Scientometrics
https://doi.org/10.1007/s11192-024-05038-8

Does Granger causality exist between article usage and publication counts? A topic-level time-series evidence from IEEE Xplore

Wencan Tian[1] · Yongzhen Wang[1] · Zhigang Hu[2] · Ruonan Cai[3] · Guangyao Zhang[1,4] · Xianwen Wang[1]

Received: 7 August 2023 / Accepted: 19 April 2024

Abstract
In this study, employing the IEEE Xplore database as the data source, articles on different topics (keywords) and their usage data generated from January 2011 to December 2020 were collected and analyzed. The study examined the temporal relationships between these usage data and publication counts at the topic level via Granger causality analysis. The study found that almost 80% of the topics exhibit significant usage-publication interactions from a time-series perspective, with varying time lag lengths depending on the direction of the Granger causality results. Topics that present bidirectional Granger causality show longer time lag lengths than those exhibiting unidirectional causality. Additionally, the study found that the direction of the unidirectional Granger causality was influenced by the significance of a topic. Topics with a greater preference for article usage as the Granger cause of publication counts were deemed more important. The findings' reliability was confirmed by varying the maximum lag period. This study provides strong support for using usage data to identify hot topics of research.

Keywords Article usage data · Publication counts · IEEE Xplore · Time-series · Granger causality test

Introduction

With the advent of the digital era, electronic publishing has replaced print publishing and changed scientists' reading habits drastically (Schlögl et al., 2014). Increasingly, individuals' reading behaviors are being recorded in real-time, and their usage information

✉ Xianwen Wang
xianwenwang@dlut.edu.cn

[1] WISE Lab, Institute of Science of Science and S&T Management, Dalian University of Technology, Dalian, China

[2] Institute for Science Technology and Society, South China Normal University, Guangzhou, China

[3] School of Business, Shandong University, Weihai, China

[4] UNU-MERIT, Maastricht University, Maastricht, The Netherlands

Published online: 10 May 2024

of scientific literature is archived in cyberspace. Moreover, the open science movement has led to an increase in the number of academic publishers and databases of scholarly resources that display usage metrics for publications, which provide new possibilities for using usage data to conduct more in-depth and extensive bibliometrics studies. Thus, in the past few years, usage analysis has attracted the interest of numerous researchers in scientometrics (Chen et al., 2020; Chi & Glänzel, 2018; Chi et al., 2019; Khan & Younas, 2017; Kurtz & Henneken, 2017; McGillivray & Astell, 2019; Wang et al., 2013, 2014).

A paper's usage information, which includes the footprint left by the researchers when they read or download the paper online, can reflect its immediate influence (Glaenzel & Gorraiz, 2015). In research evaluation, usage data have two advantages over citation data. First, usage data is much more abundant than citation data. For instance, between 2011 and 2020, IEEE Xplore published 2.58 million publications, which generated 10.24 million citations and amassed a massive 150 million usage records. Second, usage data is more time-sensitive, as it can characterize an article's immediate scientific influence, while citation counts cannot provide timely feedback due to the citation delay phenomenon (Wang et al., 2013). It usually takes months for an article to start attracting citations. Usage statistics may be available for viewing and downloading within days, hours, or even minutes after an article is published online.

Usage data can be seen as implicit feedback from scientists at the early stages of topic selection. That is also the reason why usage data can predict hot topics of research. As a reflection of the topic selection behavior of scientists, digital footprint usage can provide an earlier indication of possible future research directions than either publication or citation behavior (Wang et al., 2014). Currently, there is plenty of research on the relationship between publication behavior and citation behavior, or that between usage behavior and citation behavior (Breitzman, 2021; Uddin & Khan, 2016; Zahedi & Haustein, 2018; Zong et al., 2020). However, few papers have addressed the relationship between usage and publication data.

