one can measure the contribution of each factor to the position of a research in the collaboration network. Using data of the collaboration network among Canadian researchers, the regression formula is designed as

$$\begin{aligned} centrality_i = {} & \beta_1 \times avgFund3_{i-1} + \beta_2 \times noArt3_{i-1} + \beta_3 \times aveIf3_{i-1} + \beta_4 \times avgCit3_{i-1} + \beta_5 \times dc_i \\ & + \beta_6 \times careerAge_i + \beta_7 \times dAcademia_i + \beta_8 \times dProvince_i + \beta_9 \times dFundProgram_i + \alpha_i, \end{aligned} \tag{70}$$

where $centrality_i$ can be the betweenness, closeness, clustering coefficient, or eigenvector centrality of researcher i in the collaboration network. The variables are many. $avgFund3_{i-1}$ is the average amount of funding received by a researcher over the past three years, $noArt3_{i-1}$ is the number of articles over a three-year time window, $aveIf3_{i-1}$ is the average impact factor of the journals in which the author has published articles for a three-year time interval, $avgCit3_{i-1}$ is the average number of citations of the articles in the past three years, dc_i is the degree centrality of a researcher in the collaboration network, $careerAge_i$ represents the time difference between the date of a researcher's first article in the database and the given year, $dAcademia_i$ indicates whether the researcher is affiliated with academic or non-academic environments, $dProvince_i$ represents Canadian provinces to distinguish the location and geographical impact in possessing various network roles, and $dFundProgram_i$ represents different NSERC funding programs. After fitting the parameters β, several patterns are uncovered. The past productivity of the researchers, along with their available funding, largely determines their positions in the co-authorship network. Highly productive researchers not only have more important connections (high eigenvector centrality) but also play critical roles in connecting other researchers (high betweenness). In addition, it is found that young researchers also play an important role in connecting different communities and promoting information transmission through the network.

In addition to the structural position, the strength of the collaboration ties also matters for a scientist's career. Focusing on the longevity and intensity of scientific collaboration, Petersen analyzed the records of collaboration between 473 researcher profiles, seeking to understand the sustainability and efficiency of teamwork as well as the relation to scientists' career achievements [69]. Between two scientists, the collaboration tie is defined as the number of co-authored papers, and the collaboration lifetime is defined as the time between their first co-authored publication and the last observed joint publication. The distribution of these two quantities shows that collaboration profiles are dominated by weak ties and short lifetimes. However, the collaboration networks contain numerous super ties as well, indicating that some scientists intensively collaborate with each other for a long time. The role of super ties is quantified based on the analysis of the corresponding scientists' productivity and the impact of their individual publications. A regression model is set up to measure the effect of collaboration on productivity and impact. It is shown that the scientists with super ties generally have above-average productivity and that citations referring to each of their publications are boosted by 17%, indicating a remarkable contribution of super ties to the career success of scientists.

6.3. Pivotal role of research funds

Pursuing research funding is essential for the academic career of scientists, as it provides resources to support various activities including purchasing equipment, hiring postdocs, and attending conferences. Obtaining some large grants, to some degree, can also be considered as a recognition of one's scientific achievement by the academic community. In this sense, research funding is support as well as an indicator of one's success in science. The increasing availability of bibliometric datasets makes it possible to quantitatively study the collective patterns in research funding.

Funding bias. A primary question asked in the literature is whether there is any bias regarding where funding goes. Serious bias in funding may lead to a noneffective spending of research money and promising talents leaving academia because of the lack of support. The bias of research funds is analyzed at different levels in the literature. Parisi shows that the national research funds of Italy are too little, making the country much less competitive in receiving European research funds [367]. Domenico et al. further noted that in addition to domestic funds, winning European research money depends on a country's ability to retain their own scientists and attract others from abroad [368]. The strong bias in research funds may hinder the development of science in some countries; thus, scientists are required to reform related policies to balance the distorted resource allocation.

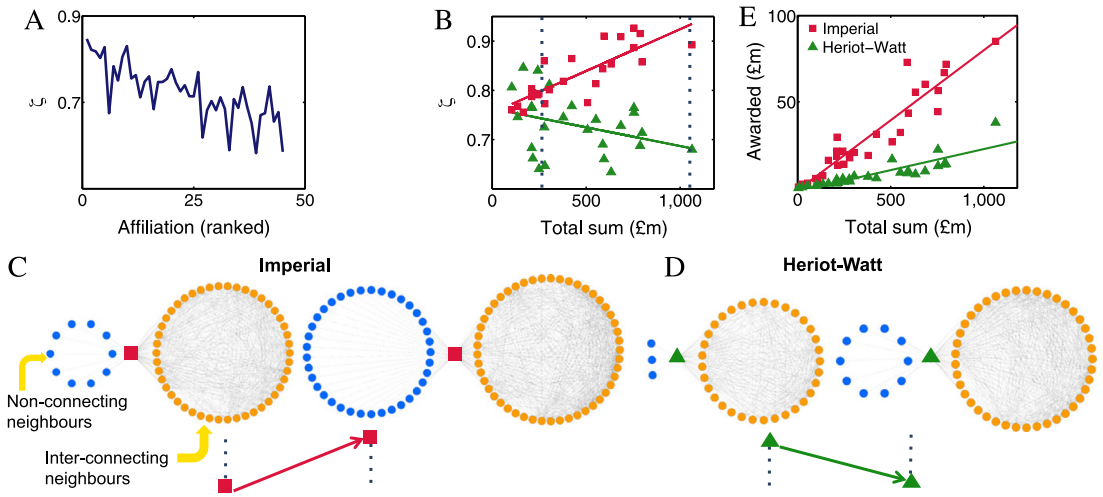

Fig. 16. (A) The yearly average effective size ζ, as a measure of the level of brokerage, versus the rank of affiliations according to their total awarded research funding. (B) The comparison between Imperial College and Heriot-Watt University with respect to their levels of brokerage. (C) The connections between Imperial College and its partners when the total funding from the EPSRC was £250 million and over £1 billion as indicated by the dotted lines in (B). (D) The connections between Heriot-Watt University and its partners when the total funding is the same as in (C). (E) Research funds obtained by the two universities, which increase linearly with the total funding.
Source: This figure is reprinted from Ref. [366].

The bias of research funding is especially significant at the institute level. Ma et al. examined the scientific project data of the Engineering and Physical Sciences Research Council (EPSRC) and identified a significant bias at the university level [366]. Over the years, the total amount of funding awarded has increased, but the total number of grants awarded has decreased. In parallel, the distribution of funds for principle investigators (PIs) and universities exhibits increasing heterogeneity, based on the Gini coefficient. The collaboration networks between investigators and affiliations are constructed using the data. The links here represent a funded partnership between two nodes. To understand the collaboration networks, the broker nodes are defined as those occupying an advantageous location in the network for detecting and developing opportunities through its connections to nonoverlapping clusters. The significance of a node i as a broker can be measured based on $\zeta_i = 1 - (k_i - 1/k_i)C_i$, where C_i is its clustering coefficient. The rise in the volume of total funding has resulted in a more significant brokerage behavior of the most-funded affiliations, while less-funded affiliations show the opposite trend. Such phenomenon is illustrated by the two universities used as examples in Fig. 16. In the collaboration networks between affiliations, the leading universities formed a rich club. This phenomenon can be interpreted as the tendency of scientists collaborate with outstanding peers in reputed universities. As a result, it was also found that the total research funds received, the relative citations and h-index tend to increase if an affiliation appears frequently in the rich club, which indicates that being in the rich club is a critical factor of the research success of an affiliation. This paper reveals that a large part of research funding goes to a limited number of affiliations in the rich club, which may cause concern among scientists and policy makers. This issue was later discussed in a commentary from the perspective of biased funding resource allocation [369]. A similar effect was also detected in other works. Murray et al. conducted a case study of the applications to Canada's Natural Sciences and Engineering Research Council (NSERC) Discovery Grant program and showed that both the success rates and grant amounts are lower for applicants from small institutes [370].

Research funds are also unbalanced across disciplines. Because interdisciplinary research is widely considered as an effective way to create novel ideas, numerous empirical works have revealed that funding plays an important role in creating and spreading knowledge in interdisciplinary research [371–374]. Bromham et al. examined whether interdisciplinary projects are more likely to be approved [375]. They made use of the Australian Research Council's Discovery Programme covering fundamental research in all disciplines. The first challenge is to quantify the degree of interdisciplinarity of each proposal. To this end, a metric called phylogenetic species evenness was employed from evolutionary biology to compute the interdisciplinary distance (IDD) score for all proposals. The IDD score combines the relative contributions of disciplines within a proposal and the collaborations between distant disciplines. As the data contains both successful and unsuccessful proposals, the relations between projects' IDD scores and their success rates can be studied. It is shown that IDD is consistently negatively correlated with funding success and independent of other factors such as affiliation, year of application, number of research codes selected and primary research field. Though interdisciplinary research has its advantage in innovation, the low success rate in funding is attributed to the substantial costs and low outputs measured by the publication-dominated evaluation systems. Apart from this work, Zhao et al. focused on the field of economics and found that the global funding ratio of economics is much lower than the average level of social sciences [376], indicating that imbalance may also exist among the primary disciplines.

Peer review. The peer review process seeks to evaluate research proposals and select the most promising ones to support. A number of studies have questioned the ability of peer review panels to predict the productivity of applications [377]. Li et al. investigated the relationship between a proposal's peer review scores and its final research outcomes [378]. The percentile score of research project grants funded by the U.S. National Institutes of Health (NIH) from 1980 to 2008 is considered. The percentile score of a proposal is provided by reviewers (the lower the better), based on which the decision regarding whether to fund a project is made. For each project, the outcomes data including publications, citations, and patents are collected. The results from Poisson regressions of future outcomes regarding peer review scores show that they are significantly related. A worse peer review score is associated with fewer publications, lower citations of these publications and fewer patents. This effect may attenuate but is still present when controlling numerous variables in the regression models, such as the principle investigator's publication history, grant history, institutional affiliations, career stage, and degree types. Fang et al. reanalyzed the same NIH data, focusing only on the best scored projects (< 20), and found that the percentile score performs poorly as a predictor for the productivity of these projects [379]. This finding suggests that despite the ability of reviewers to distinguish extremely strong grant proposals from the rest, their ability to accurately predict the future productivity of meritorious projects is somewhat limited.

Instead of the predictive power, Gallo et al. focused on the decision-making processes of peer review [380]. A retrospective multilevel regression analysis was conducted on unblinded evaluations data of biomedical research funding applications to uncover the influence of reviewer expertise on the evaluation of research proposals. They found that reviewers with higher levels of self-assessed expertise tended to be harsher in their evaluations, which supports a similar conclusion made based on data from an experimental blinded review. This relation is also found to be affected by the seniority of the reviewers and applicants, implying the subtle influences of social and professional networks on the decision-making processes of peer review.

Collaboration. Collaboration is essential for research projects, especially when the projects are large and interdisciplinary. To understand the influences of different factors on the amount of allocated funding, Ebadi et al. set up a temporal non-linear multiple regression model considering variables quantifying past productivity and impact, scientific collaboration and the career age of researchers [381]. Specifically, past productivity and impact are measured using $avgIf3$ (average impact factor of the journals that a researcher has published in within the past three years), $avgCit3$ (average citations for the articles of the past three years), and $noArt3$ (number of articles published in the past 3 years). Scientific collaboration is quantified based on the betweenness centrality (bc), degree centrality (dc), eigenvector centrality (ec), and clustering coefficient (cc) of the PI in a collaboration network. Career age $carAge$ is simply the time difference between the date of the first article of a researcher in the database and the given year. The results indicate that the past productivity and impact positively influence the amount of funding, with the impact of $avgIf3$ being much higher than that of $avgCit3$. The network centrality is also shown to play a significant role. While most centrality measures, such as bc, dc and cc, have positive coefficients in the regression model, the coefficient of ec is negative. The results can be interpreted as showing that researchers connected to many collaborators who are tightly embedded within the dense network of research partnerships are more likely to get a higher amount of funding. However, having few important collaborators is not associated with more funding.

The basic structural properties of collaboration networks among researchers with respect to funding were reported by Zinilli for an Italian competitive project fund [382]. Four hypotheses were raised and confirmed by real data: the degree distribution is not guaranteed to follow a power-law form; geographical proximity is an important driving force of collaboration formation; there is no tendency to collaborate with scholars with different sets of knowledge; no significant h-index assortativity among connections is detected. Another analysis of Italian research funding was carried out by Nicotri et al., with a special focus on the collaboration network between affiliations [383]. The network exhibits properties including a power-law degree distribution, negative degree assortativity, high clustering coefficient, relatively small average shortest path length, and obvious community structure. The node centralities are discussed in detail, with some major players identified. A funded scientific collaboration network of the main countries was constructed by Tan et al. [384]. An h-degree metric was used to quantify the importance of countries in this network.

Outcome. The aim of research funding is to accelerate scientific discoveries. Therefore, a natural question to ask is as follows: what is the relation between grant size and its outcome? To tackle this question, Fortin et al. gathered the data of individual researchers in three disciplines funded by the Natural Sciences and Engineering Research Council of Canada [385]. To estimate the influence of funding on researchers, four indices measuring scientific impact were computed over a four-year period: number of papers published, citations to those papers, the most cited paper and the number of highly cited papers. By studying the relation between these impact metrics and grant size, they found that larger grants do not necessarily lead to larger discoveries. In addition, an increase in funding given to a researcher cannot guarantee an increase in the impact of researchers. The results suggest that diversity-oriented funding strategies are more likely to have more prominent outcomes.

Remarks for this section. In this section, we have reviewed some metrics regarding the high-impact of a paper and the success of a scientific career. Some of them, such as the age of a researcher and high quality scientific collaborations, are truly connected to the scientific productivity. Some other reviewed metrics, however, are not directly connected to the scientific productivity of a researcher. Those metrics are included in this section due to the following two reasons. First, some metrics do not determine the scientific productivity of a researcher himself/herself but influence his/her scientific impact which may accelerate one's success by increasing his/her future collaborators and also bringing more attention to his/her future papers. The metrics regarding the cumulative advantage are of this kind. Second, some metrics have quantitative correlation with success from empirical data analysis but their direct relation to success is not intuitive. For instance, some

papers investigated the relation of researchers' gender and their productivity. The features of a paper such as the number of equations and title length are found to be correlated with the paper's final citations. Listing these metrics as factors does not mean that there is causality between them and scientific success, but only indicates a statistical correlation from data. We remark that detailed causality analysis of these metrics asks for future investigation.

7. Innovation and knowledge propagation

Advances in science and technology refer essentially to the processes of innovation and knowledge propagation both of which are highly complex. In addition to the direct creation of new knowledge/technologies, the combination of existing knowledge/technologies is a primary method that is used for innovation. Therefore, innovation is considered as a searching process for all combinational possibilities. Numerous studies that span a variety of disciplines elaborate on the recombinant nature of innovation using databases for scientific publications and patents. Furthermore, the spread of knowledge has been investigated from the perspective of idea flows across disciplines and scientists' mobility and collaborations. In this section, we review related studies that uncover statistical patterns for the complex process of innovation.

7.1. Knowledge creation in scientific research

Scientific publications document the accumulated knowledge of humans. Therefore, it is natural to investigate the creation of knowledge by analyzing the citation relationships between research articles. In science, numerous new ideas are inspired by prior studies. One example is the connection between recent complex network theory and classic graph theory [6]. It is quite easy to identify similar examples in citation networks; however, it is more complicated to systematically extract the intrinsic hierarchical structure of citations representing the crucial dependence of innovations, i.e., the backbone, because of the various reasons why scientists cite a specific paper. Based on a citation network created using APS data, Gualdi et al. identified the most relevant reference of each publication, which results in a backbone of the original network, similar to a family tree [88]. Specifically, the impact of cited paper α (regarded as parent) on citing paper β (regarded as child) is quantified by the similarity among β and other citing papers of α via a two-step random walk. An influential parent will share a similar focus among its offspring; conversely, a less relevant parent will have more heterogeneous descendants. The final backbone is obtained by only retaining the most influential parent for each publication. The backbone largely simplifies the structure of the citation networks; however, it reveals the complex structure, including the hierarchical and fractal nature of scientific development. As a result, fields and subfields in physics can be efficiently classified using the backbone.

Based on scientific publication data, Uzzi et al. classified the references of each paper as conventional reference combinations or atypical reference combinations [326]. Conventional and atypical reference combinations refer to whether the articles that are cited by a publication are generally co-cited or not, respectively. The atypical reference combination indicates that cited papers are rarely cited together and are distant in cognitive space. Uzzi et al. noted that high-impact papers generally include both types of reference combinations, the majority are conventional reference combinations and minority are atypical combinations. Stephan et al. found that papers with high novelty as measured by atypical combinations tend to be less cited at the beginning, yet more likely to be highly cited after it has been published for three or more years [386,387]. The results suggest quantitative long-term measures when evaluating scientists, in order to reward research with high potential to shift the frontier. Mukherjee et al. analyzed the combinations in time dimension, revealing that high-impact papers cites literature with a low mean age (i.e. relatively recent works) and high age variance (i.e. contains also some old and classic papers) [388].

