IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 23, 2019, accepted December 1, 2019, date of publication December 10, 2019,
date of current version December 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2958649

API: An Index for Quantifying a Scholar's Academic Potential

JING REN[1], LEI WANG[1], KAILAI WANG[1], SHUO YU[1], MINGLIANG HOU[1], IVAN LEE[2], (Senior Member, IEEE), XIANGJIE KONG[1], (Senior Member, IEEE), AND FENG XIA[1,3], (Senior Member, IEEE)

[1]School of Software, Dalian University of Technology, Dalian 116620, China
[2]School of Information Technology and Mathematical Sciences, The University of South Australia, Adelaide, SA 5001, Australia
[3]School of Science, Engineering and Information Technology, Federation University Australia, Ballarat, VIC 3353, Australia

Corresponding author: Xiangjie Kong (xjkong@ieee.org)

This work was supported in part by the National Natural Science Foundation of China under Grant 61872054, and in part by the Fundamental Research Funds for the Central Universities under Grant DUT19LAB23 and Grant DUT18JC09.

ABSTRACT In the context of big scholarly data, various metrics and indicators have been widely applied to evaluate the impact of scholars from different perspectives, such as publication counts, citations, *h*-index, and their variants. However, these indicators have limited capacity in characterizing prospective impacts or achievements of scholars. To solve this problem, we propose the *Academic Potential Index* (*API*) to quantify scholar's academic potential. Furthermore, an algorithm is devised to calculate the value of *API*. It should be noted that *API* is a dynamic index throughout scholar's academic career. By applying *API* to rank scholars, we can identify scholars who show their academic potentials during the early academic careers. With extensive experiments conducted based on the Microsoft Academic Graph dataset, it can be found that the proposed index evaluates scholars' academic potentials effectively and captures the variation tendency of their academic impacts. Besides, we also apply this index to identify rising stars in academia. Experimental results show that the proposed *API* can achieve superior performance in identifying potential scholars compared with three baseline methods.

INDEX TERMS Scholarly big data, scholarly data analysis, academic potential, rising stars.

I. INTRODUCTION

There are many perspectives to profile a scholar, such as academic age, research field and academic reputation. In recent years, many methods have been proposed to evaluate scholars' academic impacts. It is widely recognized that the number of publications or citations is the most direct and simple way to reflect a scholar's productivity and achievement [1]. In spite of its simplicity and popularity, there are limitations on adopting citation count as the only indicator for assessing scholar's influence [2]. For example, senior scholars tend to have more citations than juniors. Besides, previous studies mainly focus on quantifying scholars' present impacts, without reflecting future potentials. Taking the well-known *h*-index [3] as an example, the index only considers the quantity of papers in terms of citations, while ignoring scholars' contributions to these papers and the quality of citations. For instance, being cited by scholars with high academic reputation is likely more significant than by ordinary individuals.

Extensive studies have been conducted on academic impact assessment, which can be used for job matching, collaborator recommendation, and research funding allocation, etc. While a wide range of indicators have been proposed to evaluate scholar's academic impact or prestige, it is significant to propose an indicator reflecting scholar's academic potential, for both scholars themselves and academic institutions. Scholars' academic potentials are reflected by predicting the papers' qualities they published in the future, which is distinguished from indicators for evaluating scholar's current impact. In other words, the higher the quality of published papers in the future, the higher the academic potential of scholars. Evaluating academic potential can be applied in many situations, such as research funding, advisor recommendation and position evaluation. Therefore, this paper proposes *Academic Potential Index (API)* to quantify scholars' academic potential.

The associate editor coordinating the review of this manuscript and approving it for publication was Vlad Diaconita.

According to Tessa Lansu, a psychologist from Radboud University Nijmegen [4], "*popularity has to do with being the middle point of a group and having influence on it.*" The literal meaning of activeness means that a person is engaged in active work, which can also be described as "liveliness," "briskness," or "diligence". Motivated by [5], the indicator popularity in this paper is related to the citations a scholar has received. Similarly, activeness corresponds to the behavior of citing others. Based on this concept, several metrics are considered to quantify scholar's academic potential in terms of papers and citations. First, scholar's contribution to their papers and the impact of theses papers are considered as their academic achievements, which are used to initialize the value of popularity and activeness. Then, citations are used to quantify the degree that a scholar attracts others in the academic social network. Therefore, based on the scholar-citation network, we propose an indicator named *Attractiveness* to calculate the value of influence of each scholar on others. *Attractiveness* between any two scholars is related to the following three factors: popularity, activeness, and the similarity between their research fields. Based on a well-known algorithm named HITS [6], we present scholar's popularity analogous to the authority, and activeness to the hub in HITS. The similarity between scholars' research fields can be quantified by the relative positions in author-citation network. Finally, the value of a scholar's *API* can be calculated by his/her *Attractiveness* and others' *h*-indices.

API is a dynamic value throughout a scholar's academic career, which is different from year to year. Besides, another advantage of *API* is that it is not affected by the academic age of scholars. This characteristic makes the proposed algorithm universally applicable to assess academic potential of scholars at any scientific stages. From the perspective of input-output ratio, we can conclude that scholars who have higher academic potentials are worthy of cultivating and investing, especially at the beginning of their careers. As a consequence, it will be significant to profile a scholar's academic potential, which is not affected by his/her academic age.

The contributions of this paper includes:

1) Propose a new indicator named *Academic Potential Index* (*API*) to quantify scholar's academic potential regardless of academic age.
2) Devise an algorithm to calculate the proposed *API*, which can be easily implemented on any academic social networks.
3) Examine the effectiveness of the proposed algorithm in different academic fields.
4) Identify academic rising stars using the proposed algorithm.

The rest of this paper is structured as follows. Section II presents some related studies and Section III describes the framework of the proposed algorithm. Experimental results are introduced in Section IV. Finally, we conclude this work in Section V.

II. RELATED WORK

Recent years have witnessed the development of applying academic social network to evaluate the impact of academic entities, such as authors, publications and journals [7]–[9]. As for ranking of authors, the methods based on the structure of academic social network can be divided into iterative methods and non-iterative methods [10].

A. ITERATIVE METHODS

Most iterative methods mainly extend PageRank and HITS algorithms to evaluate the impact of academic entities [7], [11], [12]. Initially, PageRank [13] is an iterative algorithm used by search engine, which can rank webpages according to their importance. Subsequently, these algorithms have been applied to evaluate scientific impact in many studies. By modifying the PageRank algorithms, papers can be more suitable to be ranked in various academic networks [2]. In order to find milestone papers, Mariani *et al.* [14] proposed a metric by combining PageRank centrality with the explicit requirement that paper score is not biased by paper age. Zhang *et al.* [15] proposed a topic-dependent model to evaluate academic impact of scientific papers. According to the citation relationships of papers and authors, Zhou *et al.* [16] constructed a directed author–paper interactive bipartite network. They propose an iterative algorithm to quantify the scientists' reputation and the quality of their publications via their inter-relationship on this network. Besides, Bai *et al.* [17] developed a higher-order weighted quantum PageRank algorithm to quantify the impact of scholarly papers.

