



Can we predict ESI highly cited publications?

Fenghua Wang¹ · Ying Fan¹  · An Zeng¹ · Zengru Di¹

Received: 21 March 2018 / Published online: 13 December 2018
© Akadémiai Kiadó, Budapest, Hungary 2018

Abstract

The highly cited papers defined by Clarivate Analytics' Essential Science Indicators (ESI) have been widely used to measure the scientific performance of scientists, research institutions, universities and countries. However, researchers have seldom studied which factors can affect a paper to be an ESI highly cited paper. The prediction of ESI highly cited papers is much less studied, too. According to the existing researches about factors influencing paper's citations, four classical papers' factors are chosen in this study, which are scientific impact of the first author, scientific impact of the potential leader, scientific impact of the team and the relevance of authors' existing papers. Similar to the definition of ESI highly cited papers, we develop a new measure of papers' scientific impact. Firstly, we get statistics properties of four factors with APS data and Nobel data in order to study four factors' performance of ESI highly cited papers. Then, Spearman correlation and Logistic regression are applied to explore the relationship between four factors and papers' scientific impact. At last, we try to predict highly cited papers by NN algorithms incorporating four factors. The results show that the potential leader factor plays a more important role in the short term than in the long term, while the team factor is on the contrary, more important in the long term. Interestingly, the first author factor doesn't have an obvious effect on papers' scientific impact among top 1%. The prediction results are better than random.

Keywords ESI · Citation network · Scientific impact · Prediction

Introduction

Nowadays with the development of science, the number of published papers has increased exponentially. As a result it becomes impossible for researchers to investigate all papers related to their research field to find useful ones. To identify papers' scientific impact makes both theoretical and practical significance for scientists whose time and energy are limited (Wang et al. 2013). A lot of studies have been conducted quantifying papers' scientific

✉ Ying Fan
yfan@bnu.edu.cn

✉ An Zeng
anzeng@bnu.edu.cn

¹ School of Systems Science, Beijing Normal University, Beijing 100875, People's Republic of China

impact in citation networks. The simplest and most straightforward measure is the number of citations which are the number of links of a node (Redner 1998). Based on citations, we can classify papers in specific categories (Glänzel and Schubert 1988). Then some variants about citations arise (Radicchi et al. 2008; Chatterjee et al. 2016). Afterwards, PageRank and its variants are commonly implemented in citation networks to rank scientific impact of publications (Chen et al. 2007; Mariani et al. 2016; Yao et al. 2014; Zhou et al. 2016). In addition to these measurements, growing models are also utilized to measure publications' scientific impact (Jeong et al. 2003; Bianconi and Barabási 2001; Wang et al. 2013). But, these definitions of papers' scientific impact haven't taken into account the influence of other papers published in the same year like ESI highly cited papers have.

Essential Science Indicators (ESI) was proposed in 2002. According to highly cited papers defined by ESI, ESI highly cited papers need to receive enough citations to place them in the top 1% of their academic fields from each of 10 database years. Since its introduction, ESI highly cited papers have been widely used to evaluate research institutions, universities and countries (Zhu et al. 2004; Pouris 2007; Soteriades and Falagas 2005; Hu and Rousseau 2009). As we know, the highly cited papers have been attached great importance to the science of science (Zeng et al. 2017).

Scientists have tried to explore the characteristics of highly cited papers due to their importance. Compared with "ordinary" papers, Aksnes finds that the number of authors of highly cited papers is large and authors usually internationally collaborate (Aksnes 2003). In addition, Noorhidawati et al. observe that ESI highly cited papers in Malaysia have nine characteristics such as lowly self-cited and publication in the First Quartile of the science discipline publications (Noorhidawati et al. 2017). Besides these features of papers, some other factors of publication are also important. The first-mover advantage proposed by Newman indicates that if a paper is the first paper in the field, it will get enormously higher citations than other papers published after it, regardless of its content (Newman 2009). Zhang and Guan (2017) find papers with lower scientific relatedness (extract from knowledge element coupling network) and higher intellectual base (extract from citation network) are more likely to be highly cited. More importantly, the initial citation of a paper is highly correlated with the long-term citation of the paper and can be used as an indicator for the prediction of highly cited papers (Adams 2005; Ponomarev et al. 2014; Winnink and Tijssen 2015; Winnink et al. 2016). The prediction of highly cited papers can help scientists track research trends and identify significant work in advance. Regardless of the definition of highly cited papers, these papers have the same characteristic: high citations. With this feature, scientists have tried to predict citations of papers to solve the problem of predicting highly cited papers. There are three main ways to predict papers' citations.

The first way is to use regression models. Hurley et al. get author-level attributes and article-level attributes of collaborations to build a logistic regression model. They find the strongest predictors are the language and journal of the paper (Hurley et al. 2013). Yan et al. extract some features (topic rank, diversity, recency, *h*-index, author rank, productivity, sociality, authority, venue rank and venue centrality) to build several predictive models including linear regression, *k*-nearest neighbor, support vector regression, classification and regression tree. According to their results, author expertise and venue impact are regarded as strongest predictors. Author rank and venue rank can make papers become more attractive, however content features in isolation are not valuable to prediction (Yan et al. 2011). Similarly, Yu et al. consider the external features of a paper, features of authors, features of the journal of publication, and features of citations to set up a stepwise regression model to predict papers' citations after the first 5 years of publication. Their model is effective within Information Science & Library Science (Yu et al. 2014).

The second way is to use machine learning which has a similar research pattern with regression models. The method of their research starts by selecting effective factors among papers' characteristics. After that, those factors are applied to regression models or machine learning to predict papers' citations. Wang et al extract the external features (sixteen features coming from three aspects: the authors, the journals, and the external features of the paper itself) and quality features (nine features coming from papers' citing diffusion properties and the information of papers' first citation) to develop a case-based classifier (CBC) to predict highly cited papers. The result shows that the paper's quality, the journal factors and the research ability of the first author are most predictive (Wang et al. 2012). Furthermore, Xiao et al. propose a machine learning prediction model considering not only papers' intrinsic popularity (paper-specific covariates) but also impact decaying over time and recency-sensitive citation triggering and this model has better performance than previous methods (Xiao et al. 2016).

