From Zero to One: A Perspective on Citing

Yong Huang †

Information Retrieval and Knowledge Mining Laboratory, School of Information Management, Wuhan University, Wuhan, Hubei, China

Yi Bu †

Center for Complex Networks and Systems Research, School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, U.S.A.

Ying Ding

School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, U.S.A.

School of Information Management, Wuhan University, Wuhan, Hubei, China

School of Management, Tianjin Normal University, Tianjin, China

School of Management, Jilin University, Changchun, Jilin, China

School of Management, Shanxi Medical University, Taiyuan, Shanxi, China

Wei Lu *

Information Retrieval and Knowledge Mining Laboratory, School of Information Management, Wuhan University, Wuhan, Hubei, China

†: Equal contribution.

***: Correspondence concerning this article should be addressed to Dr. Wei Lu**, Email: weilu@whu.edu.cn.

From Zero to One: A Perspective on Citing

Abstract: This paper investigates the lengths of time that publications with different numbers of citations take to receive their first citation (the beginning stage), and then compares the lengths of time to receive two or more citations after receiving the first citation (the accumulative stage) in the field of computer science. We find that in the beginning stage, i.e., from zero to one citation, highly, medium-, and lowly cited publications do not obviously exhibit different lengths of time. However, in the accumulative stage, i.e., from one to N citations, highly cited publications begin to receive citations much more rapidly than medium and lowly cited publications. Moreover, as N increases, the difference in receiving new citations among highly-, medium-, and lowly-cited publications increases quite significantly.

Keywords: incremental time; response time; beginning stage; accumulative stage; scientific impact; citation count; success; science of science.

INTRODUCTION

In the field of science of science, much extant literature has focused on the temporal process of scientific publications' receiving citations and research on citation distributions. It is worth noting that a considerable number of publications exist that have never been cited, regardless of discipline or date of publication. Thus, studies related to citation distribution mainly focus on publications that have been cited at least once. For instance, by proposing a simple model with a random selection process, Wallace *et al.* (2009) found that the proportion of cited publications is correlated to three variables: 1) the number of competing publications; 2) the number of citing publications; and 3) the number of references that publications contain. A scientific article's first citation *might* occur, if at all, shortly after its publication (Barnett, Fink, & Debus, 1989; Rousseau, 1994; Wallace, Larivière, & Gingras, 2009). Several researchers have modeled the first citation distribution of publications mathematically, among which Rousseau (1994) proposed two exponential models to fit the first citation processes and the response distribution of publications. Egghe (2000) combined an exponentially-decreasing aging function of publications' citations and a Lotka function, and demonstrated how to estimate two important parameters, namely, aging rate and Lotka's exponent, in the model. A further discussion of the relationship between first citation and general citation age-distribution has been presented in his following work (Egghe & Rao, 2001). Similarly, Burrell (2001) employed a non-homogeneous Poisson process and provided a stochastic model for simulating first citations. Although these studies provided mathematical interpretations of receiving the first citation, they failed to compare the beginning stage of different-impact publications' citations or to discuss additional implications.

An important branch of study on the beginning stage of scientific publications' citations is delayed recognition (e.g., Burrell, 2005; Glänzel. Schlemmer, & Thijs, 2003). This has been termed a "sleeping beauty in science" by van Raan (2004), referring to publications that are not cited (or cited at a very low rate) and then suddenly become highly-cited. He also defined several measurements, such as depth of "sleep," length of "sleep," and "awake" intensity, and identified some such "sleeping beauties." This scenario had actually been discussed previously by Garfield (1989a, 1989b, 1990), but he used a small volume of citation data. More recently, Redner (2005) analyzed the sleeping beauty phenomenon in the field of physics, yet his selection of sleeping beauties was arbitrary. By using almost the entire Web of Science data, Ke, Ferrara, Radicchi, and Flammini (2015) defined sleeping beauties by importing a beauty coefficient quantitatively and highlighted that the phenomenon of the sleeping beauty is not exceptional.

After a publication receives its initial citation, it is likely to be cited more, which constitutes an accumulative stage of receiving citations (from one to *n* citations). Extant research related to temporal-based citation distribution could be categorized into two sets, with retrospective and prospective perspectives, respectively (Yin & Wang, 2017). The retrospective studies first target a specific (set of) publication(s) and analyze the citation distribution of the references of this (these) publication(s) (e.g., Pan, Petersen, Pammolli, & Fortunato, 2018; Price, 1965; Stinson & Lancaster, 1987), while prospective studies consider the citation distribution of a publication with a forward perspective (e.g., Parolo *et al.*, 2015; Redner, 2005; Sanyal, 2006). Research concentrating on the accumulative stage of publications' citations mainly comprises the latter view. In this branch of inquiry, Burrell (2002) investigated the n^{th} citation distribution with a stochastic model with a Gamma distribution for a latent rate based on his previous models (Burrell, 2001). Egghe and Rousseau (2000) examined the effects of growth on citation distribution, and concluded that more items exist to be referenced with growth. Min *et al.* (2018) employed the Bass diffusion model to understand the citation process of scientific publications. Wang, Song, and Barábasi (2013) quantitatively modeled publications' long-term citation distribution and proposed a mathematical formula to predict publications' citation counts in the future. Some other studies concerning the accumulative stage of publications' citations focused on cumulative advantage at this stage (e.g., Allison and Stewart, 1974; Allison, Long, and Krauze, 1982; Nakamoto, 1988; Price, 1976; Rousseau, 1988).