The combination process can be regarded as an innovation process that links knowledge with different distances in the knowledge base. In most investigations of innovation dynamics, the knowledge base is modeled as uniformly distributed in a multiple dimension space. However, numerous recent studies noted that knowledge space can also by modeled by using complex networks [389,390]. In the evolving knowledge network, nodes represent scientific concepts and edges represent their relationships. Innovation may be regarded as a combination process that more often occurs to adjacent possibilities. Foster et al. categorized the research strategies that are used by scientists facing the knowledge network [391]: jump, new consolidation, repeat consolidation, new bridge and repeat bridge. These strategies are explained in Fig. 17. Jump, new consolidation and new bridge represent innovation at different levels (adding new relationships to concepts); repeat consolidation and repeat bridge correspond to tradition (reproducing established relationships). A statistical analysis of the research impact (as measured by citation count) and research strategy reveals that high-risk innovation strategies are positively correlated with impact.

Rzhetsky et al. also adopted the knowledge network, where innovation is investigated by analyzing the discovery of chemical relationships in biomedicine [392]. They constructed a knowledge network where the concept nodes represent molecules that are linked by physical interactions or shared clinical relevance. Research progress is modeled as a sequence of experiments that uncover the knowledge network and the performance of the scientific community is accordingly measured as the number of experiments that are conducted to reveal a specific portion of the knowledge graph. At the individual level, a scientist's selection of a research problem is often influenced by its importance and difficulty, which can be mapped, respectively, as the node degree and distance between two concepts. A large node degree indicates that the topics are well-studied and a result regarding this topic will be highly relevant to other scientists' work. When the concepts that are studied are more distant, the scientist should take more effort to figure out their potential combinations and take more risks.

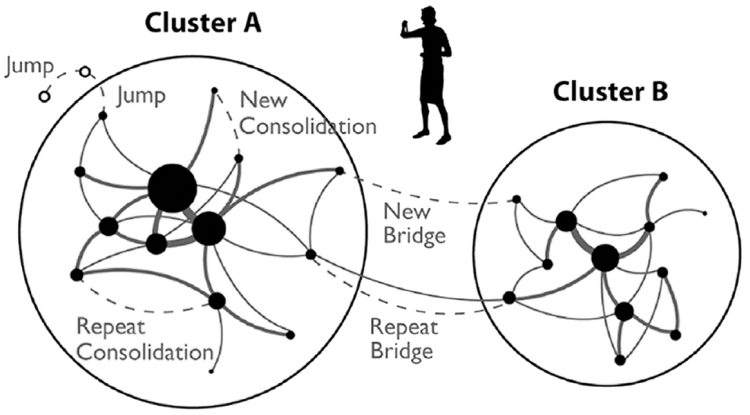

Fig. 17. Illustration of five research strategies of a scientist in the knowledge network where nodes represent chemicals and links represent chemical relationships.
Source: The figure is reprinted from Ref. [391].

Rzhetsky et al. developed a model that characterizes scientists' strategy of selecting a topic/problem to research [392]. In this model, the scientific discovery strategy defines the probability of selecting a pair of entities as a function of the importance (degree) of each entity and the difficulty associated with combining them (distance). Specifically, the probability function of selecting entity i and entity j is determined as follows:

$$p_{i,j}^{t} \propto \max\left(r_i^{t-1}, r_j^{t-1}\right)^{\alpha_\mu} \times \min\left(r_i^{t-1}, r_j^{t-1}\right)^{\alpha_l} \times \begin{cases} \left(\dfrac{d_{ij}^{t-1}}{d_{\max}}\right)^{-\beta}\left(1-\dfrac{d_{ij}^{t-1}}{d_{\max}}\right)^{-\gamma} & : d_{ij}^{t-1} < \infty \\ e^{\delta} & : d_{ij}^{t} = \infty \end{cases} \tag{71}$$

where α_μ and α_l control the preference for the more central node and lower less central node, respectively. β and γ control the preference for short and long distances between two entities if they are mutually reachable. δ defines the preference for selecting entities in distinctly connected components. Each configuration of these five parameters determines a specific research strategy.

With millions of biomedical articles published from 1976 to 2010, the underlying research strategies are inferred by fitting the model with the empirical data. Results indicate that biomedical scientists tend to adopt a conservative research strategy, i.e., they primarily explore the local neighborhood of high degree nodes (important molecules). The underlying reason is that this strategy can produce steady output and is less risky. However, it is inefficient for exploring the entire knowledge network and is therefore a disadvantage for the development of science. By searching the parameter space, strategies that are more efficient for exploring the knowledge networks are identified and discussed. A comparison of optimal research strategies and the real strategy in the empirical data is illustrated in Fig. 18. A typical finding is that risk-taking attempts, and even certain failed experiments, considerably accelerate exploring the network. This analysis is a valuable reference for policy makers when developing evaluation regulations and grant decisions to ensure more efficient research strategies, which will benefit the entire scientific community and society.

Shi et al. described knowledge space by using a hyper-graph where nodes represent the biomedical variables (i.e., people, methods, diseases, chemicals) and a hyper-edge will form among these variables if they appear in the same article [393]. By analyzing millions of abstracts from MEDLINE, they found that the network distance between these biomedical variables is surprisingly small, and the hyper-graph representation clearly facilitates link prediction. In a hyper-graph of MEDLINE, the chemicals, diseases and methods that are enclosed by a hyper-edge belong to a paper in MEDLINE. The hyper-graph is potentially useful for studying the innovation process because it includes more complete information. We must note that despite the effort made in constructing knowledge networks, the definition and extraction of knowledge conceptualization is still under investigation. In computer science, one research branch focuses on knowledge extraction and graph construction [394].

7.2. Technology and patent invention

7.2.1. Innovation processes

Innovation in technology generally refers to inventions that include "exploitation" and "exploration", both of which are also fundamental processes of other adaptive systems [395]. For innovation, exploration refers to activities that involve the creation of new knowledge, technologies and products. Conversely, exploitation generally refines and improves existing knowledge, technologies, and products. Prior studies agree that the precise definitions of exploitation and exploration are

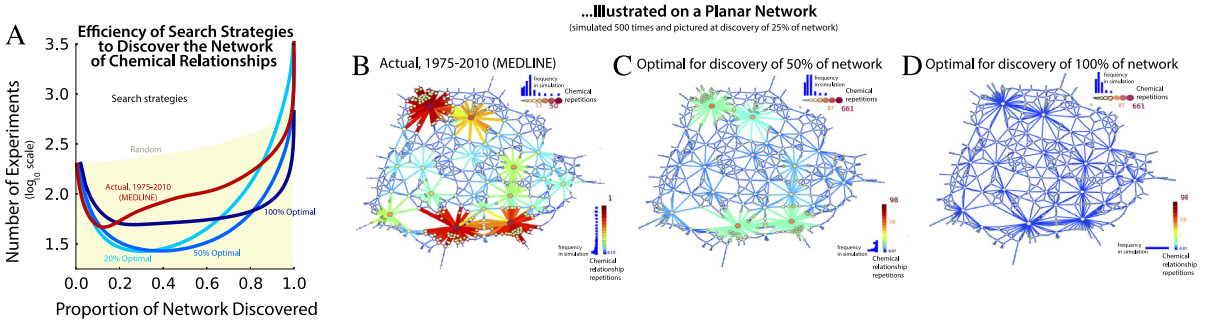

Fig. 18. (A) Comparison of the efficiency of different research strategies in discovering the knowledge network in which nodes are chemical and two chemicals are links if they appear in the same article or patent abstract. Efficiency is defined as the estimated number of experiments required to discover from 1% to 100% of the network. The strategies include random choice, the inferred MEDLINE strategy, and optimal strategies for discovering 20%, 50%, and 100% of the network. (BCD) illustrate the realizations of the real strategy and the optimal strategies in the knowledge network.
Source: The figure is reprinted from Ref. [392].

still lacking because they both attach importance to learning, improvement, and the acquisition of new knowledge. The primary ambiguity is whether these two activities are differentiated by the type of learning or the existence of learning. Gupta et al. argued that it is more logical to differentiate between exploration and exploitation by using the former criteria [396]. The trade-off between exploration and exploitation is essential for solving numerous problems. Specifically, a choice must be made between searching for new solutions and improving existing solutions [395–400]. Because exploration is a riskier process, it generally leads to more uncertain and distant benefits than exploitation. A proper balance between exploration and exploitation is considered to be critical for adaptive behavior in humans and other animals [395,400,401].

Combination may be a primary method for innovation during the process of exploration and exploitation. Youn et al. analyzed US patent records from 1790 to 2010 and characterize invention as a combinatorial process because it identifies distinct technologies and their combinations with patent technology codes [402]. In the context of combination, because time evolves, the larger the pool of inventions that are accumulated, the higher the probability of generating new inventions by using combination processes, which ultimately leads to a singularity that the innovation system transitions to super explosive growth and the total number inventions diverges. To better understand the evolutionary dynamics of innovation combination and the condition for singularity, Sood [403] modeled combinational innovation dynamics as an interacting branching process where each mating pair of new inventions and old inventions produces a certain number of new inventions that follow a Poisson distribution with mean p, and old inventions expire with a characteristic probability p/τ. Using theoretical and numerical analysis, they reported that no phase transition occurs and when $p \ll 1$ and $\tau = \infty$, the surviving processes that occur prior to a super explosive phase follow a long period of a quiescent state. Soléet al. [404] extended the pairwise combination model to a multiple combination case and consider different choices of aging functions, e.g., power-law aging and exponential aging. Based on mean-field theory analysis, they found that singularity emerges when the long-range memory mechanism is used; however, if the aging has a characteristic time scale, instead of singularity, a black hole of old invention presents and slows down the rate of invention creation.

In certain situations, we cannot derive an analytic expression for innovation dynamics; agent-based modeling (ABM) [405] is an alternative for analyzing the basic rules that govern the innovation system and their relationships among these rules and for proposing effective guides and strategies to optimize the system. Berger-Tal [406] developed an agent-based model with time varying exploration strategy to pursue certain goals (e.g., energy, money or prestige) to investigate the trade-off between exploration and exploitation. In their model, four distinct phases are produced: knowledge establishment, knowledge accumulation, knowledge maintenance, and knowledge exploitation. During a subject's life-span, the four knowledge phases are mutually transitioned and occur multiple times according to the changing environment, which implies that the optimal solution to the exploration and exploitation trade-off depends on the subject's life-stage and current environmental conditions.

For firms, the Research and Development (R&D) process is crucial for preserving their competitiveness and increasing market share. Forming R&D alliances is one important development strategy where the engaged firm can gain access to different assets more quickly, enlarge their technology pool more effectively and incur fewer costs and less risk than they could individually [407]. The allied R&D process is often modeled as agents moving in a knowledge space to search for potential technology and exchanging acquired knowledge among alliances. Using ABM, Tomasello et al. [408] studied the exploration process with partner selection and rewiring. They found that firms tend to form clusters with the number of clusters depending on the rewiring rate and interaction radius. The exploration performance was defined as the total move distance. It has an inverted U-shaped dependence on the two parameters.

7.2.2. Innovation networks

Communication networks. The trade-off between exploitation and exploration is particularly important when individuals engage in social learning to collectively solve problems [409–413]. The exploration process introduces novel solutions

in the population, while the exploitation process diffuses good solutions to increase the overall performance of the group. The interacting social networks of individuals will highly affect how the information of novel solutions is disseminated in global groups. This situation is often modeled as a group of individuals searching for optimal solutions regarding rugged landscapes [414–418]. The primary difficulty of this type of search process is the presence of local optima. As a result, groups seek to reach the optimal solution by avoiding local optima with various communication structures.

Generally, the effect of communication structures on innovation creation and propagation continues to be debated. In certain early studies [409,414,419], they found that when facing complex problems, networks of agents that exhibit lower efficiency can outperform more efficient networks, where "efficiency" refers to the speed that information regarding trial solutions can spread throughout the network. For example, Lazer and Friedman [414] conducted an agent-based simulation and demonstrated that a "locally connected lattice" outperforms a "fully connected network" for determining the optimal solutions in the long run. Mason and Watts [415] conducted a series of 256 online behavioral experiments where groups of individuals solve complex tasks with eight different interacting networks. They reported a conflicting result: efficient networks (that promote faster information flow in the population) outperform inefficient networks as related to the average success of group members. In addition, network efficiency affects the distribution of individual success. Derex and Boyd [416] modeled the innovation problem in a more realistic manner. Each individual is provided with certain ingredients to create a new ingredient; the created ingredient can then be used to create a high-level ingredient. Different combinations of ingredients lead to the failure or success of ingredients with different scores. Their experiment demonstrates that when individuals learn from other successful individuals, fully connected groups strongly reduce cultural diversity, while partially connected groups can provide more diverse solutions. This diversity is critical because it allows groups to develop complex solutions that are impossible for fully connected groups to develop.

In addition to the effect of special communication structures on collective performance for social problems, the influence of social learning strategies that are used by individuals and organizations are valuable because disregarding these strategies might not produce the desired effects of a given structure. The agent-based simulation that was conducted by Barkoczi and Galesic [420] revealed that the social learning strategies of individual members significantly affect the effectiveness of certain communication networks on group performance. Specifically, by applying two social learning strategies (i.e., best member and conformity) and eight network structures that include a broad range of possible topologies, they found that inefficient network structures outperform efficient network structures when individuals adopt the best member strategy, and the opposite occurs when the conformity strategy is adopted. In addition, they found that groups that rely on the best member strategy perform well for simple tasks, while complex tasks need groups that follow the conformity strategy based on a small sample of other individuals.

Patent citation networks. Compared to scientific knowledge innovation, technology innovation is more clearly defined. Each citation for technology is interest related; therefore, the redundant citing links are much less significant than for the scientific domain. In addition, the patent office plays a key role in issuing new patents and the patent codes fully encapsulate the novelty that is clearly delineated in the claims. In a patent citation network where each patent serves as the vertex, the presented citation between two patents, e.g., patent *A* and *B*, indicates that patent *A* is partly built on patent *B*. Therefore, the entire patent citation network may resemble a network of idea production, combination and propagation and subsequently can be used to map the technological trajectories that record incremental innovation and breakthroughs.

Verspagen [421] extended the method used in Ref. [422] and analyzed the scientific citation network of publications on the discovery of DNA. He discovered the primary flows of ideas by using the extracted backbone of the patent citation network. The application of this method to fuel cells reveals the historical dynamics of fuel cell research beginning with broad exploration in various directions followed by a focus on persistent and evolutionary interpretable results for further exploitation. This selective and persistent nature is shared by other technological trajectories. Acemoglu, Akcigit and Kerr [423] constructed the technology innovation network from 1.8 million US patents and their citation properties. In this network, each node represents a technology field, and the citation links are weighted by relative citations. A regression model is proposed to investigate the effect of network structures and patent growth in upstream technology fields on a certain technology field's future development. It is found that innovation advances in one section of the network significantly impact nearby technology fields. If a technology class includes more prior upstream innovations, it tends to have more innovations in the future. However, this effect is limited to local areas within the network. To clarify, innovation of technologies will not accelerate the innovation of distant technologies in the network.

Software networks. Software systems are one of the most influential systems in modern society and recently have been utilized to study the evolution of technology. A software system includes numerous units that interact with each other when designing, coding and executing software. Although software engineering is purpose-driven, it shares numerous similar characteristics with patents. One of most important common characteristics is the use of a combination of existing codes and new codes to achieve new functional requirements of software systems. The benefit of combination is a reduction in the duplication of effort and acceleration in the progress of development. Because rich empirical data are available, it is possible to understand the innovation processes of software systems.

A software system can be modeled using complex networks where nodes represent classes, motifs, patterns, libraries, packages, subsystems and components, and edges represent dependency relationship between the nodes. Complex networks have been used to analyze software systems at different scales that range from software motifs [424] to collaboration graphs [425]. Empirical studies regarding the networks that are extracted from software systems reveal several key structural properties. Generally, these networks follow power-law degree distributions [426–428], exhibiting small-world

properties [426], community structures [426,429] and various other statistical topological features [430–432]. The evolution of software networks has attracted widespread attention. Most software systems obey certain structural evolution laws [425,430,433–435]. In addition, certain models have been proposed for software network evolution and describe the evolutionary mechanisms from different perspectives such as refactoring processes [426], software patterns [436], modular attachment [437] and multi-level networks [438].

Distributed development makes the design of large-scale programs possible. However, reusing existing code may result in the emergence of incompatibilities among software packages, which leads to failures in the functionality of certain packages and makes the system unstable [433]. The perspective of complex systems encourages researchers to analyze systems from a local and global aspect, which helps to identify key nodes in the system that ensure the code combination, encapsulation and maintenance and enhance the system's robustness.

7.3. The spread of knowledge and scientific ideas

The spread of knowledge and scientific ideas occurs through various manners, including the production (writing) and consumption (citing) of scientific papers, coauthorships and other social relationships and the physical migration of researchers through geographical space.

7.3.1. Knowledge epidemics and citation flows

Once knowledge is created, the next process is diffusion, which is reflected by the adaptation and application of novel techniques and ideas, first by a few individuals and then by the entire scientific and engineering community. In the scientific community, scientific publications (via reading, adoption and citing) are the primary channels for knowledge diffusion. In addition, scientific meetings and other informal communications are channels where knowledge diffusion occurs. Currently available big data make it possible to trace the diffusion of knowledge and quantify the depth and breadth of the diffusion process.