B. NON-ITERATIVE METHODS

For non-iterative methods, scholars are usually ranked by considering topological features. With the popularity of using impact factor to evaluate the authority and popularity of journals [18], [19], using impact factors to quantify the impact of scholars is gradually emerging. For example, Pan and Fortunato [20] proposed a dynamic indicator named author impact factor (AIF) to evaluate the scholars' impact currently, which is the extension of the impact factor [21] to authors. Among all the statistic analysis of evaluating academic impact, both homogeneous and heterogeneous academic networks provide a simple and direct way to identify the influential and popular scholars. Zhu *et al.* [22] evaluated scholar's scholarly impact with diverse racial/ethnic groups. Recently, by analyzing the highly-cited papers in different kinds of journal, Antonoyiannakis [23] analyzed the effect that a single paper has on the impact factor of this journal.

III. DESIGN OF API

This section first introduces the networks used in the experiments. Moreover, the framework of calculating the scholars' *API* and relevant variables are presented.

A. CONSTRUCTION OF THE SCHOLAR-CITATION NETWORK

By regarding scholars as nodes and the citation relationship between scholars as edges, a weighted and direct network $G = (V, E)$ can be derived, where V is the set of scholars and $E \subseteq V \times V$ is the set of relationships. Specifically, the directed edge from scholar i to scholar j denotes that i has cited j's publications and the weight w_{ij} denotes the times scholar i has cited scholar j.

B. CALCULATION OF RELEVANT VARIABLES

1) CALCULATION OF SCHOLAR'S POPULARITY AND ACTIVENESS

Scholar's popularity and activeness are key factors that evaluate *Attractiveness* between two scholars. The calculation formula of popularity is defined as

$$pop_k = \sum_{i \in A_k} \frac{w_{ik}}{\sum_{l \in B_i} w_{il}} \cdot act_i, \tag{1}$$

where A_k and B_i are two sets of nodes whose elements are the nodes pointing to k and the nodes pointed by i respectively. w_{ik} is the number of scholar i having cited scholar k and $\sum_{l \in B_i} w_{il}$ is the total number of scholar i having cited others. The calculation formula of activeness is defined as

$$act_k = \sum_{i \in B_k} \frac{w_{ki}}{\sum_{l \in A_i} w_{li}} \cdot pop_i, \tag{2}$$

where B_k and A_i are two sets of nodes whose elements are the nodes pointed by k and the nodes pointing to i respectively. w_{ki} is the number of scholar k having cited scholar i and w_{li} is the total number of scholar i having been cited by others. These two equations are inspired from HITs [6].

As the above two equations alternately operate, pop_k and act_k converge to fixed values gradually.

2) INITIALIZATION OF SCHOLAR'S POPULARITY AND ACTIVENESS

In the process of initializing the popularity and activeness of scholars, we choose the paper-citation network, which is different from scholar-citation network. This is a single-directed, acyclic and unweighted network, where the nodes represent papers and arrows represent reference relationships.

The initialization of scholar k's popularity and activeness are both calculated by Eq.3 and get the same original value represented by R_k uniformly. Based on the paper-citation network, some related factors of every publication i in paper collection P_k of scholar k are used: the contribution of scholar k in paper i, which is denoted by $ORD_{k,i}$; the rank of paper i in all articles of scholar k, which is denoted by PR_i and the journal impact factor of the journal in which the paper i is published, which is denoted by J_i. The calculation formula of R_k and related factors are defined as follows:

$$R_k = \sum_{i \in P_k} (ORD_{k,i} \cdot PR_i \cdot J_i) \tag{3}$$

$$ORD_{k,i} = \frac{1/r_{k,i}}{\sum_{j=1}^{n_i} 1/j} \tag{4}$$

$$PR_i = \frac{1-s}{N} + s \sum_{j \in C_i} \frac{PR_j}{O_j} \tag{5}$$

$$J_i = \frac{Cit_{y-1} + Cit_{y-2}}{Pub_{y-1} + Pub_{y-2}}. \tag{6}$$

In the above definitions of variables, we denote $r_{k,i}$ and n_i as the order of scholar k in paper i and the total number of the authors in paper i respectively [24]. Inspired by PageRank [13], [25] for bibliographic networks, we use Eq.5 to calculate the value of PR_i, denoting $s = 0.85$ as a constant damping factor. N is the total number of nodes in the paper-citation network, j is one node of the nodes set C_i who has cited paper i and O_j is the outdegree of node j. In the process of calculating PR_i, we first initialize the value of PR_i of every paper i, and get the preliminary value of PR_i through many times of iterations. It should be noted that the preliminary value of PR_i needs to be normalized at every step. Besides, the value of J_i in every year can be calculated with Eq.6 [21], where y is the year of publication i. Cit_{y-1} and Cit_{y-2} is the citation counts journal received in year $y-1$ and $y-2$ respectively, and Pub_{y-1} and Pub_{y-2} is the number of publications journal published in year $y-1$ and $y-2$ respectively.

3) CALCULATION OF SIMILARITY BETWEEN TWO SCHOLARS

The value of sim_{ij} suggests the similarity of the research field between two scholars, which is also an important factor that influences the *Attractiveness* between them. The way of calculating similarity between two scholars has mainly two perspectives, i.e. publications and locations in a network. From the perspective of publications, we extract the abstract information by doc2vec [26]. By concatenating all the abstracts of the papers as the vector of the scholar, we calculate the cosine values of these two scholars' vectors as their similarity. From the perspective of locations in networks, we utilize a method called Jaccard Similarity Coefficient [27]. The calculation of similarity between scholars is also based on the scholar-citation network. The final value of these two methods is between 0 and 1, and the closer this final value is to 1, the higher similarity these two scholars have. In this paper, Pearson Correlation Coefficient between the value of *API* and the citation counts accumulated in the next 5, 10 and 15 years is selected as the metric to measure the performance of different similarity methods. Besides, running time is another important evaluation metric for comparison.