However, few studies have explored differences in the periods required to receive one or more citations between publications that receive different numbers of citations at last retrospectively, or have investigated whether differences exist in the "difficulty" in receiving the first citation among publications with different numbers of citations. Therefore, the current paper fills this gap by investigating the lengths of time between the years that highly, medium, and lowly cited articles are published and when they receive their first citations, as well as the lengths of time needed to receive second or more citations relative to when they received their first.

People might think that highly cited publications possess certain “inherent” advantages and receive new citations much more rapidly than lowly-cited publications, even regarding being cited for the very first time—a seemingly “inherent” advantage that highly-cited publications might need a shorter time to accumulate one more citation. However, our findings show that publications with different numbers of citations take a similar length of time to receive their first citation, and thus highly-cited publications did not receive obviously inherent advantages in obtaining their first citations since they were published. Moreover, significant differences in the time required to receive the second and more citations for highly, medium-, and lowly cited publications exist, and these differences are augmented in later stages. Overall, our findings identify, counter-intuitively, the non-existence of inherent advantages of highly-cited publications in receiving their first citations.

The remainder of this paper proceeds as follows. We detail the data processing and methods utilized for our analysis. Next, we present and discuss our findings, and their interpretations. Finally, we conclude with a summary of our findings, implications, limitations, and directions for future research.

METHODOLOGY

Data and processing

The data used in this article are derived from the ArnetMiner dataset, which covers the most important articles in conferences and journals from the domain of computer science (Bu *et al.*, 2018a, 2018c; Li *et al.*, 2008; Tang *et al.*, 2007, 2008). There are approximately one million articles published between 1936 and 2014 in the dataset. Figure 1 shows the publication number distribution over these years, in which the years after 1980 witnessed the largest number of publications. This dataset also includes approximately eight million citation relationships within these publications collected at the end of 2014. Among all publications, 940,974 (73.2%) have received at least one citation. To eliminate the effect that recently published articles have lower possibilities to be cited, we analyze all articles published prior to 2005. The reason why 2005 is selected as a criterion is detailed in **Appendix 1**.

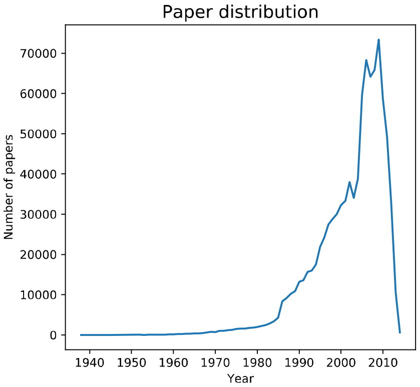


Figure 1. Distribution of the number of publications in different years.

We also partition all publications into three groups, lowly, medium-, and highly cited publications, based on their numbers of citations. To achieve this, the citation distribution is plotted and analyzed. Finally, those publications that are cited fewer than 14 times are defined as lowly cited publications, while publications that received at least 1,000 citations are defined as highly cited publications. The remainder are classified as medium cited publications. The details of how the thresholds are determined can be found in **Appendix 2**. As shown in Figure 2, lowly, medium-, and highly cited publications comprise 83.87%, 16.10%, and 0.03% of all publications, respectively. In addition, all of the publications that did not receive any citations have been removed in advance.

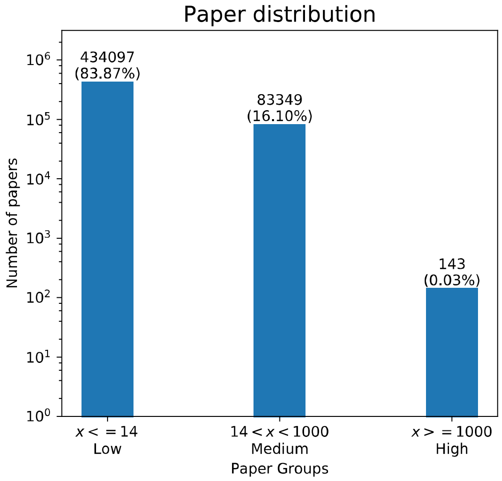


Figure 2. Descriptive statistics on highly (right), medium (middle), and lowly cited (left) publications.

From Figure 2, one can see that the differences between the highly cited publications and the other two are extremely large, and thus a process of random sampling in lowly and medium cited publications is requisite. This random sampling procedure should ensure that the sampled data follow a similar distribution to the original data. We therefore employ two normal distributions to select lowly and medium cited publications, which is detailed in **Appendix 3**. Finally, 1,000 lowly-cited publications, 1,000 medium-cited publications, and 143 highly-cited publications from the original dataset are chosen for further analyses.

Methods

Beginning stage and accumulative stage

The process of a publication receiving citations can be divided into two stages: 1) a *beginning stage*, in which the number of citations increases from zero to one; and 2) an *accumulative stage*, in which the number of citations accumulates after reaching one.

