Published in "Journal of Informetrics 13(2): 717–725, 2019"
which should be cited to refer to this work.

Discoverers in scientific citation data

Gui-Yuan Shi[a], Yi-Xiu Kong[a], Guang-Hui Yuan[b], Rui-Jie Wu[a], An Zeng[c,*], Matúš Medo[d,a,e,**]

[a] Department of Physics, University of Fribourg, Fribourg 1700, Switzerland
[b] Fintech Research Institute, Shanghai University of Finance and Economics, Shanghai 200433, PR China
[c] School of Systems Science, Beijing Normal University, Beijing 100875, PR China
[d] Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054, PR China
[e] Department of Radiation Oncology, Inselspital, University Hospital of Bern and University of Bern, Bern 3010, Switzerland

Keywords:
Citation networks
Preferential attachment
Microscopic mechanisms
Prediction

Identifying the future influential papers among the newly published ones is an important yet challenging issue in bibliometrics. As newly published papers have no or limited citation history, linear extrapolation of their citation counts—which is motivated by the well-known preferential attachment mechanism—is not applicable. We translate the recently introduced notion of discoverers to the citation network setting, and show that there are authors who frequently cite recent papers that become highly-cited in the future; these authors are referred to as discoverers. We develop a method for early identification of highly-cited papers based on the early citations from discoverers. The results show that the identified discoverers have a consistent citing pattern over time, and the early citations from them can be used as a valuable indicator to predict the future citation counts of a paper. The discoverers themselves are potential future outstanding researchers as they receive more citations than average.

1. Introduction

Preferential attachment, also known as the rich-get-richer phenomenon, cumulative advantage, or the Matthew effect, is omnipresent in bibliometrics. This phenomenon was first systematically studied in 1965 (Price, 1965), and later modeled under the name cumulative advantage process (Price, 1976). At the beginning of the surge of interest in complex networks, preferential attachment was independently proposed to explain the power-law distribution of the number of links connecting to web pages (Barabási & Albert, 1999). See Mitzenmacher (2004), Perc (2014) and Zeng et al. (2017) for reviews of preferential attachment and related models.

In the context of bibliometric research, the preferential attachment model assumes that the rate at which a paper receives new citations is proportional to the number of citations that the paper already has. The simplistic original model has been later generalized by introducing node fitness (Bianconi & Barabási, 2001; Caldarelli, Capocci, De Los Rios, & Munoz, 2002), aging (Dorogovtsev & Mendes, 2000; Wu, Fu, & Chiu, 2014), node similarity (Papadopoulos, Kitsak, Serrano, Boguná, & Krioukov, 2012). The introduction of aging is particularly important as in its absence, preferential attachment naturally leads to strong first-mover advantage (Newman, 2009). Without aging, the oldest papers will constistenly be the most cited

* Corresponding author.
** Corresponding author.
E-mail addresses: anzeng@bnu.edu.cn (A. Zeng), matus.medo@unifr.ch (M. Medo).

(Berset & Medo, 2013; Mariani, Medo, & Zhang, 2016). In particular, aging in combination with heterogeneous node fitness (Medo, Cimini, & Gualdi, 2011) produces various degree distributions that are commonly found in real systems. In Medo and Cimini (2016), this model was used to compare researcher citation impact metrics in a controlled setting. At the same time, there are many more factors that can be included to make the models more realistic (Golosovsky & Solomon, 2017; Wang, Yu, & Yu, 2009).

A particular example of extending the basic preferential paradigm is due to Medo, Mariani, Zeng, and Zhang (2016) who show that in e-commerce data represented by bipartite user-item networks, there are users who are repeatedly among the first in collecting items that become popular much later. These users are referred to as discoverers (see Cervellini, Menezes, and Mago (2016) for a related concept of trendsetters.) One simple and fundamental explanation of this observation is that besides the common users who are driven by item popularity (in agreement with preferential attachment), there are also users who are driven by intrinsic item fitness. Here fitness is the intrinsic item quality as viewed by a given audience; high fitness items are likely to become popular. A user who is sensitive to item fitness can thus link with a high-fitness item despite its lower popularity. As shown in Medo et al. (2016), discoverers are ubiquitous across various e-commerce systems. Another mechanism that, albeit fundamentally different, can lead to a similar pattern has been put forward by Wu and Holme (2009) who studied the process of the citations made by influential researchers being copied by their followers, and its impact on the network clustering coefficient.

While scholarly citation data can be naturally represented as an author-paper bipartite networks, and it is thus natural to apply the framework of discoverers on them, this has not been done before. We use the much-studied American Physical Society (APS) data (Chen & Redner, 2010; Martin, Ball, Karrer, & Newman, 2013; Medo et al., 2011) to show that discoverers are common also in citation data. Upon identifying the discoverers, we track the papers first cited by them in the future and find that they are above-average cited. Finally, we focus on the discoverers themselves and assess the success of their own publications and their overall impact (as compared with the impact of other authors).