As the final experimental results shown in Fig. 1a and 1b, it can be demonstrated that Jaccard Similarity Coefficient can greatly reduce the running time, with similar performance compared with cosine similarity. The method of cosine similarity needs all abstract information of the papers between each two scholars, which took up lots of storage space and consumed abundant operation times. The calculation of Jaccard Similarity Coefficient only needs the common neighbors

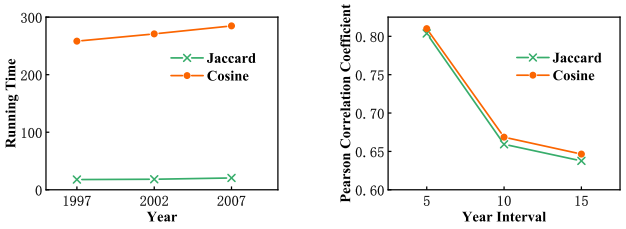


FIGURE 1. (a) The running time of two different similarity algorithms for calculating the similarity between two scholars in year 1997, 2002 and 2007. (b)The Pearson Correlation Coefficient between two different similarity algorithms and total citations in the next 5, 10, 15 years.

between two scholars in a network, so the similarity between two scholars whose shortest path is greater than 2 is 0. This characteristic can reduce the quantity of calculation to a great extent. Therefore, Jaccard Similarity Coefficient is finally selected as the method of calculating similarity between two scholars.

C. ATTRACTIVENESS

Attractiveness is an indicator measuring scholars' ability of attracting citations from others. Every scholar will be attractive to all other scholars, and the *Attractiveness* from scholar i to scholar j is totally different from that from scholar j to scholar i, which is denoted by At_{ij} and At_{ji} respectively. The equation of *Attractiveness* At_{ij} from scholar i to scholar j is defined as

$$At_{ij} = pop_i \cdot act_j \cdot sim_{ij}. \tag{7}$$

D. ACADEMIC POTENTIAL INDEX

Academic Potential Index is an author-level index for measuring scholars' academic potential. The index is dynamic from year to year. The specific computational formula of *API* is defined as

$$API_i = \sum_{j \in S_i} (At_{ij} \cdot h_j), \tag{8}$$

where scholar j belongs to S_i, the set of all remaining scholars in the scholar-citation network except scholar i. At_{ij} is the *Attractiveness* from scholar i to scholar j, and h_j is the h-index of scholar j.

Algorithm 1 presents the process of calculating API_i. pop_i and act_i were initialized with R_i first (Line 1), and their values were calculated iteratively according to Eq. 1 and Eq. 2 until the convergence condition is satisfied (Line 2, 4 and 7). Then, for every two scholars i and j, we can get the *Attractiveness* between them in Eq. 7 (Line 12). Finally, API_i can be calculated according to Eq.8 (Line 15). Besides, the flow chart of our algorithm is shown in Fig. 2.

IV. EXPERIMENTS

In this section, we elaborate on the dataset adopted in our experiments and some filtered conditions in detail. The program language used in the process of data cleasing is Python, and the software is PyCharm. Besides, experimental settings

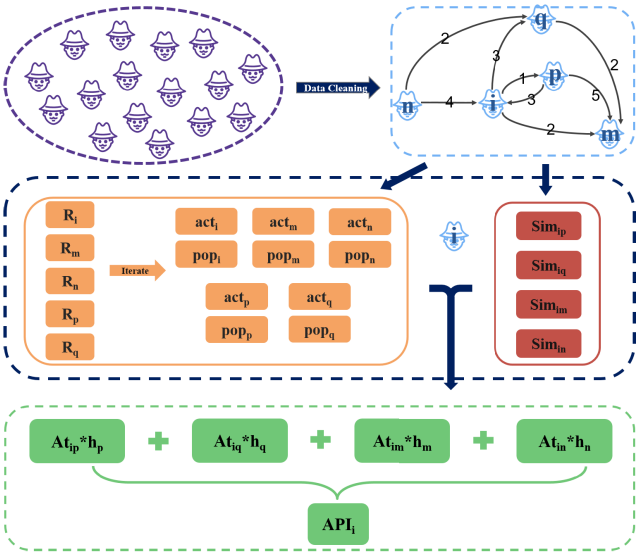


FIGURE 2. The flow chart illustrating the execution of the algorithm for calculating scholar *i*'s *API*.

Algorithm 1 Potential Scholars Mining Algorithm

Require: initial value R_i for each node $i \in V$, similarity sim_{ij} for each pair $(i, j) \in E$, convergence condition: relative error $< 10^{-4}$

Ensure: API_i for $i \in V$

1: $pop_i = act_i = R_i$ for each node $i \in V$
2: **while** convergence condition is not satisfied: **do**
3: **for** each node $i \in V$ **do**
4: calculate pop_i according to Eq. 1
5: **end for**
6: **for** each node $i \in V$ **do**
7: calculate act_i according to Eq. 2
8: **end for**
9: **end while**
10: **for** each node $i \in V$ **do**
11: **for** each node $j \in V$ **do**
12: calculate At_{ij} according to Eq. 7
13: **end for**
14: **end for**
15: calculate API_i for each node $i \in V$ according to Eq. 8

in each experiment are also described separately. Finally, the experimental results in each field are presented.

A. DATASET

In this paper, we choose scholars from the Microsoft Academic Graph (MAG) dataset. It covers not only information of scholars and papers from various fields, but also detailed matching relationships such as author-paper, paper-field, paper-venue. Besides, MAG classified the papers into 6 levels, and each level contains different number of subfields. More detailed information about the field division is shown in Table 1. Table 1 shows the number of subfields in

TABLE 1. Detailed information of field division in MAG.

Field Level	1	2	3	4	5	6
art	1	7	3921	1351	306	104
biology	1	32	47735	28753	11571	2407
business	1	13	2381	1411	623	228
chemistry	1	21	14940	12384	4349	909
computer science	1	34	12737	7028	2784	1301
economics	1	40	5483	3316	1091	361
engineering	1	44	9941	6884	2281	750
environmental science	1	8	1139	1158	254	38
geography	1	11	4630	2727	673	207
geology	1	18	15658	7578	3156	542
history	1	7	3532	1643	307	82
materials science	1	7	3380	3940	824	119
mathematics	1	20	10060	5658	2438	1063
medicine	1	47	18135	15936	5817	1238
philosophy	1	7	5578	2368	539	136
physics	1	27	10569	8777	3049	934
political science	1	3	4551	3412	822	224
psychology	1	14	6194	3885	1113	179
sociology	1	13	3502	1423	257	39

different levels of 19 fields. Considering that the scale of this dataset is large, we first select scholars from three fields: computer science, biology and psychology. The technical span of this three areas are relative large, and their citation behaviors are different. Therefore, the experimental results in these three fields can cover most scholars in a large part. Because each scholar in the latest MAG dataset has been assigned with a unique ID number, the problem of name disambiguation can, to a great extent, be eliminated. Due to the consideration of journal impact factor (JIF), all conference publications are excluded.