Accumulative time, incremental time, and response time

Mathematically, a scientific publication P, published in the year of y_0, was cited by publications c_1, $c_2, \ldots,$ c_n in accordance with the time that it received these citations. As a result, c_1 is the first citation, and was published in year y_1, and c_i is the i^{th} citation published in year y_i, and the last citation c_n was published in year y_n ($y_n \geq y_i \geq y_1 \geq y_0, 1 \leq i \leq n$). Here, we define the *accumulative time* as the length of time required to receive the first i citations for P, which is equal to $(y_i - y_0)$ and denoted as Y_i.

The *incremental time* is defined as the length of time required between receiving the i^{th} citation and the $(i-1)^{th}$ citation, which is equivalent to $(Y_i - \ Y_{i-1})$ and is represented as T_i. If two adjacent citations, c_{i-1} and c_i, were received in the same year, T_i is equal to zero. Publications might take different periods of time to receive one more citation in which the number of citations increases from zero to one, from one to two, …, and from $(n-1)$ to n. The incremental time, i.e., the time interval between two adjacent citations, is used to evaluate the “difficulty” in different stages of citation processes.

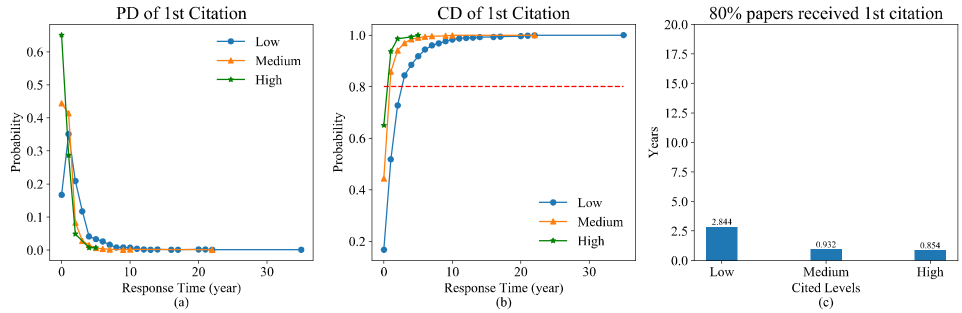
Obviously, $Y_1 = T_1 = y_1 - y_0$, and both Y_1 and T_1 serve as the time required to receive the first citation for P; Y_1 and T_1 are defined as the *response time* of P. Essentially, the response time expresses the length of time that a scientific publication takes to receive its first citation, which constitutes a critical indicator because, after the response time, this publication will shift its status from “unseen” to “seen” (Egghe, 2000). The length of response time reflects the initial “difficulty” that a publication experiences in the beginning stage.

The empirical study in this paper investigates the probability and cumulative distributions (PD and CD, respectively) of the time of receiving the first citation ($N = 1$). For accumulative stages, we examine the PD and CD of the time of receiving the fifth and the tenth citations ($N = 5$ and $N = 10$), respectively, as two typical examples. Note that here N is defined as the index of the citation received—for instance, $N = 1$ means the first citation received, $N = 5$ refers to the fifth citation received, etc.

RESULTS AND DISCUSSION

Beginning stage: From zero to one

Figures 3(a) and 3(b) show the probability distribution (PD) and cumulative distribution (CD) of highly-, medium-, and lowly-cited publications' response time in the beginning stage, shown as green, orange, and blue lines, respectively. From Figure 3(a), one can see that the curves representing highly- and medium-cited publications exhibit clear decreasing trends as response time increases, which means that most of these publications receive their first citation within one year after they have been published. For instance, Figure 3(a) shows that ~65% and 44% of the highly- and medium-cited publications, respectively, receive their first citation within their published years. Moreover, in our dataset, all highly-cited publications received their first citation within four years after their publication. This indicates the infrequency of "sleeping beauties" (Van Raan, 2004) among Association for Computing Machinery (ACM) publications, as ArnetMiner mainly covers publications from ACM conferences and journals (Bu, Ding, Liang, & Murray, 2018b; Tang *et al.*, 2008). The lowly-cited publications, however, show a slightly different pattern, in which the curve representing lowly-cited publications increases and then decreases when the response time is greater than one year. Specifically, approximately 18% and 35% of the lowly-cited publications received their first citation exactly in the year that they were published and one year after their publishing, respectively.



(a)

(b)

(c)

Figure 3. First citation distribution over response time of publications in three groups.

The cumulative distribution curves of the three groups of publications (Figure 3(b)) exhibit steep slopes, demonstrating that most of the publications with cited records receive their first citation quickly. When comparing the three curves in Figure 3(b), it can be seen that the curve representing highly-cited publications increases faster than the other two, while that representing medium-cited publications increases more rapidly than that for lowly-cited publications. Particularly, more than 70% of the lowly-cited publications are able to receive their first citations within two years after they are published; the corresponding numbers for medium- and highly-cited publications are 93% and 99%, respectively. However, such differences are not very obvious when the response time is greater. For example, Figure 3(b) shows that 95% and 99% of the lowly- and medium-cited publications, respectively, received their first citations within 10 years after their publication, and such difference is quite small.