2. Data and model

2.1. Data description and author name disambiguation

We use the citation data made available by the American Physical Society (APS). The data span from 1893 to 2016 with the time resolution of one day. There are 593443 papers, and 5553199 citations among them in total. Note that the papers published by Reviews of Modern Physics are excluded from the analysis, because review articles usually have many more citations, so they are in principle "easy targets" which should not contribute to the authors' record of early citing papers that later become very popular. We also exclude self-citations, i.e., the citing paper and the cited paper have at least one author in common. It is conventional that researchers cite their own papers, so we believe this is largely deviated from how we define a "discovery".

We consider two authors to be the same individual if they satisfy the following two conditions: (1) Their last names, and the initials of both first names and given names are exactly the same, (2) One has been cited by or citing the other one, or they have co-authors whose name satisfies the first condition. After the name disambiguation, we finally obtain 559940 authors in total.

2.2. Model

For a given time point T, the papers published between $T-K-10$ and $T-K$ are used as the training set (here 10 represents 10 years). To identify the discoverers, i.e. the researchers who often early cite papers that are eventually highly-cited, the first step is to set the thresholds for "cite early" and "highly-cited". Since in reality we cannot know any information after time T, we use a short-term citation count C_K (the citation count K years after publication) as a proxy for the paper's eventual popularity. As K increases, C_K becomes closer to the eventual popularity. At the same time, large K means that the training set is far from the current time T which limits the predictive power of the identified discoverers for two main reasons: (1) Authors active in the training set may become inactive until time T, (2) Authors' discovery ability does not have to be stationary, so the authors who were discoverers in the training set need not to be discoverers at time T. The papers whose C_K is in the top f_p% among all papers published between 1970 and 2006 form the target group of highly-cited papers. For each training set paper, its first F citations received within one year after the publication date are denoted as explorations of the paper's potential. If a paper turns out to be a highly-cited paper, those F explorations are also labeled as discoveries. The authors are assessed by the number of explorations and discoveries that they have made (see below). Fig. 1 shows a schematic diagram of the evaluation scheme. Note that our setting is more controlled than that used in Medo et al. (2016) where the current popularity of items at time T was used instead of C_K. The use of C_K is advantageous when analyzing citation data where citation counts of highly-cited papers are known to accumulate citations over extremely long periods of time (Golosovsky & Solomon, 2013).

Using the above-described procedure, we determine the total number of explorations, E_i, and discoveries, D_i, for each researcher i. Now the null hypothesis is that there are no differences between the authors in how they cite papers that later become highly-cited. In our terminology, each "exploration" has the same chance to become a "discovery" (i.e., each early citation has the same chance to be a citation of a future highly-cited paper). The overall probability that an exploration turns

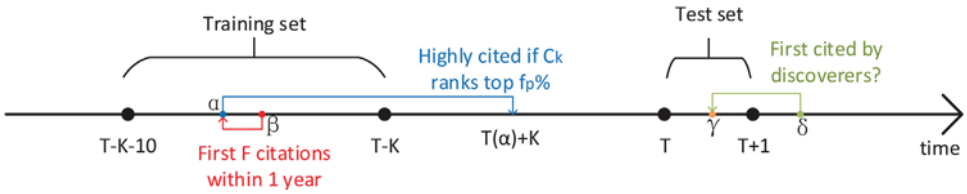

Fig. 1. A schematic plot of the evaluation of discoverers. **1)** α is a paper in the training set. If β is one of α's first F citations within 1 year, β's authors' number of explorations increase 1. If α's citation after K years ranks the top f_p%, β's authors' number of discoveries increase 1. **2)** γ is a paper in the test set, and γ is first cited by δ, if one of δ's authors is a discoverer, γ is predicted to be a highly-cited paper.

out to be a discovery is $P_d = \sum D_i / \sum E_i$. Under the null hypothesis, the number of discoveries a researcher makes follows the binomial distribution

$$P(D_i; E_i, P_d, H_0) = \binom{E_i}{D_i} P_d^{D_i} (1 - P_d)^{E_i - D_i}. \tag{1}$$

We shall test whether the null hypothesis is supported by the data. If not, how much the result disagree with the null hypothesis?

3. Results

3.1. Statistical significance of discoverers

With a large number of researchers, some of them can achieve many discoveries even if the null hypothesis actually holds. To be able to focus on really notable authors, we compare the found empirical numbers of explorations and discoveries with their expected behavior under the null hypothesis.