In these three fields, we choose scholars who published their first paper in 1970. In order to find scholars who keep researching in academic for a long time, we select scholars whose academic age at least 10, and the number of papers and citations are more than 20. This process of data cleaning can filter schoalrs who have published few papers, and those who publish papers only for master's or bachelor's degrees. After filtering the scholars whose values of *API* are equal to 0, we finally keep 878 scholars in computer science, 3515 in biology and 907 in psychology.

B. EXPERIMENTAL SETTINGS

1) DEMONSTRATION OF ACADEMIC POTENTIAL

Since variables such as *pop* and *act* are normalized during the calculation process, the preliminary value of the *API* is so tiny that is indistinguishable to compare. Therefore, the preliminary values are first rescaled into a number between 0 and 100 uniformly, and then take the logarithm of them, which makes it explicit to observe the relationship between scholars' *API* in every year.

Naturally, as the number of citations changes with the number of papers and the time of publication, the *API* of scholars will change over time. In other words, the *API* of a scholar is a dynamic indicator whose value is different from year to year. Accordingly, after filtering out scholars whose value of *API* equal to 0 (scholars who are not cited by others), the values of *API* of the rest were analyzed. In order to highlight the characteristic of *API*, we compare the distribution of the *h*-index [3] and Author Impact Factor (AIF) [20] of these qualified scholars during 1970-2017. *H*-index is a common metric evaluating the scholar's academic achievement, which is calculated as scholar has h publications with at least h citations in each publication. AIF is an extentinon of Journal Imapct Factor, whose equation in shown in Eq. 6. This indicator can only charaterize the current academic impacts, instead of potentials. Among the experiments, scholars with high journal impact factor (JIF) of their papers in the next 5 years are regarded as high potentials. Therefore, a curve on the actual scatter diagram was fitted, and an indicator R^2 was introduced to quantify the error between the fitted curve and the actual distribution. R^2 is the denotation of the coefficient of determination ranging from 0 to 1. The closer R^2 to 1, the better the fitted regression equation.

TABLE 2. Comparison of Pearson Correlation Coefficient between each indicator and the citation counts in the next # years.

Time Interval	AIF	*h*-index	*API*
5	0.47208	0.63962	**0.80370**
10	0.41122	0.57867	**0.65928**
15	0.37867	0.52392	**0.63769**

It is considered that the higher the quality of papers published in the next few years, the higher academic potential scholars have. In order to verify the inference obtained above, the Pearson Correlation Coefficient was further calculated to measure the correlation between the citation counts and the above methods (*h*-index and AIF) during different years for comparison. The value of Pearson Correlation Coefficient varies from -1 to 1 with correlation ranging from total negative linear to total positive linear. As shown in Table 2, the value of *API* in each time interval is the highest among the three methods. Therefore, it is easy to distinguish *API* from *h*-index and AIF. Our proposed indicator can largely reflect the scholars' academic potential in the next few years, whereas other methods cannot.

Moreover, an indicator was provided to quantify highly-cited papers whose citations exceed Y. The equation of calculating indicator Y is defined in Eq. 9, where i denotes a scholar in scholar set N, and P_i denotes the set of scholar i's papers whose citations should in top 30% throughout the first 15 years of his academic career path. k is one of the papers in set P_i, and C_{ik} denotes the citation of paper k received by scholar i, $|P_i|$ denoting the number of papers in set P_i and $|N|$ denoting the number of scholars in set N.

$$Y = \frac{\sum_{i \in N} \frac{\sum_{k \in P_i} C_{ik}}{|P_i|}}{|N|} \tag{9}$$

In general, the total number of scholars' citations will accumulate with the growing of academic age [28]. Therefore, the academic ages of all scholars are obtained and the relationship between scholars' age and their value of *API* is analyzed.