We further compare the average time period that 80% of publications in each group received their first citation. As the red dashed line in Figure 3(b) and the blue bars in Figure 3(c) show, the length of time required by 80% of the lowly-cited publications to receive their first citations is 2.844 (years, the same below), while that of medium- and highly-cited publications is 0.932 and 0.854, respectively. Again, these findings indicate that highly-cited publications receive their first citation a little bit more rapidly than medium-cited publications, which are faster than lowly-cited publications. However, such differences are not as large as traditional wisdom asserts, i.e., the difference between medium- and highly-cited publications, as well as that between highly- and medium-cited publications, is quite small based on Figure 3. Regardless of the stages of an academic career, such as Ph.D. or faculty, 1.924 years is not a very long period of time to receive more citations. In other words, the process of receiving the first citation is found to constitute a relative period of time for highly-, medium-, and lowly-cited publications, given that these publications finally receive at least one citation.

Based on these findings, one can see that although minor differences exist among highly-, medium-, and lowly-cited publications, such differences are not very apparent in the beginning stage. In other words, publications with different numbers of citations need relatively similar periods of obtaining their first citation.

Accumulative stage: From one to N

To investigate the process of the accumulative stage, we first provide two case studies, receiving the first to the fifth citations ($N = 5$), and from the first to the tenth citations ($N = 10$). The reason why 10 is selected as a criterion is that the average number of citations of the lowly-cited publications in the current paper is approximately 10 (see details in **Appendix 2**). Figures 4(a), 4(b), 4(d), and 4(f) show the probability distribution (PD) and cumulative distribution (CD) of the time required to receive five or 10 citations, in which highly-, medium-, and lowly-cited publications are shown as green, orange, and blue lines, respectively. In Figures 4(a) and 4(d), all six curves exhibit increasing-decreasing trends, although the peak values occur in various years. If comparing the performances of the same group in these two figures, one can find that highly- and medium-cited paper groups are similar, but lowly-cited publications show more difference. Specifically, in the fourth year after a lowly-cited article is published, approximately 18% and 5% of them are able to receive five and 10 citations, respectively.

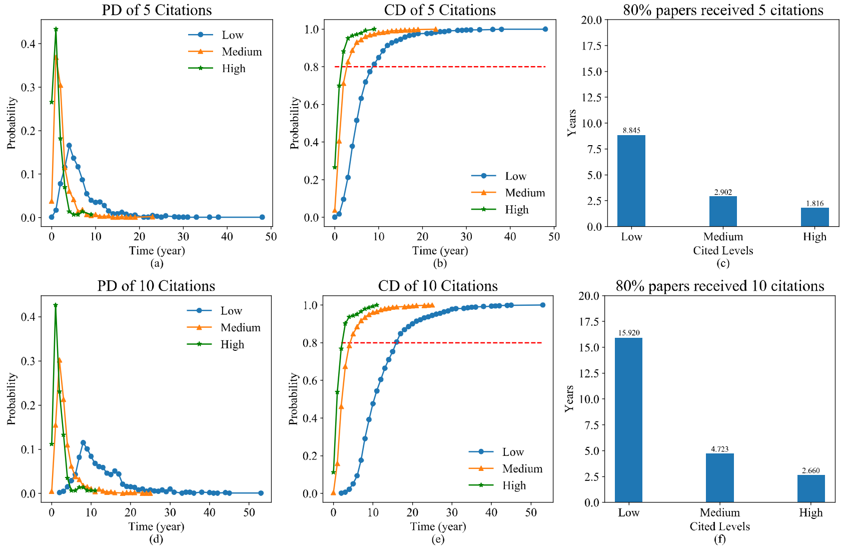


Figure 4. Distribution of the time required to accumulate citations from the first to the fifth citations (a and b); distribution of the time required to accumulate citations from the first to the tenth citations (d and e); and comparisons of the time required for 80% of publications in each group to receive five or 10 citations (c and f).

Although all curves in Figures 4(b) and 4(e) exhibit a clear increasing trend, the difference among curves in Figure 4(b) is smaller than that in Figure 4(e). Within two years after publication, for example, approximately 90%, 75%, and 21% of the highly-, medium-, and lowly-cited articles, respectively, are expected to receive five citations. The corresponding numbers are 78%, 45%, and 3% in the case of 10 citations. These findings indicate that the speed of highly-cited publications receiving additional citations is faster than that of medium- and lowly-cited publications in the accumulative stage.

A comparison of the time required for 80% of publications in each group to receive five or 10 citations is shown as red dotted lines in Figures 4(b) and 4(d) or Figures 4(c) and 4(f). In Figure 4(c), for instance, to acquire five citations, 80% of the highly-cited publications need 1.816 years, while the same percentage of lowly-cited publications is 8.845 years; the numbers of years of highly- and lowly-cited publications for receiving 10 citations are, respectively, 2.660 and 15.920, as shown in Figure 4(e).

The results presented in Figures 4(c) and 4(f) are also illuminating. Almost nine years are required for 80% of the lowly-cited publications to receive five citations, but highly-cited publications just need ~ 1.8 years. When examining the length of time publications needed to receive 10 citations, it is shown that the numbers for highly- and lowly-cited publications are 2.660 and 15.920 years, respectively. Indeed, 15 years is a very long period of time for researchers—some of which will be promoted to be a full professor from a beginning assistant professor during this time span—, and there are *not* many sets of 15 years within a typical academic career. Since highly visible publications are often measured by their large number of citations, the current result highlights the obvious difference in the "difficulty" of receiving more citations between highly and lowly visible publications. Our results show that highly visible publications receive more citations much more easily in the accumulative stage, which suggests that scholars should try their best to show and introduce their work to increase the visibility of their publications, as those with high visibility tend to receive citations more quickly. We know that those who are initially appointed as new faculty members in the U.S. normally have fewer than seven years to secure a tenured position at a university, and citation count is often regarded as an important indicator in tenure evaluation (Long, 1978); consequently, during this period, scholars should produce highly visible publications in order to obtain citations rapidly for promotion.