Since preliminary literature retrieval and review are necessary steps before paper writing and publication in the research process, it is believed that a certain degree of correlation, or even causality, exists between usage counts and publication counts on a research topic. From an economic perspective, the relationship between publication behavior and usage behavior is similar to the production and consumption of knowledge (Baker & Mayernik, 2020; Borner et al., 2006; Clarkson et al., 2013), with a dynamic interaction effect at the time series level. The production of knowledge provides the conditions for active consumption, which in turn promotes the reproduction of knowledge. In this interactive process, some research areas or topics become hot because of the virtuous circle of publication and usage behavior, while others decline because of the vicious circle created by the two. At present, whether there is a temporal causal relationship between the two has yet to be supported or verified by the data.

Therefore, this study chooses the Granger causality test to verify the logical relationship between usage behavior and publication behavior from a time series perspective. The Granger causality test is a statistical method that originated in the field of econometrics (Granger, 1969), primarily to test whether one set of time series is the cause of another set of time series. The basic idea of Granger's causality test is that for time series X and Y, if X is the cause of the change in Y, then the change in X should occur before the change in Y, and the past value of X should help predict the future value of Y. In this paper, the Granger causality test is used to examine the dynamic interaction between usage data and publication data, that is, whether the historical information of usage data can be used to predict future changes in publication data

and vice versa. If all the above relationships exist, there is a two-way Granger causality between usage data and publication data; if only one is true, there is a one-way Granger causality between the two.

Specifically, this study answers the following research questions: For the topics,

RQ1 Does Granger causality exist between article usage and publication counts in the past ten years (2011–2020)? If so, which direction of the causal relationship is most common? Usage to publication, or publication to usage?
RQ2 What are the lengths of time lags for Granger causality of different directions? Which direction is the longest, and which one is the shortest?
RQ3 Does the topic's importance influence the direction of Granger causality?

To answer these questions, we examined 56,343 topics involving almost 4 million pieces of literature data and 300 million usage data points in IEEE Xplore. Taking a thematic perspective and using publication counts as an indication of research hotness, the usage-publication interactions in the time series were confirmed by applying Granger causality tests, adding weight to the finding that article usage might serve as an early indicator of what areas of study will be popular in the near future.

Theoretical exposition and related work

Attention economy theory

In the digital age, we are faced with a dramatically increasing amount of information, making people's attention more precious and limited. This has given rise to an innovative economic concept-the attention economy. Under this framework, attention is defined as the continuity of people's focus on specific things, which forms the basis for cognition, decision-making, and action (Lanham, 2007). The core idea of this theory is that people's attention has transformed into a scarce resource, becoming the capital that various companies, brands, and platforms compete for. For scientists, the scarcity of attention resources requires them to allocate their time and energy reasonably (Zhang et al., 2023).

For scientists' usage and publishing behaviors, to complete a research task or write a paper, they need to expend attention resources (i.e., downloading and viewing literature) to absorb scientific knowledge. Then, based on a deep understanding and critical analysis of existing scientific knowledge, scientists will propose new research hypotheses or theories, design experiments or models to test these new ideas, and reproduce knowledge (i.e., publish articles). The process from absorbing knowledge to reproducing knowledge is cyclical, where new knowledge becomes the basis for subsequent research, continuously being received, integrated, and reproduced by later scientists, forming an ever-advancing system of scientific knowledge. In this context, usage data can be seen as an observable metric in the early stages of knowledge production, representing, to some extent, the research interests and future topic directions of scientists. Therefore, from a topic perspective, using usage data to predict the number of publications or to judge whether it will become a research hotspot is theoretically supported and a valuable research direction.

Related work

Research related to usage data mainly covers three aspects. First, a significant body of research from different levels, such as disciplines (Chi, 2020; Gorraiz et al., 2014), journals (Schloegl & Gorraiz, 2010; Schloegl et al., 2014; Vaughan et al., 2017), and articles (Brody et al., 2006; Guerrero-Bote & Moya-Anegón, 2014; Lippi & Favaloro, 2013; Markusova et al., 2018), explored the correlation between usage counts and citation counts, aiming to seek a novel academic evaluation metric to more scientifically and reasonably measure the research capabilities of researchers as well as academic communities like journals and research institutions. This has led to the development of metrics such as the "usage impact factor" (Bollen & Sompel, 2008; Schloegl & Gorraiz, 2010), the "usage immediacy index" (Rowlands & Nicholas, 2007), or the "download immediacy index" (Wan et al., 2010). Second, from an article-level perspective, some scholars explored the usage advantages brought by funding (Dorta-González & Dorta-González, 2023; Zhao et al., 2018), international collaboration (Chi & Glänzel, 2018; Thelwall & Maflahi, 2015; Tian et al., 2024), and open access (Wang et al., 2015; Zhang et al., 2021). Lastly, it is feasible to utilize usage data to identify research hotspots (Bollen & Sompel, 2006; Fang et al., 2020; Wang et al., 2013). In line with the research question of this paper, this section primarily focuses on the last point.