Quantitatively, knowledge diffusion of a specific concept can be measured by the evolving number of published papers and researchers related to that concept. As documented by the spread of numerous historical examples of knowledge, ideas and innovations, the propagation of knowledge has many properties that are similar to an epidemic that spreads through a population and was triggered from a seed. Based on the assumption of an epidemic model, the population is divided into multiple categories, including adopters (that have been explicitly diffused), non-adopters (that have not been diffused) and other intermittent species. Goffman [439] treated the transmission of ideas as a standard SIR model with three categories: S (non-adopter), I (adopter) and R (leave the idea or topic for various reasons) to investigate the spread of knowledge regarding mast cells. Bettencourt et al. [440] add E (idea incubators) and Z (skeptics) to the population model to quantify the spread of Feynman diagrams throughout the theoretical physics community. Other models that characterize the diffusion of ideas, knowledge and technologies were summarized in Ref. [441], which discusses the evolution and relationships between these models in detail.

In addition to the abovementioned well-mixed dynamics, diffusion is also modeled as occurring on the networks that describe the multiple relationships among researchers, i.e., collaboration relationships and citing relationships. These networked models consider the heterogeneity of individual researchers and their interaction patterns. Kiss et al. [442] proposed an individual-based model to describe the diffusion of a specific research topic through the citation-based contact network of disciplines and report a good fit between the model and empirical data of protein class kinesin. Gao and Guan [443] used the citation network of individual papers as the contact network where an epidemic occurs and study the spreading dynamics of research regarding h-index.

Citation flow, which is mapped using the paper-level citation network, is the proxy for knowledge diffusion at higher levels such as fields, institutions, cities and countries. This network-based approach provides a new research perspective for understanding global-level knowledge diffusion. In addition, citation flow has been applied to the patent citation network [444]. Numerous studies have analyzed the citation flow of journals, affiliations, cities and countries. For example, Yan [445] studied disciplinary knowledge flows using Scopus' journal-level citation dataset via a three-step approach: examining the citation characteristics of disciplines through scientific trading dimensions; analyzing citation flows between two disciplines; and assessing individual disciplinary citation flow diversity through Shannon entropy. They found that most disciplines become more diversified in knowledge exchange. Based on papers that were published in the Proceedings of the National Academy of Sciences from 1982 to 2001, Börner et al. [446] constructed an institution-to-institution network and treated institution as a node that assumes a dual role of both producer and consumer. They studied knowledge diffusion among institutions and identify the highest producers and their consumers, including the highest consumers and their producers on the basis of the normalized out-degree or in-degree of nodes. Zhang et al. [236] regarded cities as knowledge producers or consumers and studied the flow of Physics knowledge among cities based on the knowledge diffusion proxy algorithm that describes a biased random walk process and is applied on the network of net trade flows. They identified the top four producer cities in 2009 and their top ten consumers using a knowledge diffusion proxy algorithm in 1990–2009 in the US and Europe. In addition, Mazloumian et al. [235] measured knowledge flows using the Web of Science data for 2000–2009 at the continent, country and city level by analyzing excess scientific production and consumption.

The advisor–student relationship is another channel that conveys the flow of ideas. During mentoring, supervisors may pass their ideas to their students and inspire creative ideas. In addition, mentorship leads to the transfer of knowledge

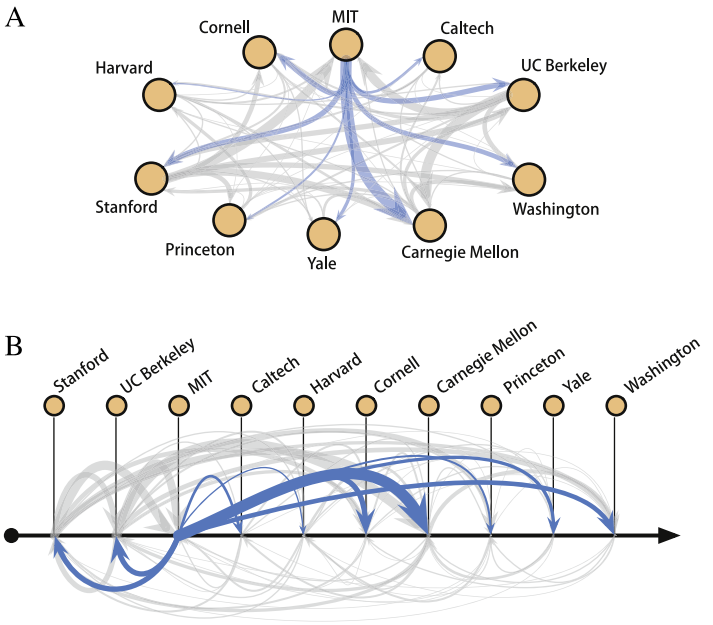

Fig. 19. (A) The hiring relations of 267 computer science faculties among 10 universities, with the width of a directed link uv representing the number of faculty members at university u who received their doctoral degree at university v. The outgoing links of MIT are highlighted. (B) The ranking of 10 universities that minimizes the total weight of "upward" links from lower ranked universities to higher ranked ones.
Source: The figure is reprinted from ref. [239].

among countries and disciplines, e.g., foreign students may bring ideas back to their home countries, and students may use the ideas that they borrowed from their mentors' disciplines in their own research. Gargiulo et al. [447] extensively analyzed the mentor–student relationship using data obtained from *The Mathematics Genealogy Project*, including country and disciplinary information. *The Mathematics Genealogy Project*, one of the largest academic genealogy datasets, includes approximately 200,000 mathematicians and their demographics over several centuries. Using the aggregated country network and discipline network that was mapped from the genealogy data, these scholars analyzed patterns of mathematical concepts that spread in temporal and spatial scales, e.g., the evolution of the central countries (The US, Germany, Russia and the UK) and their surrounding countries, including the emerging scientific paradigms at different time periods. Similarly, ideas spread along kinship lines. Ideas are frequently exchanged among family members. We have witnessed numerous outstanding families in the scientific community, e.g., the Curies and the Bohr family. Prosperi et al. [448] systematically examined surname patterns in health science and identified the evolving county-specific patterns of family-tied authorships. Kinship may affect resource allocation and opportunities for talented new researchers; however, the direction (good or bad) and degree of these effects need additional investigation.

7.3.2. Idea propagation in physical and virtual spaces

Scientists seldom remain in one place for their entire career. Generally, after obtaining their Ph.D., researchers move to several other groups and work as a postdoc prior to finding a permanent position [60]. Mobility is generally accompanied with the flow and exchange of scientific ideas. Generally, an inflow of scientists will substantially benefit the scientific development of a country [449]. From this perspective, a better understanding of the statistical patterns of the mobility of scientists can help us understand the spatial spread of knowledge and scientific ideas. Human mobility is an emerging research field of complexity science that has been studied for more than ten years. Thus far, numerous models have been developed to reproduce the statistical patterns that are extracted from empirical human mobility data [20,22,23,25]. Recently, it has been noted that local mobility decisions are driven by opportunity seeking, which leads to a radiation model that captures the mobility fluxes of actual data [23]. The mobility of scientists follows a similar rule, i.e., following research funding [368]. However, this rule is also remarkably influenced by culture [450].

In regard to movements of researchers between universities, Clauset et al. constructed a hiring network with a directed link from university i to j representing the j is hiring a researcher who obtains Ph.D. from i [239]. This model reveals an obvious hierarchical structure in the network, which leads to a natural ranking of universities as exhibited in Fig. 19. More detailed analyses of the patterns that govern scientists' movements are conducted with the APS data where the affiliation information of authors can be extracted. Deville et al. constructed scientists' career trajectory by linking the affiliations in all his other publications [30]. Similarly, quantities, such as the number of authors, publications, and total citations, are computed for each affiliation; each of the quantities follows a heterogeneous distribution. The number of authors that an institute is positively

correlated with has an impact on the institute (i.e., average citation per paper); however, the number of authors has little influence on average productivity (i.e., average paper per author). The distribution of this mobility distance has a fat-tail that can be fitted by a power-law, which indicates career movements are spatially localized. The probability of movement decays quickly with career age, which implies that most movements occur during the early stages of scientists' careers. Interestingly, moving to lower-ranked institutions (ranked by the total number of citations) slightly decreases one's scientific performance, but moving to a prestigious institution is not associated with performance enhancement. One possible explanation is that scientists who are capable of obtaining a position in higher-ranked institutions may already be very outstanding and have relatively little room to improve their performance. Gargiulo et al. investigated the mobility of scientists using network theory by constructing weighted and directed mobility networks between universities and countries [451]. A link from a university i to university j naturally indicates scientists move from i to j, and the link weight records the number of scientists who made this movement. The motif analysis reveals that researchers with short careers primarily visit different universities, and researchers with longer careers tend to move between two universities. This phenomenon is consistent at the country level. Consistent with Ref. [450], the mobility of scientists is strongly influenced by the linguistic and historical similarity between two countries. The mobility process also has a memory effect, i.e., the first one or two affiliations play an important role in determining scientists' career.

Scientific ideas propagate spatially through citations and collaboration. The rapid development of Internet and communication technologies has made long-distance collaboration much more convenient, which accelerates the spatial transmissibility of information. Using the affiliation data of scientific publications, Pan et al. analyzed the role of geography in global citations and collaborations [453]. The results indicate that the probability of citations and collaborations follow gravity laws. The total number of collaboration relationships between two cities i and j can be estimated as $p_{ij} = s_i s_j / d_{ij}^{\alpha}$, where d_{ij} represents the geographic distance between these two cities and s_i is the node strength that represents the number of collaborations of city i. For citation networks, the citation flow from i to j is $p_{ij} = s_i^{out} s_j^{in} / d_{ij}^{\alpha}$, where s_i^{out} represents the total number of outgoing citations from i and s_j^{in} represents the total number of incoming citations to j. α is the exponent that represents the effect of distance, which is estimated to be approximately 1.16 for collaborations and 0.77 for citations.

Another major way to spread scientific ideas is through the evolution of scientists' research interests, which can be modeled as scientists moving on a virtual space that is formed by research topics. When a scientist changes his/her research topic, the scientific concepts from the prior topics he/she worked on will naturally spread to the new field. Jia et al. constructed a research topic space using the PACS codes of scientific publications [454]. By tracing the publication records of scientists, they found that the percentage of scientists within a given range of change in interest follow an exponential distribution. A random-walk-based model is developed to reproduce this empirical observation. Based on a similar concept, De Domenico investigated the evolution of human knowledge by modeling changes in the research interests of academics across time by using multilayer networks for each topic or area of knowledge as a layer [452]. A feature referred to as "diaspora of the knowledge" is revealed, where disciplines act as sources or sinks of academics' interest. The flow network of knowledge diaspora is shown in Fig. 20. The attractiveness of each topic evolves with time and reveals fundamental periods where there was an increase of interest in specific areas of human knowledge.

8. Conclusions and outlook

The increasingly available scholarly big data provides an unprecedented opportunity to quantitatively explore research activities as well as their outcomes. Although considerable effort has been devoted to analyzing large-scale academic data, various key aspects of scholarly big data have remained underexplored until recently. This is because the accelerating growth of the scholarly data recording scientific publications, preprints, patents and grants requires not only much more computing power but also novel, suitable theoretical frames in which the science system can be better analyzed, modeled and predicted. Along with the expanding data size, science is increasingly being recognized as a complex system, with various types of components interacting with each other according to multiple relations. The persistent growth of scientists and research teams, together with the connections between them via multiple channels such as co-authored publications, international conferences, projects, citations, e-mail, and social media, have largely advanced the development of science and meanwhile significantly multiplied the complexity of the system. Using the approaches from complex networks, human dynamics and statistical physics, many emergent collective phenomena in the science system have been revealed, and a variety of quantification algorithms combining different types of information have been developed. In this context, there is a recent trend of studying scholarly big data from the perspective of complex systems, which forms an emerging new research domain called the science of science (SOS).

In this review, we summarize recent major progress in SOS, with a special emphasis on the works regarding science as a complex system. The topics cover the classic issues of evaluating papers and scientists, new research directions such as understanding and modeling structural and dynamical patterns, and more practical issues, such as predicting the evolution of a system and the early identification of promising components. The progress of other hot topics such as the paths to success in science and the creation and diffusion of knowledge are reviewed. In all of these topics, the approaches of complexity science play a unique role in providing new insights. The most prominent approach is the tool of complex networks. Describing scholarly data by using complex networks reveals several significant structural properties, including the high clustering tendency, rich-club effect and community structure, with which one can predict future links (interactions) and develop network growth models capturing the evolution mechanisms of a system. In addition, the citation networks

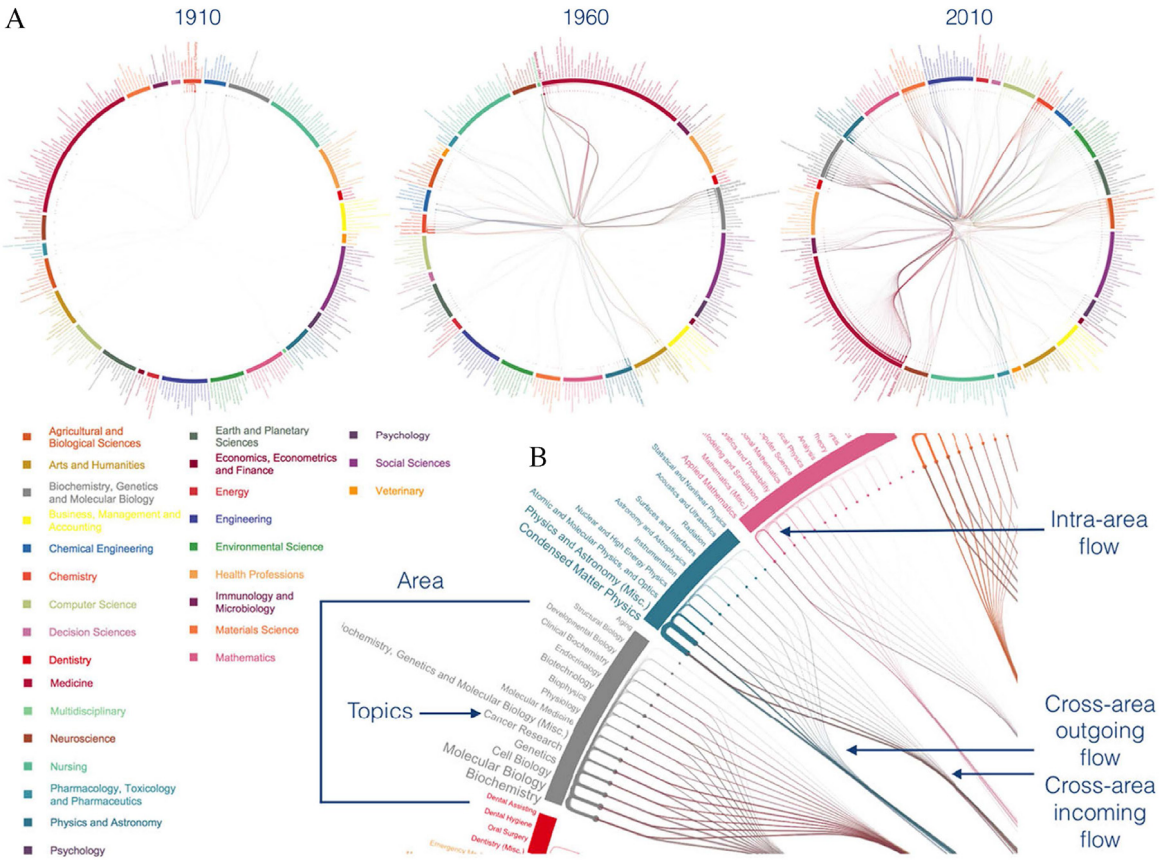

Fig. 20. (A) Flow of authors' research activity moving from one topic to others across time. Points on the circle represent topics which are further colored according to their coarse-grained area. A link between two topics indicates that at least one author switches from one to another 5 years later. (B) A local enlarged picture of the flow network illustrating the definition of topic and area, as well as different type of flows including "intra-area flows" and "cross-area flow".
Source: The figure is reprinted from Ref. [452].

and collaboration networks allow quantifying the impact of papers and scientists via iterative algorithms that incorporate their neighborhood information, leading to more objective evaluations. The aforementioned multiple interactions between scientists can also be naturally modeled by the multilayer networks. The other useful approach is human dynamics analysis. Scientific research is conducted by scientists, and their activities, such as their spatial distribution, mobility, and productivity, can be analyzed with the help of the methods developed in human dynamics. As such, numerous scaling laws and spatial–temporal patterns are identified. More importantly, several mechanistic models are proposed to reproduce these empirical observations, which are also used as powerful tools for prediction. Apart from these two major approaches, various techniques from statistical physics are employed to analyze scholarly data. One representative example is the rescaling and data collapsing techniques, which are used for a wide range of purposes, including correcting the bias in time, research domains, journals and universities.