To demonstrate our results more explicitly, we calculate the difference between the lengths of time for highly-, medium-, and lowly-cited publications to receive one, five, and 10 citations, as shown in Table 1. It is indicated that the speed of accumulating citations for differently-cited publications is not obviously large in the beginning stage (from zero to one); however, differences gradually appear as time passes (from one to N), and such differences become increasingly apparent as more citations accumulate. The difference essentially shows the diverse degrees of "difficulty" of receiving more citations for various publications.

Table 1. Differences between the lengths of time of receiving the first, fifth, and tenth citations ($\boldsymbol{N = 1, 5, \mathbf{and}\ \ 10}$) for highly-, medium-, and lowly-cited publications.

	H			**M**			**L**		
	BS	**AS (N=5)**	**AS (N=10)**	**BS**	**AS (N=5)**	**AS (N=10)**	**BS**	**AS (N=5)**	**AS (N=10)**
H	-			0.078	1.086	2.063	1.990	7.029	13.260
M	0.078	1.086	2.063	-			1.192	5.943	11.197
L	1.990	7.029	13.260	1.192	5.943	11.197	-		

Note: BS = beginning stage, AS = accumulative stage, H = highly-cited paper group, M = medium-cited paper group, L = lowly-cited paper group.

The aforementioned section used $N = 5$ and $N = 10$ as two cases to elucidate how publications in different groups dynamically accumulate their citations over time. To achieve a detailed understanding of this, we divide all received citations of a publication into five citation zones: 1) 0~20%; 2) 21%~40%; 3) 41%~60%; 4) 61%~80%; and 5) 81%~100%. For example, suppose that a publication has received 50 citations; we then divide this into five zones, each of which, respectively, contains the 1-10th, 11-20th, 21-30th, 31-40th, and 41-50th citations. For each zone, we calculate the mean, median, and mode of the incremental time, T_i, of citations and include them into a plot, as shown in Figure 5. In the figure, the extreme values are plotted with two "-" at the top and the bottom, and the blue lines represent the distributed span of incremental time for receiving the corresponding number of citations. For better visualization, we apply a logarithmic scale in the vertical axes. Note that the real value shown in Figures 5(a), 5(b), and 5(c) are $\log(T_{i+1}) + 1$ instead of $\log(T_{i+1})$ to avoid zero T_i s. The distributions of mean, median, and mode for each group are shown in Figures 5(d), 5(e), and 5(f), respectively.

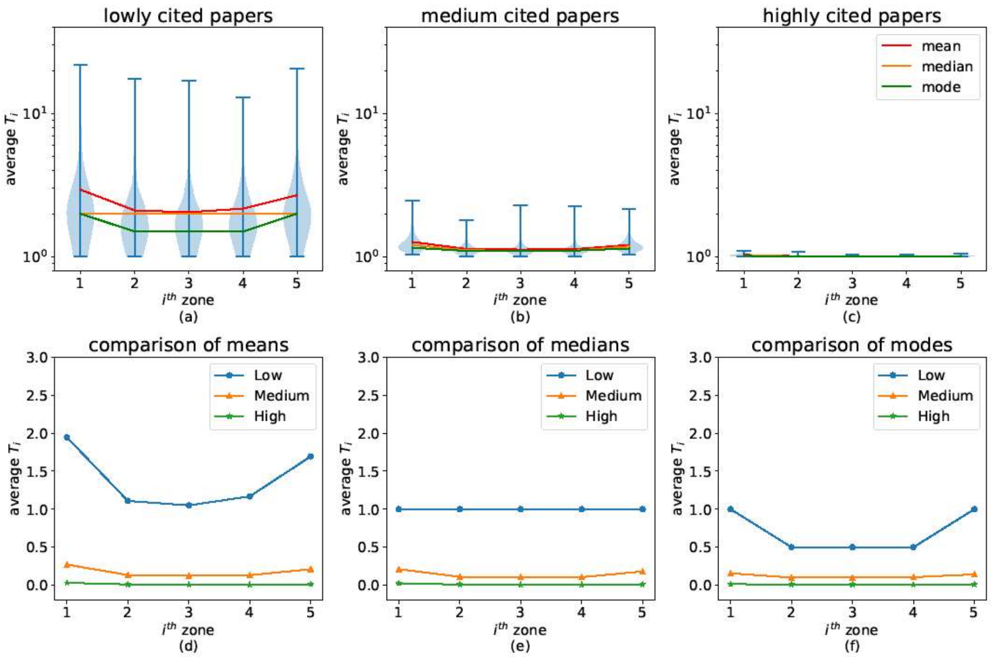


Figure 5. Plots of publications receiving citations for different groups of publications (a-c). Mean curves (d), median curves (e), and mode curves (f) are plotted together for comparison.