Furthermore, traditional methods for detecting research hotspots include citation analysis-based (such as bibliographic coupling and co-citation methods) identification methods (Boyack & Klavans, 2010; Chen, 2006; Glänzel & Thijs, 2012; Small et al., 2014), knowledge unit-based (usually employing text mining methods like LDA to extract keywords from literature titles and abstracts as the basic knowledge units of the literature, and exploring field hotspots by examining the absolute frequency of keywords, average frequency per article, and rate of frequency change on an annual and periodic basis) identification methods (Blei et al., 2003; Ding & Chen, 2014; Jeong & Song, 2014; Lee, 2008; Miao et al., 2020; Wu et al., 2021; Xu et al., 2021), time series-based identification methods (Liang et al., 2021; Porter et al., 2019), and multi-source data-based (such as integrating patent literature and journal articles data or combining funding data with journal articles data) identification methods (Park et al., 2015; Bai et al., 2020; Chen & Chen 2022; Ye et al., 2023). However, as usage data increasingly garners the interest of scientists, some scholars have also begun to experiment with utilizing usage data to identify research hotspots.

Specifically, Wang and Fang (2016) revealed what is currently trending in computational neuroscience based on the frequency of keywords and usage data in the Web of Science. In another study, Wang (2013) used the ratio of article keyword downloads to the number of articles using that keyword to determine what topics were trending in *Scientometrics*. Fang et al. (2020) built a framework to locate research hubs by mining the Web of Science for 12.3 million publications and the 12 altimetric events corresponding to them. Bollen and Van De Sompel (2006) created scientific maps by deriving journal relationships from usage data to determine the current research trends. The usage data of topics are used to predict publication counts of topics and then forecast the research hotness of topics in advance, which gives the usage data the meaning and value of scientific foresight and helps researchers better grasp the law of scientific development (Tahamtan & Bornmann, 2019; Waltman, 2016).

As mentioned in the above review, utilizing usage data to detect research hotspots represents another valuable manifestation of usage data and is a worthwhile research

direction. However, the causal relationship between the two has not yet been established. Based on this, this study utilizes the unique monthly usage data generated by the IEEE Xplore database to address this challenge, aiming to provide strong empirical support and theoretical backing for identifying research hotspots using usage data.

Data and method

Dataset

The data on usage and publication behavior was obtained from IEEE Xplore, a professional resource in computers and electronic communications (Breitzman, 2021; Khan & Younas, 2017; Tian et al., 2019). Unlike the commonly used Web of Science or Scopus databases, which only provide static usage data in the form of cumulative usage counts for each item, IEEE Xplore can offer dynamic usage statistics in the form of article utilization numbers that are updated monthly. For each item in the repository, including those published before 2011, IEEE Xplore has provided monthly usage data (total number of HTML views and PDF downloads) since 2011. The unique feature of providing dynamic usage data over a long period allows us to construct a sufficiently lengthy time series of usage data on one topic, which perfectly matches the dataset required to answer the research questions of this paper. Additionally, some scholars have already conducted valuable research using the usage data provided by the IEEE Xplore database. For example, Khan and Younas (2017) investigated the downloading behaviors of readers towards *IEEE Transactions on Learning Technologies* and *IEEE Transactions on Education*, two IEEE journals. Tian et al. (2024) utilized the IEEE Xplore database to examine the mediating role of Mendeley readership between usage counts and citations.

