Solving the cold-start problem in scientific credit allocation

Xing, Yanmeng; Wang, Fenghua; Zeng, An; Ying, Fan

DOI
10.1016/j.joi.2021.101157

Publication date
2021

Document Version
Final published version

Published in
Journal of Informetrics

Journal of Informetrics 15 (2021) 101157

Contents lists available at ScienceDirect

Journal of Informetrics

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

Solving the cold-start problem in scientific credit allocation

Yanmeng Xing[a], Fenghua Wang[b], An Zeng[a,*], Fan Ying[a,*]

[a] School of Systems Science, Beijing Normal University, Beijing, PR China
[b] Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, The Netherlands

ARTICLE INFO

Article history:
Received 24 June 2020
Received in revised form 7 March 2021
Accepted 9 March 2021

Keywords:
Credit allocation
Co-citing network
Authorship byline
Scientific impact

ABSTRACT

A nearly universal trend in science today is the prominence of ever-increasing collaborative teams. Hence, identifying the relative credit due to each collaborator of published studies is of high significance. Although numerous methods have been employed to address this issue, allocating credit to all co-authors of new papers remains challenging. To address this cold-start issue, we introduce a credit allocation algorithm based on the co-citing network that captures the co-authors' shared credit of a multi-authored publication. Using the American Physical Society publication data, we validate the method by examining papers by Nobel laureates. Accordingly, we perform many experiments to demonstrate that the proposed method can be implemented on academic papers in any period after publication with a significantly higher degree of accuracy and robustness than the existing algorithms applied to new papers. This method enables us to explore the universal credit evolution pattern of scientific elites. Importantly, by testing the relation between an author's credit and authorship byline, we observe that the first authors of papers are currently assigned less credit than in the early days with respect to physics. With collaboration and a large team set to dominate the agenda of the current science system, our study provides a more effective method for allocating early credit to co-authors of a paper, which may be beneficial to various academic activities, including faculty hiring, funding, and promotion decisions.

1. Introduction

The increasing ascendancy of collaboration is one of the most common trends observed in all domains of modern science and technology along with the disappearance of solo scientific discoveries (Guimerà, Uzzi, Spiro, & Lus, 2005; Wu, Wang, & Evans, 2019). Hence, the synergy of collaboration is an essential component in complex scientific projects that require multidisciplinary solutions (Falk-Krzesinski, 2011; Milojević, 2014). Collaboration allows for the integration of knowledge and the mandate of research, both of which require comprehensiveness and diversity. Solitary works generally lead to lower impact publications relative to collaborative science. Moreover, high-quality published papers are the products of science activity and bear crucial effects on scientists' academic reputations and stances (Carpenter, Cone, & Sarli, 2014; Li, Fortunato, Yin, & Wang, 2020; Wuchty, Jones, & Uzzi, 2007). Given that many researchers independently developed the academic community in the old era, the current community continues to reward self-sufficient researchers based on individual scholastic achievements. In this sense, the sole author is credited with all of the contributions of papers only with a single author, which was the commonly accepted norm in science decades ago. However, as that rule fails for co-authored publications,

* Corresponding authors.
E-mail addresses: anzeng@bnu.edu.cn (A. Zeng), yfan@bnu.edu.cn (F. Ying).

https://doi.org/10.1016/j.joi.2021.101157
1751-1577/

it has created a situation that may become even worse when co-authors from varied domains implement different contribution assignment criteria in multidisciplinary projects (Lehmann, Jackson, & Lautrup, 2006). Meanwhile, switching more frequently between topics has recently become an increasing trend (Zeng et al., 2019) Nevertheless, it is expected that talented, ethical, well-prepared individuals be rewarded for their hard-earned accomplishments. This expectation is beneficial to the long-term development of the system of science (Pavlidis, Petersen, & Semendeferi, 2014). As such, identifying the relative credit of each collaborator to the co-authored domain-specific work is of much significance and is therefore fundamental to the academic appointment and promotion process of institutions (Juhász, Tóth, & Lengyel, 2020; Shen & Barabási, 2014).

Considerable research attention has been given to assigning credit fairly for multi-authored publications, and as a result, the scientific community has recently called for increasing concern regarding a subjective evaluation of the author's contribution combined with assessments of their co-authors' contributions (Herz, Dan, Censor, & Bar-Haim, 2020). On the one hand, scientific journals have developed guidelines that recognize the contributions of each author to promote more reasonable credit allocation (Herz et al., 2020; Mohammad Tariqur Rahman, J.M.R.B., & A., N.H., 2017; Radicchi, Fortunato, Markines, & Vespignani, 2009). On the other hand, quantitative algorithms for discriminating scientific and intellectual contributions between individuals or scientific institutions were invented that ranged from the simple to the more elaborate. The simplest algorithm involves assigning each author equivalent contribution recognition, such as either full counting or fractional counting (Zeng et al., 2017). The full counting algorithm regards every author as a single author and thus, every author is awarded full credit, whereas the fractional counting algorithm calculates every author's credit as reciprocal to the total number of authors. However, since authors' contributions to papers differ, the full counting algorithm inflates some authors' contributions, while the fractional counting algorithm dilutes the principal contributors' involvement in the papers (Waltman & van Eck, 2015). Thus, methods based primarily on the authorship are proposed, such as the geometric method (Egghe, Rousseau, & Van Hooydonk, 2000), the arithmetic method (Trueba & Guerrero, 2004), the harmonic method (Hagen, 2008) as well as the method based on network(Kim & Diesner, 2014). However, these types of algorithms cannot be used in all research fields as the rules of authorship bylines vary substantially. For example, in mathematics, the authorship is alphabetic; whereas in biology, the first and the last authors contribute the most to the article. Another way to allocate author credit is by declaring the contributions of each author in the article, thereby clarifying all authors' roles in the research (Foulkes & Neylon, 1996; Mohammad Tariqur Rahman et al., 2017). Currently, the collective process perspective method to allocate author credit has become popular (Bao & Wang, 2020; Radicchi et al., 2009; Shen & Barabási, 2014). The main hypothesis of this method is that the citing process of the paper and other papers written by the same authors regarding the same research topic encodes the informal credit allocation, indicating that the main contributors to the paper are experienced in the research topic. The improved algorithms further consider the aging effect and the importance of citing sources during the collective process (Bao & Zhai, 2017; Wang, Fan, Zeng, & Di, 2019).

Typical state-of-the-art quantitative algorithms for allocating shared credit to authors of a paper have been recently designed and are, in one form or another, ultimately built on the direct citations of the target papers. Nevertheless, these algorithms neglect that each paper accumulates an unequal number of citations and that a relatively high proportion of all papers has only a few citations, a factor that results in less effective identification due to the extremely sparsely populated co-cited networks. This problem is more prominent in newly published papers, as they have insufficient time to accumulate citations. Although many previous studies have focused on the contribution allocating issues of scientific community collective methods, the intellectual contribution allocation of individual authors of papers during the early period has not been emphasized or systematically studied in the literature, which means this is a typical cold-start problem. Hence, we consider it significant to develop a more comprehensive and universal algorithm that appropriately characterizes the scientific credit of each author of a co-authored paper, wherein the credit to authors of papers during their early careers, as well as that in their late careers, can be appropriately allocated using our algorithm.

This paper is organized into four distinct sections. The first section is the introduction. This section is followed by a brief description of the dataset used in the article and statistical analyses of the dataset to demonstrate the various limitations of the existing quantitative algorithms of credit allocation in the second section. Next, we propose a new method based on referenced studies. In the third section, we select papers by Nobel laureates to validate the proposed algorithm's effectiveness and then apply the algorithm to ordinary papers in the early period following their publication to test the robustness of the proposed algorithm. This analysis is followed by an illustration of the credit share evolution of co-authors and an exploration of the universal credit share evolution pattern of scientific elites. Finally, we discuss the relation between credit share and position in the authorship bylines in the field of physics. Section 4 presents a discussion of the results and outlines the paper's conclusions.