It is found that the lowly-cited publication group has the largest range of incremental time, shown by the length of the blue lines, and thus a larger variation is indicated. Although Figures 5(a) to 5(c) do not present obvious trends of the incremental time range for the lowly-cited group, it is shown that for medium- and highly-cited publications, the number at each bar decreases as T_N increases.

In Figure 5(d), the blue line, representing the mean of the incremental time of the corresponding received citation of low-impact publications, is higher than the orange line, which is also higher than the green line. This shows that highly-cited publications receive citations more easily than do medium-cited publications, which receive citations more easily than lowly-cited publications. The three median curves in Figure 5(e) demonstrate that when receiving each citation, more than half of the lowly-cited publications take more time to receive one more citation than highly- and medium-cited publications, as the median of the lowly-cited publications is obviously lower than that of other publications. The mode curves shown in Figure 5(f) are also illustrative. Specifically, all modes for medium- and highly-cited publications are approximately zero, which means that at each citation position, most publications that cited them are published in the same year as them. We also find that the mode values of time intervals of lowly-cited publications are 0.5 or one, indicating that even most of the lowly-cited publications receive their citations in a short time period.

One of the implications of this finding is that scholars should focus on enhancing the visibility of publications in order to receive more citations quickly, especially those who have recently been appointed as new faculty members, as the rate of accumulating citations is widely disparate for highly- and lowly-cited publications. Yet, the current analysis did not highlight any causality between these factors, as ours is simply a descriptive analysis without any causal inference.

CONCLUSIONS

This paper investigates the lengths of time that publications with different numbers of citations need to receive their first citation (the beginning stage), and compares the lengths of time needed to receive more citations after receiving the first citation (the accumulative stage) in the field of computer science. We find that in the beginning stage, i.e., from zero to one citation, highly-, medium-, and lowly-cited publications do not exhibit obviously different lengths of time. However, in the accumulative stage, i.e., from one to N citations, highly-cited publications begin to receive citations much more rapidly than do medium- and lowly-cited publications. Moreover, as N increases, the difference in receiving new citations between highly-, medium-, and lowly-cited publications increases.

One of the limitations of this article is that scientific publications might be cited before officially published, due to the increasingly heavier use of preprint repositories, which is not considered in this paper, but can be a future topic to explore. Several other related studies can be performed in the future to expand and deepen the generalizability of the present findings. First, we simply targeted *publications* and examined the incremental, response, and accumulative time of publications' receiving citations; future work might focus on determining precisely how *authors* accumulate citations in their career. Second, this research fails to consider who cited the targeted publications. Thus, additional research could investigate the relationship between the citation distribution of citing publications and targeted publications, e.g., whether the first several citations of a given publication were from self-citations. Third, the publications analyzed in this study are limited to the field of computer science. Future studies could apply these techniques to other disciplines, examining the various patterns with which different-impact publications accumulate their citations over time. Finally, the citation count used in this paper is actually a local citation count, which might bias the current results by excluding citations from outside fields. To address this, future work should also comprise global citation counts from various fields. A lack of theoretical underpinning is another limitation of the current study. In the future, we will conduct qualitative studies (e.g., survey and interview) to explore more about the motivation of citing—for instance, why authors cite a specific publication shortly after this publication was published, and whether authors like to cite articles that have more citations and are related.

ACKNOWLEDGMENTS

This article is financially supported by the major program of the Social Science Foundation of China (No. 17ZDA292). Part of this paper was presented by Ying Ding in several seminars held in China in the summers of 2017 and 2018. The authors would like to thank Pei-Ying Chen, Chao Lu, and Lili Miao for their valuable suggestions, and are also grateful to two anonymous reviewers for improving the quality of this paper.

REFERENCES

Allison, P. D., Long, J. S., & Krause, T. K. (1982). Cumulative advantage and inequality in science. *American Sociological Review, 47*(5), 615-625.

Allison, P. D., & Stewart, J. A. (1974). Productivity differences among scientists: Evidence for accumulative advantage. *American Sociological Review, 39*(4), 596-606.

Barnett, G. A., Fink, E. L., & Debus, M. B. (1989). A mathematical model of academic citation age. *Communication Research, 16*(4), 510-531.

Bu, Y., Ding, Y., Xu, J., Liang, X., Gao, G., & Zhao, Y. (2018a). Understanding success through the diversity of collaborators and the milestone of career. *Journal of the Association for Information Science and Technology, 69*(1), 87-97.

Bu, Y., Ding, Y., Liang, X., & Murray, D. S. (2018b). Understanding persistent scientific collaboration. *Journal of the Association for Information Science and Technology, 69*(3), 438-448.

Bu, Y., Murray, D. S., Ding, Y., Huang, Y., & Zhao, Y. (2018c). Measuring the stability of scientific collaboration. *Scientometrics, 114*(2), 463-479.

Burrell, Q. L. (2001). Stochastic modelling of the first-citation distribution. *Scientometrics, 52*(1), 3-12.

Burrell, Q. L. (2002b). Will this paper ever be cited? *Journal of the American Society for Information Science and Technology, 53*(3), 232-235.

Burrell, Q. L. (2002a). The nth-citation distribution and obsolescence. *Scientometrics, 53*(3), 309-323.