Together with the development of the theoretical works on SOS, there are also many applications in practice. Some traditional metrics have been incorporated into web portals for users to browse and search. To keep those metrics up-to-date, these web services are updated regularly in an incremental, automated manner. The biggest such web site is Google Scholar (www.scholar.google.com), where one can set up a personal home page and many metrics, such as total citations, total publications, and *h*-index, are automatically computed. The authors of some papers also establish a web page, where the complete rankings of papers or authors are made available for users in a visual and customizable format. Two examples of this type of web site are as follows: www.sciencenow.info from paper [179]; https://icite.od.nih.gov from paper [165]. In addition, some software has been developed to help scientists analyze scholarly data. Some widely used ones include *CiteSpace* [455], which is used for visualizing and analyzing emerging trends and temporal patterns embedded in scientific literature, *VOSviewer* [456], which is used for constructing and visualizing bibliographic networks and for the further visualization and analysis of research areas and research trends, and *Sci* ^{2} *Tool* [457], which is for the temporal, geospatial, topical, and network analysis and visualization of scholarly datasets at different levels.

The practical use of research metrics may cause negative consequences if administrators overly rely on the selected indicator as a basis of tenure, promotion, and resource decisions. This issue has become increasingly significant in recent

years. For instance, universities are obsessed to achieve a higher position in the global ranking, so the resource and research funding are biasedly allocated; researchers struggle to have a high h-index and more top journal papers in short time, so many long-standing challenges in science are left for the last choice. The problem has been widely realized in scientific communities. Two key moves to correct this trend are the "San Francisco Declaration on Research Assessment" [458] by biologists in 2012 and the "Leiden Manifesto for research metrics" [459] by scientometricians in 2014. The San Francisco Declaration aims to stop the use of the journal impact factor in judging an individual scientist's work. It clearly states that "Journal-based metrics, such as journal impact factors, must not be used as a surrogate measure of the quality of individual research articles, to assess an individual scientist's contributions, or in hiring, promotion, or funding decisions". The Leiden Manifesto is more general. It proposes ten principles to guide the practical use of research evaluation metrics. These principles can be roughly summarized in four key aspects. First, quantitative evaluation is only a complementary to expert assessment. Second, the differences in research missions, disciplines, locally relevant research, individual portfolio should be taken into account when evaluating research. Third, data and evaluation must be open and verified. Finally, indicator system must be reviewed and updated. We believe that with persistent effort of scientists, the quantitative and qualitative evaluation of research will be more properly combined.

Despite considerable efforts, numerous issues in SOS remain challenging. Ranking is a major topic in SOS research. In the literature, countless algorithms designed to rank papers and scientists have been proposed. Each of them is based on distinct motivations and claimed to outperform the previous methods. However, relatively few of the works provide concrete evidence for these claims. A strict validation will not only reveal the limitation of a ranking method but also provide clues for further improvement. Thus, a systematic examination of the existing algorithms within a standard validation framework is needed. The popular validation approaches include prediction, scientific awards, dynamic processes (spreading coverage), robustness analysis, and models with the ground truth. The following questions should be addressed in future research: Are these validation methods meaningful for evaluating the performance of a metric? Are there other neglected validation methods? Will a new metric outperform the existing ones in one or more validation methods?

Predicting the future performance of a scientist is a very practical issue, with valuable applications to the faculty hiring process and grant decision process [252]. Though the predictions of some algorithms can be very accurate for papers and scientists with a large amount of historical data, it is still hard to identify the promising ones early on. In this article, we review some related efforts. However, there is much more to do in the future. The major challenge is to determine how to extract the useful predictors from the limited information from various data sources, including publications, citations, collaborations, and affiliations. In addition, an ultimate goal for the prediction would be not only identifying promising nodes but also quantitatively predicting their long-term impact. In recent years, there has been a trend of developing mechanistic models for the science system. Two representative examples are the model characterizing the citation dynamics of papers [32] and the model describing the publishing dynamics of scientists [55]. These models usually capture the essential driving mechanism of a system's growth. Thus, they can be used to address various issues, ranging from quantitatively explaining an empirical phenomenon to predicting the long-term evolution of a system. However, these models need to be further improved for early prediction. Additionally, mechanistic models for other science activities remain to be developed. For instance, mechanistic models of citation behavior and collaboration behavior are still missing. In addition to scientists, one can also focus on modeling the publishing dynamics of entities at higher levels, such as journals, funding groups, research teams and countries.

Another issue in SOS is to reveal the hidden casuality between different factors. Numerous past works in SOS involve correlation analysis, especially papers investigating the factors leading to success. Though various factors with positive influences have been uncovered, these factors are not completely independent. Further analyses are required for dimension reduction and to extract the key factors. In addition, some of the factors may be consequences of success rather than the causes. Even though they will appear to be positively correlated with success, they should be eliminated. For instance, both reputation and cumulative advantage are considered as important factors of success. However, they are intrinsically entangled in the sense that scientists with great reputations usually have high cumulative advantages.

The complex network as a tool has already remarkably advanced the research of SOS. With the rapid development of the network science, some latest analysis frameworks and approaches can be applied to SOS for better understanding and modeling the system. For instance, multilayer networks have been intensively studied in network science recently. Issues include the centrality measures [460], community detection [461], structural reducibility [462], and coupled dynamics (e.g., spreading and synchronization) [463] and so on. As the science system can be naturally described by multilayer networks, these approaches are very valuable for deepening our understanding of the structural organizations and evolution mechanisms of the system. In addition, most existing multilayer networks in the SOS only contain citation and co-authorship links. In fact, many other connections are still neglected, such as the interactions between scientists via conferences, projects and social media. Various other layers can also be included, such as the layer of research domains, of key words, of affiliations, of journals, and of research teams. Including these layers will lead to a more complete understanding of the structure and evolution of the science system.

SOS research has significantly benefited from the complexity science, it is actually also pushing forward the development of complexity science by providing general analysis methodologies as well as high-quality empirical databases. Specifically, some methods in SOS can inspire novel tools for other problems such as online information filtering, critical part identification, algorithm robustness enhancement, and trend prediction. For instance, the evaluation of publications and scientists is actually closely related to the problem of critical node identification in complex networks. Numerous metrics

can certainly be used as network centrality measures. For example, the h-index has recently been extended to ranking nodes in complex networks, and its relations with respect to the traditional degree centrality and k-core centrality have been revealed [189]. More importantly, the evaluation of publications and scientists usually requires taking into account not only the network structure but also the temporal information, resulting in numerous methods for node ranking in evolving networks, which can be applied to other real networks. One important issue in SOS is predicting the future evolution of a system, which is also one major challenge in information filtering. The mechanistic models of SOS provide accurate prediction of the citation dynamics of papers and the production dynamics of scientists. Based on a similar framework, mechanistic models can be developed for online users and products to predict user activities and product popularity. Finally, science systems provide large-scale, high-quality data consisting of multiple types of complex interactions with highly accurate time information. Such data can be analyzed from various perspectives, including human dynamics, complex networks, and agent-based modeling. With the persistent dedication of researchers in complexity science and other fields in SOS, various further breakthroughs are expected in the near future.

Acknowledgments

The authors would like to thank Zengru Di, Chi Ho Yeung, Linyuan Lü, Bertrand Roehner, Liying Yang for their helpful suggestions and advice to improve this manuscript. This work was supported by the National Natural Science Foundation of China (Grant Nos. 61603046, 61374175 and 61773069), the Natural Science Foundation of Beijing (Grant No. 16L00077) and the Fundamental Research Funds for the Central Universities (Grant No. 2015 kJJCA06). The Boston University work was supported by NSF Grants PHY-1505000, CMMI-1125290, and CHE-1213217, and by DTRA Grant HDTRA1–14-1–0017 and DOE Contract DE-AC07–05Id14517.

References

[1] R.P. Light, D.E. Polley, K. Börner, Open data and open code for big science of science studies, Scientometrics 101 (2) (2014) 1535–1551.
[2] J.E. Hirsch, An index to quantify an individual's scientific research output, Proc. Natl. Acad. Sci. USA 102 (46) (2005) 16569–16572.
[3] E. Garfield, Citation analysis as a tool in journal evaluation, Science 178 (4060) (1972) 471–479.
[4] L. Leydesdorff, S. Milojevic, Scientometrics, Int. Encyclopedia of the Soc. & Behav. Sci. 21 (2012) 322–327.
[5] R. Van Der Hofstad, Random Graphs and Complex Networks, Cambridge University Press, Cambridge, 2016.
[6] D.J. Watts, S.H. Strogatz, Collective dynamics of small-worldnetworks, Nature 393 (6684) (1998) 440–442.
[7] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509–512.
[8] R. Albert, A.-L. Barabási, Statistical mechanics of complex networks, Rev. Modern Phys. 74 (1) (2002) 47–97.
[9] M.E.J. Newman, The structure and function of complex networks, SIAM Rev. 45 (2) (2003) 167–256.
[10] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Complex networks: Structure and dynamics, Phys. Rep. 424 (4) (2006) 175–308.
[11] F. Giannotti, L. Pappalardo, D. Pedreschi, D. Wang, A complexity science perspective on human mobility, in: Mobility Data: Modeling, Management, and Understanding, Cambridge University Press, United Kingdom, 2012, pp. 297–314.
[12] V.D. Blondel, A. Decuyper, G. Krings, A survey of results on mobile phone datasets analysis, EPJ Data Sci. 4 (1) (2015) 10.
[13] F. Asgari, V. Gauthier, M. Becker, A survey on human mobility and its applications, 2013. ArXiv preprint arXiv:1307.0814.
[14] T. Zhou, X.P. Han, X.Y. Yan, Z.M. Yang, Z.D. Zhao, B.H. Wang, Statistical mechanics on temporal and spatial activities of human, J. Univ. Electron. Sci. Technol. China 42 (4) (2013) 481–540.
[15] J.G. Oliveira, A.-L. Barabási, Human dynamics: Darwin and Einstein correspondence patterns, Nature 437 (7063) (2005) 1251.
[16] A.-L. Barabási, The origin of bursts and heavy tails in human dynamics, Nature 435 (7039) (2005) 207–211.
[17] A. Kentsis, Correspondence patterns: Mechanisms and models of human dynamics, Nature 441 (7092) (2006) E5–E5.
[18] A. Vázquez, J.G. Oliveira, Z. Dezsö, K.-I. Goh, I. Kondor, A.-L. Barabási, Modeling bursts and heavy tails in human dynamics, Phys. Rev. E 73 (2006) 036127.
[19] D. Brockmann, L. Hufnagel, T. Geisel, The scaling laws of human travel, Nature 439 (7075) (2006) 462–465.
[20] M.C. Gonzalez, C.A. Hidalgo, A.-L. Barabási, Understanding individual human mobility patterns, Nature 453 (7196) (2008) 779–782.
[21] C. Song, Z. Qu, N. Blumm, A.-L. Barabási, Limits of predictability in human mobility, Science 327 (5968) (2010) 1018–1021.
[22] C. Song, T. Koren, P. Wang, A.-L. Barabási, Modelling the scaling properties of human mobility, Nat. Phys. 6 (10) (2010) 818–823.
[23] F. Simini, M.C. Gonzalez, A. Maritan, A.-L. Barabási, A universal model for mobility and migration patterns, Nature 484 (7392) (2012) 96–100.
[24] X.-Y. Yan, X.-P. Han, B.-H. Wang, T. Zhou, Diversity of individual mobility patterns and emergence of aggregated scaling laws, Sci. Rep. 3 (2013) 2678.
[25] X.-Y. Yan, C. Zhao, Y. Fan, Z. Di, W.-X. Wang, Universal predictability of mobility patterns in cities, J. R. Soc. Interface 11 (100) (2014) 20140834.
[26] D. Brockmann, D. Helbing, The hidden geometry of complex, network-driven contagion phenomena, Science 342 (6164) (2013) 1337–1342.
[27] A. Bogomolov, B. Lepri, R. Larcher, F. Antonelli, F. Pianesi, A. Pentland, Energy consumption prediction using people dynamics derived from cellular network data, EPJ Data Sci. 5 (2016) 13.
[28] J. Alonso-Mora, S. Samaranayake, A. Wallar, E. Frazzoli, D. Rus, On-demand high-capacity ride-sharing via dynamic trip-vehicle assignment, Proc. Natl. Acad. Sci. USA 114 (3) (2017) 462–467.
[29] J. Ladyman, J. Lambert, K. Wiesner, What is a complex system? Eur. J. Philos. Sci. 3 (1) (2013) 33–67.
[30] P. Deville, D. Wang, R. Sinatra, C. Song, V.D. Blondel, A.-L. Barabási, Career on the move: Geography, stratification, and scientific impact, Sci. Rep. 4 (2014) 4770.
[31] F. Radicchi, S. Fortunato, C. Castellano, Universality of citation distributions: Toward an objective measure of scientific impact, Proc. Natl. Acad. Sci. USA 105 (45) (2008) 17268–17272.
[32] D. Wang, C. Song, A.-L. Barabási, Quantifying long-term scientific impact, Science 342 (6154) (2013) 127–132.
[33] S. Redner, Citation statistics from 110 years of physical review, Phys. Today 58 (6) (2005) 49–54.
[34] R. Sinatra, P. Deville, M. Szell, D. Wang, A.-L. Barabási, A century of physics, Nat. Phys. 11 (10) (2015) 791–796.
[35] M.E.J. Newman, The structure of scientific collaboration networks, Proc. Natl. Acad. Sci. USA 98 (2) (2001) 404–409.
[36] S. Redner, How popular is your paper? An empirical study of the citation distribution, Europhys. J. B 4 (2) (1998) 131–134.
[37] Q. Ke, Y.-Y. Ahn, Tie strength distribution in scientific collaboration networks, Phys. Rev. E 90 (2014) 032804.
[38] M.E.J. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E 74 (3) (2006) 036104.