Similarly to the discovery probability P_D, one can introduce also the exploration probability $P_e = \sum E_i / \sum L_i$ where L_i is the number of citations made by author i in the training set. The expected distribution of the number of explorations is then closely interwound with the distribution of the number of citations. Under the null hypothesis, the expected number of authors who made E explorations is

$$N_e(E) = \sum_{x=1}^{\infty} A(x) \binom{x}{E} P_e^E (1 - P_e)^{x-E}, \tag{2}$$

where $A(x)$ is the number of authors who cited x papers. The expected number of authors who made D discoveries, $N_d(D)$, can be obtained similarly. The results in Fig. 2**a,b** demonstrate there are many researchers who make many more explorations and discoveries than expected under the null hypothesis.

Similarly to Medo et al. (2016), we use the p-value to quantify the statistical significance of author i making D_i discoveries out of E_i explorations. For a given author, the p-value, i.e. the probability of making at least D_i discoveries under the null hypothesis H_0 is

$$P_v(D_i, E_i) = \sum_{D=D_i}^{E_i} P(D; E_i, P_d, H_0). \tag{3}$$

Author achieve small p-value P_v only if they cite new papers sufficiently often (when E_i is small or zero, even the maximal possible $D_i = E_i$ may be insufficient for a statistically significant observation, see Fig. 2**d** for an illustration), and a disproportionate number of them turns out to be highly-cited in the future. To identify the statistically significant discoverers, we control the False Discovery Rate (FDR) at the level of 5% by applying the Benjamini-Hochberg procedure. The procedure works as follows: (1) rank all the active N authors who made at least one exploration in ascending order of P_v; (2) find the largest x such that $P_v(x) \leq x\alpha/N$, here α is the FDR level; (3) these x researchers who have the minimal P_v are identified as discoverers. Benjamini and Hochberg proved that the expected proportion of false discoveries (identified as discoverers because of pure luck) is α (Benjamini & Hochberg, 1995). See Fig. 2**c** for an example.

We find that the results of the our algorithm are little sensitive to the choice of parameters. To identify as many discoverers as possible, we use multiple parameter choices and mark the authors who are identified by the statistical procedure for at least one parameter setting as discoverers. In particular, we used $F \in \{1, \ldots, 5\}$, $K \in \{1, \ldots, 10\}$, and $f_P \in \{0.2, 0.4, \ldots, 2.0\}$, which corresponds to $5 \times 10 \times 10 = 500$ different parameter settings.

We identify the discoverers from 1990 to 2006 with the method described above. The total number of discoverers in each year, shown in Fig. 3**a**, wavers around 900. In Fig. 3**b**, we show the proportion of papers first cited by discoverers in each year. About 8% of all papers are first cited by discoverers and thus predicted to be potential highly-cited papers.

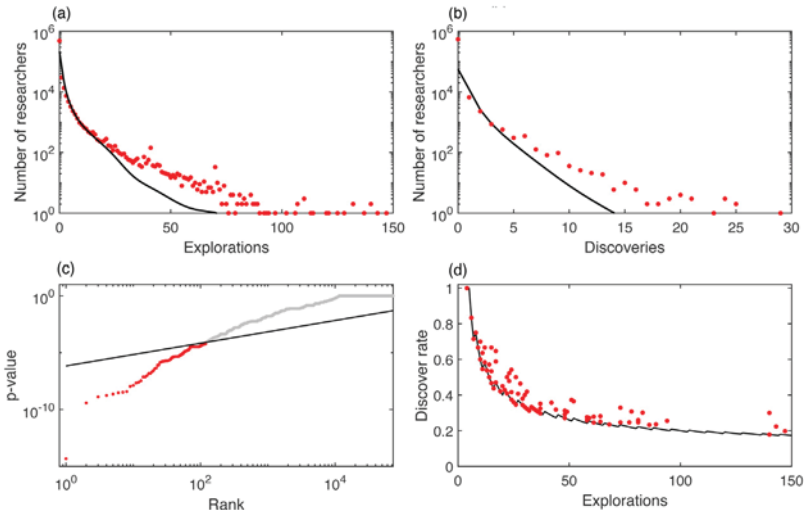

Fig. 2. Statistical significance of discoverers. We use the parameters $F=5$, $K=3$, $f_p=1.0(\%)$, and $T=2006$ as an example, that imply $P_D=7.6\%$ and $P_E=5.4\%$. **(a)** Number of researchers who make different number of explorations (red dots); the black line represents the result under the null hypothesis, see Eq. (2). **(b)** Number of researchers who make different number of discoveries (red dots); the black line represent the result under the null hypothesis. **(c)** Author p-values obtained with Eq. (3) and the Benjamini-Hochberg procedure to control the False Discovery Rate. The solid line is $r\alpha/N$ where N is the number of all active authors who made at least one exploration under the above parameter setting at T (here $N=72931$), r is the author rank (the x-axis), and α is the FDR level (we choose 5%). The dots below the line mark the authors who significantly deviate from the null hypothesis (red symbols); in this case, 124 discoverers are identified whose p-values are below $p_m=8.9\times10^{-5}$. **(d)** The discovery rate (D_i/E_i) of the identified discoverers. The solid line shows the smallest discovery rate that would suffice to achieve the p-value under p_m at a given number of explorations.