When constructing a time series of monthly article usage for a topic, it is discovered that the closer the current month is, the more previous articles are included in the usage counts. Thus, to minimize data bias, after careful consideration, we added usage data from 2011 to 2020 generated by articles related to the topic published before 2011. Eventually, we collected the original data of 3,975,602 publications included in IEEE Xplore between 2001 and 2020 and the 313,099,560 usage data points created between 2011 and 2020 over eight months (February 2021 to September 2021).

This study uses the topic level as our entry point for analysis. For the sake of simplicity, topics refer to keywords directly, including both "Author Keywords" (the keywords added by the article's authors) and "IEEE Terms" (the keywords added by IEEE Xplore). These two topic subsets were elaborately merged into a single set and subjected to text cleanings, including removing special characters, performing stemming extraction (such as "3d models," "3d model," and "3D modeling" all become "3d model" after undergoing stem processing) with Python's *NLTK* library,[1] and topic filtering (removing topics with a frequency of less than ten). Eventually, 64,633 topics were retained.

To determine the start and end times for each topic, we established the following two guidelines: (1) For topics that appeared in 2001–2010, we set their start time to January 2011. For other topics, we used the publication date of the earliest article published on the topic from 2011 to 2020 as the start time. (2) We set the end time of all topics uniformly

[1] https://www.nltk.org/

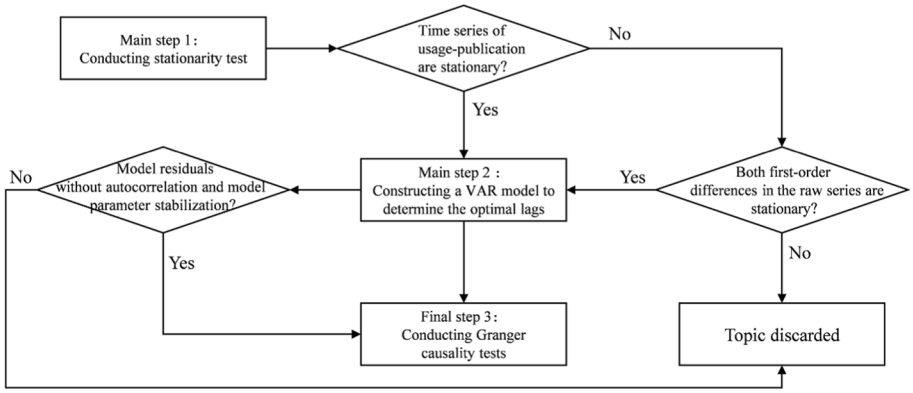

Fig. 1 A flowchart for the Granger causality test

to the update time of the most recent usage data available at the time of data acquisition, December 2020. Using these guidelines, we constructed a time series of article usage and a corresponding time series of publication counts for each topic. In total, we collected 7,447,358 time points of usage data for 64,633 topics and the corresponding 7,447,358 time points of publication counts.

Analysis method

There are two main steps when using the Granger causality test to verify the causal relationship between two time series. First, the non-stationary time series must be transformed into stationary time series. Second, a Vector Autoregression (VAR) model needs to be constructed utilizing the stationary time series to determine the optimal lag order. Subsequently, the smooth time series and the optimum lag order can be fed into the Granger causality test model to obtain the logical relationship between the two time series. The aforementioned process can be seen in Fig. 1. Additionally, the first part of this subsection introduces the principles of Granger causality testing, while the second and third parts present the two main data processing procedures described above, respectively.

Principle of Granger causality test

The Granger causality test is a statistical hypothesis test for determining whether one time series is useful in forecasting another. The method is now being applied in the field of scientometrics (Ding et al., 2021; Hu et al., 2021; Lee et al., 2011; Luan et al., 2022). For example, Lee et al. (2011) used Granger causality tests to determine the causal relationship between economic productivity and research output. Hu et al. (2021) studied the Granger causality relationship between download counts and citation counts using 7552 articles published in *the Lancet*. Luan et al. (2022) used Granger causality tests to examine the interaction effects between scientific equipment and scientific publications.