[39] J. Leskovec, J. Kleinberg, C. Faloutsos, Graphs over time: densification laws, shrinking diameters and possible explanations, in: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, ACM, 2005, pp. 177–187.
[40] J. Gehrke, P. Ginsparg, J. Kleinberg, Overview of the 2003 KDD cup, ACM SIGKDD Explorations Newslet. 5 (2) (2003) 149–151.
[41] R.K. Pan, J. Saramäki, The strength of strong ties in scientific collaboration networks, Europhys. Lett. 97 (1) (2012) 18007.
[42] Y.-H. Eom, S. Fortunato, Characterizing and modeling citation dynamics, PLoS One 6 (9) (2011) e24926.
[43] M. Li, Y. Fan, J. Chen, L. Gao, Z. Di, J. Wu, Weighted networks of scientific communication: the measurement and topological role of weight, Physica A 350 (2) (2005) 643–656.
[44] Y.-B. Zhou, L. Lü, M. Li, Quantifying the influence of scientists and their publications: distinguishing between prestige and popularity, New J. Phys. 14 (3) (2012) 033033.
[45] C. Schulz, A. Mazloumian, A.M. Petersen, O. Penner, D. Helbing, Exploiting citation networks for large-scale author name disambiguation, EPJ Data Sci. 3 (1) (2014) 11.
[46] P. Erdös, A. Rényi, On the evolution of random graphs, Publ. Math. Inst. Hung. Acad. Sci. 5 (1) (1960) 17–60.
[47] M.L. Goldstein, S.A. Morris, G.G. Yen, Group-based Yule model for bipartite author-paper networks, Phys. Rev. E 71 (2005) 026108.
[48] Y. Fan, M. Li, J. Chen, L. Gao, Z. Di, J. Wu, Network of econophysicists: a weighted network to investigate the development of econophysics, Internat. J. Modern Phys. B 18 (17n19) (2004) 2505–2511.
[49] S. Lehmann, B. Lautrup, A.D. Jackson, Citation networks in high energy physics, Phys. Rev. E 68 (2003) 026113.
[50] M. Girvan, M.E.J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99 (12) (2002) 7821–7826.
[51] Z.-X. Wu, P. Holme, Modeling scientific-citation patterns and other triangle-rich acyclic networks, Phys. Rev. E 80 (2009) 037101.
[52] S. Boccaletti, G. Bianconi, R. Criado, C. del Genio, J. Gómez-Gardenes, M. Romance, I. Sendina-Nadal, Z. Wang, M. Zanin, The structure and dynamics of multilayer networks, Phys. Rep. 544 (1) (2014) 1–122.
[53] M. De Domenico, A. Solé-Ribalta, E. Omodei, S. Gómez, A. Arenas, Ranking in interconnected multilayer networks reveals versatile nodes, Nat. Commun. 6 (2015) 6868.
[54] D.F. Klosik, S. Bornholdt, M.-T. Hütt, Motif-based success scores in coauthorship networks are highly sensitive to author name disambiguation, Phys. Rev. E 90 (3) (2014) 032811.
[55] R. Sinatra, D. Wang, P. Deville, C. Song, A.-L. Barabási, Quantifying the evolution of individual scientific impact, Science 354 (6312) (2016) aaf5239.
[56] V.I. Torvik, M. Weeber, D.R. Swanson, N.R. Smalheiser, A probabilistic similarity metric for medline records: A model for author name disambiguation, J. Am. Soc. Inf. Sci. Tec. 56 (2) (2005) 140–158.
[57] A.A. Ferreira, M.A. Gonçalves, A.H. Laender, A brief survey of automatic methods for author name disambiguation, ACM SIGMOD Record 41 (2) (2012) 15–26.
[58] J. Kim, J. Diesner, Distortive effects of initial-based name disambiguation on measurements of large-scale coauthorship networks, J. Assoc. Inf. Sci. Technol. 67 (6) (2016) 1446–1461.
[59] D.R. Amancio, O.N. Oliveira, L. D.F. Costa, On the use of topological features and hierarchical characterization for disambiguating names in collaborative networks, Europhys. Lett. 99 (4) (2013) 48002.
[60] J. Bohannon, K. Doran, Introducing ORCID, Science 356 (6339) (2017) 691–692.
[61] M.E.J. Newman, Scientific collaboration networks. I. Network construction and fundamental results, Phys. Rev. E 64 (2001) 016131.
[62] M.E.J. Newman, Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality, Phys. Rev. E 64 (2001) 016132.
[63] L. Krumov, C. Fretter, M. Müller-Hannemann, K. Weihe, M.T. Hütt, Motifs in co-authorship networks and their relation to the impact of scientific publications, Europhys. J. B 84 (4) (2011) 535–540.
[64] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (3) (2010) 75–174.
[65] M.E.J. Newman, Assortative mixing in networks, Phys. Rev. Lett. 89 (2002) 208701.
[66] V. Colizza, A. Flammini, M.A. Serrano, A. Vespignani, Detecting rich-club ordering in complex networks, Nat. Phys. 2 (2) (2006) 110–115.
[67] T. Opsahl, V. Colizza, P. Panzarasa, J.J. Ramasco, Prominence and control: The weighted rich-club effect, Phys. Rev. Lett. 101 (2008) 168702.
[68] J.J. Ramasco, S.A. Morris, Social inertia in collaboration networks, Phys. Rev. E 73 (2006) 016122.
[69] A.M. Petersen, Quantifying the impact of weak, strong, and super ties in scientific careers, Proc. Natl. Acad. Sci. USA 112 (34) (2015) E4671–E4680.
[70] Y.-H. Eom, H.-H. Jo, Generalized friendship paradox in complex networks: The case of scientific collaboration, Sci. Rep. 4 (2014) 4603.
[71] L. Wardil, C. Hauert, Cooperation and coauthorship in scientific publishing, Phys. Rev. E 91 (1) (2015) 012825.
[72] M.-G. Hâncean, M. Perc, Homophily in coauthorship networks of East European sociologists, Sci. Rep. 6 (2016) 36152.
[73] C.K. Fatt, E.A. Ujum, K. Ratnavelu, The structure of collaboration in the journal of finance, Scientometrics 85 (3) (2010) 849–860.
[74] H. Hou, H. Kretschmer, Z. Liu, The structure of scientific collaboration networks in scientometrics, Scientometrics 75 (2) (2008) 189–202.
[75] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, Graph structure in the web, Comput. Netw. 33 (1) (2000) 309–320.
[76] L. Šubelj, D. Fiala, M. Bajec, Network-based statistical comparison of citation topology of bibliographic databases, Sci. Rep. 4 (2014) 6496.
[77] S. Bilke, C. Peterson, Topological properties of citation and metabolic networks, Phys. Rev. E 64 (2001) 036106.
[78] B. Karrer, M.E.J. Newman, Random acyclic networks, Phys. Rev. Lett. 102 (2009) 128701.
[79] S. Gualdi, M. Medo, Y.-C. Zhang, Influence, originality and similarity in directed acyclic graphs, Europhys. Lett. 96 (1) (2011) 18004.
[80] M. Rosvall, C.T. Bergstrom, Maps of random walks on complex networks reveal community structure, Proc. Natl. Acad. Sci. USA 105 (4) (2008) 1118–1123.
[81] Y. Kim, S.-W. Son, H. Jeong, Finding communities in directed networks, Phys. Rev. E 81 (2010) 016103.
[82] E.A. Leicht, M.E.J. Newman, Community structure in directed networks, Phys. Rev. Lett. 100 (2008) 118703.
[83] V. Palchykov, V. Gemmetto, A. Boyarsky, D. Garlaschelli, Ground truth? Concept-based communities versus the external classification of physics manuscripts, EPJ Data Sci. 5 (1) (2016) 28.
[84] L. Šubelj, N.J. van Eck, L. Waltman, Clustering scientific publications based on citation relations: A systematic comparison of different methods, PLoS One 11 (4) (2016) e0154404.
[85] J.R. Clough, T.S. Evans, What is the dimension of citation space? Physica A 448 (2016) 235–247.
[86] M. Bertin, I. Atanassova, Weak links and strong meaning: The complex phenomenon of negational citations, in: Proc. of the 3rd Workshop on Bibliometricenhanced Information Retrieval (BIR2016), 2016, pp. 14–25.
[87] S. Kumar, Structure and dynamics of signed citation networks, in: Proceedings of the 25th International Conference Companion on World Wide Web, International World Wide Web Conferences Steering Committee, 2016, pp. 63–64.
[88] S. Gualdi, C.H. Yeung, Y.-C. Zhang, Tracing the evolution of physics on the backbone of citation networks, Phys. Rev. E 84 (4) (2011) 046104.
[89] M.C. Waumans, H. Bersini, Genealogical trees scientific papers, PLoS One 11 (3) (2016) e0150588.
[90] J.R. Clough, J. Gollings, T.V. Loach, T.S. Evans, Transitive reduction of citation networks, J. Complex Netw. 3 (2) (2013) 189–203.
[91] F. Radicchi, S. Fortunato, B. Markines, A. Vespignani, Diffusion of scientific credits and the ranking of scientists, Phys. Rev. E 80 (2009) 056103.
[92] M.L. Wallace, V. Lariviére, Y. Gingras, A small world of citations? The influence of collaboration networks on citation practices, PLoS One 7 (3) (2012) e33339.

[93] I. Fister Jr., I. Fister, M. Perc, Toward the discovery of citation cartels in citation networks, Front. Phys. 4 (2016) 49.
[94] H.-N. Su, P.-C. Lee, Mapping knowledge structure by keyword co-occurrence: a first look at journal papers in technology foresight, Scientometrics 85 (1) (2010) 65–79.
[95] T. Van Holt, J.C. Johnson, S. Moates, K.M. Carley, The role of datasets on scientific influence within conflict research, PLoS One 11 (4) (2016) e0154148.
[96] I. Rafols, M. Meyer, Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience, Scientometrics 82 (2) (2010) 263–287.
[97] V.P. Guerrero-Bote, F. Moya-Anegón, A further step forward in measuring journals scientific prestige: The SJR2 indicator, J. Informetr. 6 (4) (2012) 674–688.
[98] B. González-Pereira, V.P. Guerrero-Bote, F. Moya-Anegón, A new approach to the metric of journals scientific prestige: The SJR indicator, J. Informetr. 4 (3) (2010) 379–391.
[99] K.W. Boyack, R. Klavans, K. Börner, Mapping the backbone of science, Scientometrics 64 (3) (2005) 351–374.
[100] L. Leydesdorff, Betweenness centrality as an indicator of the interdisciplinarity of scientific journals, J. Am. Soc. Inf. Sci. Tec. 58 (9) (2007) 1303–1319.
[101] G. Cimini, A. Gabrielli, L.F. Sylos, The scientific competitiveness of nations, PLoS One 9 (12) (2014) e113470.
[102] G. Menichetti, D. Remondini, P. Panzarasa, R.J. Mondragón, G. Bianconi, Weighted multiplex networks, PLoS One 9 (6) (2014) e97857.
[103] S. Uddin, L. Hossain, K. Rasmussen, Network effects on scientific collaborations, PLoS One 8 (2) (2013) e57546.
[104] C. Biscaro, C. Giupponi, Co-Authorship and bibliographic coupling network effects on citations, PLoS One 9 (6) (2014) e99502.
[105] T. Martin, B. Ball, B. Karrer, M.E.J. Newman, Coauthorship and citation patterns in the physical review, Phys. Rev. E 88 (1) (2013) 012814.
[106] Y. Ding, Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks, J. Informetr. 5 (1) (2011) 187–203.
[107] P.L. Krapivsky, S. Redner, F. Leyvraz, Connectivity of growing random networks, Phys. Rev. Lett. 85 (2000) 4629–4632.
[108] S.N. Dorogovtsev, J.F.F. Mendes, A.N. Samukhin, Structure of growing networks with preferential linking, Phys. Rev. Lett. 85 (2000) 4633–4636.
[109] P.L. Krapivsky, S. Redner, Network growth by copying, Phys. Rev. E 71 (2005) 036118.
[110] P. Sen, Directed accelerated growth: application in citation network, Physica A 346 (1) (2005) 139–146.
[111] K. Klemm, V.M. Eguíluz, Highly clustered scale-free networks, Phys. Rev. E 65 (2002) 036123.
[112] A. Vázquez, Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations, Phys. Rev. E 67 (2003) 056104.
[113] Z. Xie, Z. Ouyang, Q. Liu, J. Li, A geometric graph model for citation networks of exponentially growing scientific papers, Physica A 456 (2016) 167–175.
[114] F. Papadopoulos, M. Kitsak, M.Á. Serrano, M. Boguná, D. Krioukov, Popularity versus similarity in growing networks, Nature 489 (7417) (2012) 537–540.
[115] Z. Xie, Z. Ouyang, P. Zhang, D. Yi, D. Kong, Modeling the citation network by network cosmology, PLoS One 10 (3) (2015) e0120687.
[116] H. Zhu, X. Wang, J.-Y. Zhu, Effect of aging on network structure, Phys. Rev. E 68 (2003) 056121.
[117] M.E.J. Newman, The first-mover advantage in scientific publication, Europhys. Lett. 86 (6) (2009) 68001–68006.
[118] S.N. Dorogovtsev, J.F.F. Mendes, Evolution of networks with aging of sites, Phys. Rev. E 62 (2000) 1842–1845.
[119] K.B. Hajra, P. Sen, Aging in citation networks, Physica A 346 (1) (2005) 44–48.
[120] K.B. Hajra, P. Sen, Modelling aging characteristics in citation networks, Physica A 368 (2) (2006) 575–582.
[121] M. Wang, G. Yu, D. Yu, Effect of the age of papers on the preferential attachment in citation networks, Physica A 388 (19) (2009) 4273–4276.
[122] S. Lehmann, A.D. Jackson, B. Lautrup, Life, death and preferential attachment, Europhys. Lett. 69 (2) (2005) 298–303.
[123] X. Geng, Y. Wang, Degree correlations in citation networks model with aging, Europhys. Lett. 88 (3) (2009) 38002.
[124] F.-X. Ren, H.-W. Shen, X.-Q. Cheng, Modeling the clustering in citation networks, Physica A 391 (12) (2012) 3533–3539.
[125] G. Bianconi, A.-L. Barabási, Bose-Einstein condensation in complex networks, Phys. Rev. Lett. 86 (2001) 5632–5635.
[126] M.C.V. Medo, G. Cimini, S. Gualdi, Temporal effects in the growth of networks, Phys. Rev. Lett. 107 (2011) 238701.
[127] E. Garfield, Premature discovery or delayed recognition-why, Current Contents (21) (1980) 5–10.
[128] A.F. Van Raan, Sleeping beauties in science, Scientometrics 59 (3) (2004) 467–472.
[129] Q. Ke, E. Ferrara, F. Radicchi, A. Flammini, Defining and identifying sleeping beauties in science, Proc. Natl. Acad. Sci. USA 112 (24) (2015) 7426–7431.
[130] M. Golosovsky, S. Solomon, Stochastic dynamical model of a growing citation network based on a self-exciting point process, Phys. Rev. Lett. 109 (2012) 098701.
[131] G.J. Peterson, S. Pressé, K.A. Dill, Nonuniversal power law scaling in the probability distribution of scientific citations, Proc. Natl. Acad. Sci. USA 107 (37) (2010) 16023–16027.
[132] A.M. Petersen, S. Fortunato, R.K. Pan, K. Kaski, O. Penner, A. Rungi, M. Riccaboni, H.E. Stanley, F. Pammolli, Reputation and impact in academic careers, Proc. Natl. Acad. Sci. USA 111 (43) (2014) 15316–15321.
[133] S. Wuchty, B.F. Jones, B. Uzzi, The increasing dominance of teams in production of knowledge, Science 316 (5827) (2007) 1036–1039.
[134] M.E.J. Newman, Clustering and preferential attachment in growing networks, Phys. Rev. E 64 (2001) 025102.
[135] A. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, T. Vicsek, Evolution of the social network of scientific collaborations, Physica A 311 (3) (2002) 590–614.
[136] T. Zhou, B.-H. Wang, Y.-D. Jin, D.-R. He, P.-P. Zhang, Y. He, B.-B. Su, K. Chen, Z.-Z. Zhang, J.-G. Liu, Modelling collaboration networks based on nonlinear preferential attachment, Internat. J. Modern Phys. C 18 (02) (2007) 297–314.
[137] M. Li, J. Wu, D. Wang, T. Zhou, Z. Di, Y. Fan, Evolving model of weighted networks inspired by scientific collaboration networks, Physica A 375 (1) (2007) 355–364.
[138] P. Zhang, J. Wang, X. Li, M. Li, Z. Di, Y. Fan, Clustering coefficient and community structure of bipartite networks, Physica A 387 (27) (2008) 6869–6875.
[139] J.J. Ramasco, S.N. Dorogovtsev, R. Pastor-Satorras, Self-organization of collaboration networks, Phys. Rev. E 70 (2004) 036106.
[140] K. Börner, J.T. Maru, R.L. Goldstone, The simultaneous evolution of author and paper networks, Proc. Natl. Acad. Sci. USA 101 (Suppl. 1) (2004) 5266–5273.
[141] M. Peltomäki, M. Alava, Correlations in bipartite collaboration networks, J. Stat. Mech. Theory Exp. 2006 (01) (2006) P01010.
[142] M.C.V. Medo, G. Cimini, Model-based evaluation of scientific impact indicators, Phys. Rev. E 94 (3) (2016) 032312.
[143] B.F. Jones, S. Wuchty, B. Uzzi, Multi-university research teams: Shifting impact, geography, and stratification in science, Science 322 (5905) (2008) 1259–1262.
[144] A. Gazni, C.R. Sugimoto, F. Didegah, Mapping world scientific collaboration: Authors, institutions, and countries, J. Am. Soc. Inf. Sci. Tec. 63 (2) (2012) 323–335.
[145] M. Coccia, L. Wang, Evolution and convergence of the patterns of international scientific collaboration, Proc. Natl. Acad. Sci. USA 113 (8) (2016) 2057–2061.
[146] D. Hsiehchen, M. Espinoza, A. Hsieh, Multinational teams and diseconomies of scale in collaborative research, Sci. Adv. 1 (8) (2015) e1500211.
[147] S. Milojević, Principles of scientific research team formation and evolution, Proc. Natl. Acad. Sci. USA 111 (11) (2014) 3984–3989.
[148] R. Guimera, B. Uzzi, J. Spiro, L. Amaral, Team assembly mechanisms determine collaboration network structure and team performance, Science 308 (5722) (2005) 697–702.
[149] A.-L. Barabási, C. Song, D. Wang, Publishing: Handful of papers dominates citation, Nature 491 (7422) (2012) 40–40.
[150] T. Wei, M. Li, C. Wu, X. Yan, Y. Fan, Z. Di, J. Wu, Do scientists trace hot topics? Sci. Rep. 3 (2013) 2207.