The third way is to set up dynamic models based on citation dynamics to predict papers' citations. Ponomarev et al. (2014) propose two curve-fitting models about papers' citations dynamics to predict highly cited papers. Wang et al. build a citation dynamic model through a set of parameters by employing maximum likelihood estimation and three factors (fitness term representing paper's intrinsic value; the aging effect over time; the reinforcement term showing highly cited papers). This citation dynamic model performs well in predicting papers' citations (Wang et al. 2013). In addition to these three main methods, there are other ways to predict impact of papers' citations. For example, Cao et al. propose a model by matching the appropriate database and finding the similar citation pattern of a paper to predict citations of a paper. The model performs well (Cao et al. 2016). Newman (2014) uses first-mover advantage to predict highly cited papers with high prediction accuracy. As we all know, there are some studies on ESI highly cited papers. However, we did not find any paper about the predictions of ESI highly cited papers.

In this study, we select four factors, which are scientific impact of first author, scientific impact of potential leader, scientific impact of team and the relevance of authors' existing papers. Firstly, we put forward a new measure of papers' scientific impact similar to ESI. Then we apply spearman rank correlation analysis and logistic regression analysis to papers in APS database which is about physics field from 1980 to 1989 to investigate the relationship between papers' scientific impact and these four factors in 3 years and in 10 years after publication respectively. At last, we try to predict ESI highly cited papers by a combination of four factors using Neural networks algorithm.

Method

Data

The database used in this paper covers all papers published in American Physical Society (APS) journals from 1893 to 2009, which are Physical Review series and Reviews of Modern Physics. There are 198,476 authors, 458,863 papers and 4,658,283 citations in this dataset.

In APS database, there is no unique author identifier. Some authors used different name forms in their papers. In order to assign each paper to each author, we first give a unique identifier to each author. In the APS data, the author names with the same surname and the same first letter of the given name is considered as from one author. For example, Kikuji

Hirose and K Hirose are identified as the same author while F Green and A Green are identified two. In particular, we get Nobel authors and papers from wiki website, each of them is identified manually.

Measure of papers' scientific impact

Based on the definition of highly cited papers in ESI, we put forward a new measure of papers' scientific impact, considering not only the number of citations but also the rank of citations in all papers published in that year. For paper i , if its citation rank among all papers published in the same year is N and the total number of papers published in that year is R , then the scientific impact of paper i is N/R . According to such scientific impact measure, papers whose rank are in the top 1% are highly cited papers. Papers published in different time have different features. Those published earlier must have more time to accumulate citations. The traditional way to select highly cited papers makes decision based on citations, which is unfair to new papers. However, our highly cited papers selection only takes papers published in the same year into consideration, thus more reasonable. Three years and 10 years are two kind of time window which represent short term and long term (Abramo et al. 2011; Wang et al. 2013). In this article, we choose papers' scientific impact in 3 years and in 10 years after papers publish as two kind of representative values.

We apply this measure to the Nobel prize-winning papers. There are 28 Nobel prize-winning papers from 1946 to 2000. Their scientific impact value and total citations each year from the publication year to 2009 are shown in Fig. 1. Citations of papers increase monotonically with variant growth rate all the time. However, the measure we proposed manages to show changes in scientific impact. We can know whether papers become more

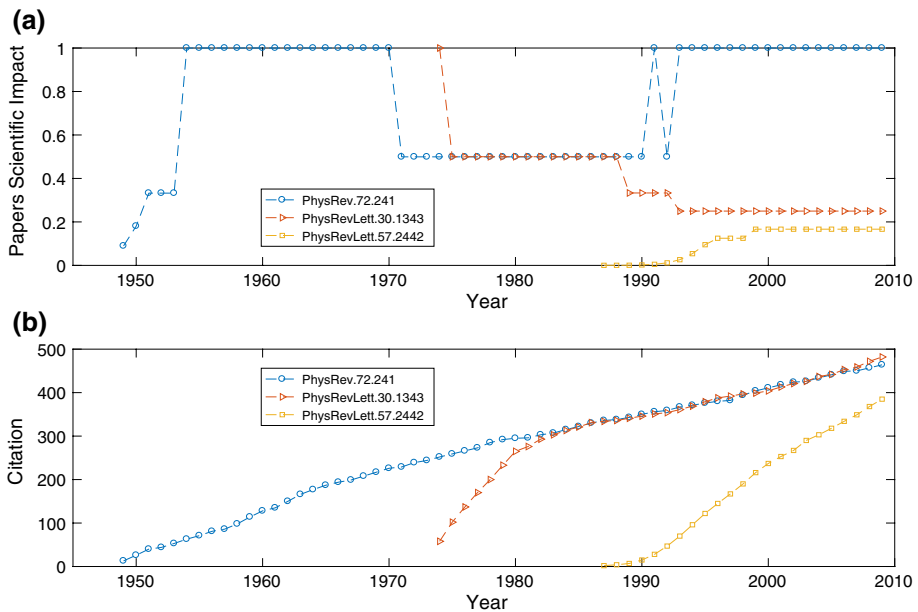


Fig. 1 **a** The scientific impact value of three Nobel prize-winning papers in each year from the year published to 2009. **b** The total citations

influential compared with papers published in the same year. Table 1 shows Nobel prize-winning papers' scientific impact in 3 years after publication, in 10 years after publication and in the awarded year, as well as whether they are highly cited papers in these years respectively. We could see 13 Nobel prize-winning papers are ESI highly cited papers in 3 years after publication. This number increases to 15 in 10 years after publication and reaches a maximum of 16 in the awarded year. People usually think of Nobel prize-winning papers as high quality papers (Niu et al. 2016; Qi et al. 2017). Not all Nobel prize-winning papers are highly cited papers in the short term, but most of them gradually become highly cited papers in the long term.

Factors influencing papers' scientific impact

Based on the way that we measure the paper's scientific impact, the number of papers published in a year and the paper's citation rank among all papers published in that year are two direct factors influencing papers' scientific impact. As we know, the number of published papers has increased exponentially. Present studies show that paper citations are influenced by a lot of characteristics, such as author characteristics, field characteristics, article characteristics, journal characteristics and so on (Danell 2011; Biscaro and Giupponi 2014; Yu et al. 2014; Wang et al. 2012).