2. Method

2.1. Data

The database used in this study is obtained from the American Physical Society (APS) journals for the period 1893 to 2009 and includes journals of the physical review series and the reviews of modern physics. To avoid the problem of author name ambiguity, we use the author name dataset obtained from Sinatra et al., which has been processed using a comprehensive disambiguation method in the APS dataset (Sinatra, Wang, Deville, Song, & Barabási, 2016). The dataset is comprised of

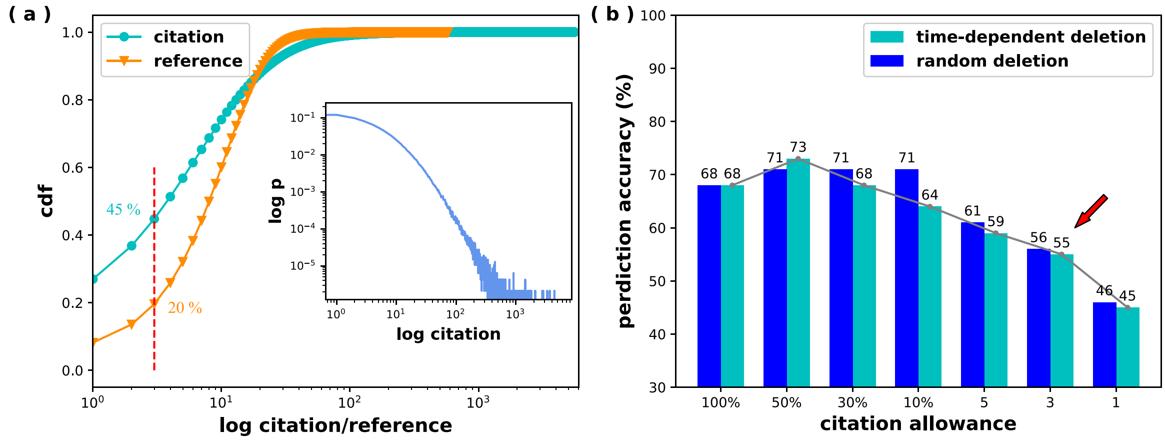

Fig. 1. The citation or reference distribution of APS dataset and changes in identification accuracy of the co-cited network credit allocation algorithm under perturbations. (*a*) The log cumulative distribution of citations and references within APS publications from 1893 to 2009. (*Inset*) The *loglog* distribution of citations for all publications. (*b*) The identification accuracy of the co-cited network algorithm declines when randomly deleting citations (dark blue bar) or deleting according to citing time of citation (dark green bar).

458,584 papers with 4,620,025 citations from 236,884 distinct scientists, allowing us to systematically analyze citation records as well as the scientific community. In addition, by merging the Nobel Prize data from two previous studies (Bao & Zhai, 2017; Shen & Barabási, 2014), we identify 32 Nobel physics laureates to demonstrate our results in the main text. More tested groups of prize winners are included in the Appendix. Furthermore, Li et al. collected more comprehensive Nobel data in their paper published in Scientific Data in 2019 (Li, Fortunato, Yin, & Wang, 2019; Li, Yin, Fortunato, & Wang, 2019). Thus, we downloaded this data and identified 98 additional Nobel papers in physics. According to the same selection rules previously mentioned, we manually collected 34 award winners of the Max Planck Medal and Boltzmann Award.

2.2. *Basic statistics of citation relation*

A published paper commonly has references and citations. As presented in Fig. 1(a), the distribution of citations has a long tail, which indicates that numerous papers have few citations. For example, papers with fewer than three citations account for 45% of the database. However, the papers with fewer than three references account for only 20% (Fig. 1(a)). As mentioned in the introduction, the existing collective credit allocation methods of a target paper are highly dependent on its citations, which implies that if there is a target paper with no citation or with only a few citations, the credit allocation of coauthors may be very noisy.

By applying the collective credit allocation method (*CCD*) proposed by Shen and Barabási (Shen & Barabási, 2014) to Nobel Prize-winning papers, we find that, from Fig. 1(b), the degree of accuracy with respect to identifying the Noble laureates is low when Nobel Prize-winning papers have fewer three citations. This suggests that the *CCD* based on citing papers should be improved. Thus, when we investigated the relationship between the number of citations and the credit allocation, we defined the identification accuracy as the ratio of Nobel Prize-winning papers for which Nobel laureates obtained the greatest credit as calculated by the algorithm. In Fig. 1(b), the set of light blue bars represents the accuracy of papers whose citations were deleted according to the citing time. For example, the numeral "1" means there is a first paper citing the Nobel Prize-winning paper and "30%" means that we retain the first 30% citing papers among all citing papers of Nobel Prize-winning papers based on ascending order of the citing time. The blue bars represent the accuracy of Nobel Prize-winning papers whose citing papers are randomly deleted. As evidenced in Fig. 1(b), the accuracy of the laureates with the highest credit drops dramatically from more than 70% when preserving 50% of the citations to approximately 50% when preserving only three citations. Moreover, the identification accuracy is lower if we delete citations by time (dark green bars in Fig. 1(b)).

2.3. *Co-citing network algorithm for credit allocation*

Based on the above results, we attempted to incorporate references into the algorithm to measure the credit of co-authors. Therefore, we proposed a co-citing network algorithm for credit allocation (*COCD*). The co-citing network of the target paper is mapped by papers (nodes) if they share at least one reference with the target paper. In the following, we present the process of applying *COCD* to allocate the credit share of every author of a published paper. There are M coauthors $a_i (1 \leq i \leq M)$ of the target paper P_0. All references of P_0 form the set $R \equiv \{R_1, R_2, \ldots, R_n\}$ and all citing papers of set R comprise the set $P \equiv \{P_0, P_1, \ldots, P_m\}$. We developed a strength vector s to measure how many papers in the set R are cited by each paper in the set P. For instance, in Fig. 2, the paper P_0 cites four references in set R, specifically, paper R_1, paper R_2, paper R_3 and paper R_n. Thus, the element of the strength vector s for P_0 is 4. Similarly, the strength for paper P_1 is 1; the strength for paper P_2 is 3 and the strength for paper P_2 is 2. For the target paper P_0, the matrix A is used to denote the authors' credit

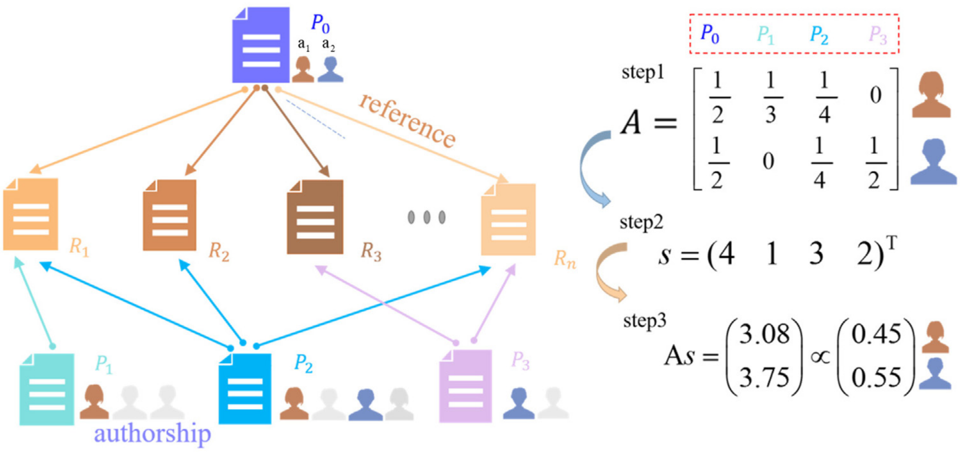

Fig. 2. The illustration of collective credit allocation method *COCD* based on the co-citing the co-citing network. The target paper P_0 has two authors, colored in orange and blue, respectively. We also show the cited references R_n and the co-citing papers $P_j (0 \le j \le 3)$ that cite the same reference together with P_0. At step 1, we obtain the credit allocation matrix A from the authorship lists of the co-citing papers. The matrix A provides the authors' share for each co-citing paper. For example, because the first author (orange female) in the target paper is one of three authors in P_1 but the second author (blue male) in the target paper is not the author of P_1, it votes 1/2 for the female author and 0 for the male author. At step 2, the strength vector s denotes the co-citing links among the paper P_0 and the co-citing papers. At step 3, after obtaining strength vector s and matrix A, we can calculate the relative credit share of the two authors of P_0 according to Eq. (1) or (2) by normalizing C.