[151] M. Li, L. Yang, H. Zhang, Z. Shen, C. Wu, J. Wu, Do mathematicians, economists and biomedical scientists trace large topics more strongly than physicists? J. Informetr. 11 (2) (2017) 598–607.
[152] R. Pan, S. Sinha, K. Kaski, J. Saramäki, The evolution of interdisciplinarity in physics research, Sci. Rep. 2 (2011) 551.
[153] M. Perc, Self-organization of progress across the century of physics, Sci. Rep. 3 (2013) 1720.
[154] M. Herrera, D.C. Roberts, N. Gulbahce, Mapping the evolution of scientific fields, PLoS One 5 (5) (2010) e10355.
[155] A.H. Shirazi, A. Badie Modiri, S. Heydari, J.L. Rohn, G.R. Jafari, A.R. Mani, Evolution of communities in the medical sciences: Evidence from the medical words network, PLoS One 11 (12) (2016) e0167546.
[156] X. Sun, K. Ding, Y. Lin, Mapping the evolution of scientific fields based on cross-field authors, J. Informetr. 10 (3) (2016) 750–761.
[157] D. Chavalarias, J.-P. Cointet, Phylomemetic patterns in science evolutionthe rise and fall of scientific fields, PLoS One 8 (2) (2013) e54847.
[158] S.P. Jr, R.S. Mendes, L.C. Malacarne, E.K. Lenzi, Scaling behavior in the dynamics of citations to scientific journals, Europhys. Lett. 75 (4) (2006) 673.
[159] M. Stanley, L. Amaral, S. Buldyrev, S. Havlin, H. Leschhorn, P. Maass, M. Salinger, H. Stanley, Scaling behaviour in the growth of companies, Nature 379 (6568) (1996) 804.
[160] O. Mryglod, R. Kenna, Y. Holovatch, Is your EPL attractive? Classification of publications through download statistics, Europhys. Lett. 108 (108) (2014) 50011.
[161] T. Kuhn, M. Perc, D. Helbing, Inheritance patterns in citation networks reveal scientific memes, Phys. Rev. X 4 (4) (2014) 041036.
[162] A. Chatterjee, A. Ghosh, B.K. Chakrabarti, Universality of citation distributions for academic institutions and journals, PLoS One 11 (1) (2016) e0146762.
[163] F. Radicchi, C. Castellano, Rescaling citations of publications in physics, Phys. Rev. E 83 (4) (2011) 046116.
[164] C. Castellano, F. Radicchi, On the fairness of using relative indicators for comparing citation performance in different disciplines, Arch. Immunol. Ther. Exp. 57 (2) (2009) 85–90.
[165] B.I. Hutchins, X. Yuan, J.M. Anderson, G.M. Santangelo, Relative Citation Ratio (RCR): A new metric that uses citation rates to measure influence at the article level, PLoS Biol. 14 (9) (2016) e1002541.
[166] M.E.J. Newman, Prediction of highly cited papers, Europhys. Lett. 105 (2) (2014) 28002.
[167] P. Stephan, R. Veugelers, J. Wang, Reviewers are blinkered by bibliometrics, Nature 544 (7651) (2017) 411.
[168] F. Radicchi, In science there is no bad publicity: Papers criticized in comments have high scientific impact, Sci. Rep. 2 (2012) 815.
[169] J.P. Ioannidis, A generalized view of self-citation: Direct, co-author, collaborative, and coercive induced self-citation, J. Psychosom. Res. 78 (1) (2015) 7–11.
[170] X. Zhu, P. Turney, D. Lemire, A. Vellino, Measuring academic influence: Not all citations are equal, J. Assoc. Inf. Sci. Technol. 66 (2) (2015) 408–427.
[171] M. Valenzuela, V. Ha, O. Etzioni, Identifying meaningful citations, in: AAAI Workshop: Scholarly Big Data, 2015.
[172] O. Etzioni, Artificial intelligence: Ai zooms in on highly influential citations, Nature 547 (7661) (2017) 32.
[173] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, Comput. Netw. ISDN Syst. 30 (1–7) (1998) 107–117.
[174] P. Chen, H. Xie, S. Maslov, S. Redner, Finding scientific gems with googles pagerank algorithm, J. Informetr. 1 (1) (2007) 8–15.
[175] N. Ma, J. Guan, Y. Zhao, Bringing pagerank to the citation analysis, Inf. Process. Manage. 44 (2) (2008) 800–810.
[176] D.F. Gleich, PageRank beyond the web, SIAM Rev. 57 (3) (2015) 321–363.
[177] L. Ermann, K.M. Frahm, D.L. Shepelyansky, Google matrix analysis of directed networks, Rev. Modern Phys. 87 (4) (2016) 1261–1310.
[178] D. Walker, H. Xie, K.-K. Yan, S. Maslov, Ranking scientific publications using a model of network traffic, J. Stat. Mech. Theory Exp. 2007 (06) (2007) P06010.
[179] M.S. Mariani, M. Medo, Y.-C. Zhang, Identification of milestone papers through time-balanced network centrality, J. Informetr. 10 (4) (2016) 1207–1223.
[180] Q. Mei, J. Guo, D. Radev, Divrank: the interplay of prestige and diversity in information networks, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2010, pp. 1009–1018.
[181] C. Su, Y. Pan, Y. Zhen, Z. Ma, J. Yuan, H. Guo, Z. Yu, C. Ma, Y. Wu, PrestigeRank: A new evaluation method for papers and journals, J. Informetr. 5 (1) (2011) 1–13.
[182] L. Yao, T. Wei, A. Zeng, Y. Fan, Z. Di, Ranking scientific publications: the effect of nonlinearity, Sci. Rep. 4 (2014) 6663.
[183] J. Zhou, A. Zeng, Y. Fan, Z. Di, Ranking scientific publications with similarity-preferential mechanism, Scientometrics 106 (2) (2016) 805–816.
[184] G. Salton, M.J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, Inc., New York, NY, USA, 1986.
[185] H. Jeong, Z. Néda, A.-L. Barabási, Measuring preferential attachment in evolving networks, Europhys. Lett. 61 (4) (2003) 567.
[186] J. Wang, Y. Mei, D. Hicks, Comment on quantifying long-term scientific impacti, Science 345 (6193) (2014) 149.
[187] L. Waltman, A review of the literature on citation impact indicators, J. Informetr. 10 (2) (2016) 365–391.
[188] S. Redner, On the meaning of the h-index, J. Stat. Mech. Theory Exp. 2010 (03) (2010) L03005.
[189] L. Lü, T. Zhou, Q.-M. Zhang, H.E. Stanley, The h-index of a network node and its relation to degree and coreness, Nat. Commun. 7 (2016) 10168.
[190] S. Alonso, F.J. Cabrerizo, E. Herrera-Viedma, F. Herrera, h-Index: A review focused in its variants, computation and standardization for different scientific fields, J. Informetr. 3 (4) (2009) 273–289.
[191] L. Wildgaard, J.W. Schneider, B. Larsen, A review of the characteristics of 108 author-level bibliometric indicators, Scientometrics 101 (1) (2014) 125–158.
[192] L. Egghe, Theory and practise of the g-index, Scientometrics 69 (1) (2006) 131–152.
[193] S. Alonso, F.J. Cabrerizo, E. Herrera-Viedma, F. Herrera, hg-index: A new index to characterize the scientific output of researchers based on the h-and g-indices, Scientometrics 82 (2) (2010) 391–400.
[194] B. Jin, L. Liang, R. Rousseau, L. Egghe, The R-and AR-indices: Complementing the h-index, Chin. Sci. Bull. 52 (6) (2007) 855–863.
[195] S. Dorogovtsev, J. Mendes, Ranking scientists, Nat. Phys. 11 (11) (2015) 882–883.
[196] L. Egghe, R. Rousseau, An h-index weighted by citation impact, Inf. Process. Manage. 44 (2) (2008) 770–780.
[197] J. Smart, A. Bayer, Author collaboration and impact: A note on citation rates of single and multiple authored articles, Scientometrics 10 (5–6) (1986) 297–305.
[198] P.D. Batista, M.G. Campiteli, O. Kinouchi, Is it possible to compare researchers with different scientific interests? Scientometrics 68 (1) (2006) 179–189.
[199] E. Yan, Y. Ding, Applying centrality measures to impact analysis: A coauthorship network analysis, J. Am. Soc. Inf. Sci. Tec. 60 (10) (2009) 2107–2118.
[200] E. Otte, R. Rousseau, Social network analysis: a powerful strategy, also for the information sciences, J. Inf. Sci. 28 (6) (2002) 441–453.
[201] R. Guns, Y.X. Liu, D. Mahbuba, Q-measures and betweenness centrality in a collaboration network: a case study of the field of informetrics, Scientometrics 87 (1) (2011) 133–147.
[202] A. Abbasi, L. Hossain, L. Leydesdorff, Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks, J. Informetr. 6 (3) (2012) 403–412.
[203] J. Bar-Ilan, Informetrics at the beginning of the 21st century a review, J. Informetr. 2 (1) (2008) 1–52.
[204] X. Liu, J. Bollen, M.L. Nelson, H. Van de Sompel, Co-authorship networks in the digital library research community, Inf. Process. Manage. 41 (6) (2005) 1462–1480.
[205] E. Yan, Y. Ding, Discovering author impact: A PageRank perspective, Inf. Process. Manage. 47 (1) (2011) 125–134.
[206] Y. Ding, E. Yan, A. Frazho, J. Caverlee, PageRank for ranking authors in co-citation networks, J. Am. Soc. Inf. Sci. Tec. 60 (11) (2009) 2229–2243.

[207] Y. Ding, Applying weighted pagerank to author citation networks, J. Am. Soc. Inf. Sci. Tec. 62 (2) (2011) 236–245.
[208] Y. Ding, B. Cronin, Popular and/or prestigious? Measures of scholarly esteem, Inf. Process. Manage. 47 (1) (2011) 80–96.
[209] D. Fiala, F. Rousselot, K. Ježek, PageRank for bibliographic networks, Scientometrics 76 (1) (2008) 135–158.
[210] D. Fiala, Time-aware pagerank for bibliographic networks, J. Informetr. 6 (3) (2012) 370–388.
[211] M. Nykl, K. Ježek, D. Fiala, M. Dostal, PageRank variants in the evaluation of citation networks, J. Informetr. 8 (3) (2014) 683–692.
[212] M. Nykl, M. Campr, K. Ježek, Author ranking based on personalized pagerank, J. Informetr. 9 (4) (2015) 777–799.
[213] H. Wang, H.-W. Shen, X.-Q. Cheng, Scientific credit diffusion: Researcher level or paper level? Scientometrics 109 (2) (2016) 827–837.
[214] G. Van Hooydonk, Fractional counting of multiauthored publications: Consequences for the impact of authors, J. Am. Soc. Inf. Sci. 48 (10) (1997) 944–945.
[215] L. Egghe, R. Rousseau, G. Van Hooydonk, Methods for accrediting publications to authors or countries: Consequences for evaluation studies, J. Am. Soc. Inf. Sci. 51 (2) (2000) 145–157.
[216] F.J. Trueba, H. Guerrero, A robust formula to credit authors for their publications, Scientometrics 60 (2) (2004) 181–204.
[217] N.T. Hagen, Harmonic allocation of authorship credit: Source-level correction of bibliometric bias assures accurate publication and citation analysis, PLoS One 3 (12) (2008) e4021.
[218] J. Kim, J. Diesner, A network-based approach to coauthorship credit allocation, Scientometrics 101 (1) (2014) 587–602.
[219] J. Kim, J. Kim, Rethinking the comparison of coauthorship credit allocation schemes, J. Informetr. 9 (3) (2015) 667–673.
[220] J. Stallings, E. Vance, J. Yang, M.W. Vannier, J. Liang, L. Pang, L. Dai, I. Ye, G. Wang, Determining scientific impact using a collaboration index, Proc. Natl. Acad. Sci. USA 110 (24) (2013) 9680–9685.
[221] H.-W. Shen, A.-L. Barabási, Collective credit allocation in science, Proc. Natl. Acad. Sci. USA 111 (34) (2014) 12325–12330.
[222] L. Lü, M. Medo, C.H. Yeung, Y.-C. Zhang, Z.-K. Zhang, T. Zhou, Recommender systems, Phys. Rep. 519 (1) (2012) 1–49.
[223] Q. Niu, J. Zhou, A. Zeng, Y. Fan, Z. Di, Which publication is your representative work? J. Informetr. 10 (3) (2016) 842–853.
[224] D.A. Pendlebury, The use and misuse of journal metrics and other citation indicators, Arch. Immunol. Ther. Exp. 57 (1) (2009) 1.
[225] V. Lariviere, V. Kiermer, C.J. MacCallum, M. McNutt, M. Patterson, B. Pulverer, S. Swaminathan, S. Taylor, S. Curry, A simple proposal for the publication of journal citation distributions, Biorxiv (2016) 062109.
[226] C.J. Bradshaw, B.W. Brook, How to rank journals, PLoS One 11 (3) (2016) e0149852.
[227] H.F. Moed, Measuring contextual citation impact of scientific journals, J. Informetr. 4 (3) (2010) 265–277.
[228] E.S. Vieira, J.A. Gomes, The journal relative impact: an indicator for journal assessment, Scientometrics 89 (2) (2011) 631–651.
[229] S. Milojević, F. Radicchi, J. Bar-Ilan, Citation success index- An intuitive pair-wise journal comparison metric, J. Informetr. 11 (1) (2017) 223–231.
[230] L. Leydesdorff, L. Bornmann, Integrated impact indicators compared with impact factors: An alternative research design with policy implications, J. Am. Soc. Inf. Sci. Tec. 62 (11) (2011) 2133–2146.
[231] J. Bollen, M.A. Rodriquez, H. Van de Sompel, Journal status, Scientometrics 69 (3) (2006) 669–687.
[232] C.T. Bergstrom, J.D. West, M.A. Wiseman, The eigenfactor^{TM} metrics, J. Neurosci. 28 (45) (2008) 11433–11434.
[233] D.A. King, The scientific impact of nations, Nature 430 (6997) (2004) 311–316.
[234] R. Fairclough, M. Thelwall, More precise methods for national research citation impact comparisons, J. Informetr. 9 (4) (2015) 895–906.
[235] A. Mazloumian, D. Helbing, S. Lozano, R.P. Light, K. Brner, Global multi-level analysis of the 'scientific food web', Sci. Rep. 3 (2013) 1167.
[236] Q. Zhang, N. Perra, B. Gonçalves, F. Ciulla, A. Vespignani, Characterizing scientific production and consumption in physics, Sci. Rep. 3 (4) (2013) 1640.
[237] J.-F. Molinari, A. Molinari, A new methodology for ranking scientific institutions, Scientometrics 75 (1) (2008) 163–174.
[238] A. Kinney, National scientific facilities and their science impact on nonbiomedical research, Proc. Natl. Acad. Sci. USA 104 (46) (2007) 17943–17947.
[239] A. Clauset, S. Arbesman, D.B. Larremore, Systematic inequality and hierarchy in faculty hiring networks, Sci. Adv. 1 (1) (2015) e1400005.
[240] J.A. Crespo, I. Ortuño-Ortín, J. Ruiz-Castillo, The citation merit of scientific publications, PLoS One 7 (11) (2012) e49156.
[241] Z. Shen, L. Yang, J. Pei, M. Li, C. Wu, J. Bao, T. Wei, Z. Di, R. Rousseau, J. Wu, Interrelations among scientific fields and their relative influences revealed by an input–output analysis, J. Informetr. 10 (1) (2016) 82–97.
[242] H. Sayyadi, L. Getoor, Futurerank: Ranking scientific articles by predicting their future pagerank, in: Proceedings of the 2009 SIAM International Conference on Data Mining, SIAM, 2009, pp. 533–544.
[243] D. Zhou, S.A. Orshanskiy, H. Zha, C.L. Giles, Co-ranking authors and documents in a heterogeneous network, in: Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on, IEEE, 2007, pp. 739–744.
[244] H. Liao, R. Xiao, G. Cimini, M. Medo, Network-Driven reputation in online scientific communities, PLoS One 9 (12) (2014) e112022.
[245] E. Yan, Y. Ding, C.R. Sugimoto, P-rank: An indicator measuring prestige in heterogeneous scholarly networks, J. Am. Soc. Inf. Sci. Tec. 62 (3) (2011) 467–477.
[246] X. Jiang, X. Sun, Z. Yang, Z. Hai, J. Yao, Exploiting heterogeneous scientific literature networks to combat ranking bias: Evidence from the computational linguistics area, J. Assoc. Inf. Sci. Technol. 67 (7) (2016) 1679–1702.
[247] D. Yu, W. Wang, S. Zhang, W. Zhang, R. Liu, A multiple-link, mutually reinforced journal-ranking model to measure the prestige of journals, Scientometrics 111 (1) (2017) 521–542.
[248] J.M. Kleinberg, Authoritative sources in a hyperlinked environment, J. Assoc. Comput. Mach. 46 (5) (1999) 604–632.
[249] A. Halu, R.J. Mondragn, P. Panzarasa, G. Bianconi, Multiplex pagerank, PLoS One 8 (10) (2013) e78293.
[250] J. Iacovacci, G. Bianconi, Extracting information from multiplex networks, Chaos 26 (6) (2016) 065306.
[251] J. Iacovacci, C. Rahmede, A. Arenas, G. Bianconi, Functional multiplex pagerank, Europhys. Lett. 116 (2) (2016) 28004.
[252] A. Clauset, D.B. Larremore, R. Sinatra, Data-driven predictions in the science of science, Science 355 (6324) (2017) 477–480.
[253] A. Clauset, C. Moore, M.E.J. Newman, Hierarchical structure and the prediction of missing links in networks, Nature 453 (7191) (2008) 98.
[254] Y.R. Wang, H. Huang, Review on statistical methods for gene network reconstruction using expression data, J. Theoret. Biol. 362 (2014) 53–61.
[255] T. Hao, W. Peng, Q. Wang, B. Wang, J. Sun, Reconstruction and application of protein–protein interaction network, Int. J. Mol. Sci. 17 (6) (2016) 907.
[256] P. Jaccard, Étude comparative de la distribution florale dans une portion des Alpes et des Jura, Bull. Soc Vaudoise Sci. Nat. 37 (1901) 547–579.
[257] T. Sørensen, A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons, Biol. Skr. 5 (1948) 1–34.
[258] E.A. Leicht, P. Holme, M.E.J. Newman, Vertex similarity in networks, Phys. Rev. E 73 (2) (2006) 026120.
[259] L. Katz, A new status index derived from sociometric analysis, Psychometrika 18 (1) (1953) 39–43.
[260] L. Lü, C.-H. Jin, T. Zhou, Similarity index based on local paths for link prediction of complex networks, Phys. Rev. E 80 (4) (2009) 046122.
[261] Z. Liu, Q.-M. Zhang, L. Lü, T. Zhou, Link prediction in complex networks: A local naïve Bayes model, Europhys. Lett. 96 (4) (2011) 48007.
[262] Y.-X. Zhu, L. Lü, Q.-M. Zhang, T. Zhou, Uncovering missing links with cold ends, Physica A 391 (22) (2012) 5769–5778.
[263] T. Tylenda, R. Angelova, S. Bedathur, Towards time-aware link prediction in evolving social networks, in: Proceedings of the 3rd Workshop on Social Network Mining and Analysis, SNA-KDD '09, ACM, New York, NY, USA, 2009, pp. 9:1–9:10.
[264] Z. Huang, D.K. Lin, The time-series link prediction problem with applications in communication surveillance, INFORMS J. Comput. 21 (2) (2009) 286–303.