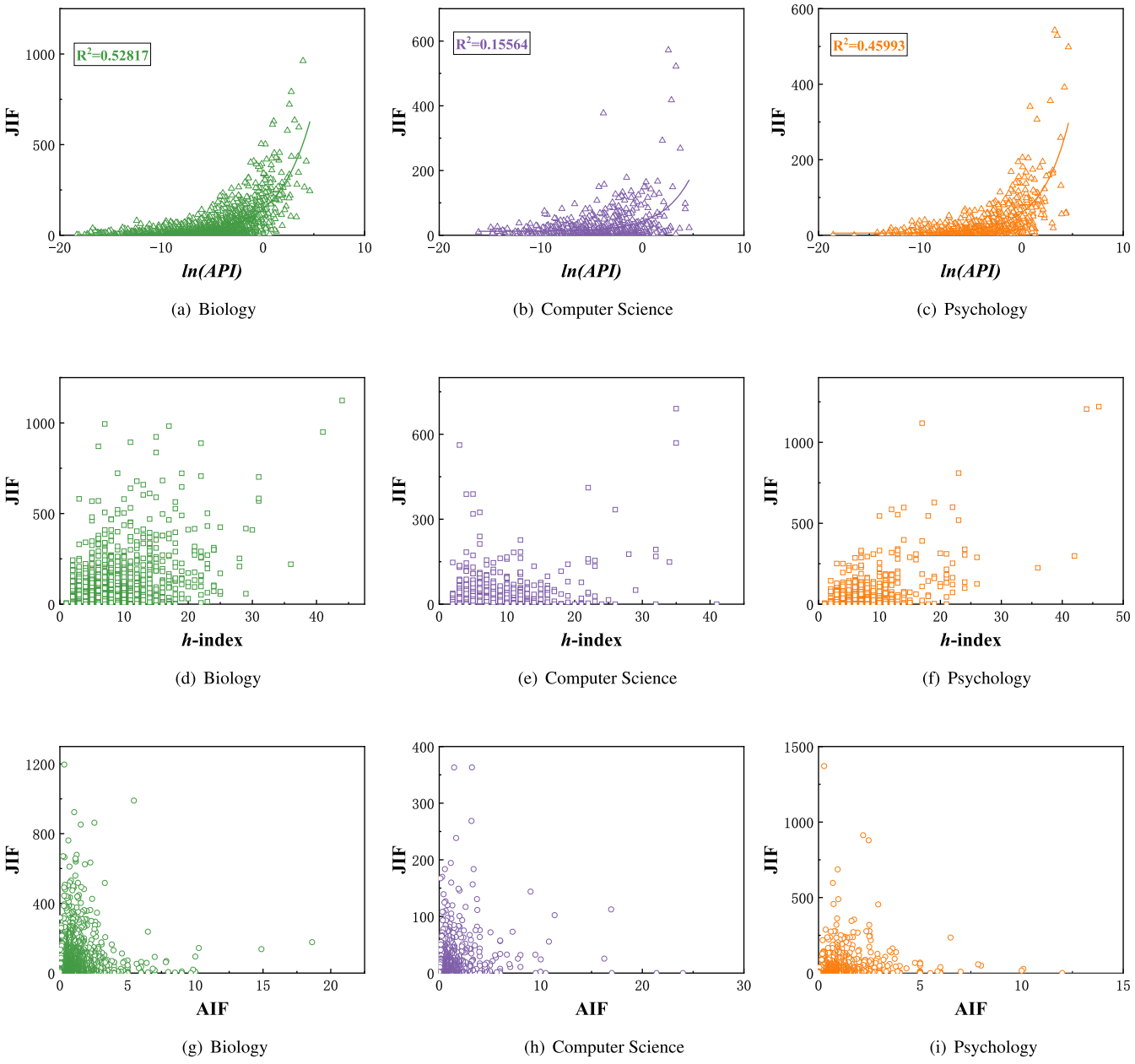


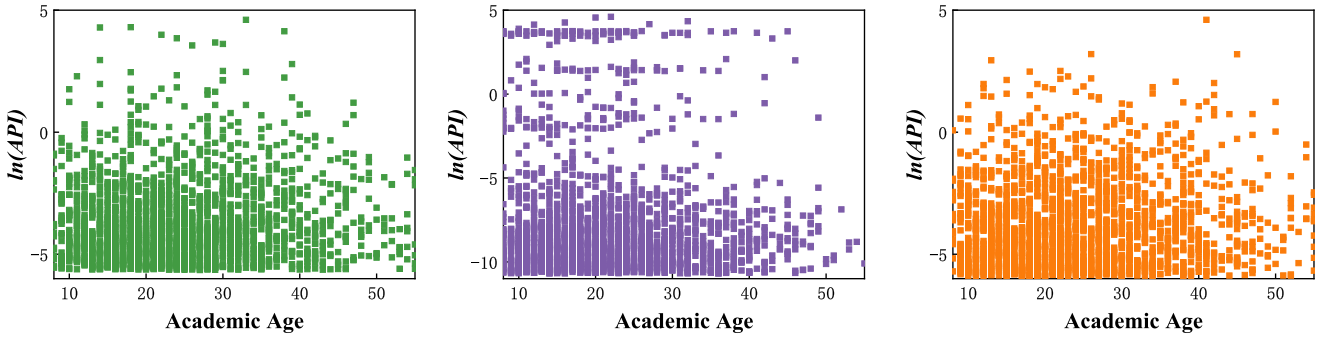
FIGURE 3. Distribution of the journal impact factor (JIF) scholar has received till now against the logarithm to base e of *API*, *h*-index and AIF of all scholars in biology, computer science and psychology.

Finally, to compare the scholars with different *API*, the scholars are divided into three categories according to their value of *API* in 2002: low *API* (bottom 70%, $0 < A_i < 1$), medium *API* (middle 20%, $1 < A_i < 10$), and high *API* (top 10%, $10 < A_i < 100$).

2) DISTRIBUTION OF ACADEMIC POTENTIAL INDEX

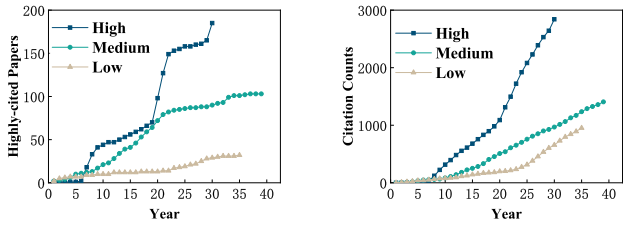
As the *API*s of scholars vary from year to year, the values of their *API*s during their academic careers from 1970 to 2017 are calculated to observe the variation trends. The maximum value of their *API*s from 1970 to 2017 is selected as the representative value, and the distribution of scholars with different values of *API* can be observed.

Despite that scholars all start their academic careers in 1970, their *API*s start (unequal to 0) from different years. The first year of having their nonzero values of *API* can be simplified as the term "first-*API* year" in the later. If a scholar's *API* is equal to 0 in a year, it denotes that he/she has not shown his/her potential in that year. As for the scholars whose first-*API* year is in their later academic career, it is considered that these scholars are inactive and unattractive. By classifying the scholars according to the first-*API* year, the scholars are divided into two categories:



(a) Biology (b) Computer Science (c) Psychology

FIGURE 4. Distribution of the different *API* across scholars of all academic ages.



(a) Number of highly-citted papers (b) Number of citations

FIGURE 5. (a) Numbers of highly-cited papers with high, medium and low *API* scholars across their whole academic career path. (b) Citation counts received by high, medium and low *API* scholars across their whole academic career path.

one category consists of scholars whose first-*API* year are between 1971 and 2000, and the other includes the remainings whose first-*API* year are after 2000. In order to analyze most representative scholars, this research mainly focuses on the scholars in the first category.

Due to the fact that different scholars vary greatly in the values of *API*, from 10^{-15} to 10^{-2}, the logarithm of the final value of *API* is taken. Then, statistics are done on the number of scholars with different values, and it is found that these scholars follow a normal distribution.

3) IDENTIFICATION OF ACADEMIC RISING STARS

To further prove the validity of the proposed index, *API* is applied to identify the academic rising stars. In order to evaluate the performance of our proposed method, we compare the performances of the state-of-the-art methods, which are chosen as the baseline methods for comparison. The details of the above methods are as follows:

- StarRank [29]. This method finds rising stars based on authors' contribution oriented mutual influence and dynamic publication venue scores.
- CocaRank [25]. This hybrid method integrates both the statistical indicators and the topological features to calculate the impact of scholars. The relevant factors include: the value of Coca, citation counts and the importance calculation results on heterogeneous academic networks.
- ScholarRank [30]. This method considers three factors: the citation counts of authors, the mutual influence among coauthors and the mutual reinforce process among different entities in heterogeneous academic networks.

First, scholars who start their academic careers in 1970 are selected and the rank of these four methods in 1974 is obtained. The citation counts and the number of published papers during 1990-2000 were calculated. Then, according to the rank of citation counts and paper numbers, we choose top 5%, 10%, and 20% scholars in each rank. After selecting these top scholars in different ranks, we take the intersection of sets composed by different methods and citation counts sets, paper number sets respectively. The more scholars in the intersection, the more academic rising stars can be found.

C. RESULTS AND ANALYSIS

1) DEMONSTRATION OF ACADEMIC POTENTIAL

As shown in Figs. 3a, 3b, 3c, the distribution between journal impact factor and *API* all follow power law functions with R^2 approximately equaling to 0.53, 0.16, and 0.46 in biology, computer science, and psychology respectively. However, the points in Figs. 3d, 3e, 3f and Figs. 3g, 3h, 3i are randomly scattered. It is indicated from the results that scholars who have relative high values of *API* will get more citations in the near future, maybe more later. While the value of *h*-index and AIF cannot reflect their academic impacts in the future.