shares in the co-citing papers from the set P. Specifically, A_{ij} represents the credit share of a_i in the paper P_{j-1} and the credit share is computed using the fractional credit allocation method which is the reciprocal of the total number of authors. For example, the number of co-authors in P_0 in Fig. 2 is two. According to the fractional credit allocation method, the credit share of author a_1 equals to the credit share of author a_2 and the value of their credit share for paper P_0 is $\frac{1}{2}$, so $A_{11} = A_{21} = \frac{1}{2}$. Similarly, there are three co-authors in P_1 and the credit share of each author is $\frac{1}{3}$. Female author a_1 is one of the co-authors of the paper P_1 but the male author a_2 is not one of the co-authors in the paper P_1, which yields $A_{21} = \frac{1}{3}$ and $A_{22} = 0$. After the matrix A and the strength vector s are acquired, the credit shares of the co-authors in the target paper P_0 is defined as

$$c_i = \sum_{j=1}^{m+1} A_{ij} s_j, \tag{1}$$

or a matrix

$$\mathbf{C} = \mathrm{A}s. \tag{2}$$

The relative credit of co-authors of the paper P_0 is based on the vector C. Next, the credit share among the co-authors can be obtained by normalizing C.

3. Result

3.1. Validation

To quantitatively validate the effectiveness of *COCD*, first and foremost, we first test it by examining Nobel Prize-winning papers, where the Nobel committee has decided who the Nobel prize is awarded (Turki, Hadj Taieb, & Aouicha, 2020). A widely accepted consensus is that the Nobel winner is the author who contributes most to the Nobel Prize-winning paper. Hence, he/she should be allocated greater credit shares than other collaborators. As the Nobel committee decides to whom the Nobel Prize is awarded, we assume that the laureates of Nobel Prize papers are regarded as ground truth in the next validation. Accordingly, the algorithm identifying accuracy is defined as the ratio of Nobel Prize-winning publications for which the author with the greatest credit allocated by our method is exactly the laureate awarded by the Nobel committee. Table 1 lists the normalized credit share of four Nobel Prize-winning publications in various time spots after the target paper was published, thus displaying the time-dependent credit marks. The laureate is marked with an asterisk, the main competitor of the laureate is underlined and the largest credit share in 2009 is highlighted in bold. The results indicate that in three of the four cases, the prize winners were allocated the most credits in 2009 (i.e., the time cap of the dataset used in this work). It is evident that the performance of laureates seems to be slightly superior to that of his main competitors within our credit allocation algorithm. We still successfully identify the accurate winner among the candidates in consensus with authorities and consider the result of the Nobel committee, which indicates that our methodology may draw the same conclusion as subjective and comprehensive evaluations with respect to credit allocation for outstanding scientific projects. Nevertheless, one can observe that an exception occurs in the last case in Table 1, where the laureate's credit share is less than

Table 1
Credit shares are calculated by Eq. (1) or Eq. (2) in the Publish year, Award year, Award year+1(i.e., one year late after Award year) and 2009 year (maximum number of years in our data set) of papers, employing the *APS* data set. The rank of co-authors is shown consisting of their orders in the authorship byline. The greatest credit shares in 2009 are highlighted in bold and the laureates are marked with asterisks. The main competitors of laureates are underlined. For papers awarded after 2009, we put *NA* for credit shares.

Papers (publish/award year)	Credit share Authors	Publish year	Award year	Award year+1	2009
PhysRev.83.333 (1951/1994)	W. A. Strauser	0.24	0.19	0.19	0.19
	E. O. Wollan	0.32	0.38	0.38	0.38
	Clifford G. Shull*	**0.44**	**0.43**	**0.43**	**0.43**
PhysRevLett.55.48 (1985/1997)	L. W. Hollberg	0.12	0.08	0.07	0.07
	A. A. Ashkin	0.48	0.36	0.31	0.31
	Alex Cable	0.12	0.21	0.18	0.18
	S. Y. Chu*	**0.12**	**0.20**	**0.30**	**0.32**
	J. E. Bjorkholm	0.15	0.15	0.13	0.13
PhysRevLett.84.5102 (2000/2005)	Steven T. Cundiff	0.04	0.13	0.11	0.11
	Jinendra K. Ranka	0.04	0.03	0.03	0.03
	Theodor W. Hänsch*	**0.20**	**0.21**	**0.25**	**0.25**
	John I. Hall	0.17	0.13	0.11	0.11
	Scott A. Diddams	0.04	0.07	0.06	0.07
	Thomas Udem	0.04	0.02	0.02	0.02
	David J. Jones	0.04	0.02	0.02	0.02
	Ronald Holzwarth	0.18	0.11	0.10	0.09
	Robert S. Windeler	0.04	0.10	0.08	0.09
	Jinendra K. Ranka	0.17	0.20	0.23	0.23
PhysRevLett.77.4887 (1996/2009)	Jens dreyer	0.06	0.03	*NA*	0.03
	Abdelhamid Maali	0.09	0.05	*NA*	0.05
	Christof Wunderlich	0.05	0.04	*NA*	0.04
	M. Brune	0.22	0.25	*NA*	0.25
	Serge Haroche*	**0.24**	**0.26**	*NA*	**0.26**
	X. Maître	0.05	0.04	*NA*	0.04
	J. M. Raimond	0.22	0.29	*NA*	0.29
	E. W. Hagley	0.06	0.05	*NA*	0.05

Table 2
Accuracy of algorithms with different priors. The co-citing network method computed by references is compared with other integrated collective algorithms and the cases when the algorithms are implemented into the networks with links missing.

Algorithms	CCD	COCD	CCD3	COCDB	CCD3+COCD	CCD+COCD
Accuracy	68%	82%	56%	68%	82%	72%

the score of his main competitor when he was awarded the Nobel Prize in 2009. Fig. 3 presents the identification accuracy of the proposed method with respect to determining the prize winners from the authorship bylines of all 22 multi-authored prized papers in physics. We observe that the authors who have the greatest credit share correspond to the laureates in 18 target prized papers (i.e., 82% of identification accuracy), despite the diverse rank of the prize winners place on the authorship list. When we expand the tested sample size of the Nobel Physics Prize, the identification accuracy is as high as 82%, as presented in Table 3 of the Appendix A. Table 4 illustrates that other tested groups of the Max Planck Medal and the Boltzmann Award also exhibited high accuracy with a rate of 68% in Appendix B.

It is beneficial to analyze the underlying reasons for a few failures of our method. Similar to the last case in Table 1, where we fail to discern the laureate in the award year, there is the same result in the example of PhysRevLett.55.48 (1985/1997, i.e., the second case in Table 1). Given the lack of a long career trajectory preceding the publication of the prized paper, Chu had been allocated a lower credit share than Ashkin, even though Chu received the top credit many years after the award year in this field. Specifically, in this case, regarded by many physicists as the founder of the research area of optical tweezers, Ashkin had published many impactful papers before collaborating with Chu in this domain. These works developed the knowledge that resulted in the Nobel Prize discovery. Afterwards, Chu started to generate a more dramatic contribution to this important research finding, thus suggesting that authors require time to prove their actual contributions to the research field. The detailed credit share evolution that may account for the award results will be further discussed herein.