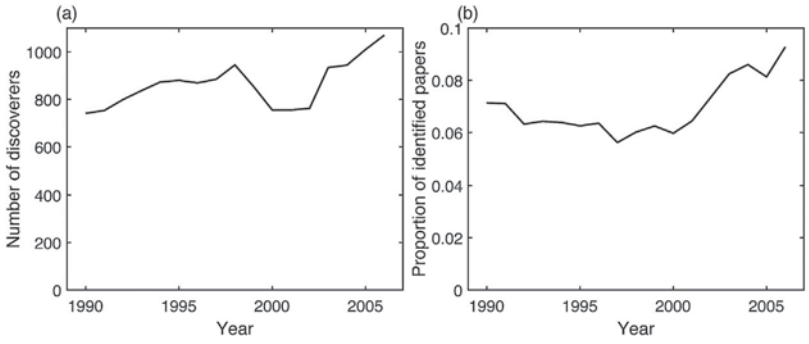

Fig. 3. **(a)** Number of discoverers in each year. **(b)** The proportion of papers first cited by discoverers among all the papers with $C_{10}\geq1$.

3.2. An example of discoverers

Previous studies (Jeong, Néda, & Barabási, 2003; Redner, 2004; Wang, Yu, & Yu, 2008)show that most of the researchers cite papers according to the preferential attachment mechanism, and there exists a strong tendency for those researchers to cite popular papers. However, we find that there exists a tiny fraction of researchers who repeatedly cite new papers that will become highly-cited in the future. Here we give an example of a typical discoverer identified in year 1990, with his/her publishing time series illustrated in Fig. 4. The discoverer is marked as discoverer 301 times among all the 500 parameter settings at $T=1990$, which is the highest of all researchers. The discoverer's first paper was published in July, 1973.

4. The behavior of discoverers

4.1. Future success of papers

Since the discoverers have historically high success in early citing highly-cited papers, it is natural to investigate whether they also have the ability to discover highly-cited papers in the future. To this end, we employ a procedure illustrated in Fig. 1: using the described procedure, we identify the discovers at the beginning of a year (time T, from 1990 to 2006) and consider all papers published between T and $T+1$ that have $C_{10}\geq1$. Among those papers, if a paper's first citation comes from another paper that has at least one discoverer among the authors, we say that the paper is "first cited by discoverers".

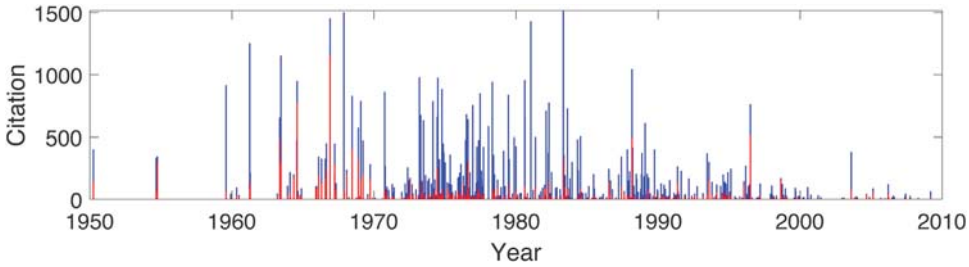

Fig. 4. The most outstanding discoverer identified in the year of 1990. Blue bars show the total citation counts of all the papers the discoverer has cited, and the red bars show the citation count of the paper when the discoverer cited it.

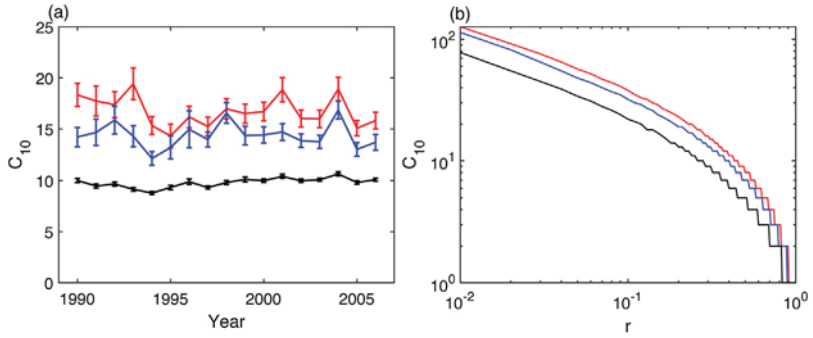

Fig. 5. **(a)** A comparison between C_{10} of the papers first cited by the discoverers (red), high h-index authors (blue), and all researchers (black), each dot represents the mean C_{10} of papers published in the year. The error bars represent the standard error of papers' C_{10}. **(b)** C_{10} of the papers that ranks r by C_{10}, r is the normalized ranking of the paper in the group. The red line, blue line and black line represent the group first cited by discoverers, high h-index authors and all authors respectively.