Burrell, Q. L. (2005). Are "sleeping beauties" to be expected? *Scientometrics, 65*(3), 381-389.

Egghe, L. (2000). A heuristic study of the first-citation distribution. *Scientometrics, 48*(3), 345–359.

Egghe, L., & Rao, I. R. (2001). Theory of first-citation distributions and applications. *Mathematical and Computer Modelling, 34*(1-2), 81-90.

Egghe, L., & Rousseau, R. (2000). Aging, obsolescence, impact, growth, and utilization—Definitions and relations. *Journal of the American Society for Information Science, 51*(11), 1004-1017.

Garfield, E. (1989a). Delayed recognition in scientific discovery: Citation frequency analysis aids the search for case histories. *Current Contents*, 23, 3-9.

Garfield, E. (1989b). More delayed recognition. Part 1. Examples from the genetics of color blindness, the entropy of short-term memory, phosphoinositides, and polymer rheology. *Current Contents*, 38, 3-8.

Garfield, E. (1990). More delayed recognition. Part 2. From inhibin to scanning electron microscopy. *Current Contents*, 39, 3-9.

Glänzel, W., Schlemmer, B., & Thijs, B. (2003). Better late than never? On the chance to become highly cited only beyond the standard bibliometric time horizon. *Scientometrics, 58*(3), 571-586.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceeding of the National Academy Sciences of the United States of America, 102*(46), 16569-16572.

Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and identifying

Sleeping Beauties in science. *Proceedings of the National Academy of Sciences of the United States of America, 112*(24), 7426-7431.

Li, J., Tang, J., Zhang, J., Luo, Q., Liu, Y., & Hong, M. (2008). ArnetMiner: Expertise oriented search using social networks. *Frontiers of Computer Science in China, 2*(1), 94-105.

Long, J. S. (1978). Productivity and academic position in the scientific career. *American Sociological Review, 43*(6), 889-908.

Min, C., Ding, Y., Li, J., Bu, Y., Pei, L., & Sun, J. (2018). Innovation or imitation: The diffusion of citations. *Journal of the Association for Information Science and Technology, 69*(10), 1271-1282.

Nakamoto, H. (1988). Synchronous and diachronous citation distribution. In L. Egghe & R. Rousseau (Eds.), *Informetrics 87/88* (pp. 157-163). Amsterdam: Elsevier.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review, 45*(2), 167-256.

Pan, R. K., Petersen, A. M., Pammolli, F., & Fortunato, S. (2018). The memory of science: Inflation, myopia, and the knowledge network. *Journal of Informetrics, 12*(3), 656-678.

Parolo, P. D. B., Pan, R. K., Ghosh, R., Huberman, B. A., Kaski, K., & Fortunato, S. (2015). Attention decay in science. *Journal of Informetrics, 9*(4), 734-745.

Price, D. J. (1965). Networks of scientific publications. *Science, 149*(3683), 510-515.

Price, D. J. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science, 27*(5), 292-306.

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.

Redner, S. (2005). Citation statistics from 110 years of physical review. *Physics Today, 58*(6), 49-54.

Rousseau, R. (1988). Citation distribution of pure mathematics journals. In L. Egghe & R. Rousseau (Eds.), *Informetrics 87/88* (pp. 249-262). Amsterdam: Elsevier.

Rousseau, R. (1994). Double exponential models for first-citation processes. *Scientometrics, 30*(1), 213-227.

Sanyal S. (2006). Effect of citation patterns on network structure. Retrieved from https://arxiv.org/pdf/physics/0611139.pdf

Stinson, E. R., & Lancaster, F. (1987). Synchronous versus diachronous methods in the measurement of obsolescence by citation studies. *Journal of Information Science, 13*(2), 65-74.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). ArnetMiner: Extraction and mining of academic social networks. In *Proceedings of the fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp.990-998), August 24-27, 2008, Las Vegas, NV., U.S.A.

Tang, J., Zhang, J., Zhang, D., Yao, L., Zhu, C., & Li, J. (2007). ArnetMiner: An Expertise Oriented Search System for Web Community. In Semantic Web Challenge.

Van Raan, A. F. (2004). Sleeping beauties in science. *Scientometrics, 59*(3), 467-472.

Wallace, M. L., Larivière, V., & Gingras, Y. (2009). Modeling a century of citation distributions. *Journal of Informetrics, 3*(4), 296-303.

Wang, D., Song, C., & Barabasi, A.-L. (2013). Quantifying long-term scientific impact. *Science, 342*(6154), 127-132.

Yin, Y., & Wang, D. (2017). The time dimension of science: Connecting the past to the future. *Journal of Informetrics, 11*(2), 608-621.

APPENDIX 1: PAPER PERIOD SELECTION

Some publications published in recent years have had a limited number of opportunities to receive all of their citations, and this might bias our data analysis. To eliminate this effect, we first calculate the average citation age of publications grouped by publication year, in which the citation age is defined as the difference between the year that an article is published and that of obtaining the last citation recorded in our dataset. Figure A1 shows the average citation age distribution of all publications in our dataset, in which the solid blue line exhibits an overall descending trend of the average citation age of publications ranging from 1936 to 2014. We find that the average citation age of all publications is ~9.7, shown as a red dotted line in Figure 1A. This indicates that publications published after 2005 are not likely to have received all of their citations. Hence, 517,589 publications published in 2005 or earlier are selected and targeted for the following analyses.