The number of co-authored publications has increased substantially in science. First authors should get more credits in scientific credit allocation because of more contributions in some disciplines (Shen and Barabási 2014). There are some evidences suggesting that the scientific impact of first author has a positive effect on paper citations. Yu et al. (2014) showed that citation impact can be predicted by some features including the total number of citations to the papers previously published by the first author. The Matthew Effect suggests that people are more willing to cite reputable scholars (Merton 1968). Garfield and Welljamsdorof (1992) observe that use of a high citation list of authors can significantly predict Nobel Prize laureates. Another study conducted by Maske et al. (2003) shows expert author with a deeper ability in publishing ideas. In scientific credit allocation, the corresponding authors get higher allocation credit in disciplines, such as physics. For group leaders in scientific research, their names often appear at the bottom (or top) of the list of authors in the co-author's paper (Kosmulski 2012). Based on these studies, we select two factors: the scientific influence of the first author and the scientific influence of potential leaders. There are several ways to measure the scientific impact of scientists, such as the total number of papers published, the total number of citations obtained or the mean number of citations per paper and *h*-index (Hirsch 2007). Stewart (1983) found that the average author quality, measured by the average citations of his published papers, is an important author characteristic that influences the citations of the paper. In this article, the average citations of the author's publications is used as the author's scientific influence. Collaborations are universal among scientists. A rapidly increasing trend in co-authorship has been observed in both social science and natural science. Through collaboration, scientists exchange and refine ideas, producing higher quality publications. It's revealed that cooperation can promote production efficiency (Maske et al. 2003). Previous research has shown that the number of co-authors is a predictor of research productivity (Yu et al. 2014). Experience in the field has a positive effect on the paper's impact. Biscaro and Giupponi (2014) quantified experience as the maximum of number of articles published by each co-author prior to the paper. Specialized authors are better known within the field. In the expertise field, scientists can choose their representative work. A research conducted by

Table 1 Nobel prize-winning papers' scientific impact in 3 years after publication, in 10 years after publication and in the awarded year, as well as whether they are highly cited papers in these years respectively

No.	DOI	Publishing year	Awarding year	Three year SI	Ten year SI	Awarding year SI	Y-3 or N-3	Y-10 or N-10	Y-Awa or N-Awa
1	https://doi.org/10.1103/PhysRev.69.37	1946	1952	264.00	66.00	132.00	Y	N	Y
2	https://doi.org/10.1103/PhysRev.72.241	1947	1955	313.00	50.08	52.17	Y	N	N
3	https://doi.org/10.1103/PhysRev.73.679	1948	1981	289.00	867.00	433.50	Y	Y	Y
4	https://doi.org/10.1103/PhysRev.83.333	1951	1994	82.89	213.14	165.78	N	Y	Y
5	https://doi.org/10.1103/PhysRev.122.345	1961	2008	53.78	135.52	564.67	N	Y	Y
6	https://doi.org/10.1103/PhysRevLett.9.439	1962	2002	4.66	2.64	2.44	N	N	N
7	https://doi.org/10.1103/PhysRevLett.10.84	1963	2005	34.22	41.06	114.06	N	N	Y
8	https://doi.org/10.1103/PhysRevLett.13.508	1964	2013	1.99	41.50	173.17	N	N	-
9	https://doi.org/10.1103/PhysRevLett.13.321	1964	2013	4.49	44.84	193.15	N	N	-
10	https://doi.org/10.1103/PhysRevLett.20.1205	1968	2002	6.78	15.60	148.21	N	N	Y
11	https://doi.org/10.1103/PhysRevLett.28.885	1972	1996	97.47	43.27	34.81	Y	N	N
12	https://doi.org/10.1103/PhysRevLett.29.1227	1972	2003	64.76	21.34	14.29	N	N	N
13	https://doi.org/10.1103/PhysRevLett.30.1343	1973	2004	2406.50	2406.50	1203.25	Y	Y	Y
14	https://doi.org/10.1103/PhysRevLett.30.1346	1973	2004	4813.00	4813.00	1604.33	Y	Y	Y
15	https://doi.org/10.1103/PhysRevLett.35.1489	1975	1995	1539.33	769.67	288.63	Y	Y	Y
16	https://doi.org/10.1103/PhysRevLett.48.1559	1982	1998	2748.50	2748.50	1832.33	Y	Y	Y
17	https://doi.org/10.1103/PhysRevLett.50.1395	1983	1998	5773.00	5773.00	5773.00	Y	Y	Y
18	https://doi.org/10.1103/PhysRevLett.55.48	1985	1997	64.08	46.29	46.92	N	N	N
19	https://doi.org/10.1103/PhysRevLett.57.2442	1986	2007	6.30	947.00	1262.67	N	Y	Y
20	https://doi.org/10.1103/PhysRevLett.58.1490	1987	2002	353.96	256.90	261.11	Y	Y	Y
21	https://doi.org/10.1103/PhysRevLett.61.169	1988	2013	38.98	78.86	66.12	N	N	-
22	https://doi.org/10.1103/PhysRevLett.61.2472	1988	2007	91.45	4298.00	4298.00	N	Y	Y
23	https://doi.org/10.1103/PhysRevLett.61.826	1988	1997	14.33	152.14	140.92	N	Y	Y
24	https://doi.org/10.1103/PhysRevLett.75.3969	1995	2001	12,990.00	12,990.00	12,990.00	Y	Y	Y

Table 1 (continued)

No.	DOI	Publishing year	Awarding year	Three year SI	Ten year SI	Awarding year SI	Y-3 or N-3	Y-10 or N-10	Y-Awa or N-Awa
25	https://doi.org/10.1103/PhysRevLett.76.1796	1996	2012	1299.60	684.00	684.00	Y	Y	–
26	https://doi.org/10.1103/PhysRevLett.77.4887	1996	2012	553.02	1181.45	1181.45	Y	Y	–
27	https://doi.org/10.1103/PhysRevLett.84.3232	2000	2005	9.58	6.21	8.70	N	–	N
28	https://doi.org/10.1103/PhysRevLett.84.5102	2000	2005	31.59	30.61	27.27	N	–	N

“–” means that because of the data limit, we can't calculate the value

Niu et al. (2016) argued that the representative work of a scientist should be an important publication in his area of expertise and they proposed a self-avoiding preferential diffusion (SPD) process to identify one's representative work. Based on SPD, we put forward the relevance of authors' existing papers. This factor focuses on measuring the relevance between the paper's topic and topics of authors' existing papers through the paper's reference list.

In this paper, we pick four factors according to present studies as our main analytical factors.

(1) Scientific impact of the first author

We calculate the first author's average number of citations among all his papers published before the year when this certain paper was published as our first factor named The First Author Factor.

(2) Scientific impact of the potential leader

We calculate every author's average number of citations among all his papers published before the year when this certain paper was published and choose the biggest one as our second factor named The Potential Leader Factor.

(3) Scientific impact of the research team

We compute the average number of all authors' previous paper citations as our third factor named The Team Factor.