In Fig. 4, the scholars' *API* against their academic ages were plotted, where the points are evenly scattered in biology, computer science and psychology. This phenomenon indicates that the algorithm is equitable and cannot be affected by scholars' academic age and their total number of citations.

Finally, according to the classification of the value of *API*, we randomly choose one scholar as a representative from each of the three classifications, namely low, medium, and high *API*. Then, the variation tendency of the number of highly-cited papers, and citation counts of scholar *i* against

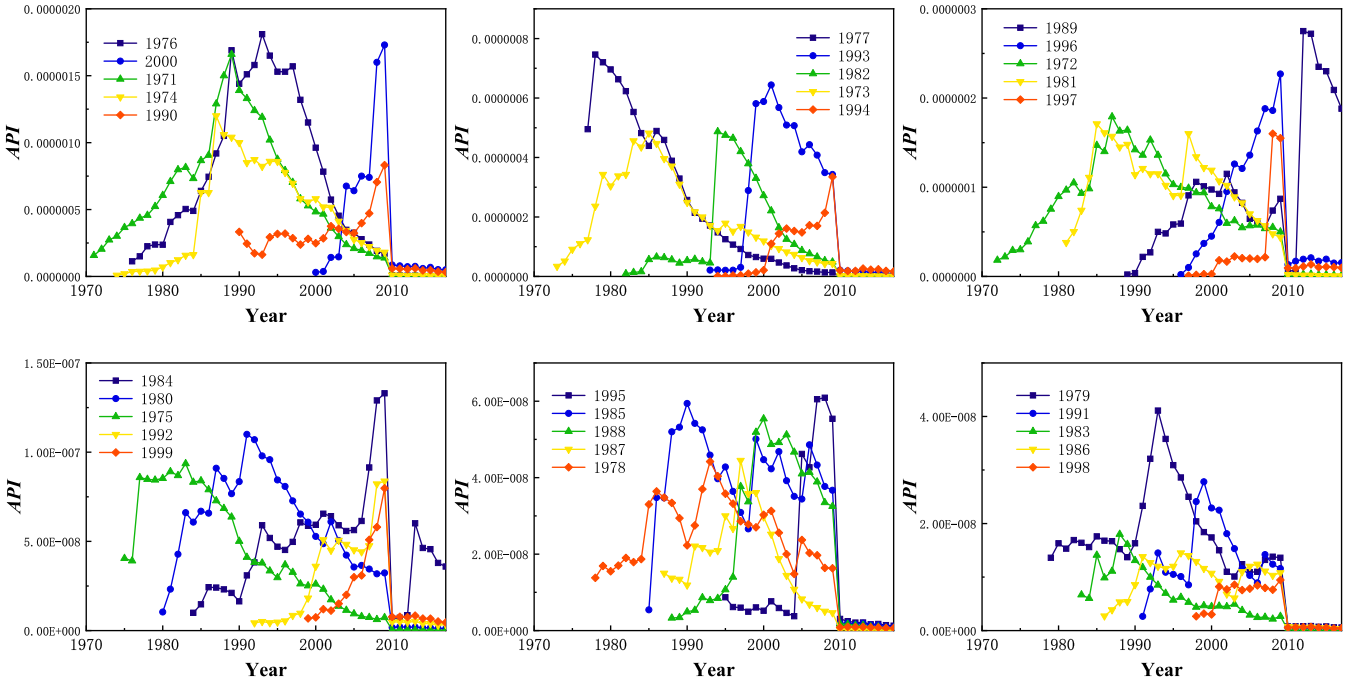


FIGURE 6. The variation trend of scholar's *API* from the first-*API* year to 2017. The scholars are classified by their first-*API* year. The year showed in the legend is the first-*API* year.

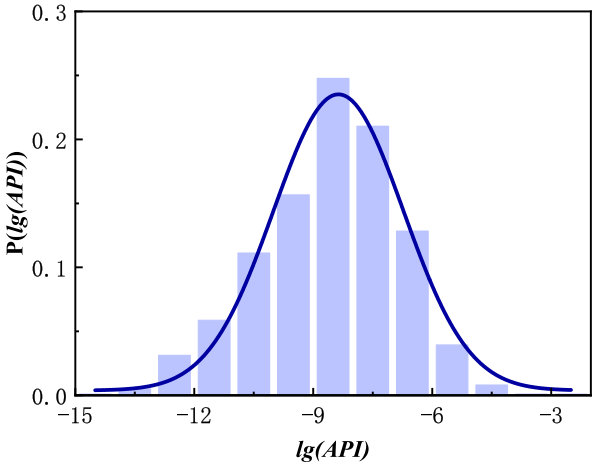


FIGURE 7. The distribution of the logarithm to base 10 of scholars' *API* in computer science, which obeys normal distribution with $\mu = -8.365$ and $\sigma = 1.639$.

scholar's academic age t are plotted, respectively. The experimental results show that if a scholar has relatively high *API*, the productivity and quality of papers are generally high (see Fig. 5).

2) DISTRIBUTION OF API

From the statistical analysis on scholars who have their first-*API* year from 1971 to 2000, it can be concluded that the value of *API* is almost irrelevant to their first-*API* year (see Fig. 6). This is because scholars with similar value range of *API* are placed in a subfigure. Besides, another interesting phenomenon is that most scholars' values of *API* show a rising tendency in the first few years, and then falls later, despite that they have different first-*API*s years. In Fig. 6, we can see clearly that nearly all curves of scholar's *API* follow normal distribution in general.

In Fig. 7, it is found that the number of scholars with different values of *API* obey normal distribution, where the expectation μ and standard deviation σ are approximately equal to -8.365 and 1.639 respectively.

3) IDENTIFICATION OF ACADEMIC RISING STARS

In the end, by applying *API* into the identification of academic rising stars, the experimental results on real datasets demonstrate that our method can find more rising stars in each percent of top, compared with other three baseline methods (see Figs. 8). Therefore, the results can sufficiently demonstrate that the proposed indicator *API* can be applied into many

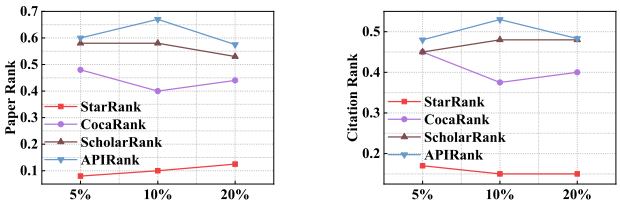


FIGURE 8. (a) The number of top 5%, 10% and 20% scholars of the intersection ranked by four different methods by the number of published papers. (b) The number of top 5%, 10% and 20% scholars of the intersection ranked by four different methods by citation counts.

other practical applications, such as award evaluation, agency employment and allocation of research funds.