To demonstrate the identifying strength of the *COCD* algorithm in credit allocation, we compare the identifying accuracy of the *COCD* method with that of the *CCD* method in Table 2. To investigate the accuracy of the *CCD* method in papers with few citations, we randomly delete the citing papers of the Nobel papers until they have three citations left. This test is denoted as *CCD*3. The accuracy of *COCD* is determined to be 82%, which is significantly higher than the accuracy of *CCD*3, i.e. 56%. Analogously, we also examine the performance of the *COCD* in papers with limited information. We only use co-citing papers which published before the publishing year of the target paper to implement the *COCD* method. This test is denoted

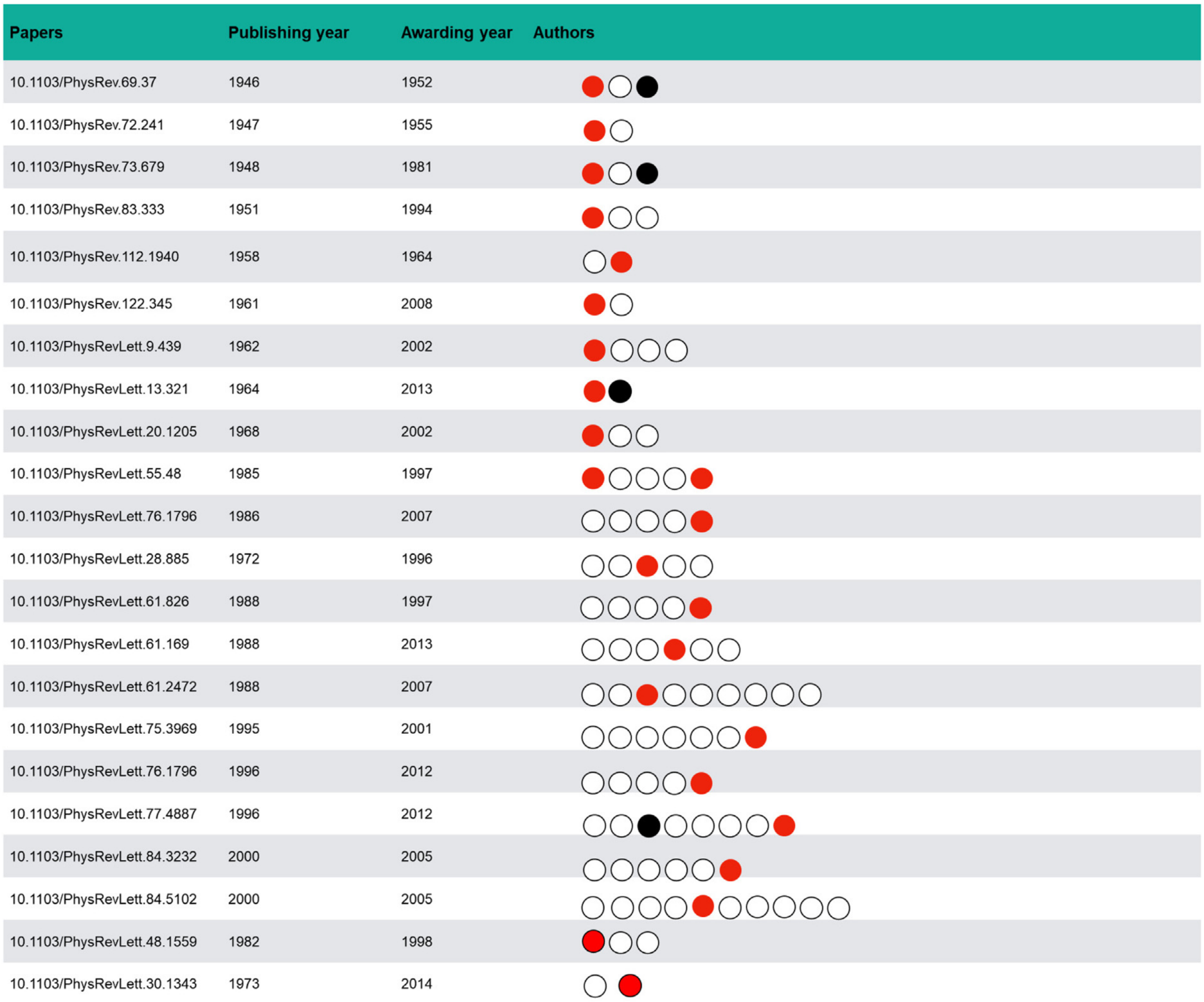

Fig. 3. The result of identifying Nobel laureates of Nobel-Prize publications base on the *COCD* algorithm. For each publication, the laureate is shown in red-filled circles. By contrast, the author who is allocated the greatest credit share is shown as a black-filled circle when he/she is not the real laureate. Other co-authors are shown as empty circles. Thus, the presence of black-filled circles indicates that the result of credit allocation obtained by the *COCD* algorithm is not consistent with the official list of the prize winners.

as *COCDB*. Note that this test is equivalent to allocate credit for co-authors in a new paper. The *COCDB* test yields a 68% accuracy rate, which is as high as the accuracy rate of the *CCD* method with complete citation information. Nonetheless, we cannot apply the *CCD* method to a newly published paper because it does not have any citation to map a co-cited network, in which each node is a paper citing the target paper and two papers are linked if they are cited together by at least one paper.

The above comparison indicates that our method, which is based on co-citing networks, can more precisely identify who receives the most credit for a paper. Importantly, even if applied to a new paper that has not accumulated any citations, our method can still effectively allocate credit shares to each author. We further combine the co-citing and co-cited paper pool to calculate credit share and find that the identifying accuracy increases respectively to 82% and 72% for $CCD3 + COCD$ and $CCD + COCD$, respectively. In fact, the *CCD* and *COCD* methods allocate credits according to the same principle, i.e., the authors whose research is more closely related to the target paper are allocated with greater credits. This relation is measured via co-cited networks (citations of the papers) for the *CCD* method while it is measured via co-citing networks in *COCD* (references of the papers). *CCD* faces a cold-start problem because the reference information is complete for papers with low citations and new papers, meaning that the *COCD* method thus demonstrates a higher degree of accuracy than the *CCD* method for these papers. When combining the two, i.e., *CCD* and *COCD*, one can obtain a better measure of the relation of an author's research to the target paper, thereby resulting in a higher credit allocation accuracy.

In the previous analyses above, we focus on the effectiveness validation of the proposed method with respect to outstanding scientific achievements. Another means of validating effectiveness is to test the robustness of these methods using an extensive sample. To this end, we apply credit allocation algorithms in different years to generic papers whose citation

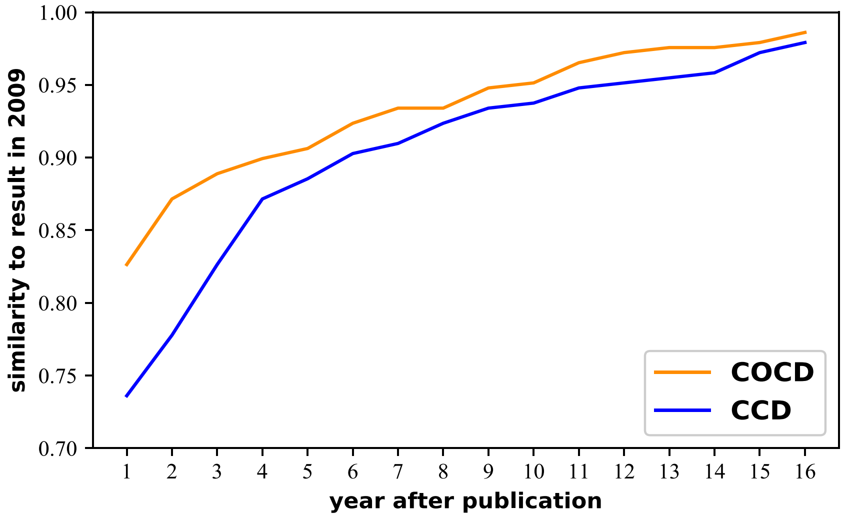

Fig. 4. The robustness of *the CCD* and *COCD* algorithms against time-dependent perturbations. The similarity is the proportion of evaluated papers. The author of them with the greatest credit share identified in each following year after publication does not change compared to the one identified in 2009. The citations of evaluated papers are spanned from 10 to 1000 in our data set.

counts range from 10 to 1000. As we treat the dataset of 2009 as complete, the dataset prior to 2009 can be treated as incomplete. In addition, the ordinary paper has no specific laureate to directly calculate the identification accuracy used in any previous validation. Therefore, herein we define identification consistency as verifying whether the authors with the greatest credit share are the same as those identified in 2009 when using the same method. Specifically, in each year following publication, the similarity is calculated by the proportion of papers in which the author with the greatest credit is the same as that detected by *CCD* in 2009. Similar to the analyses for works of science elites in Fig. 1(a), we compare the evolution of the allocation consistency of *CCD* and *COCD* after the publication of a paper. In other words, in each year following publication, we calculate the proportion of papers in which the author with the highest credit is the same as that detected by *CCD* in 2009. Fig. 4 indicates that the consistency of the *COCD* method is always higher than that of *CCD*, especially in the early period following publication of a paper. This result indicates that the *COCD* method is more robust than the *CCD* method against incomplete information.