(4) Relevance of authors' existing papers

First, we gather all papers previously written by each author (PLIST), all papers citing these previously written papers (CLIST) and all papers cited by these previously written papers (RLIST) to get three lists and their union set. Then, we get the reference list of this certain paper, PCLIST. Finally, the number of papers in the intersection between PCLIST and the union set is our fourth factor named relevance of authors' existing papers factor (RAEP Factor).

Spearman correlation analysis

To investigate the relationship between four factors and scientific impact, we apply Spearman correlation analysis to four factors and scientific impact in different categories. When we calculate the correlation coefficients between each factor and the scientific impact, we would delete samples whose factor are zero at first.

Logistic regression

In order to explore the relationship between the four factors and papers' scientific impact synthetically, we apply a stepwise logistic regression to post-treatment data where samples with at least one zero-value factor from 1980 to 1989 are deleted both in the short and long term cases.

Neural-net algorithms

The multilayer perceptron is one type of artificial neural networks. It is widely used as a tool for prediction, function approximation and classification. The advantages of the multilayer perceptron approach are obvious in applications where a full theoretical model cannot be established, and particularly when handling with non-linear systems (Gardner and

Dorling 1998). The multilayer perceptron is known as a feed-forward neural networks and consists of several layers of neurons. The input layer is only used to pass the input vector to the network. There are one or more hidden layers and finally an output layer in a multilayer perceptron. In this paper, the training data are all papers from 1980 to 1986 and test data are all papers from 1987 to 1989.

Results

Statistics properties of four factors

We classify papers from 1980 to 1989 by scientific impact compared with papers published in the same year into four categories, papers among top 1%, papers among top 5%, papers among top 25% and all papers. Then we get the statistical properties of four factors within each category respectively. There are 695 samples among top 1%, 3626 samples among top 5%, 19,001 samples among top 25% and 67,474 samples in total. The range of first author scientific impact factor is from 0 to 147.33, same as the potential leader scientific impact factor. The team scientific impact factor has a range from 0 to 116 and the relevance of authors' existing papers has a range from 0 to 71. The mean, median and standard deviation of papers in the top 1% category among four factors are the biggest, followed by papers among top 5%, papers among top 25% and all papers both in 3 years and in 10 years after papers published. If we delete samples as long as one of the four factors is zero, there are 387 samples left among top 1%, 1964 samples among 5%, 10,104 samples among 25% and 37,252 samples in total. Their mean, median and standard deviation in four categories satisfy the same pattern as before.

We focus on the statistical properties of highly cited papers, which are among top 1% in each year from 1980 to 1989. The total number of papers published from 1980 to 1989 are 4688, 5140, 5497, 5773, 6073, 6921, 7576, 7964, 8596, 9246. In 3 years and in 10 years after highly cited papers published, the mean and standard deviation of the four factors are shown in Figs. 2 and 3. The mean of four factors just fluctuate a little both in 3 years and in 10 years after publication. However, the standard deviation of four factors changes more obviously in different years from 1980 to 1989 in both cases.

We also calculate the above four factors for each Nobel prize-winning paper. We classify Nobel prize-winning papers into six classes, which are papers in top 1% in 3 years after publication (Y-3), papers not in top 1% in 3 years after publication (N-3), papers in top 1% in 10 years after publication (Y-10), papers not in top 1% in 10 years after publication (N-10), papers in top 1% in the awarded year (Y-Awa), papers not in top 1% in the awarded (N-Awa). Then we compute the mean and standard deviation of these four factors in each class. Results are shown in Fig. 4. For all four factors, the in-top-1% classes all exhibit higher mean and standard deviation.

The relationship between papers' scientific impact and four factors

From Spearman correlation analysis, four factors all have positive correlations with papers' scientific impact no matter in 3 years or in 10 years after publication for all papers. However, papers in the top 1% category demonstrate a different pattern. In 3 years after publication, the first author factor doesn't have significant correlation with scientific impact, and the potential leader factor in 1983 has negative relationship with scientific impact while in

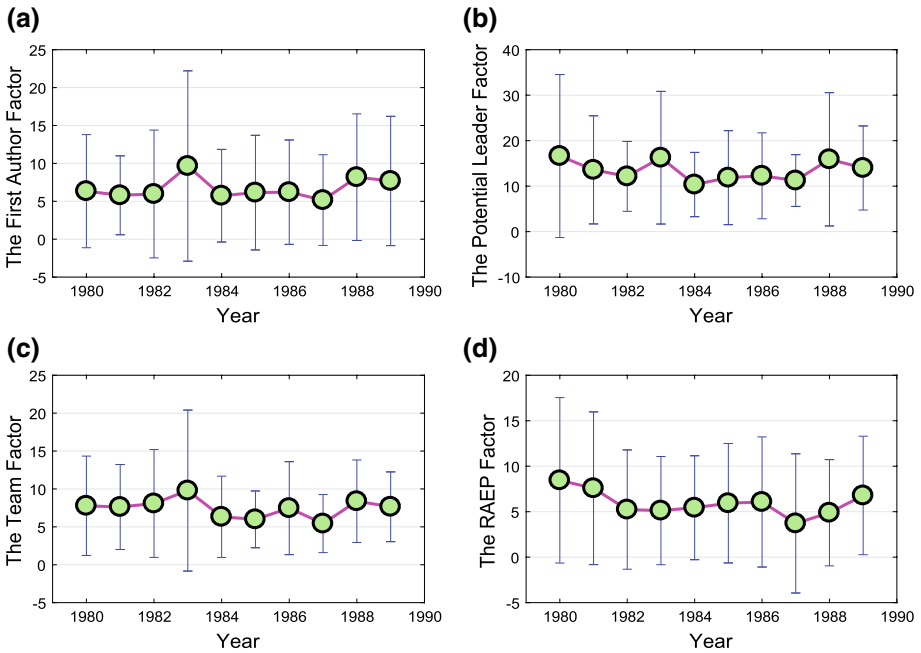


Fig. 2 The erro bar of four factors in the top 1% in 3 years after papers published. **a** The first author factor. **b** The potential leader factor. **c** The team factor. **d** The relevance of authors' existing papers factor