In the APS data, from the beginning of 1990 to the end of 2006, 186567 papers have been published and have $C_{10} \geq 1$ citation. Among them, 13157 papers have been first cited by discoverers. On average, the papers first cited by discoverers have C_{10} of 16.7, while the overall average (for papers with $C_{10} \geq 1$) is 9.85. As an additional benchmark, we compute the h-index Hirsch (2005) of all authors at time T (here the beginning of each year), select the authors with high h-index, and evaluate the papers that are first cited by high h-index authors. The h-index thresholds are 14–20, different from year to year so that the number of the selected papers is close to the number of papers cited by discoverers. The resulting 14001 papers first cited by high h-index researchers have on average C_{10} of 14.4. Fig. 5**a** shows C_{10} of the papers chosen by the three criteria (all papers with $C_{10} \geq 1$, first cited by discoverers, and first cited by high h-index authors) by paper publication year. We see that the papers cited by the discoverers have the highest C_{10} in each year. As shown in Fig. 5**b**, if we sort the papers by its ranking of C_{10} in each group, the papers first cited by discoverers outperforms the papers which have the same rankings in the other two groups.

In our procedure, the papers first cited by the discoverers are found to be highly-cited in the future. To measure how many highly-cited papers are missed by our prediction, we calculate the False Positive Rate (FPR) and the True Positive Rate (TPR) in our method. The FPR is defined as $FPR = FP/N_n$, where FP is the number of false positives, and N_n is the total number of negatives. Meanwhile the TPR is defined as $TPR = TP/N_p$, where TP is the number of true positives, and N_p is the total number of positives.

We define a paper that has $C_{10} > 68$ (ranks in the top 1% of all papers with $C_{10} \geq 1$ from 1990 to 2006) as highly-cited. As a consequence, a true positive event is defined as a paper first cited by discoverers having $C_{10} > 68$ in the future. We find the number of occasions of true positive events $TP_d = 368$. A false positive event is defined as a paper first cited by discoverer having $C_{10} < 68$ in the future. We obtain in total $FP_d = 12789$ false positive events. These values imply the FPR $FPR_d = 6.9\%$ and the TPR $TPR_d = 20.3\%$ for the prediction of highly-cited papers using discoverers. For comparison, we also study the papers first cited by high h-index authors, finding the FPR $FPR_h = 7.4\%$ and the TPR $TPR_h = 17.0\%$. Discoverer-based method outperforms the method based on high h-index authors in both FPR and TPR. We further tested other thresholds of defining the positive events and find that the conclusion holds for all thresholds.

4.2. Discoverers and social influence

The observation that the papers first cited by the discoverers are more likely to be cited in the future calls for possible explanations. In line with Medo et al. (2016), one explanation is that the discoverers are fitness-driven researchers who are sensitive to the intrinsic quality of papers, thus the papers that they cite have overall higher quality which in turn leads to the more citations. Another possible explanation is that the discoverers themselves are influential researchers, which means

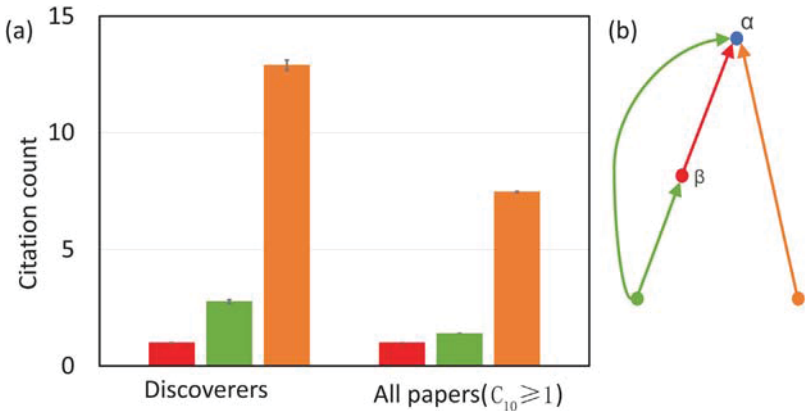

Fig. 6. **(a)** A comparison between the papers first cited by discoverers and all papers ($c_{10} \geq 1$). The red bars indicate the first citation, green bars and orange bars represent those citations from papers which cited the first citation paper as well, or not, respectively. See **(b)** for illustration.

they have many followers who frequently cite their papers and also possibly copy their references (Wu & Holme, 2009). Such copying of references is known to be an important element of the growth of citation data (Simkin & Roychowdhury, 2005, 2007). In addition, when entering a new research field, a possible way to start is to review the milestone papers by outstanding researchers in the field (and the references therein). In summary, it is plausible that researchers are more exposed to the papers that were cited by famous researchers.