3.2. Credit share evolution

The co-authors' credit share evolution is examined by controlling the publishing year of papers in the citation network. The author's credit in t year is calculated using the *COCD* method and the papers in the co-citing pool before t year. In the co-citing pool of the target paper, each paper shares at least one same reference with the target paper. The results indicate that for the 1994 prize-winning paper (Fig. 5(a)), Clifford G. Shull received the majority of the credit after the article was published in 1952. He continuously maintained the leading position in contribution even though there was a slight decline during the early period. With respect to the 1997 prize-winning paper in Fig. 5(b), Ashkin received the dominant part of the credit for the discovery immediately after the publication, whereas Chu's credit share is slight due to a lack of previous publication records in the research area, which, in turn, explains Ashkin's higher score in the earlier years. Notwithstanding, Ashkin did not publish papers after 1986 and was retired in 1992 while Chu began to pursue endeavors and make continuous contributions in the scientific field. Hence, Chu's credit share has increased over time, while Ashkin's credit share has decreased. When Chu won the Nobel Prize in 1997, and his credit share finally surpassed that of A. Ashkin in 2001. This case demonstrates that if the junior colleagues have taken essential independent contributions to the field, their credit shares may overwhelm the senior scientists' credit shares over time.

These two intriguing examples of Chu's and Strauser's careers prompt us to systematically investigate the evolution pattern of laureates' relative credit in a Nobel Prize-winning paper at the individual-career level (Li et al., 2020; Li, Fortunato, et al., 2019; Li, Yin, et al., 2019). To this end, Fig. 5(c) delineates wide-ranging evolution trajectories followed by real careers. We investigate scientific careers over time and calculate the normalized credit share of each Nobel laureate, c_t/c_0, defined as the quotient between subsequent credit share in t_{th} year and constant credit share in the year of publication t_0. Here we find that the normalized credit share per laureate reveals a significant increase or maintenance at the highest levels after publication. This 'jumping' effect has occurred primarily among Nobel laureates in recent years. One may conclude that the laureates continue to make insightful exploitation and valuable contributions related to the prize-winning topic, and thus, eventually they are awarded the prize, given their substantially elevated contributions and high visibility compared to others.

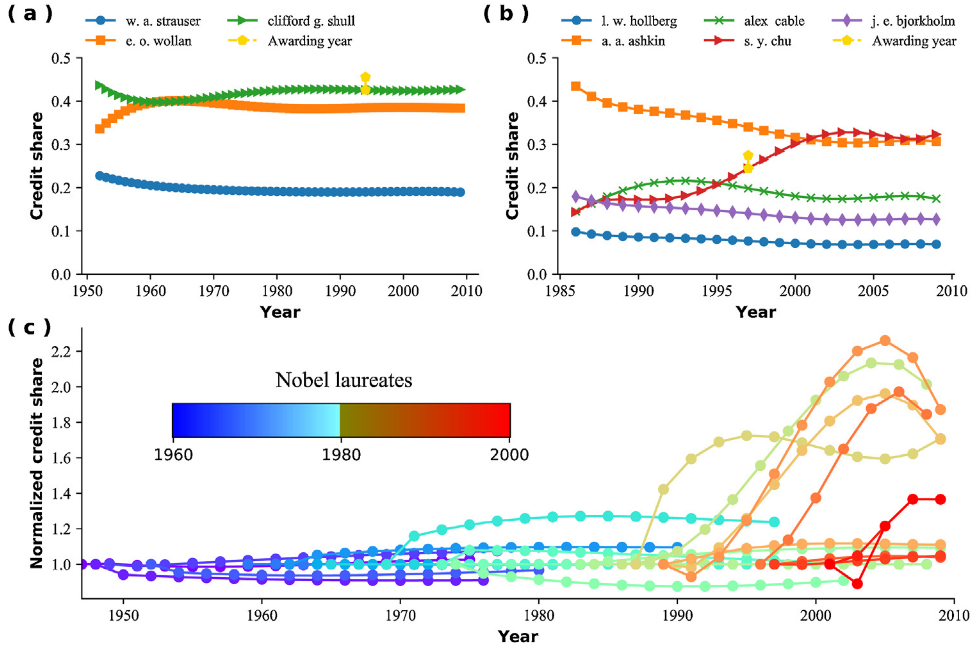

Fig. 5. Credit share evolution. (a) The relative credit shares of co-authors of the Nobel Prize-winning publication '10.1103/*PhysRev*.83.333' in 1951. (b) The relative credit share of co-authors of the Nobel Prize-winning publication '10.1103/*PhysRevLett*.55.48' in 1997. (c) Normalized credit share for laureates of all Nobel prize-winning papers. For each laureate, we use the increasing ratio c_t/c_0 to quantify the change of her/his credit share after she/he published the paper. Color corresponds to a paper's publishing year t_0.

3.3. Authorship byline and author's credit

We explored the relation between an author's credit and his/her position in the authorship byline by using the *COCD* algorithm in this part. To this end, we use the allocated credit to approximate the contribution of an author made to a paper (Jung & Yoon, 2019). Generally, it is recognized that the first author and the corresponding author who is normally the last author, make main contributions to an article in physics. Therefore, we focus on two main questions. The first question is whether the first author and the last author are allocated higher credits than other authors, and the second question is to which position is the author with the greatest credit allocated.

To investigate the first question, we classified papers into several categories based on the number of authors. We then chose five types of papers whose number of authors ranged from three to seven to perform the analyses. For each type of paper, we obtain the distribution of the first authors, the credit distribution of the last authors and the credit distribution of other authors (i.e., the authors are not located at either the first or last position in the byline). By comparison of the credit distributions, we find that credit differences indeed exist in Fig. 6. Moreover, it is interesting that the results before 2000 and after 2000 exhibit different phenomena. Specifically, in the papers published before 2000, Fig. 6(a) shows that credit distributions of the first authors and other authors have similar longer and thicker tails, and the median of the first authors' credit share does not significantly differentiate from the others. Intuitively, the credit distributions of the last authors in the papers published before 2000, as shown in Fig. 6(c), have longer and thicker tails in all categories, and the median of the last authors' credit share is significantly higher, which implies that the last author has a higher probability of having a larger credit share than other authors. However, in the papers published after 2000, Fig. 6(b) illustrates that the credit distribution of first authors has a significant difference from the credit distribution of other authors in all categories, whose quantile is remarkably smaller than that of others. The results of the distribution comparison of the last authors and other authors after 2000 (Fig. 6(d)) are consistent with the results before 2000. In summary, the last author commonly is allocated greater credit than other authors, while the first author possesses a smaller contribution than other authors in the papers published after the year 2000. In the papers published before the year 2000, the first authors and other authors have similar contribution distributions. A possible reason for the differences in contribution distribution comparisons between the first authors and other authors may be that, in recent years, most of the first authors started as junior authors.