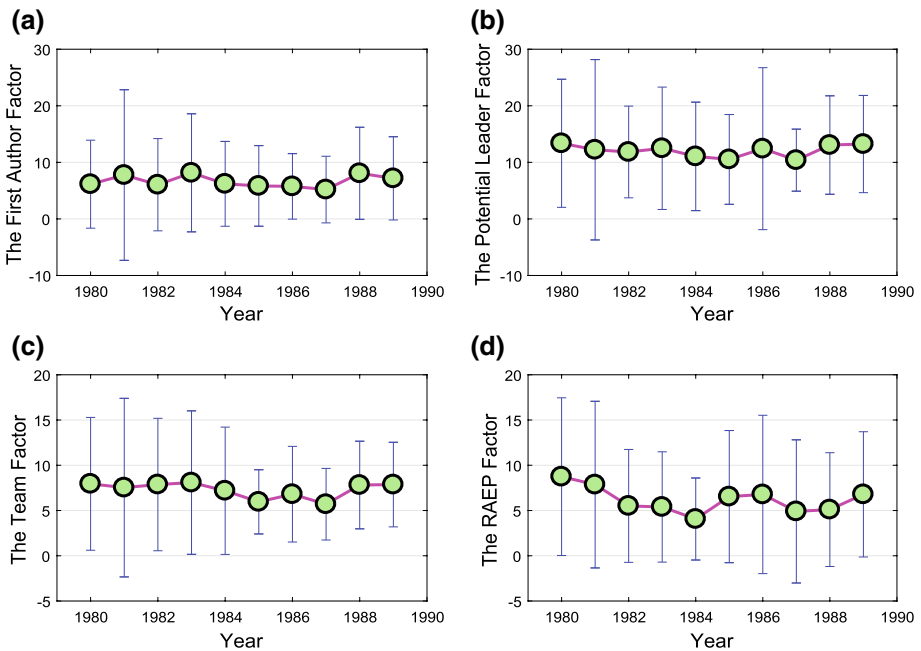


Fig. 3 The erro bar of four factors in the top 1% in 10 years after papers published. **a** The first author factor. **b** The potential leader factor. **c** The team factor. **d** The relevance of authors' existing papers factor

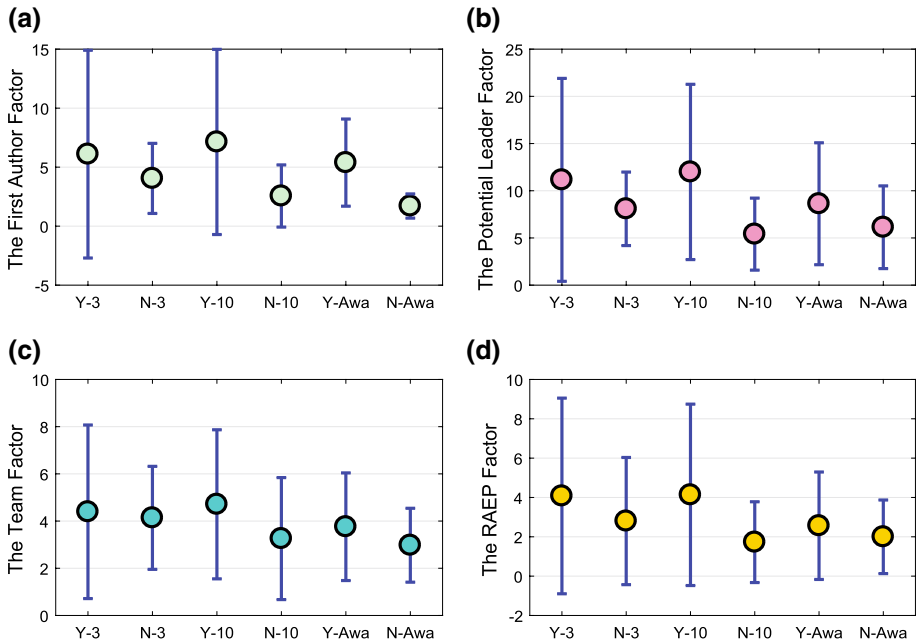


Fig. 4 The erro bar of four factors in six kinds of Nobel prize-winning papers. **a** The first author factor. **b** The potential leader factor. **c** The team factor. **d** The relevance of authors’ existing papers factor

1986 the relationship between them became positive. The pattern of RAEP factor is similar to that of the potential leader factor and the team factor have positive effect on scientific impact only in 1988. In 10 years after publication, the first author factor and the team factor have a positive effect on papers’ scientific impact in 1987. The RAEP factor has negative effect on papers’ scientific impact in 1984. Additionally, papers in the top 5% category have little correlation coefficients passing the significance test among the first author factor and the RAEP factor. But the correlation coefficients passing the significance test of the potential leader factor and the team factor are more. Situation has changed in papers in the top 25% category, among four factors, in most years, correlations coefficients are significant in both 3 years and 10 years after publication (Figs. 5 and 6).

Results of stepwise logistic regression to post-treatment data are displayed in Tables 2 and 3. In 3 years after publication, papers’ scientific impact in the top 1% category has a positive relationship with the potential leader factor, the team factor and the RAEP factor. What’s more, the coefficient of the potential leader factor is the highest. Situations coincide when looking into papers in the top 5% category and in the top 25% category. In 10 years after publication, papers’ scientific impact in the top 1% category has a positive relationship with the potential leader factor, the team factor and the RAEP factor where the coefficient of the team factor is the highest. There are same three factors positively related to papers’ scientific impact in the top 5% and in the top 25%. In the top 5%, the highest coefficient is the team factor while in the top 25% the highest coefficient is the potential leader factor.

Are logistic regression results stable in each year from 1980 to 1989? To answer this question, we apply logistic regression to data in each year respectively. Results in 3 years and in 10 years are showed in Figs. 7 and 8. We find that for papers in the 3 years after