V. CONCLUSION

In this paper, we have proposed *API* to profile scholars from the perspective of *Attractiveness*. Unlike conventional indicators that only reflect the achievement and impact of scholars in the past, *API* is capable of quantifying scholars' potential in the near future. Also, *API* is independent to the academic age of scholars. Experimental results on the MAG dataset demonstrate that the proposed algorithm can effectively quantify the academic potential of scholars. By applying the proposed index *API* to the identification of academic rising stars, it has shown up to 10% improvement in performance compared with other baseline methods. Therefore, our algorithm is particular useful for funding agencies, peer reviewers, and hiring committees who have to deal with a wide range of applicants.

In order to further verify the universality of *API*, several heterogeneous academic social networks will be applied to our algorithm, such as paper-author network. As for applications in the real world, *API* will be applied to recommendation of academic collaborators and scholar-institution matching in the future.

APPENDIX

AN EXAMPLE OF CALCULATING API

To better understand the process of our algorithm, we construct a scholar-citation network with only five scholars (shown in Fig. 2). For simpicity, values of *pop* and *act* are initialized as 1 for each scholar. Final values of *pop* and *act* for each scholar are calculated as:

$$pop_i = \frac{3}{8} \cdot act_p + \frac{4}{6} \cdot act_n, \tag{10}$$

$$pop_m = \frac{2}{6} \cdot act_i + \frac{5}{8} \cdot act_p + act_q, \tag{11}$$

$$pop_n = 0, \tag{12}$$

$$pop_p = \frac{1}{6} \cdot act_i, \tag{13}$$

$$pop_q = \frac{2}{6} \cdot act_n + \frac{3}{6} \cdot act_i, \tag{14}$$

$$act_i = \frac{3}{5} \cdot pop_q + \frac{2}{9} \cdot pop_m + pop_p, \tag{15}$$

$$act_m = 0, \tag{16}$$

$$act_n = \frac{2}{5} \cdot pop_q + \frac{4}{7} \cdot pop_i, \tag{17}$$

$$act_p = \frac{3}{7} \cdot pop_i + \frac{5}{9} \cdot pop_m, \tag{18}$$

$$act_q = \frac{2}{9} \cdot pop_m. \tag{19}$$

After several iterations, $pop_i = 0.318$, and $act_m = 0$, $act_n = 0.273$, $act_p = 0.364$, $act_q = 0.091$. Then, we let sim_{im}, sim_{in}, sim_{ip}, sim_{iq} equal to 0.3, 0.5, 0.1 and 0.7. According to Eq. 7, Values of $At_{im}, At_{in}, At_{ip}, At_{iq}$ can be calculated as:

$$At_{im} = pop_i \cdot act_m \cdot sim_{im}, \tag{20}$$

$$At_{in} = pop_i \cdot act_n \cdot sim_{in}, \tag{21}$$

$$At_{ip} = pop_i \cdot act_p \cdot sim_{ip}, \tag{22}$$

$$At_{iq} = pop_i \cdot act_q \cdot sim_{iq}. \tag{23}$$

Then, we can get $At_{im} = 0$, $At_{in} = 0.043$, $At_{ip} = 0.012$ and $At_{iq} = 0.02$. If we want to calculate the API of scholar i, the specific calculation fomular is:

$$API_i = At_{im} \cdot h_m + At_{in} \cdot h_n + At_{ip} \cdot h_p + At_{iq} \cdot h_q. \tag{24}$$

Assuming that h_m, h_n, h_p, h_q equal to 5, 7, 20, 12, respectively. Finally, we can calculate that $API_i = 0.778$.

REFERENCES

[1] H. F. Moed, *Citation Analysis in Research Evaluation*, vol. 9. Berlin, Germany: Springer, 2006.

[2] A. Zeng, Z. Shen, J. Zhou, J. Wu, Y. Fan, Y. Wang, and H. E. Stanley, "The science of science: From the perspective of complex systems," *Phys. Rep.*, vol. 714, no. 715, pp. 1–73, 2017.

[3] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proc. Nat. Acad. Sci. USA*, vol. 102, no. 46, pp. 16569–16572, 2005.

[4] T. A. Lansu and A. H. Cillessen, "Peer status in emerging adulthood: Associations of popularity and preference with social roles and behavior," *J. Adolescent Res.*, vol. 27, no. 1, pp. 132–150, 2012.

[5] T. F. Frandsen and J. Nicolaisen, "Effects of academic experience and prestige on researchers' citing behavior," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 63, no. 1, pp. 64–71, 2012.

[6] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46, no. 5, pp. 604–632, 1999.

[7] T. Amjad, Y. Ding, A. Daud, J. Xu, and V. Malic, "Topic-based heterogeneous rank," *Scientometrics*, vol. 104, no. 1, pp. 313–334, 2015.

[8] M. Timilsina, B. Davis, M. Taylor, and C. Hayes, "Towards predicting academic impact from mainstream news and weblogs: A heterogeneous graph based approach," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, 2016, pp. 1388–1389.

[9] P. Su, C. Shang, T. Chen, and Q. Shen, "Exploiting data reliability and fuzzy clustering for journal ranking," *IEEE Trans. Fuzzy Syst.*, vol. 25, no. 5, pp. 1306–1319, Oct. 2017.

[10] X. Kong, Y. Shi, S. Yu, J. Liu, and F. Xia, "Academic social networks: Modeling, analysis, mining and applications," *J. Netw. Comput. Appl.*, vol. 132, no. 7, pp. 86–103, 2019.

[11] S. Brin and L. Page, "Reprint of: The anatomy of a large-scale hypertextual Web search engine," *Comput. Netw.*, vol. 56, no. 18, pp. 3825–3833, 2012.

[12] M.-F. Chiang, J.-J. Liou, J.-L. Wang, W.-C. Peng, and M.-K. Shan, "Exploring heterogeneous information networks and random walk with restart for academic search," *Knowl. Inf. Syst.*, vol. 36, no. 1, pp. 59–82, 2013.

[13] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the Web," Stanford InfoLab, Stanford Univ., Stanford, CA, USA, Tech. Rep. 1999-66, Nov. 1999.

[14] M. S. Mariani, M. Medo, and Y.-C. Zhang, "Identification of milestone papers through time-balanced network centrality," *J. Inf.*, vol. 10, no. 4, pp. 1207–1223, Nov. 2016.