[265] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, A.-L. Barabási, Hierarchical organization of modularity in metabolic networks, Science 297 (5586) (2002) 1551–1555.
[266] T. Zhou, L. Lü, Y.-C. Zhang, Predicting missing links via local information, Europhys. J. B 71 (4) (2009) 623–630.
[267] L. Lü, T. Zhou, Link prediction in complex networks: A survey, Physica A 390 (6) (2011) 1150–1170.
[268] V. Ciotti, M. Bonaventura, V. Nicosia, P. Panzarasa, V. Latora, Homophily and missing links in citation networks, EPJ Data Sci. 5 (1) (2016) 7.
[269] Q.-M. Zhang, L. Lü, W.-Q. Wang, T. Zhou, Potential theory for directed networks, PLoS One 8 (2) (2013) e55437.
[270] N. Shibata, Y. Kajikawa, I. Sakata, Link prediction in citation networks, J. Am. Soc. Inf. Sci. Tec. 63 (1) (2012) 78–85.
[271] D. Liben-Nowell, J. Kleinberg, The link prediction problem for social networks, in: Twelfth International Conference on Information and Knowledge Management, 2003, pp. 556–559.
[272] X. Sun, H. Lin, K. Xu, K. Ding, How we collaborate: characterizing, modeling and predicting scientific collaborations, Scientometrics 104 (1) (2015) 43–60.
[273] L. Backstrom, J. Leskovec, Supervised random walks: predicting and recommending links in social networks, in: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, ACM, 2011, pp. 635–644.
[274] Q. Zhang, H. Yu, Computational approaches for predicting biomedical research collaborations, PLoS One 9 (11) (2014) e111795.
[275] B. Perozzi, R. Al-Rfou, S. Skiena, Deepwalk: Online learning of social representations, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2014, pp. 701–710.
[276] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, Q. Mei, Line: Large-scale information network embedding, in: Proceedings of the 24th International Conference on World Wide Web, ACM, 2015, pp. 1067–1077.
[277] K.-K. Kleineberg, M. Boguñá, M.Á. Serrano, F. Papadopoulos, Hidden geometric correlations in real multiplex networks, Nat. Phys. 12 (2016) 1076–1081.
[278] A. Allard, M.Á. Serrano, G. García-Pérez, M. Boguñá, The geometric nature of weights in real complex networks, Nat. Commun. 8 (2017) 14103.
[279] M. Kitsak, F. Papadopoulos, D. Krioukov, Latent geometry of bipartite networks, Phys. Rev. E 95 (3) (2017) 032309.
[280] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Adv. Neural Inf. Process. Syst. 26 (2013) 3111–3119.
[281] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, 2013. arXiv:1301.3781.
[282] A. Grover, J. Leskovec, Node2vec: Scalable feature learning for networks, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 855–864.
[283] A. Zeng, C.H. Yeung, Predicting the future trend of popularity by network diffusion, Chaos 26 (6) (2016) 063102.
[284] V. Mahajan, E. Muller, F.M. Bass, New product diffusion models in marketing: A review and directions for research, J. Market. 54 (1) (1990) 1–26.
[285] F.M. Bass, Comments on a new product growth for model consumer durables the bass modeli, Manage. Sci. 50 (Suppl. 12) (2004) 1833–1840.
[286] B. Gompertz, On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies, Philos. Trans. R. Soc. Lond. 115 (1825) 513–583.
[287] X. Cao, Y. Chen, K.R. Liu, A data analytic approach to quantifying scientific impact, J. Informetr. 10 (2) (2016) 471–484.
[288] M. Wang, G. Yu, D. Yu, Mining typical features for highly cited papers, Scientometrics 87 (3) (2011) 695–706.
[289] M. Wang, G. Yu, S. An, D. Yu, Discovery of factors influencing citation impact based on a soft fuzzy rough set model, Scientometrics 93 (3) (2012) 635–644.
[290] M. Wang, G. Yu, J. Xu, H. He, D. Yu, S. An, Development a case-based classifier for predicting highly cited papers, J. Informetr. 6 (4) (2012) 586–599.
[291] E. Sarigöl, R. Pfitzner, I. Scholtes, A. Garas, F. Schweitzer, Predicting scientific success based on coauthorship networks, EPJ Data Sci. 3 (1) (2014) 9.
[292] D.G. Brizan, K. Gallagher, A. Jahangir, T. Brown, Predicting citation patterns: defining and determining influence, Scientometrics 108 (1) (2016) 183–200.
[293] T. Yu, G. Yu, P.-Y. Li, L. Wang, Citation impact prediction for scientific papers using stepwise regression analysis, Scientometrics 101 (2) (2014) 1233–1252.
[294] K. McKeown, H. Daume, S. Chaturvedi, J. Paparrizos, K. Thadani, P. Barrio, O. Biran, S. Bothe, M. Collins, K.R. Fleischmann, L. Gravano, R. Jha, B. King, K. McInerney, T. Moon, A. Neelakantan, D. O'Seaghdha, D. Radev, C. Templeton, S. Teufel, Predicting the impact of scientific concepts using full-text features, J. Assoc. Inf. Sci. Technol. 67 (11) (2016) 2684–2696.
[295] J.E. Hirsch, Does the h index have predictive power? Proc. Natl. Acad. Sci. USA 104 (49) (2007) 19193–19198.
[296] M. Schreiber, How relevant is the predictive power of the h-index? A case study of the time-dependent Hirsch index, J. Informetr. 7 (2) (2013) 325–329.
[297] D.E. Acuna, S. Allesina, K.P. Kording, Future impact: Predicting scientific success, Nature 489 (7415) (2012) 201–202.
[298] M.A. García-Pérez, Limited validity of equations to predict the future h index, Scientometrics 96 (3) (2013) 901–909.
[299] C. Mccarty, J.W. Jawitz, A. Hopkins, A. Goldman, Predicting author h-index using characteristics of the co-author network, Scientometrics 96 (2) (2013) 467–483.
[300] O. Penner, R.K. Pan, A.M. Petersen, K. Kaski, S. Fortunato, On the predictability of future impact in science, Sci. Rep. 3 (10) (2013) 3052.
[301] O. Penner, A.M. Petersen, R.K. Pan, S. Fortunato, Commentary: The case for caution in predicting scientists? Future impact, Phys. Today 66 (66) (2013) 8–9.
[302] A. Mazloumian, Predicting scholars' scientific impact, PLoS One 7 (11) (2012) e49246.
[303] P. Dorta-González, M.I. Dorta-González, R. Suárez-Vega, An approach to the author citation potential: Measures of scientific performance which are invariant across scientific fields, Scientometrics 102 (2) (2015) 1467–1496.
[304] C. Stegehuis, N. Litvak, L. Waltman, Predicting the long-term citation impact of recent publications, J. Informetr. 9 (3) (2015) 642–657.
[305] M. Medo, M.S. Mariani, A. Zeng, Y.-C. Zhang, Identification and impact of discoverers in online social systems, Sci. Rep. 6 (2016) 34218.
[306] A.M. Petersen, M. Riccaboni, H.E. Stanley, F. Pammolli, Persistence and uncertainty in the academic career, Proc. Natl. Acad. Sci. USA 109 (14) (2012) 5213–5218.
[307] M. Qi, A. Zeng, M. Li, Y. Fan, Z. Di, Standing on the shoulders of giants: the effect of outstanding scientists on young collaborators careers, Scientometrics (2017) 1–12.
[308] T. Amjad, Y. Ding, J. Xu, C. Zhang, A. Daud, J. Tang, M. Song, Standing on the shoulders of giants, J. Informetr. 11 (1) (2017) 307–323.
[309] W.F. Laurance, D.C. Useche, S.G. Laurance, C.J. Bradshaw, Predicting publication success for biologists, Bioscience 63 (10) (2013) 817–823.
[310] N. Choudhury, S. Uddin, Time-aware link prediction to explore network effects on temporal knowledge evolution, Scientometrics 108 (2) (2016) 745–776.
[311] X. Sun, J. Kaur, S. Milojević, A. Flammini, F. Menczer, Social dynamics of science, Sci. Rep. 3 (1069) (2013) 1069.
[312] Y. Hu, S. Havlin, H.A. Makse, Conditions for viral influence spreading through multiplex correlated social networks, Phys. Rev. X 4 (2) (2014) 021031.
[313] R. Van Noorden, Interdisciplinary research by the numbers, Nature 525 (7569) (2015) 306–307.
[314] A. Stirling, A general framework for analysing diversity in science, technology and society, J. R. Soc. Interface 4 (15) (2007) 707–719.
[315] V.B. Mansilla, I. Feller, H. Gardner, Quality assessment in interdisciplinary research and education, Res. Eval. 15 (1) (2006) 69–74.

[316] A. Porter, I. Rafols, Is science becoming more interdisciplinary? Measuring and mapping six research fields over time, Scientometrics 81 (3) (2009) 719–745.
[317] J. Wang, B. Thijs, W. Glanzel, Interdisciplinarity and impact: Distinct effects of variety, balance, and disparity, PLoS One 10 (5) (2015) e0127298.
[318] L. Leydesdorff, Mapping interdisciplinarity at the interfaces between the science citation index and the social science citation index, Scientometrics 71 (3) (2007) 391–405.
[319] H. Lee, Uncovering the multidisciplinary nature of technology management: journal citation network analysis, Scientometrics 102 (1) (2015) 51–75.
[320] L. Leydesdorff, I. Rafols, Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations, J. Informetr. 5 (1) (2011) 87–100.
[321] T.W. Steele, J.C. Stier, The impact of interdisciplinary research in the environmental sciences: a forestry case study, J. Am. Soc. Inf. Sci. 51 (5) (2000) 476–484.
[322] E. Rinia, T.N. Van Leeuwen, H. Van Vuren, A. Van Raan, Influence of interdisciplinarity on peer-review and bibliometric evaluations in physics research, Res. Policy 30 (3) (2001) 357–361.
[323] J. Adams, L. Jackson, S. Marshall, Report to the higher education funding council for England bibliometric analysis of interdisciplinary research, 2007.
[324] J.M. Levitt, M. Thelwall, Is multidisciplinary research more highly cited? A macrolevel study, J. Am. Soc. Inf. Sci. Tec. 59 (12) (2008) 1973–1984.
[325] V. Larivière, Y. Gingras, On the relationship between interdisciplinarity and scientific impact, J. Am. Soc. Inf. Sci. Tec. 61 (1) (2010) 126–131.
[326] B. Uzzi, S. Mukherjee, M. Stringer, B. Jones, Atypical combinations and scientific impact, Science 342 (6157) (2013) 468–472.
[327] V. Lariviere, S. Haustein, K. Boerner, Long-distance interdisciplinarity leads to higher scientific impact, PLoS One 10 (3) (2015) e0122565.
[328] A. Yegros-Yegros, I. Rafols, P. DEste, Does interdisciplinary research lead to higher citation impact? The different effect of proximal and distal interdisciplinarity, PLoS One 10 (8) (2015) e0135095.
[329] A. Letchford, H.S. Moat, T. Preis, The advantage of short paper titles, R. Soc. Open Sci. 2 (8) (2015) 150266.
[330] A. Letchford, T. Preis, H.S. Moat, The advantage of simple paper abstracts, J. Informetr. 10 (1) (2016) 1–8.
[331] T.W. Fawcett, A.D. Higginson, Heavy use of equations impedes communication among biologists, Proc. Natl. Acad. Sci. USA 109 (29) (2012) 11735–11739.
[332] A.D. Fernandes, No evidence that equations cause impeded communication among biologists, Proc. Natl. Acad. Sci. USA 109 (45) (2012) E3057.
[333] J.E. Kollmer, T. Pöschel, J.A. Gallas, Are physicists afraid of mathematics, New J. Phys. 17 (1) (2015) 013036.
[334] J. Gibbons, Do not throw equations out with the theory bathwater, Proc. Natl. Acad. Sci. USA 109 (45) (2012) E3054.
[335] K. Siler, K. Lee, L. Bero, Measuring the effectiveness of scientific gatekeeping, Proc. Natl. Acad. Sci. USA 112 (2) (2015) 360–365.
[336] M. Pautasso, H. Schäfer, Peer review delay and selectivity in ecology journals, Scientometrics 84 (2) (2010) 307–315.
[337] S. Shen, R. Rousseau, D. Wang, D. Zhu, H. Liu, R. Liu, Editorial delay and its relation to subsequent citations: the journals nature, science and cell, Scientometrics 105 (3) (2015) 1867–1873.
[338] Z. Lin, S. Hou, J. Wu, The correlation between editorial delay and the ratio of highly cited papers in nature, science and physical review letters, Scientometrics 107 (3) (2016) 1457–1464.
[339] V. Calcagno, E. Demoinet, K. Gollner, L. Guidi, D. Ruths, M.C. De, Flows of research manuscripts among scientific journals reveal hidden submission patterns, Science 338 (6110) (2012) 1065–1069.
[340] F. Didegah, M. Thelwall, Which factors help authors produce the highest impact research? Collaboration, journal and document properties, J. Informetr. 7 (4) (2013) 861–873.
[341] P.E. Smaldino, R. Mcelreath, The natural selection of bad science, R. Soc. Open Sci. 3 (9) (2016) 160384.
[342] C. Catalini, N. Lacetera, A. Oettl, The incidence and role of negative citations in science, Proc. Natl. Acad. Sci. USA 112 (45) (2015) 13823–13826.
[343] J.A. Davis, Clustering and structural balance in graphs, Human Relations 20 (2) (1967) 181–187.
[344] C.A. Kochan, J.M. Budd, The persistence of fraud in the literature: The darsee case, J. Assoc. Inf. Sci. Technol. 43 (7) (1992) 488–493.
[345] M.P. Pfeifer, G.L. Snodgrass, The continued use of retracted, invalid scientific literature, JAMA 263 (10) (1990) 1420–1423.
[346] J.M. Budd, M. Sievert, T.R. Schultz, Phenomena of retraction: reasons for retraction and citations to the publications, JAMA 280 (3) (1998) 296–297.
[347] J.M. Campanario, Fraud: retracted articles are still being cited, Nature 408 (6810) (2000) 288.
[348] F.C. Fang, R.G. Steen, A. Casadevall, Misconduct accounts for the majority of retracted scientific publications, Proc. Natl. Acad. Sci. USA 109 (42) (2012) 17028–17033.
[349] S.F. Lu, G.Z. Jin, B. Uzzi, B. Jones, The retraction penalty: Evidence from the web of science, Sci. Rep. 3 (2013) 3146.
[350] S. Kocabas, Elements of scientific creativity, in: Working Notes of the AAAI Spring Symposium on Artificial Intelligence and Creativity, 1993, pp. 39–45.
[351] H. Stumpf, Scientific creativity: A short overview, Educ. Psychol. Rev. 7 (3) (1995) 225–241.
[352] R.S. Mansfield, T.V. Busse, The Psychology of Creativity and Discovery: Scientists and their Work, Nelson-Hall, 1981.
[353] T.V. Busse, R.S. Mansfield, Theories of the creative process: A review and a perspective, J. Creat. Behav. 14 (2) (1980) 91–132.
[354] E.P. Torrance, Guiding Creative Talent, Prentice-Hall, 1962.
[355] C.W. Taylor, A high-tech high-touch concept of creativitywith its complexity made simple for wide adaptability, Front. Creat. Res.: Beyond Basics (1987) 131–155.
[356] V. Larivière, E. Vignola-Gagné, C. Villeneuve, P. Gélinas, Y. Gingras, Sex differences in research funding, productivity and impact: an analysis of Québec university professors, Scientometrics 87 (3) (2011) 483–498.
[357] J.R. Pohlhaus, H. Jiang, R.M. Wagner, W.T. Schaffer, V.W. Pinn, Sex differences in application, success, and funding rates for NIH extramural programs, Acad. Med. 86 (6) (2011) 759–767.
[358] B.F. Jones, B.A. Weinberg, Age dynamics in scientific creativity, Proc. Natl. Acad. Sci. USA 108 (47) (2011) 18910–18914.
[359] A.M. Petersen, W.-S. Jung, J.-S. Yang, H.E. Stanley, Quantitative and empirical demonstration of the Matthew effect in a study of career longevity, Proc. Natl. Acad. Sci. USA 108 (1) (2011) 18–23.
[360] A.M. Petersen, O. Penner, Inequality and cumulative advantage in science careers: a case study of high-impact journals, EPJ Data Sci. 3 (1) (2014) 24.
[361] A. Mazloumian, Y.H. Eom, D. Helbing, S. Lozano, S. Fortunato, How citation boosts promote scientific paradigm shifts and nobel prizes, PLoS One 6 (5) (2011) e18975.
[362] D. Contandriopoulos, A. Duhoux, C. Larouche, M. Perroux, The impact of a researcher's structural position on scientific performance: An empirical analysis, PLoS One 11 (8) (2016) e0161281.
[363] R.H. Heiberger, O.J. Wieczorek, Choosing Collaboration partners. How scientific success in physics depends on network positions, 2016. arXiv:1608.03251.
[364] S. Servia-Rodríguez, A. Noulas, C. Mascolo, A. Fernández-Vilas, R.P. Díaz-Redondo, The evolution of your success lies at the centre of your co-authorship network, PLoS One 10 (3) (2015) e0114302.
[365] A. Ebadi, A. Schiffauerova, How to become an important player in scientific collaboration networks? J. Informetr. 9 (4) (2015) 809–825.
[366] A. Ma, R.J. Mondragn, V. Latora, Anatomy of funded research in science, Proc. Natl. Acad. Sci. USA 112 (48) (2015) 14760–14765.
[367] G. Parisi, Governments: Balance research funds across Europe, Nature 530 (7588) (2016) 33.
[368] M. De Domenico, A. Arenas, EU cash goes to the sticky and attractive, Nature 531 (7596) (2016) 580.