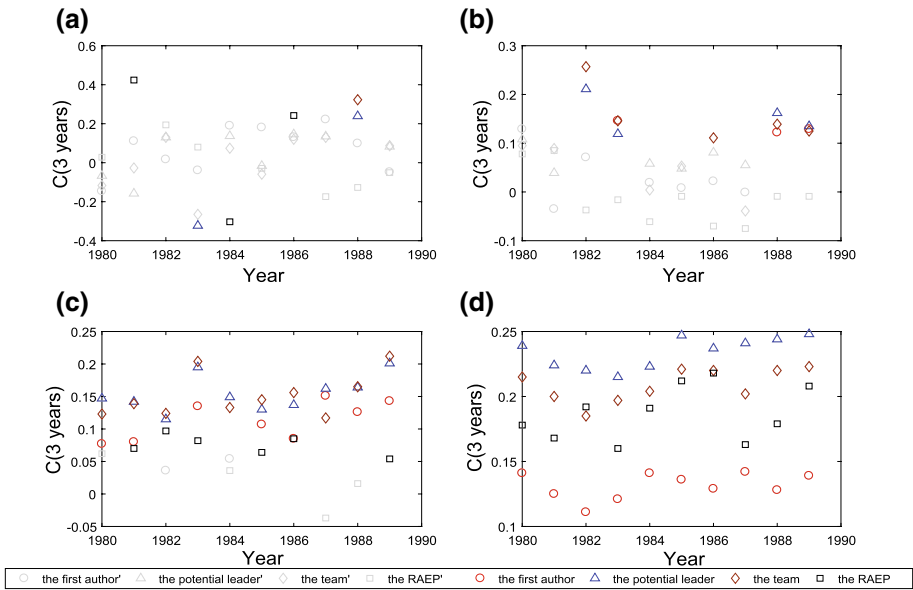


Fig. 5 Spearman correlation coefficients between four factors and the scientific impact of papers in 3 years after publication. The shaded labels represent those that haven't passed the significance test. **a** Top 1% category. **b** Top 5% category. **c** Top 25% category. **d** All papers category

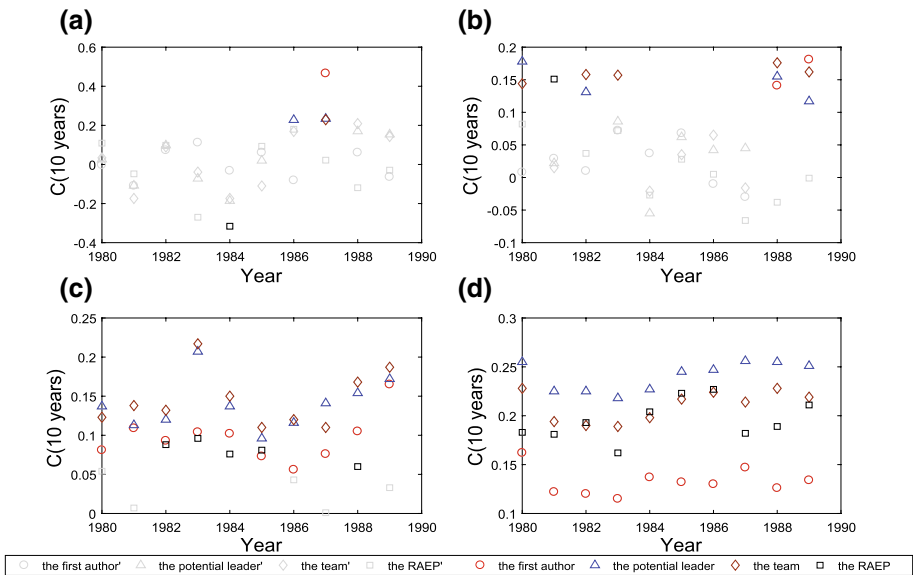


Fig. 6 Spearman correlation coefficients between four factors and the scientific impact of papers in 10 years after publication. The shaded labels represent those that haven't passed the significance test. **a** Top 1% category. **b** Top 5% category. **c** Top 25% category. **d** All papers category

Table 2 Logistic regression coefficients of four factors in 3 years after publication

	Model (1%)	Model (5%)	Model (25%)
Intercept	- 5.234	- 3.457	- 1.502
The first author scientific impact	-	-	-
The potential leader scientific impact	3.28	3.249	4.448
The team scientific impact	2.999	2.711	1.398
The relevance of authors' exsiting papers	2.45	2.467	3.194

Table 3 Logistic regression coefficients of four factors in 10 years after publication

	Model (1%)	Model (5%)	Model (25%)
Intercept	- 5.198	- 3.53	- 1.607
The first author scientific impact	-	-	-
The potential leader scientific impact	1.361	2.513	3.438
The team scientific impact	4.153	3.37	2.165
The relevance of authors' exsiting papers	2.761	2.526	3.26

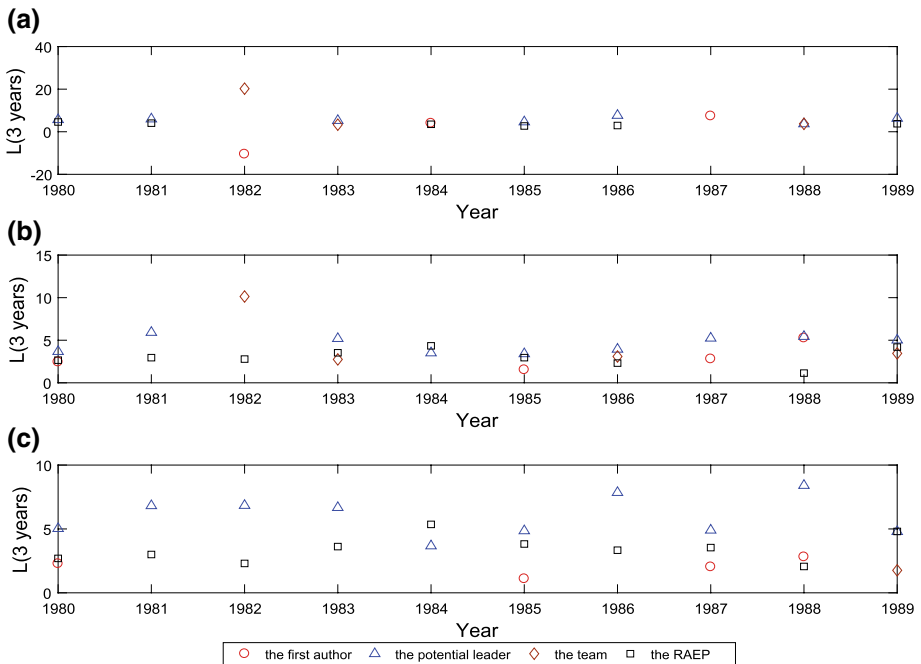


Fig. 7 Logistic regression coefficients of four factors in 3 years after publication from 1980 to 1989. **a** Top 1% category. **b** Top 5% category. **c** Top 25% category