[15] Y. Zhang, J. Ma, Z. Wang, B. Chen, and Y. Yu, "Collective topical pagerank: A model to evaluate the topic-dependent academic impact of scientific papers," *Scientometrics*, vol. 114, no. 3, pp. 1345–1372, 2018.

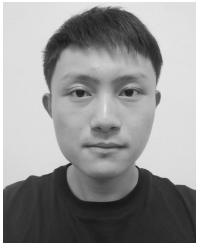
[16] Y.-B. Zhou, L. Lü, and M. Li, "Quantifying the influence of scientists and their publications: Distinguishing between prestige and popularity," *New J. Phys.*, vol. 14, no. 3, 2012, Art. no. 033033.

[17] X. Bai, F. Zhang, J. Hou, I. Lee, X. Kong, A. Tolba, and F. Xia, "Quantifying the impact of scholarly papers based on higher-order weighted citations," *PLoS ONE*, vol. 13, no. 3, 2018, Art. no. e0193192.

[18] M. Bordons, M. Fernández, and I. Gómez, "Advantages and limitations in the use of impact factor measures for the assessment of research performance," *Scientometrics*, vol. 53, no. 2, pp. 195–206, 2002.

[19] A. Nederhof, M. Luwel, and H. Moed, "Assessing the quality of scholarly journals in linguistics: An alternative to citation-based journal impact factors," *Scientometrics*, vol. 51, no. 1, pp. 241–265, 2001.
[20] R. K. Pan and S. Fortunato, "Author impact factor: Tracking the dynamics of individual scientific impact," *Sci. Rep.*, vol. 4, no. 4880, p. 1, 2014.
[21] E. Garfield, "Citation indexes for science. A new dimension in documentation through association of ideas," *Int. J. Epidemiol.*, vol. 35, no. 5, pp. 1123–1127, 1955.
[22] J. Zhu, B. H. Lee, D. Diaz, and L. Y. Flores, "Evaluating the scholarly impact of vocational research with diverse racial/ethnic groups: 1969–2017," *J. Career Develop.*, vol. 1, pp. 1–15, May 2019, Art. no. 0894845319846423.
[23] M. Antonoyiannakis, "How a single paper affects the impact factor: Implications for scholarly publishing," 2019, *arXiv:1906.02660*. [Online]. Available: https://arxiv.xilesou.top/abs/1906.02660
[24] F. J. Trueba and H. Guerrero, "A robust formula to credit authors for their publications," *Scientometrics*, vol. 60, no. 2, pp. 181–204, 2004.
[25] J. Zhang, F. Xia, W. Wang, X. Bai, S. Yu, T. M. Bekele, and Z. Peng, "CocaRank: A collaboration caliber-based method for finding academic rising stars," in *Proc. 25th Int. Conf. Companion World Wide Web*, 2016, pp. 395–400.
[26] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1188–1196.
[27] M. Levandowsky and D. Winter, "Distance between sets," *Nature*, vol. 234, no. 5323, p. 34, 1971.
[28] W. Wang, S. Yu, T. M. Bekele, X. Kong, and F. Xia, "Scientific collaboration patterns vary with scholars' academic ages," *Scientometrics*, vol. 112, no. 1, pp. 329–343, 2017.
[29] A. Daud, R. Abbasi, and F. Muhammad, "Finding rising stars in social networks," in *Proc. Int. Conf. Database Syst. Adv. Appl.* Berlin, Germany: Springer, 2013, pp. 13–24.
[30] J. Zhang, Z. Ning, X. Bai, W. Wang, S. Yu, and F. Xia, "Who are the rising stars in academia?" in *Proc. 16th ACM/IEEE-CS Joint Conf. Digit. Libraries*, Jun. 2016, pp. 211–212.

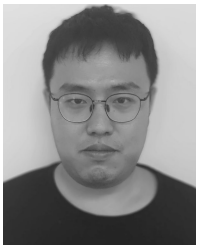
JING REN received the B.Sc. degree in software engineering from Huaqiao University, Xiamen, China. She is currently pursuing the master's degree in software engineering with the Dalian University of Technology, China. Her research interests include big scholarly data, network science, and computational social science.



LEI WANG received the B.Sc. degree in software engineering from the Dalian University of Technology, China, in 2018, where he is currently pursuing the master's degree with the School of Software. His research interests include data mining, analysis of complex networks, and machine learning.

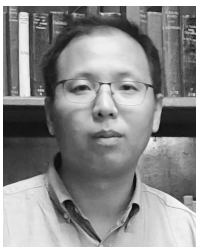
KAILAI WANG received the B.Sc. degree in software engineering from the Dalian University of Technology, China, in 2019, where he is currently pursuing the master's degree with the School of Software. His research interests include academic data mining, analysis of complex networks, and network embedding.

SHUO YU received the B.Sc. and M.Sc. degrees from the Shenyang University of Technology, Shenyang, China. She is currently pursuing the Ph.D. degree in software engineering with the Dalian University of Technology, Dalian, China. Her research interests include network science, data science, computational social science, and science of scientific team science.



MINGLIANG HOU received the B.Sc. degree from Dezhou University and the M.Sc. degree from Shandong University, Shandong, China. He is currently pursuing the Ph.D. degree in software engineering with the Dalian University of Technology, Dalian, China. His research interests include network science, data science, and urban computing.

IVAN LEE (M'05–SM'07) received the B.Eng., M.Com., M.E.R., and Ph.D. degrees from The University of Sydney, Sydney, NSW, Australia. He was a Software Development Engineer with Cisco Systems, a Software Engineer with Remotek Corporation, and an Assistant Professor with Ryerson University. Since 2008, he has been a Senior Lecturer with The University of South Australia. His research interests include smart sensor, multimedia systems, and data analytics.



XIANGJIE KONG (M'13–SM'17) received the B.Sc. and Ph.D. degrees from Zhejiang University, Hangzhou, China. He is currently an Associate Professor with the School of Software, Dalian University of Technology, China. He has published more than 100 scientific articles in international journals and conferences (with more than 70 indexed by ISI SCIE). His research interests include network science, data science, and computational social science. He is a Senior Member of CCF and a member of ACM. He has served as a Guest Editor for several international journals and as the Workshop Chair or a PC Member for a number of conferences.

FENG XIA (M'07–SM'12) received the B.Sc. and Ph.D. degrees from Zhejiang University, Hangzhou, China. He is currently an Associate Professor and the Discipline Leader with the School of Science, Engineering and Information Technology, Federation University Australia, and on leave from the School of Software, Dalian University of Technology, China, where he is also a Full Professor. He has published two books and more than 300 scientific articles in international journals and conferences. His research interests include data science, knowledge management, social computing, and systems engineering. He is a Senior Member of ACM.