To answer the second question, we calculated the credit share of each author of a paper and determined the position of the author with the highest credit value. Next, we obtained the probability of each position where the author who makes the greatest contribution is listed the papers with the same number of authors. Compared with the results of the papers whose author positions are randomly switched, the results of the papers published before 2000 and after 2000 demonstrate that the last author has the highest probability of being the author with the greatest credit share, while the probability of

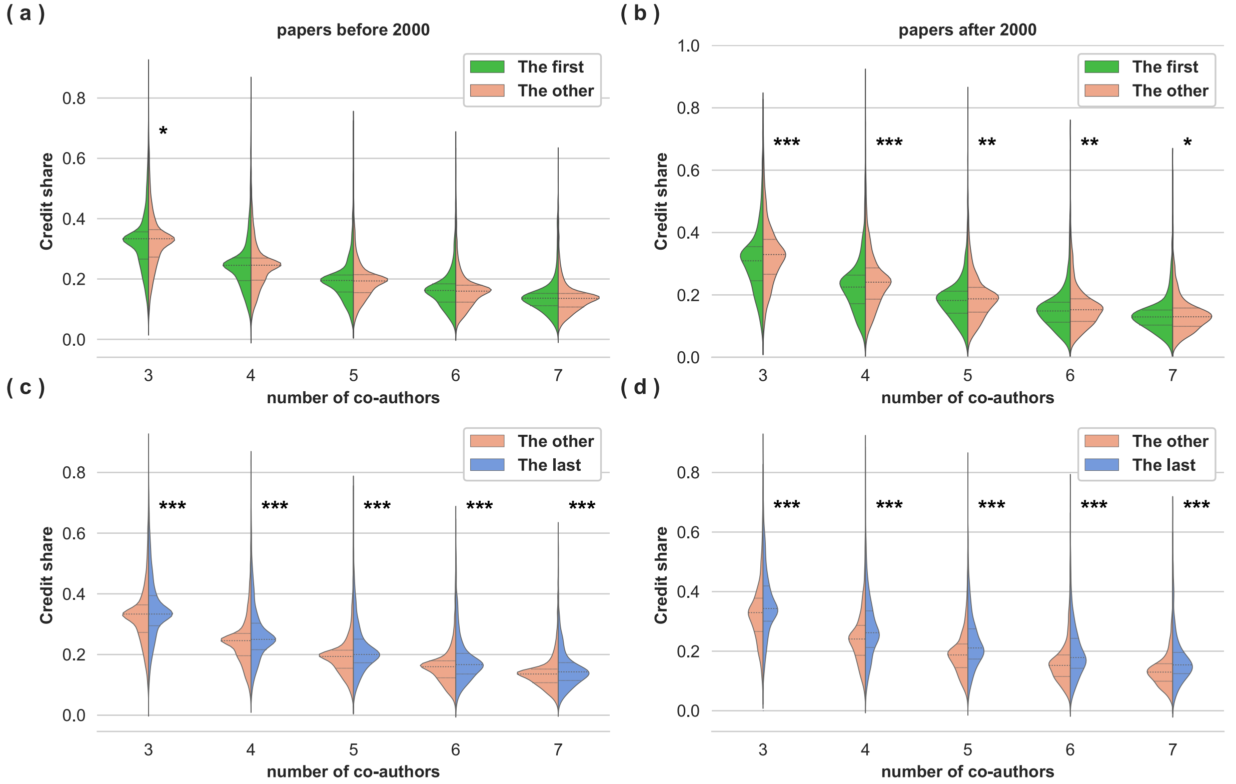

Fig. 6. Violin plots of the credit share of papers published before 2000 and after 2000. The Mann-Whitney significance test results of comparing contribution distributions are presented at the top of each subplot (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$). The green surface represents the credit share of the first authors in papers with different numbers of co-authors as *xlable* shows, and the blue surface and orange surface respectively represent the last authors and authors at other positions of authorship byline. The median and the $\pm$ interquartile range are displayed in distribution.

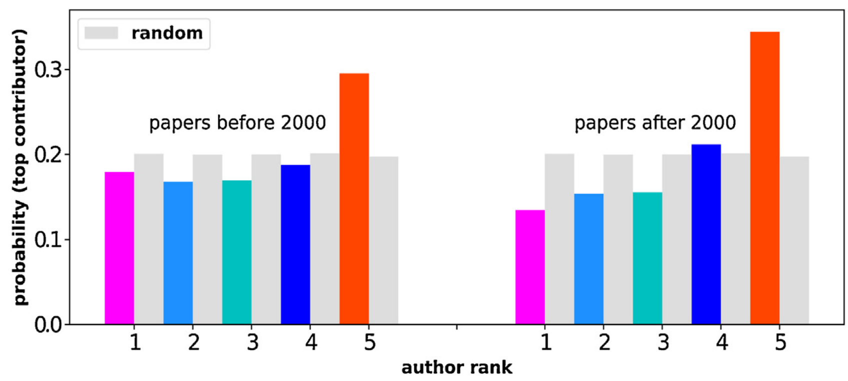

Fig. 7. Histogram of the number of the top contributor in papers with 5 co-authors published before 2000 and after 2000. The colored bars represent the real probability of top contributors in different position ranks of authorship byline. The gray bar represents the probability of the top contributor after we randomly rearrange the authorship byline.

the first author with the largest credit share in the papers published before 2000 has similar rank to that of the other listed authors. However, in the papers published after the year 2000, the probability rank of the first author becomes dramatically low, which may be explained by the fact that the first author is young and not experienced in the research field. Fig. 7 is an example of the above analyses in the papers with five authors. Meanwhile, other papers with different numbers of authors exhibit similar results.

4. Conclusions and discussion

In many research situations, such as the promoting and funding of research, researchers are usually evaluated based on their independent contributions to the academic community to which they belong. However, with today's rapid development of collaborative and multidisciplinary science today, how to allocate the relative credit share of researchers is an increasing and challenging problem, as scientific works tend to involve a remarkable collection of researchers from various groups of different fields. One of the state-of-the-art methods to quantify the relative credit of collaborators by reproducing the informal collective process has been designed and implemented by the academic community, and though effective, this method relies too heavily on accumulated citations being applied to early contribution evaluation wherein an evaluated paper received only a few citations. Such early evaluation is also of much significance. Therefore, in this paper, we propose a credit allocating algorithm based on a co-citing network to solve this problem and validate it using the APS dataset. This method depends on references rather than citations, whereby the reference relations can effectively detect topic similarity among papers, as the research topic of each paper is automatically classified by the body of subsequent citing papers. Moreover, according to this method, the authors whose research is more closely related to the target paper should be allocated more credit.

We conduct a series of experiments to validate the proposed method, and from the results, we demonstrate that the proposed method has a higher degree of effectiveness than do the existing methods with respect to identifying papers by Nobel laureates in which they are credited for major discoveries. In papers without any citations or when the papers are new, other citation-based approaches cannot be used. However, our method works effectively for these papers. Meanwhile, our method is more robust against time-dependent perturbations, indicating that the credit share assigned by the co-citing network algorithm is much less influenced by publication age. In addition, by investigating the evolution of co-authors' credit shares, we find that senior scientists gain more credit than do their junior collaborators in the early period of collaborative publications. Nevertheless, the credit share rank can change if a junior colleague makes a substantial independent contribution to the field. The evolution of the Nobel laureate's credit share follows the same pattern, i.e., they make continuous contributions to their research field at the career level. Finally, we test the acknowledged rule on authorship and contribution, i.e., the first or the corresponding authors are awarded maximum contribution, and we find distinguishable relations in the past and recent publications in the field of physics. In summary, our main contributions are threefold. (1) We propose a more robust and widely applied algorithm that can assign credit to each author of multi-author papers in the early period. (2) Credit evolution uncovers a new degree of regularity underlying individual careers such that a junior researcher can surpass collaborative established scientists in the future if he/she makes a more independent contribution to the specific field. Meanwhile, the evolution of the Nobel laureate's credit share follows the same pattern. (3) The relation between authorship byline and the author's contribution to the field has changed fundamentally over time. Our research has crucial practical significance by providing valuable advice for academic awards, identifying proper scholars for university faculty, and promoting outstanding researchers.