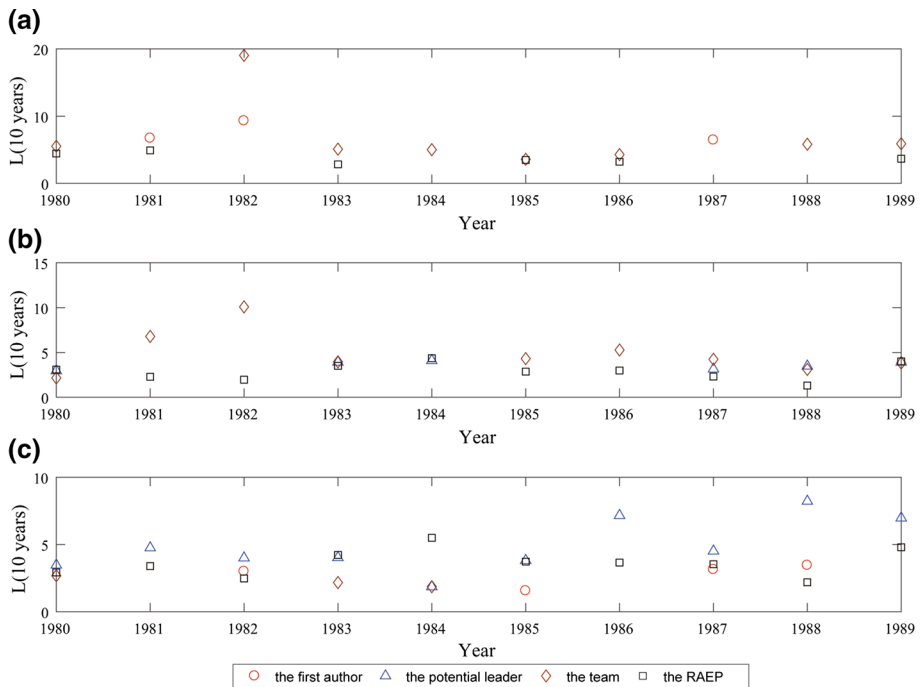


Fig. 8 Logistic regression coefficients of four factors in 10 years after publication in each year from 1980 to 1989. **a** Top 1% category. **b** Top 5% category. **c** Top 25% category

publication case, the potential leader factor and RAEP factor are two important factors of logistic regression in all categories. While in 10 years after publication, the team factor and RAEP factor dominate both in the top 1% and in the top 5%, and the potential leader factor and the RAEP factor dominate in the top 25%.

Prediction

In this paper, the training data are all papers from 1980 to 1986 and test data are all papers from 1987 to 1989. The features in neural-net algorithms are the four factors: scientific impact of first author, scientific impact of potential leader, scientific impact of team and relevance of authors' existing papers. In each category, the labels of data are only two kinds, 0 or 1. For example, for all papers in top 1% category, if papers' scientific impact rank belong to 1%, data's labels are 1 and the left are 0. So do data in 5% category and 25% category. In our data, the data whose labels are 1 are 1%, 5% and 25%. The precision is low, because most prediction values are 0. The accuracy in total is high: in top 1% category is almost 99%, in top 5% category is almost 95% and in top 25% category is almost 74%. Here, we change the way to calculate prediction accuracy. After we train the model well, we get the test data predictions probability by the trained model and sort the predictions probability values in a descending order. Then we choose top 1% dataset and calculate the size of the intersection between the predicted top 1% and the actual dataset top 1%. Prediction accuracy is the size of the intersection divided by the size of the actual dataset top 1%. Prediction accuracy is displayed in Table 4.

Table 4 The accuracy of different kinds of neural-net models

nn net (years)	Model (1%)	Model (5%)	Model (25%)
3	4.20%	14.92%	38.66%
10	6.02%	16.79%	40.06%

From Table 4, we can see the accuracy in papers among top 1% is better than choosing randomly, but still low. The prediction accuracy is higher in papers among top 5% and among top 25%. What's more, prediction accuracy is higher in 10 years after papers publications than in 3 years after publication.

Conclusions and discussion

In this paper, we mainly focus on four factors which can be used to predict ESI highly cited papers. Three factors come from the author aspects, which are the first author's scientific impact, the potential leader's scientific impact and the team's scientific impact. The last factor comes from the content of the paper, which is the relevance of the author's existing paper. We analyze the relationship between the four factors and papers' scientific impact both in short term and in long term. Our findings indicate that four factors are positively correlated with the paper's scientific impact. The potential leader factor plays a more important role in short term instead of in the long run. On the contrary, team factor is more important in the long term. When we combine the four factors to predict the top 1% of the papers through neural network algorithms, the prediction results are better than the random selection, but the prediction accuracy is still low. This means that these four factors can partially but not sufficiently explain the pattern of ESI highly cited papers. It also shows that people cannot manipulate these factors to achieve the goal of making a paper be ESI highly cited paper. We know that the possibility of becoming ESI highly cited papers depends not only on their own quality, but also on the quality of other papers and the number of papers published each year. The ESI highly cited papers are truly high quality papers. It shows that they can be used as valuable indicators for assessing the scientific performance of an organization or a scientist, helping policy makers make decisions such as funding, promotions and so on.

Because of the limit of database, in this research we cannot get information such as the corresponding author scientific impact, the key words, the abstract, the journal quality and so on. Our research is the first attempt to predict ESI high cited papers. Although we have obtained some results, we have to admit that those chosen factors are not sufficient enough. Finding appropriate and sufficient factors to predict ESI high cited papers remains a challenge. In further research, we will try to obtain more information to find more useful factors to solve the problem of predicting ESI highly cited papers. In addition, the APS data set can only represent the physical field. We can apply the models to different kinds of fields to verify the results later.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant Nos. 61603046 and 61374175) and the Natural Science Foundation of Beijing (Grant No. L160008).