[369] M. Szell, R. Sinatra, Research funding goes to rich clubs, Proc. Natl. Acad. Sci. USA 112 (48) (2015) 14749–14750.
[370] D.L. Murray, D. Morris, C. Lavoie, P.R. Leavitt, H. MacIsaac, M.E. Masson, M.-A. Villard, Bias in research grant evaluation has dire consequences for small universities, PLoS One 11 (6) (2016) e0155876.
[371] C. Lyall, A. Bruce, W. Marsden, L. Meagher, The role of funding agencies in creating interdisciplinary knowledge, Sci. Publ. Policy 40 (1) (2013) 62–71.
[372] R. Rylance, Global funders to focus on interdisciplinarity, Nature 525 (7569) (2015) 313–315.
[373] S. Jiang, Q. Gao, H. Chen, M.C. Roco, The roles of sharing, transfer, and public funding in nanotechnology knowledge-diffusion networks, J. Assoc. Inf. Sci. Technol. 66 (5) (2015) 1017–1029.
[374] Y. Huang, Y. Zhang, J. Youtie, A.L. Porter, X. Wang, How does national scientific funding support emerging interdisciplinary research: A comparison study of big data research in the US and China, PLoS One 11 (5) (2016) e0154509.
[375] L. Bromham, R. Dinnage, X. Hua, Interdisciplinary research has consistently lower funding success, Nature 534 (7609) (2016) 684–687.
[376] S.X. Zhao, S. Yu, A.M. Tan, X. Xu, H. Yu, Global pattern of science funding in economics, Scientometrics 109 (1) (2016) 463–479.
[377] N. Danthi, C.O. Wu, P. Shi, M. Lauer, Percentile ranking and citation impact of a large cohort of national heart, lung, and blood institute–funded cardiovascular r01 grants, Circ. Res. 114 (4) (2014) 600–606.
[378] D. Li, L. Agha, Big names or big ideas: Do peer-review panels select the best science proposals? Science 348 (6233) (2015) 434–438.
[379] F.C. Fang, A. Bowen, A. Casadevall, NIH peer review percentile scores are poorly predictive of grant productivity, Elife 5 (2016) e13323.
[380] S.A. Gallo, J.H. Sullivan, S.R. Glisson, The influence of peer reviewer expertise on the evaluation of research funding applications, PLoS One 11 (10) (2016) e0165147.
[381] A. Ebadi, A. Schiffauerova, How to receive more funding for your research? Get connected to the right people!, PLoS One 10 (7) (2015) e0133061.
[382] A. Zinilli, Competitive project funding and dynamic complex networks: evidence from Projects of National Interest (PRIN), Scientometrics 108 (2) (2016) 633–652.
[383] S. Nicotri, E. Tinelli, N. Amoroso, E. Garuccio, R. Bellotti, Complex networks and public funding: the case of the 2007-2013 Italian program, EPJ Data Sci. 4 (1) (2015) 8.
[384] A.M. Tan, S.X. Zhao, F.Y. Ye, Characterizing the funded scientific collaboration network, Current Sci. 103 (11) (2012) 1261–1262.
[385] J.-M. Fortin, D.J. Currie, Big science vs. little science: how scientific impact scales with funding, PLoS One 8 (6) (2013) e65263.
[386] P. Stephan, R. Veugelers, J. Wang, Blinkered by bibliometrics, Nature 544 (2017) 411–412.
[387] J. Wang, R. Veugelers, P. Stephan, Bias Against Novelty in Science: A Cautionary Tale for Users of Bibliometric Indicators, Working Paper Series 22180, National Bureau of Economic Research, 2016.
[388] S. Mukherjee, D.M. Romero, B. Jones, B. Uzzi, The nearly universal link between the age of past knowledge and tomorrows breakthroughs in science and technology: The hotspot, Sci. Adv. 3 (4) (2017) e1601315.
[389] M. Cokol, I. Iossifov, C. Weinreb, A. Rzhetsky, Emergent behavior of growing knowledge about molecular interactions, Nature Biotechnol. 23 (10) (2005) 1243–1247.
[390] E. Beam, L.G. Appelbaum, J. Jack, J. Moody, S.A. Huettel, Mapping the semantic structure of cognitive neuroscience, J. Cogn. Neurosci. 26 (9) (2014) 1949–1965.
[391] J.G. Foster, A. Rzhetsky, J.A. Evans, Tradition and innovation in scientists' research strategies, Am. Sociol. Rev. 80 (5) (2015) 875–908.
[392] A. Rzhetsky, J.G. Foster, I.T. Foster, J.A. Evans, Choosing experiments to accelerate collective discovery, Proc. Natl. Acad. Sci. USA 112 (47) (2015) 14569–14574.
[393] F. Shi, J.G. Foster, J.A. Evans, Weaving the fabric of science: Dynamic network models of science's unfolding structure, Social Networks 43 (2015) 73–85.
[394] Y. Peng, G. Kou, Y. Shi, Z. Chen, A descriptive framework for the field of data mining and knowledge discovery, Int. J. Inf. Technol. Dec. Making 7 (04) (2008) 639–682.
[395] J.G. March, Exploration and exploitation in organizational learning, Organ. Sci. 2 (1) (1991) 71–87.
[396] A.K. Gupta, K.G. Smith, C.E. Shalley, The interplay between exploration and exploitation, Acad. Manage. J. 49 (4) (2006) 693–706.
[397] J.H. Holland, Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence, MIT press, 1992.
[398] R. Radner, M. Rothschild, On the allocation of effort, J. Econ. Theory 10 (3) (1975) 358–376.
[399] M.T. Hannan, J. Freeman, Organizational Ecology, Harvard University Press, 1993.
[400] K. Mehlhorn, B.R. Newell, P.M. Todd, M.D. Lee, K. Morgan, V.A. Braithwaite, D. Hausmann, K. Fiedler, C. Gonzalez, Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures, Decision 2 (3) (2015) 191–215.
[401] D. Levinthal, Adaptation on rugged landscapes, Manage. Sci. 43 (7) (1997) 934–950.
[402] H. Youn, L.M.A. Bettencourt, D. Strumsky, J. Lobo, Invention as a combinatorial process: evidence from US patents, J. R. Soc. Interface 12 (106) (2015) 20150272.
[403] V. Sood, M. Mathieu, A. Shreim, P. Grassberger, M. Paczuski, Interacting branching process as a simple model of innovation, Phys. Rev. Lett. 105 (17) (2010) 178701.
[404] R. Solé, D.R. Amor, S. Valverde, On singularities and black holes in combination-driven models of technological innovation networks, PLoS One 11 (1) (2016) e0146180.
[405] M. Wooldridge, An Introduction to Multiagent Systems, John Wiley & Sons, 2009.
[406] O. Berger-Tal, J. Nathan, E. Meron, D. Saltz, The exploration-exploitation dilemma: a multidisciplinary framework, PLoS One 9 (4) (2014) e95693.
[407] M.D. König, S. Battiston, M. Napoletano, F. Schweitzer, The efficiency and stability of R&D networks, Games Econom. Behav. 75 (2) (2012) 694–713.
[408] M.V. Tomasello, C.J. Tessone, F. Schweitzer, A model of dynamic rewiring and knowledge exchange in R&D networks, Adv. Complex Syst. 19 (01n02) (2016) 1650004.
[409] C. Fang, J. Lee, M.A. Schilling, Balancing exploration and exploitation through structural design: The isolation of subgroups and organizational learning, Organ. Sci. 21 (3) (2010) 625–642.
[410] T. Kameda, D. Nakanishi, Cost–benefit analysis of social/cultural learning in a nonstationary uncertain environment: An evolutionary simulation and an experiment with human subjects, Evol. Hum. Behav. 23 (5) (2002) 373–393.
[411] A.R. Rogers, Does biology constrain culture? Am. Anthropol. 90 (4) (1988) 819–831.
[412] T. Kameda, D. Nakanishi, Does social/cultural learning increase human adaptability?: Rogers's question revisited, Evol. Hum. Behav. 24 (4) (2003) 242–260.
[413] L. Rendell, R. Boyd, D. Cownden, M. Enquist, K. Eriksson, M.W. Feldman, L. Fogarty, S. Ghirlanda, T. Lillicrap, K.N. Laland, Why copy others? Insights from the social learning strategies tournament, Science 328 (5975) (2010) 208–213.
[414] D. Lazer, A. Friedman, The network structure of exploration and exploitation, Adm. Sci. Q. 52 (4) (2007) 667–694.
[415] W. Mason, D.J. Watts, Collaborative learning in networks, Proc. Natl. Acad. Sci. USA 109 (3) (2012) 764–769.
[416] M. Derex, R. Boyd, Partial connectivity increases cultural accumulation within groups, Proc. Natl. Acad. Sci. USA 113 (11) (2016) 2982–2987.

[417] A. Mesoudi, An experimental simulation of the copy-successful-individualsi cultural learning strategy: adaptive landscapes, producer–scrounger dynamics, and informational access costs, Evol. Hum. Behav. 29 (5) (2008) 350–363.
[418] T.N. Wisdom, X. Song, R.L. Goldstone, Social learning strategies in networked groups, Cogn. Sci. 37 (8) (2013) 1383–1425.
[419] W.A. Mason, A. Jones, R.L. Goldstone, Propagation of innovations in networked groups, J. Exp. Psychol.-Gen. 137 (3) (2008) 422–433.
[420] D. Barkoczi, M. Galesic, Social learning strategies modify the effect of network structure on group performance, Nat. Commun. 7 (2016) 13109.
[421] B. Verspagen, Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research, Adv. Complex Syst. 10 (01) (2012) 93–115.
[422] N.P. Hummon, P. Dereian, Connectivity in a citation network: The development of DNA theory, Social Networks 11 (1) (1989) 39–63.
[423] D. Acemoglu, U. Akcigit, W.R. Kerr, Innovation network, Proc. Natl. Acad. Sci. USA 113 (41) (2016) 11483–11488.
[424] S. Valverde, R.V. Solé, Network motifs in computational graphs: a case study in software architecture, Phys. Rev. E 72 (2) (2005) 026107.
[425] W. Pan, B. Li, Y. Ma, J. Liu, Multi-granularity evolution analysis of software using complex network theory, J. Syst. Sci. Complex. 24 (6) (2011) 1068–1082.
[426] C.R. Myers, Software systems as complex networks: Structure, function, and evolvability of software collaboration graphs, Phys. Rev. E 68 (4) (2003) 046116.
[427] L. Wen, R.G. Dromey, D. Kirk, Software engineering and scale-free networks, IEEE Trans. Syst. Man Cybern. B 39 (4) (2009) 845–854.
[428] G. Concas, M. Marchesi, S. Pinna, N. Serra, Power-laws in a large object-oriented software system, IEEE Trans. Softw. Eng. 33 (10) (2007) 687–708.
[429] L. Šubelj, M. Bajec, Community structure of complex software systems: Analysis and applications, Physica A 390 (16) (2011) 2968–2975.
[430] S. Jenkins, S.R. Kirk, Software architecture graphs as complex networks: A novel partitioning scheme to measure stability and evolution, Inf. Sci. 177 (12) (2007) 2587–2601.
[431] W.-F. Pan, B. Li, Y.-T. Ma, Y.-Y. Qin, X.-Y. Zhou, Measuring structural quality of object-oriented softwares via bug propagation analysis on weighted software networks, J. Comput. Sci. Tech. 25 (6) (2010) 1202–1213.
[432] C. Roach, R. Menezes, Using networks to understand the dynamics of software development, in: Complex Networks, Springer, 2011, pp. 119–129.
[433] M.A. Fortuna, J.A. Bonachela, S.A. Levin, Evolution of a modular software network, Proc. Natl. Acad. Sci. USA 108 (50) (2011) 19985–19989.
[434] S. Koch, Software evolution in open source projectsa large-scale investigation, J. Softw. Maint. Evol.: Res. Pract. 19 (6) (2007) 361–382.
[435] K.-Y. Cai, B.-B. Yin, Software execution processes as an evolving complex network, Inf. Sci. 179 (12) (2009) 1903–1928.
[436] K. He, R. Peng, J. Liu, F. He, P. Liang, B. Li, Design methodology of networked software evolution growth based on software patterns, J. Syst. Sci. Complexity 19 (2) (2006) 157–181.
[437] H. Li, H. Zhao, W. Cai, J.-Q. Xu, J. Ai, A modular attachment mechanism for software network evolution, Physica A 392 (9) (2013) 2025–2037.
[438] H. Li, L.Y. Hao, R. Chen, Multi-Level formation of complex software systems, Entropy 18 (5) (2016) 178.
[439] W. Goffman, Mathematical approach to the spread of scientific ideas–the history of mast cell research, Nature 212 (5061) (1966) 449–452.
[440] L.M. Bettencourt, A. Cintrón-Arias, D.I. Kaiser, C. Castillo-Chávez, The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models, Physica A 364 (2006) 513–536.
[441] N.K. Vitanov, M.R. Ausloos, Knowledge epidemics and population dynamics models for describing idea diffusion, in: Models of Science Dynamics, Springer, 2012, pp. 69–125.
[442] I.Z. Kiss, M. Broom, P.G. Craze, I. Rafols, Can epidemic models describe the diffusion of topics across disciplines? J. Informetr. 4 (1) (2010) 74–82.
[443] X. Gao, J. Guan, Network model of knowledge diffusion, Scientometrics 90 (3) (2012) 749–762.
[444] C. Chen, D. Hicks, Tracing knowledge diffusion, Scientometrics 59 (2) (2004) 199–211.
[445] E. Yan, Disciplinary knowledge production and diffusion in science, J. Assoc. Inf. Sci. Technol. 67 (9) (2015) 2223–2245.
[446] K. Börner, S. Penumarthy, M. Meiss, W. Ke, Mapping the diffusion of scholarly knowledge among major US research institutions, Scientometrics 68 (3) (2006) 415–426.
[447] F. Gargiulo, A. Caen, R. Lambiotte, T. Carletti, The classical origin of modern mathematics, EPJ Data Sci. 5 (1) (2016) 26.
[448] M. Prosperi, I. Buchan, I. Fanti, S. Meloni, P. Palladino, V.I. Torvik, Kin of coauthorship in five decades of health science literature, Proc. Natl. Acad. Sci. USA 113 (32) (2016) 8957–8962.
[449] S.G. Levin, P.E. Stephen, Are the foreign born a source of strength for US science? Science 285 (5431) (1999) 1213–1214.
[450] N.R. Van, Global mobility: Science on the move, Nature 490 (7420) (2012) 326–329.
[451] F. Gargiulo, T. Carletti, Driving forces of researchers mobility, Sci. Rep. 4 (235) (2014) 4860–4860.
[452] M.D. Domenico, E. Omodei, A. Arenas, Quantifying the diaspora of knowledge in the last century, Appl. Netw. Sci. 1 (1) (2016) 15.
[453] R. Pan, K. Kaski, S. Fortunato, World citation and collaboration networks: uncovering the role of geography in science, Sci. Rep. 2 (2011) 902.
[454] T. Jia, D. Wang, B.K. Szymanski, Quantifying patterns of research-interest evolution, Nat. Hum. Behav. 1 (2017) 0078.
[455] C. Chen, Citespace II: Detecting and visualizing emerging trends and transient patterns in scientific literature, J. Am. Soc. Inf. Sci. Tec. 57 (3) (2006) 359–377.
[456] N.J. Van Eck, L. Waltman, Software survey: Vosviewer, a computer program for bibliometric mapping, Scientometrics 84 (2) (2010) 523–538.
[457] S. Team, Science of Science (Sci2) Tool, Software, Indiana University and SciTech Strategies, 2009. https://sci2.cns.iu.edu.
[458] B. Alberts, Impact factor distortions, Science 340 (6134) (2013) 787–787.
[459] D. Hicks, P. Wouters, L. Waltman, S. De Rijcke, I. Rafols, The leiden manifesto for research metrics, Nature 520 (7548) (2015) 429.
[460] M.D. Domenico, A. Solé-Ribalta, E. Omodei, S. Gómez, A. Arenas, Ranking in interconnected multilayer networks reveals versatile nodes, Nat. Commun. 6 (2015) 6868.
[461] P.J. Mucha, T. Richardson, K. Macon, M.A. Porter, J.-P. Onnela, Community structure in time-dependent, multiscale, and multiplex networks, Science 328 (5980) (2010) 876–878.
[462] M. De Domenico, V. Nicosia, A. Arenas, V. Latora, Structural reducibility of multilayer networks, Nat. Commun. 6 (2015) 6864.
[463] V. Nicosia, P.S. Skardal, A. Arenas, V. Latora, Collective phenomena emerging from the interactions between dynamical processes in multiplex networks, Phys. Rev. Lett. 118 (13) (2017) 138302.