Based on Granger's definition of causality, the model is constructed in this paper as follows.

$$\mathcal{U}^t = \sum_{j=1}^{m} \alpha_j \mathcal{U}^{t-j} + \sum_{j=1}^{m} \beta_j \mathcal{P}^{t-j} + \varepsilon_t, \tag{1}$$

$$\mathcal{P}^t = \sum_{j=1}^{m} \alpha_j \mathcal{P}^{t-j} + \sum_{j=1}^{m} \beta_j \mathcal{U}^{t-j} + \varepsilon_t, \tag{2}$$

where $\mathcal{U}^t$ and $\mathcal{P}^t$ represent, as present values, the article usage and publication counts of a topic at the current moment t, as past values, $\mathcal{U}^{t-j}$ and $\mathcal{P}^{t-j}$ represent article usage and publication counts of the topic at time t-j, respectively. m represents the lag time, which indicates the time it takes for one event to occur and begin to affect another. Equation (1) tests whether publication counts are the Granger cause of article usage, whereas Eq. (2) tests whether article usage are the Granger cause of the publication counts.

Using Eq. (1) as an illustration, the initial hypothesis of the Granger causality test is that publication counts are not the Granger cause of article usage, which is reflected in Eq. (1) as $\beta_1 = \beta_2 = \cdots = \beta_m = 0$. If there is a β that is not zero, the original hypothesis fails, and publication counts are considered the Granger cause of article usage.

Stationarity test of time series

The prerequisite for conducting Granger causality tests is that the input time series must be stationary. The smoothness of time series does not mean that the time series is static, but that the statistical patterns of time series do not change over time. If the time series is non-stationary, the estimated parameter statistics will deviate from the F distribution. Since the Granger causality test relies on the F distribution as a criterion for statistical inference, this may lead to a pseudo-regression phenomenon and affect the validity of the Granger causality test.

The stability of the monthly observation series of article usage and publication counts for each topic was tested using the Augmented Dickey-Fuller (ADF) test (Dickey & Fuller, 1979), also called the unit root test. Suppose the time series fails the unit root test. This situation shows that the series is not smooth, and transformations such as differentiation (subtracting the previous period's value from the latter period's value) are necessary to remove the unit root and get a smooth series. Notably, following differentiation, the time series must be retested for smoothness; if it does not pass, the topic is discarded.

After the smoothness test, the usage-publication time series for each topic in the dataset can be put into one of four types: Type I: the time series of both the article usage and publication counts are smooth; Type II: the time series of the article usage is smooth, but the time series of publication counts is non-smooth; Type III: the time series of the article usage is non-smooth, but the time series of publication counts is smooth; Type IV: the time series of both the article usage and publication counts are non-smooth. It is worth noting that, besides type I, all of the other three types of topics need to be differentially transformed and tested for smoothness again. The specific differential results can be seen in Table 1. It can be seen that 5.3%, or 3448 topics $(137 + 3156 + 155)$, were rejected because the usage-publication time series was still unstable after differentiation.

Table 1 Stability test results

ADF test	Types of smoothness in the "usage-publication" timeseries	Number of topics	Proportion (%)
Before differential	Type I	18,038	27.91
	Type II	890	1.38
	Type III	39,018	60.37
	Type IV	6687	10.34
	Total	64,633	100
After differential	Type I	61,185	94.67
	Type II	137	0.21
	Type III	3156	4.88
	Type IV	155	0.24
	Total	64,633	100