To better understand the role of discoverers and their contribution to boosting paper popularity, it is crucial to distinguish between impacts of these two effects. Similar effects have been rarely discussed before because most data sets lack a direct way to determine whether a choice is an independent decision or copying the behavior of others. Fortunately, the scientific citation data give us an information about the paths of knowledge propagation and thus provide us a way to further investigate this problem.

To this end, we divide paper α's citations into three groups:

1 The first citation of paper α is labeled as β. (Red bar in Fig. 6**a**)
2 Papers that cite both α and β. These papers can be further divided into three types: (i) The papers that cite α because of the information from β. (ii) The papers that cite β because of the information from α. (iii) The papers that cite both α and β just because of coincidence or other unknown causes. It is generally difficult to distinguish these types of citations. For simplicity, we consider the number of all these papers as a measure of upper limit of the social influence of β. (Green bar in Fig. 6**a**)
3 Papers that cite α but do not cite β. (Orange bar in Fig. 6**a**)

While class two corresponds to the hypothesis that the discoverers utilize their social influence and boost the citation count of a target paper by having it co-cited with their paper, class three corresponds to the hypothesis of discoverers being the first ones to spot high quality papers. As shown in Fig. 6**a**, the third class of citations is clearly the dominant component. While some social influence of discoverers still may exist, we see that copying of the discoverers' behavior is a minor influence here.

4.3. Do the discoverers act as followers?

Another question is whether the success of discoverers can possibly lie in preferentially citing (following) highly-cited authors that, in turn, increases their chance of making discoveries. To investigate this possibility, we compute the maximum h-index of each paper's authors at the time that the paper published, h_{max}. We then divide the papers in groups by h_{max}. As shown in Fig. 7, discoverers have preference in citing papers with high h_{max}, for example, 3.12% of papers first cited by discoverers have $h_{max} \geq 30$, while for all the papers with $C_{10} \geq 1$, the proportion is 1.12%. Despite the preference, in each h_{max} group, papers first cited by the discoverers receive significantly more citations than those cited by all authors, this shows that the papers first cited by the discoverers are on average more cited regardless of whether the papers' authors are highly-cited or not.

4.4. Papers authored by the discoverers

Another pertinent question is whether the discoverers' papers also receive more citations. From the 233087 papers published in 1990–2006, 16358 papers have at least one discoverer among their authors, and 17966 papers have at least one high h-index author among their authors (The h-index thresholds are 14–21, different from year to year so that the

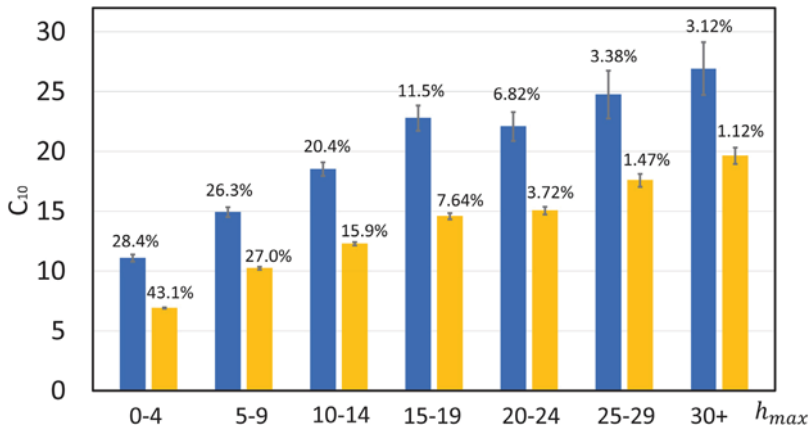

Fig. 7. A comparison of c_{10} for the papers first cited by the discoverers (blue bars), and all authors (yellow bars). Papers published in 1990–2006 are considered. The papers are divided in groups by their h_{max} (the highest h-index of their authors in the paper's publication year). The error bars represent the standard errors. The percentages above the bars indicate the proportion of the contributing papers.

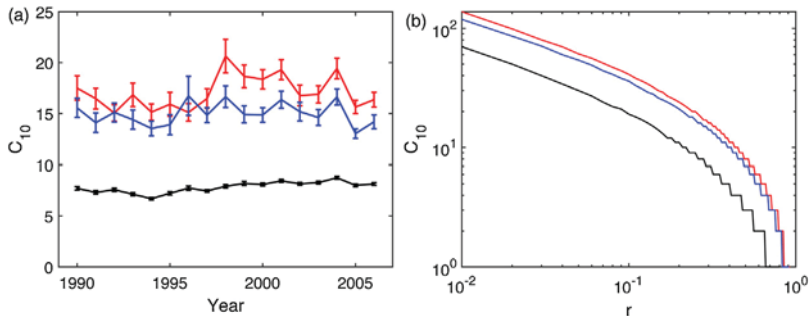

Fig. 8. **(a)** A comparison between C_{10} of the papers authored by the discoverers (red), high h-index authors (blue), and all authors (black), each dot represents the mean C_{10} of all papers published in the year. The error bars represent the standard error of papers' C_{10}. **(b)** C_{10} of the papers that ranks r by C_{10}, r is the normalized ranking of the paper in the group. The red line, blue line and black line represent the group authored by discoverers, high h-index authors and all authors respectively.