The proposed methodology indicates that there are further studies are required. When composing a pool of co-citing papers, i.e., the set of papers that share at least one common reference with the target paper, we assume that each paper is equally important, thus resulting in manipulations. Authors who publish papers which cite numerous same references of the target paper can gain a larger credit share in the beginning stage. Besides, authors can increase the credit share of the target paper by continuously publishing papers citing references to the paper. As a further improvement, we could consider an algorithm to weigh co-citing papers. For example, citations from higher impact papers would acquire more weight, and papers with zero citation would be ignored. Furthermore, while name disambiguation causes noise in the data set, the existing studies indicate that the name disambiguation process slightly influences the properties of citation networks, although Asian names result in major misidentification (Kim & Diesner, 2016; Martin, Ball, Karrer, & Newman, 2013). Accordingly, the identification results of the research are not significantly affected given the limited number of Nobel laureates who were analyzed. There are only four Asian names in the Nobel-prize papers dataset. Nonetheless, the accuracy of author name disambiguation is a long-lasting challenge and should be carefully checked whenever publication data are used. Note that the APS data set only includes the field of physics. Thus, it is necessary to validate the co-citing network algorithm results for the credit allocation (*COCD*) method using data from other fields. Promising future directions include incorporating exogenous information into the simple model proposed herein to improve its accuracy, such as mentoring relations and affiliated institution rank. Given that credit allocation has a potential impact on individuals as well as on the scientific community's standing in the long term, current science requires a more diverse and wholesome evaluation of the scientific impact of credit allocation. Hence, other directions also include combining the citation-based credit allocation method with the other available evaluation tools. Accordingly, pursuing such endeavors will not only substantially improve on the tracing, assessing, predicting, and nurturing of high-impact scientists, but it may also eliminate deleterious implications of the sole tool for contribution allocation and thus result in an altogether different picture.

Authors' contribution

Yanmeng Xing: Software, Validation, Writing - original draft, Writing - review & editing, Formal analysis.
Fenghua Wang: Software, Writing - original draft, Writing - review & editing.
An Zeng: Conceptualization, Methodology, Writing - review & editing, Formal analysis, Data curation.
Ying Fan: Conceptualization, Methodology, Supervision.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (Nos. 71843005 and 71731002).

Appendix A

Table 3

Table 3
Validation of *COCD* method in 98 prized papers of Nobel Prize in Physics. The first column is Digital Object Unique Identifier (*DOI*) of the papers. The second is real laureate of the paper listed in the same row.The ratio of Credit Rank calculated by *COCD* method and number of authors is shown in the third column. We denote them as *CR/N* for simplification. The papers that are identified wrong are highlighted in bold. We summary the total identification accuracy in the last row.

Papers	Laureates	CR/N	Papers	Laureates	CR/N	Papers	Laureates	CR/N
PhysRevL.50.1395	laughlin	1/1	PhysRev.127.1918	b.bergen	1/4	PhysRevL30.1346	politzer	1/1
PhysRevL.9.439	giacconi	1/4	PhysRev.78.699	ramsey	1/1	PhysRevL.61.2472	fert	1/9
PhysRevL.20.292	**kendall**	**3/15**	PhysRevL.19.1264	weinberg	1/1	PhysRev.128.606	b.bergen	1/2
PhysRevL.30.1343	Strauser	1/2	PhysRev.122.345	nambu	1/2	PhysRev.83.333	shull	1/3
PhysRevL.23.930	kendall	1/11	PhysRev.109.1492	anderson	1/1	PhysRev.75.1969	mayer	1/1
PhysRev.130.2529	glauber	1/1	**PhysRev.30.705**	**davisson**	**2/2**	PhysRev.28.1049	dinger	1/1
PhysRevL.2.256	brockhouse	1/1	PhysRevL.75.4714	wineland	1/5	PhysRevL61.169	phillips	1/6
PhysRevL.75.3969	ketterle	1/7	PhysRevL.5.464	giaever	1/1	PhysRev.108.1175	glauber	1/3
PhysRevL.80.72	shockley	1/2	PhysRev.140.A1869	leggett	1/1	**PhysRevL.23.935**	**kendall**	**1/9**
PhysRev.122.1101	giaever	1/2	PhysRevL.48.1559	reines	1/3	PhysRev.22.409	compton	1/1
PhysRevL.50.1153	haldane	1/1	PhysRev.117.648	nambu	1/1	**PhysRev.70.474**	**bloch**	**2/3**
PhysRev.22.409	compton	1/1	PhysRevL.50.1153	haldane	1/1	PhysRev.117.648	nambu	1/1
PhysRev.92.830	reines	1/2	PhysRevL.13.508	higgs	1/1	PhysRevL.5.147	giaever	1/1
PhysRev.100.703	townes	1/2	**PhysRevL.58.1490**	**koshiba**	**5/23**	PhysRevL.4.380	nambu	1/1
PhysRev.74.1430	feynman	1/1	PhysRev.105.1487	dehmelt	1/1	PhysRev.21.483	compton	1/1
PhysRev.74.250	kusch	1/2	PhysRevL.13.138	dinger	1/4	PhysRevL.28.885	nambu	1/3
PhysRev.40.19	lawrence	1/2	PhysRevL.61.826	claude	1/5	PhysRevL.29.1227	leggett	1/1
PhysRev.109.603	esaki	1/1	PhysRev.109.381	dehmelt	1/1	PhysRev.55.425	alfvn	1/1
PhysRev.76.769	feynman	1/1	PhysRev.125.1067	murray	1/1	**PhysRev.51.677.2**	yukawa	1/2
PhysRevL.25.1543	legget	1/1	PhysRev.76.749	feynman	1/1	PhysRevL.33.1404	ting	1/14
PhysRev.131.2766	glauber	1/1	**PhysRev.69.37**	**purcell**	**2/3**	**PhysRevL.13.321**	**englert**	**2/2**
PhysRev.53.318	rabi	1/4	PhysRevB.4.3184	wilson	1/1	PhysRev.55.434	bethe	1/1
PhysRevB.4.3174	wilson	1/1	PhysRev.75.796	robert	1/1	PhysRev.111.747	bertram	1/2
PhysRev.109.193	**murray**	**2/2**	PhysRevL.90.717.2	feynman	1/2	**PhysRevL.7.178**	**alvarez**	**3/4**
PhysRev.75.1766.2	jensen	1/3	PhysRev.73.416	schwinger	1/1	PhysRev.74.230	walter	1/2
PhysRevL.77.4887	**haroche**	**4/8**	PhysRev.74.1439	schwinger	1/1	PhysRev.82.159	fowlerquad	1/4
PhysRev.112.1940	charles	1/2	PhysRev.38.2021	lawrence	1/2	**PhysRevL.76.1800**	**haroche**	**2/7**
PhysRev.72.241	lamb	1/2	PhysRev.105.1924	dehmelt	1/1	PhysRevL.76.1796	wineland	1/5
PhysRev.60.356	landau	1/2	**PhysRevL.29.920**	lee	4/4	PhysRev.80.440	feynman	1/1
PhysRev.104.254	lee	2/2	PhysRev.74.939	feynman	1/1	PhysRev.40.749	wigner	1/1
PhysRevL.74.4043	**alferov**	**3/15**	**PhysRevL.45.494**	**klaus**	**2/3**	PhysRev.79.432	james	1/1
PhysRev.87.665	glaser	1/1	PhysRevL.57.2442	peter	1/7	**PhysRev.177.2075**	**kendall**	**5/8**
PhysRev.74.224	**tomonaga**	**2/2**	PhysRev.43.491	anderson	1/1			
				Overall Accuracy: 80/98				

Appendix B

Table 4

Table 4
Validation of *COCD* method in 34 prized papers of Max Planck Medal and Boltzmann Award. The first column is Digital Object Unique Identifier (*DOI*) of the papers. The second is real laureate of the paper listed in the same row. The ratio of Credit Rank calculated by *COCD* method and number of authors is shown in the third column. We denote them as *CR/N* for simplification. The papers that are identified wrong are highlighted in bold. We summary the overall identification accuracy in the last row.