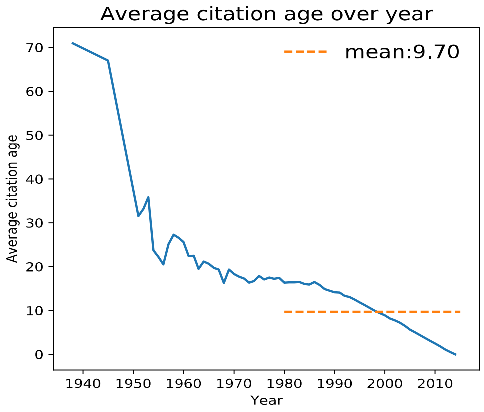


Figure A1. Distribution of average citation age over years.

APPENDIX 2: THRESHOLD DETERMINATIONS FOR PAPER PARTITIONING

Partitioning highly-, medium-, and lowly-cited publications requires two threshold values: one threshold divides the lowly- and medium-cited publication groups, while the other partitions medium- and highly-cited publications. We analyze the citation distribution of the publications in our dataset to determine these two thresholds. Figure A2 shows the citation count distribution of the publications published in 2005 or before, in which the blue dots represent the scatter plot for real data, while the red solid line serves as the fitted line under a log-log scale. Since the straight red line fits the dots quite well, it is shown that the citation distribution follows a typical power law distribution. However, we also find that there are two areas of dots that *deviate from* the fitting lines. The first area mainly includes dots representing lowly-cited publications, shown in the top left of the figure, while the second area contains more highly-cited publications, shown in the lower right corner of the figure. As pointed out by Redner (1998) and Newman (2003), dots on the fitted power law line and those deviating from it essentially reflect different mechanisms behind them. We therefore observe three potential different mechanisms: 1) the top left dots with the citation count x < 14 (deviating from the fitted line downwards); 2) the medium dots whose citation count is between 14 and 1000[1] (good fitted dots); and 3) the lower right dots with the citation count x > 1000 (deviating from the fitted line upwards). In these publications, these three groups of publications are defined as lowly-, medium-, and highly-cited publications, respectively.

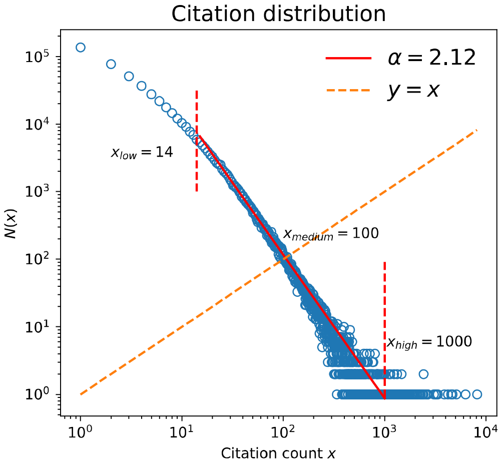


Figure A2. Distribution of citation count.

APPENDIX 3: PAPER SAMPLING

Since we have reported from Figure A2 that the dots with a small citation count (fewer than 14 times) are found to deviate from the fitted line, we choose 10 as the mean and one as the standard deviation of the normal distribution for lowly-cited paper sampling. By doing this, we can ensure that ~99.7% of the sampled lowly-cited publications feature a citation number that is less than 14 (actually 10±3). To sample medium-cited publications, similar to Hirsch (2005), we also plot $y = x$ in Figure A2 and select the horizontal coordinate of the intersection (with a value of 100) between $y = x$ and the fitted power law line as the mean of the distribution for medium-cited paper sampling.

[1] The value of 1000 here *de facto* derives from the horizontal coordinate of the intersection between y=1.0 (the least non-zero value in terms of frequency) and the fitted power law line, as shown in Figure A2.

10 is set as the standard deviation of medium-cited publications.

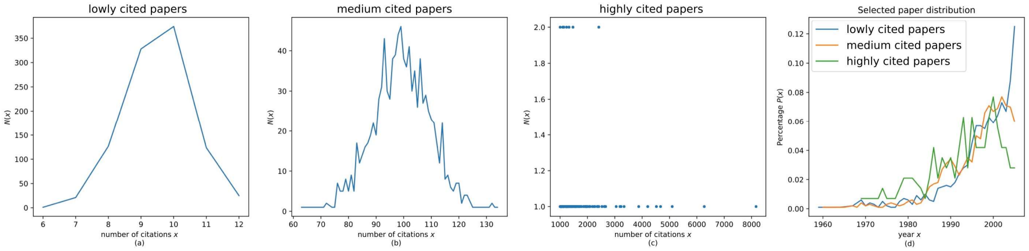


Figure A3. Citation distribution of the sampled publications in lowly- (a), medium- (b) and highly-cited (c) groups, and among all publications (d).

The citation distribution of the sampled publications in the lowly-, medium-, and highly-cited paper groups are shown in Figures A3(a), (b), and (c), respectively. The distribution of all of the three kinds of publications over publication year are plotted in Figure A3(d). Combining Figures 1 and A3, we find that the distribution of sampled publications over publication years is consistent with the paper distribution of the entire dataset.