Table 1
Mean and its standard error of C_{10} for various subsets of papers; N_s denotes the subset size. The bottom part includes the results for various intersections of paper subsets discussed in the top part.

	First cited by			Authored by		
	(A) Discoverers	(B) High h-index	All	(C) Discoverers	(D) High h-index	All
C_{10}	16.7 ± 0.3	14.4 ± 0.2	9.85 ± 0.05	17.1 ± 0.2	14.9 ± 0.2	7.88 ± 0.03
N_s	13157	14001	186567	16358	17966	233087
	$A \cap B$	$A \cap C$	$A \cap D$	$B \cap C$	$B \cap D$	$C \cap D$
C_{10}	17.2 ± 0.4	24.4 ± 0.7	24.3 ± 0.8	23.0 ± 0.8	21.5 ± 0.8	19.6 ± 0.4
N_s	5490	3172	2292	2211	2161	6822

number of the selected papers is close to the number of papers authored by discoverers.). Mean C_{10} values for these three groups of papers are 7.88, 17.1, and 14.9, respectively. Similarly to the papers first cited by the discoverers, papers authored by the discoverers also have significant citation lead over the two benchmark groups (see Fig. 8**a** for a comparison by paper publication year).

4.5. Further improving the predictions

Finally, we compare the citation counts of various subsets of papers in Table 1 . Notably, papers belonging to multiple intersections of the paper subsets discussed so far (*e.g.*, papers that are both first cited and authored by a discoverer—intersection $A \cap C$) have even higher C_{10} which suggests further possibilities for studying the predictive power of various author subsets (here we focused on the discoverers and used high h-index authors as a benchmark) in the future.

5. Discussion

This paper reveals a heterogeneity of researcher's citing behavior. We identify the researchers—the discoverers—who frequently cite among the first papers that become highly-cited in the future. We focus on these discoverers to assess their future citing behavior. Our results show that the discoverers' ability is persistent in time to the extent that the discoverers identified at time T can be used to identify future highly-cited papers. We show that the social influence plays a minor role in explaining the discoverers' behavior, for which the main hypothesis still remains that the discoverers are the authors that are driven (or at least more sensitive than the average) by the intrinsic paper fitness (as originally suggested by Medo et al. (2016)).

Despite the huge success of preferential attachment mechanism in explaining the richer-get-richer phenomenon, there is a number of citation patterns that it cannot account for: newly published papers quickly becoming popular, and highly cited old papers being not often cited anymore, for example. To solve these deviations from the preferential attachment, researchers have developed models like the fitness model (Bianconi & Barabási, 2001; Caldarelli et al., 2002) and the aging model (Dorogovtsev & Mendes, 2000; Wu et al., 2014). While these two models are able to provide possible explanations for the anomalies in the citation data diverging from the preferential attachment prediction, a detailed model that utilizes both the fitness information and temporal information to extract the fitness-driven behavior of users has been lacking for years. Our work provides a well-calibrated method assuming a mix of users with different sensitivity to paper fitness.

We divide the researchers into two groups: fitness-driven researchers and popularity-driven researchers. Fitness-driven researchers, namely the Discoverers, can find high-quality papers at early stage and increase the early citation of these papers, which makes high-quality papers easier to be noticed by popularity-driven researchers and finally stand out. This contributes to the relation between the quality of a paper and its citation count. Although the fitness model serving as the fundamental of this paper could offer a possible explanation of the phenomenon that a few researchers can constantly discover the new papers that can become popular in the future, a comprehensive understanding of why these researchers have the predictivity still await further studies.

Our findings can be used for early assessment of paper potential impact. According to our study, focusing on the discoverers and assigning more weight to their citations when evaluating new papers is helpful in finding high-quality papers short after they are published. Another interesting result is that the papers authored by discoverers tend to receive more citations than those of researchers with high h-index. For the first time, this paper reveals a correlation between researchers' citing behavior and their academic potential. This may in turn provide new inputs for the problem of identifying potential future outstanding researchers.

Author contributions

Guan-Yuan Shi: Collected the data; Performed the analysis; Wrote the paper.
Yi-Xiu Kong: Collected the data; Performed the analysis.
Guang-Hui Yuan: Collected the data; Performed the analysis.
Rui-Jie Wu: Performed the analysis; Wrote the paper.
An Zeng: Concieved and designed the analysis; Contributed data or analysis tools; Wrote the paper.
Matus Medo: Concieved and designed the analysis; Contributed data or analysis tools; Wrote the paper.