Papers	Laureates	CR/N	Papers	Laureates	CR/N	Papers	Laureates	CR/N
PhysRevL.19.700	pierre	1/2	PhysRevB.64.045103	dieter	1/3	**RevModPhys.74.425**	**cohen**	**2/3**
PhysRev.130.1605	lieb	1/2	PhysRevL.82.1987	derrida	1/3	PhysRevL.43.744	parisi	1/2
PhysRevA.70.023612	**lieb**	**3/5**	PhysRevL.177.952	pierre	1/2	PhysRevL.77.4334	giovanni	1/1
RevModPhys.36.580	david	1/1	RevModPhys.68.1125	buras	1/3	PhysRevL.68.2269	giovanni	1/3
PhysRevL.80.209	glaser	1/2	PhysRevL.20.1445	lieb	1/2	**PhysRevL.61.2582**	**cohen**	**3/3**
PhysRevA.42.3664	**kyozi**	**2/2**	PhysRevE.79.031604	kurt	1/3	**PhysRevL.85.4438**	**mukhanov**	**2/3**
PhysRev.104.1528	lebowitz	1/2	PhysRevL.62.324	dieter	1/2	PhysRevL.73.613	detlev	1/2

Table 4 (*Continued*)

Papers	Laureates	CR/N	Papers	Laureates	CR/N	Papers	Laureates	CR/N
PhysRevB.38.2297	**stanley**	**2/4**	**PhysRevE.59.977**	**lebowitz**	**4/5**	PhysRevL.84.4882	spohn	1/2
PhysRevL.74.4091	**zoller**	**2/2**	PhysRevE.56.6540	kurt	1/3	**PhysRevL.60.673**	**martin**	**2/2**
PhysRevA.8.2048	kyozi	1/2	PhysRevD.18.3998	buras	1/4	PhysRevL.57.3148	martin	1/2
PhysRevL.81.3108	**cirac**	**2/6**	PhysRevL.81.3503	frhlich	1/3	**PhysRevL.77.4322**	giovanni	2/2
PhysRevL.45.366	lebowitz	1/4						
			Overall Accuracy: 23/34					

References

Bao, P., & Zhai, C. (2017). Dynamic credit allocation in scientific literature. *Scientometrics*, *112*(1), 595–606.

Bao, P., & Wang, J. (2020). Metapath-guided credit allocation for identifying representative works. *International world wide web conference committee*.

Carpenter, C. R., Cone, D. C., & Sarli, C. C. (2014). Using publication metrics to highlight academic productivity and research impact. *Academic Emergency Medicine*, *21*(10), 1160–1172.

Egghe, L., Rousseau, R., & Van Hooydonk, G. (2000). Methods for accrediting publications to authors or countries: Consequences for evaluation studies. *Journal of the Association for Information Science and Technology*, *51*(2), 145–157.

Falk-Krzesinski, H. J., et al. (2011). Mapping a research agenda for the science of team. *Research Evaluation*, *20*, 145–158.

Foulkes, W., & Neylon, N. (1996). Redefning authorship. Relative contribution should be given after each author's name. *British Medical Journal*, *312*(7043), 1423.

Guimerà, R., Uzzi, B., Spiro, J., & Lus, A. N. A. (2005). Team assembly mechanisms determine collaboration network structure and team performance. *United States: American Association for the Advancement of Science*, *308*(2), 697–702.

Hagen, N. T. (2008). Harmonic allocation of authorship credit: Source-level correction of bibliometric bias assures accurate publication and citation analysis. *PLoS One*, *3*(12), e4021.

Herz, N., Dan, O., Censor, N., & Bar-Haim, Y. (2020). Opinion: Authors overestimate their contribution to scientific work, demonstrating a strong bias. *Proceedings of the National Academy of Sciences*, *117*(12), 6282–6285.

Pavlidis, Ioannis, Petersen, Alexander M., & Semendeferi, Ioanna. (2014). Together we stand. *Nature Physics*, *10*(2), 700–702.

Juhász, S., Tóth, G., & Lengyel, Balázs. (2020). Brokering the core and the periphery: Creative success and collaboration networks in the film industry. *PLoS One*, *15*(2), e0229436.

Jung, S., & Yoon, W. C. (2019). *Citation-based author contribution measure for byline-independency*. IEEE.

Kim, J., & Diesner, J. (2014). A network-based approach to coauthorship credit allocation. *Scientometrics*, *101*(1), 587–602.

Kim, J., & Diesner, J. (2016). Distortive effects of initial-based name disambiguation on measurements of large-scale coauthorship networks. *Journal of the Association for Information Science and Technology*, *67*(6), 1446–1461.

Lehmann, S., Jackson, A. D., & Lautrup, B. E. (2006). Measures for measures. *Nature*, *444*(7122), 1003.

Li, J., Yin, Y., Fortunato, S., & Wang, D. (2019). A dataset of publication records for Nobel laureates. *Scientific Data*, *6*(1).

Li, J., Fortunato, Y., Yin, S., & Wang, D. (2019). Nobel laureates are almost the same as us. *Nature Reviews Physics*, *1*(5), 301–303.

Li, J., Fortunato, Y., Yin, S., & Wang, D. (2020). Scientific elite revisited: Patterns of productivity, collaboration, authorship and impact. *Journal of the Royal Society, Interface*, *17*(165), 20200135.

Martin, T., Ball, B., Karrer, B., & Newman, M. E. J. (2013). Coauthorship and citation patterns in the physical review. *Physical Review E*, *88*(1), 012814.

Milojević, S. (2014). Principles of scientifc research team formation and evolution. *Proceedings of the National Academy of Sciences, USA*, *111*, 3984–4398.

Mohammad Tariqur Rahman, A., J.M.R.B, & A., N.H. (2017). The need to quantify authors' relative intellectual contributions in a multiauthor paper. *Journal of Informetrics*, *11*(3), 275–281.

Radicchi, F., Fortunato, S., Markines, B., & Vespignani, A. (2009). Diffusion of scientific credits and the ranking of scientists. *Physical Review E: Statistical, Nonlinear, Biological, and Soft Matter Physics*, *80*(5 Pt 2), 56103.

Shen, H. W., & Barabási, A. L. (2014). Collective credit allocation in science. *Proceedings of the National Academy of Sciences*, *111*(34), 12325–12330.

Sinatra, Roberta, Wang, Dashun, Deville, Pierre, Song, Chaoming, & Barabási, Albert-László. (2016). Quantifying the evolution of individual scientific impact. *Science*, *354*(6312) https://doi.org/10.1126/science.aaf5239

Trueba, F. J., & Guerrero, H. (2004). A robust formula to credit authors for their publications. *Scientometrics*, *60*(2), 181–204.

Turki, H., Hadj Taieb, M. A., & Aouicha, Mohamed B. (2020). Facts to consider when analyzing the references of Nobel Prize scientific background. *Scientometrics*, *11*(3), 275–281.

Waltman, L., & van Eck, N. J. (2015). Field-normalized citation impact indicators and the choice of an appropriate counting method. *Journal of Informetrics*, *9*(4), 872–894.

Wang, F., Fan, Y., Zeng, A., & Di, Z. (2019). A nonlinear collective credit allocation in scientific publications. *Scientometrics*, *119*(3), 1655–1668.

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, *566*(7744), 378–382.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, *316*(5827), 1036–1039.

Zeng, A., Shen, Z., Zhou, J., Wu, J., Fan, Y., Wang, Y., et al. (2017). The science of science: From the perspective of complex systems. *Physics Reports*, *714*, 1–73.

Zeng, A., Shen, Z., Zhou, J., Fan, Y., Di, Z., Wang, Y., Havlin, Shlomo, et al. (2019). Increasing trend of scientists to switch between topics. *Nature Communications*, *10*, 3439.