References

- Abramo, G., Cicero, T., & D'Angelo, C. A. (2011). Assessing the varying level of impact measurement accuracy as a function of the citation window length. *Journal of Informetrics*, *5*(4), 659–667.
- Adams, J. (2005). Early citation counts correlate with accumulated impact. *Scientometrics*, *63*(3), 567–581.
- Aksnes, D. W. (2003). Characteristics of highly cited papers. *Research Evaluation*, *12*(3), 159–170.
- Bianconi, G., & Barabási, A. L. (2001). Bose–Einstein condensation in complex networks. *Physical Review Letters*, *86*(24), 5632.
- Biscaro, C., & Giupponi, C. (2014). Co-authorship and bibliographic coupling network effects on citations. *PLoS ONE*, *9*(6), e99502.
- Cao, X., Chen, Y., & Liu, K. R. (2016). A data analytic approach to quantifying scientific impact. *Journal of Informetrics*, *10*(2), 471–484.
- Chatterjee, A., Ghosh, A., & Chakrabarti, B. K. (2016). Universality of citation distributions for academic institutions and journals. *PLoS ONE*, *11*(1), e0146762.
- Chen, P., Xie, H., Maslov, S., & Redner, S. (2007). Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics*, *1*(1), 8–15.
- Danell, R. (2011). Can the quality of scientific work be predicted using information on the author's track record? *Journal of the Association for Information Science and Technology*, *62*(1), 50–60.
- Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron)—A review of applications in the atmospheric sciences. *Atmospheric Environment*, *32*(14), 2627–2636.
- Garfield, E., & Welljamsdorff, A. (1992). Of nobel class—A citation perspective on high-impact research authors. *Theoretical Medicine*, *13*(2), 117–135.
- Glänzel, W., & Schubert, A. (1988). Characteristic scores and scales in assessing citation impact. *Journal of Information Science*, *14*(2), 123–127.
- Hirsch, J. E. (2007). Does the h index have predictive power? *Proceedings of the National Academy of Sciences*, *104*(49), 19193–19198.
- Hu, X., & Rousseau, R. (2009). A comparative study of the difference in research performance in biomedical fields among selected Western and Asian countries. *Scientometrics*, *81*(2), 475–491.
- Hurley, L. A., Ogiev, A. L., & Torvik, V. I. (2013, November). Deconstructing the collaborative impact: Article and author characteristics that influence citation count. In *Proceedings of the 76th ASIS&T annual meeting: Beyond the cloud: Rethinking information boundaries* (p. 61). American Society for Information Science.
- Jeong, H., Nda, Z., & Barabási, A. L. (2003). Measuring preferential attachment in evolving networks. *EPL (Europhysics Letters)*, *61*(4), 567.
- Kosmulski, M. (2012). The order in the lists of authors in multi-author papers revisited. *Journal of Informetrics*, *6*(4), 639–644.
- Mariani, M. S., Medo, M., & Zhang, Y. C. (2016). Identification of milestone papers through time-balanced network centrality. *Journal of Informetrics*, *10*(4), 1207–1223.
- Maske, K. L., Durden, G. C., & Gaynor, P. E. (2003). Determinants of scholarly productivity among male and female economists. *Economic inquiry*, *41*(4), 555–564.
- Merton, R. K. (1968). The Matthew effect in science: The reward and communication systems of science are considered. *Science*, *159*(3810), 56–63.
- Newman, M. E. (2009). The first-mover advantage in scientific publication. *EPL (Europhysics Letters)*, *86*(6), 68001.
- Newman, M. E. J. (2014). Prediction of highly cited papers. *EPL (Europhysics Letters)*, *105*(2), 28002.
- Niu, Q., Zhou, J., Zeng, A., Fan, Y., & Di, Z. (2016). Which publication is your representative work? *Journal of Informetrics*, *10*(3), 842–853.
- Noorhidawati, A., Aspura, M. Y. I., Zahila, M. N., & Abrizah, A. (2017). Characteristics of Malaysian highly cited papers. *Malaysian Journal of Library & Information Science*, *22*(2), 85–99.
- Ponomarev, I. V., Lawton, B. K., Williams, D. E., & Schnell, J. D. (2014). Breakthrough paper indicator 2.0: Can geographical diversity and interdisciplinarity improve the accuracy of outstanding papers prediction? *Scientometrics*, *100*(3), 755–765.
- Ponomarev, I. V., Williams, D. E., Hackett, C. J., Schnell, J. D., & Haak, L. L. (2014). Predicting highly cited papers: A method for early detection of candidate breakthroughs. *Technological Forecasting and Social Change*, *81*, 49–55.
- Pouris, A. (2007). The international performance of the South African academic institutions: A citation assessment. *Higher Education*, *54*(4), 501–509.
- Qi, M., Zeng, A., Li, M., Fan, Y., & Di, Z. (2017). Standing on the shoulders of giants: The effect of outstanding scientists on young collaborators careers. *Scientometrics*, *111*(3), 1839–1850.

- Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, *105*(45), 17268–17272.
- Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B: Condensed Matter and Complex Systems*, *4*(2), 131–134.
- Shen, H. W., & Barabási, A. L. (2014). Collective credit allocation in science. *Proceedings of the National Academy of Sciences*, *111*(34), 12325–12330.
- Soteriades, E. S., & Falagas, M. E. (2005). Comparison of amount of biomedical research originating from the European Union and the United States. *BMJ*, *331*(7510), 192–194.
- Stewart, J. A. (1983). Achievement and ascriptive processes in the recognition of scientific articles. *Social Forces*, *62*(1), 166–189.
- Wang, D., Song, C., & Barabási, A. L. (2013). Quantifying long-term scientific impact. *Science*, *342*(6154), 127–132.
- Wang, M., Yu, G., Xu, J., He, H., Yu, D., & An, S. (2012). Development a case-based classifier for predicting highly cited papers. *Journal of Informetrics*, *6*(4), 586–599.
- Winnink, J. J., & Tijssen, R. J. (2015). Early stage identification of breakthroughs at the interface of science and technology: Lessons drawn from a landmark publication. *Scientometrics*, *102*(1), 113–134.
- Winnink, J. J., Tijssen, R. J., & van Raan, A. F. (2016). Theory-changing breakthroughs in science: The impact of research teamwork on scientific discoveries. *Journal of the Association for Information Science and Technology*, *67*(5), 1210–1223.
- Xiao, S., Yan, J., Li, C., Jin, B., Wang, X., Yang, X., et al. (2016, July). On modeling and predicting individual paper citation count over time. In *IJCAI* (pp. 2676–2682).
- Yan, R., Tang, J., Liu, X., Shan, D., & Li, X. (2011, October). Citation count prediction: Learning to estimate future citations for literature. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 1247–1252). ACM.
- Yao, L., Wei, T., Zeng, A., Fan, Y., & Di, Z. (2014). Ranking scientific publications: The effect of nonlinearity. *Scientific Reports*, *4*, 6663.
- Yu, T., Yu, G., Li, P. Y., & Wang, L. (2014). Citation impact prediction for scientific papers using stepwise regression analysis. *Scientometrics*, *101*(2), 1233–1252.
- 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.
- Zhang, J., & Guan, J. (2017). Scientific relatedness and intellectual base: A citation analysis of un-cited and highly-cited papers in the solar energy field. *Scientometrics*, *110*(1), 141–162.
- Zhou, J., Zeng, A., Fan, Y., & Di, Z. (2016). Ranking scientific publications with similarity-preferential mechanism. *Scientometrics*, *106*(2), 805–816.
- Zhu, X., Wu, Q., Zheng, Y., & Ma, X. (2004). Highly cited research papers and the evaluation of a research university: A case study: Peking University 1974–2003. *Scientometrics*, *60*(2), 237–347.