Acknowledgments

The authors would like to thank Prof. Yi-Cheng Zhang for helpful discussion. This work was supported by the National Natural Science Foundation of China (Grant No. 61603046), the Natural Science Foundation of Beijing (Grant No. L160008), and the China Scholarship Council.

References

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512.
Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300.
Berset, Y., & Medo, M. (2013). The effect of the initial network configuration on preferential attachment. *European Physical Journal B*, *86*, 260.
Bianconi, G., & Barabási, A. L. (2001). Bose-Einstein condensation in complex networks. *Physical Review Letters*, *86*(24), 5632.
Caldarelli, G., Capocci, A., De Los Rios, P., & Munoz, M. A. (2002). Scale-free networks from varying vertex intrinsic fitness. *Physical Review Letters*, *89*(25), 258702.
Cervellini, P., Menezes, A. G., & Mago, V. K. (2016). *Finding Trendsetters on Yelp Dataset. 2016 IEEE Symposium Series on Computational Intelligence (SSCI).* pp. 1.
Chen, P., & Redner, S. (2010). Community structure of the physical review citation network. *Journal of Informetrics*, *4*, p278.
Dorogovtsev, S. N., & Mendes, J. F. F. (2000). Evolution of networks with aging of sites. *Physical Review E*, *62*(2), 1842.
Golosovsky, M., & Solomon, S. (2013). The transition towards immortality: Non-linear autocatalytic growth of citations to scientific papers. *Journal of Statistical Physics*, *151*, 340.
Golosovsky, M., & Solomon, S. (2017). Growing complex network of citations of scientific papers: Modeling and measurements. *Physical Review E*, *95*, 012324.
Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America*, *102*(46), 16569.

Jeong, H., Néda, Z., & Barabási, A. L. (2003). Measuring preferential attachment in evolving networks. *EPL (Europhysics Letters)*, *61*(4), 567.
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.
Martin, T., Ball, B., Karrer, B., & Newman, M. E. J. (2013). Coauthorship and citation patterns in the Physical Review. *Physical Review E*, *88*, 012814.
Medo, M., Cimini, G., & Gualdi, S. (2011). Temporal effects in the growth of networks. *Physical Review Letters*, *107*(23), 238701.
Medo, M., Mariani, M. S., Zeng, A., & Zhang, Y. C. (2016). Identification and impact of discoverers in online social systems. *Scientific Reports*, *6*, 34218.
Medo, M., & Cimini, G. (2016). Model-based evaluation of scientific impact indicators. *Physical Review E*, *94*, 032312.
Mitzenmacher, M. (2004). A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, *1*, 226.
Newman, M. E. (2009). The first-mover advantage in scientific publication. *EPL (Europhysics Letters)*, *86*(6), 68001.
Papadopoulos, F., Kitsak, M., Serrano, M. Á., Boguná, M., & Krioukov, D. (2012). Popularity versus similarity in growing networks. *Nature*, *489*(7417), 537.
Perc, M. (2014). The Matthew effect in empirical data. *Journal of The Royal Society Interface*, *11*, 20140378.
Price, D. J. D. S. (1965). Networks of scientific papers. *Science*, 510–515.
Price, D. D. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the Association for Information Science and Technology*, *27*(5), 292–306.
Redner, S. (2004). *Citation statistics from more than a century of physical review*. , arXiv preprint physics/0407137.
Simkin, M. V., & Roychowdhury, V. P. (2005). Stochastic modeling of citation slips. *Scientometrics*, *62*, 367.
Simkin, M. V., & Roychowdhury, V. P. (2007). A mathematical theory of citing. *Journal of the American Society for Information Science and Technology*, *58*, 1661.
Wang, M., Yu, G., & Yu, D. (2008). Measuring the preferential attachment mechanism in citation networks. *Physica A: Statistical Mechanics and its Applications*, *387*(18), 4692–4698.
Wang, M., Yu, G., & Yu, D. (2009). Effect of the age of papers on the preferential attachment in citation networks. *Physica A: Statistical Mechanics and its Applications*, *388*(19), 4273–4276.
Wu, Y., Fu, T. Z., & Chiu, D. M. (2014). Generalized preferential attachment considering aging. *Journal of Informetrics*, *8*(3), 650–658.
Wu, Z. X., & Holme, P. (2009). Modeling scientific-citation patterns and other triangle-rich acyclic networks. *Physical Review E*, *80*(3), 037101.
Zeng, A., Shen, Z., Zhou, J., Wu, J., Fan, Y., Wang, Y., & Stanley, H. E. (2017). The science of science: From the perspective of complex systems. *Physics Reports*, *714/715*, 1–73.