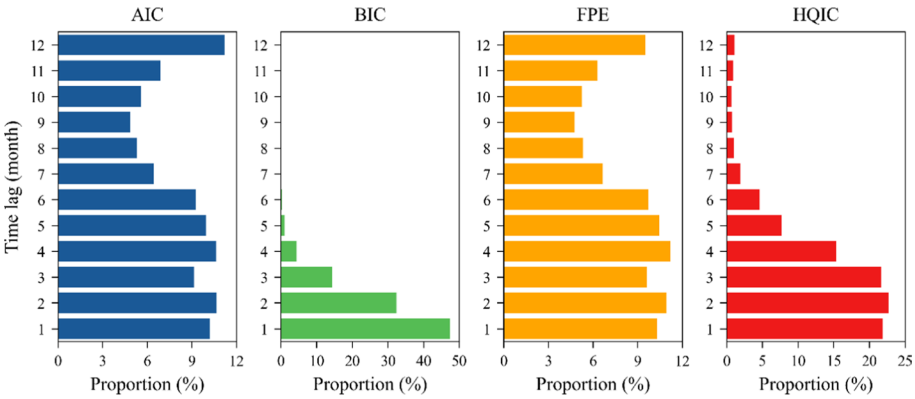

Fig. 2 Distribution of time lag length under different information criteria

Selection of information criteria

As Granger causality tests are performed in the framework of the VAR model (Luan et al., 2022), it is also necessary to input the smoothed time series of usage-publication into the VAR model to determine the optimal lags before conducting Granger causality tests. Furthermore, there are four relatively well-known information criteria for determining the optimal lag: the Akaike Information Criterion (AIC), the Bayesian Information Criteria (BIC), the Final Prediction Error Criterion (FPE), and the Hannan-Quinn Information Criterion (HQIC). In this paper, we use the AIC (Akaike, 1970; Zhang et al., 2018) to determine the best time lag length. The following two paragraphs detail the reasons for selecting this information criterion. This paper specifies a maximum lag duration of 12 months. Finally, we built VAR models for each topic where the monthly observation series of article usage and publication counts were smooth. In addition, the model's dependability is improved by excluding topics with sequence durations of less than one year, topics with model residuals that are white noise, and topics with unstable model parameters. Specifically, there were 4447 topics left out of the analysis because

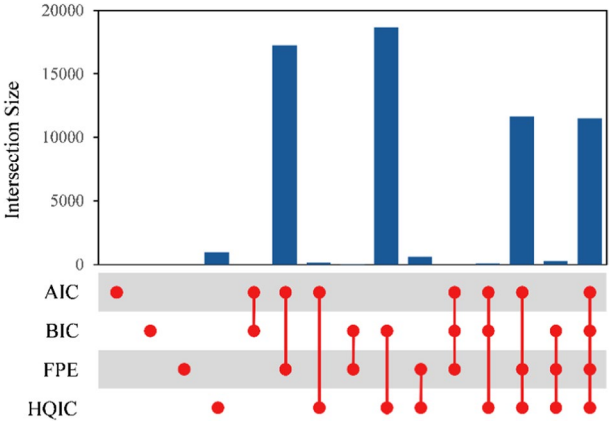

Fig. 3 Combination of the four information criteria. The bottom part shows the combinations of the four information criteria, which correspond to the columns in the upper part, and the connecting lines represent the common, i.e., the same lags are selected in the construction of the same VAR model. For example, the last column represents that when constructing the VAR model for the "usage-publication" time series of all topics, there are 11,481 topics with the same lags selected by the four information criteria

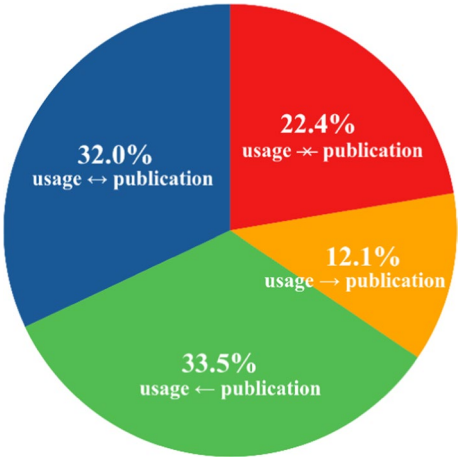

Fig. 4 Granger causality test results on article usage and publication counts for monthly. "usage ↔ publication" indicates a bidirectional Granger causality between the article usage and publication counts, "usage → publication" demonstrates that the former is the Granger cause of the latter, and vice versa, and "usage × publication" indicates that there is no bidirectional or unidirectional Granger causality between the two. Same below

of unstable model parameters; 389 topics were left out because of residual autocorrelation; and six topics were left out because the time series length was too short. In the end, 56,343 topics (87.2%) were kept.

We constructed four VAR models using different information criteria for each topic separately to determine which information criterion is most appropriate for selecting the best lag. Later, we compared the ideal lag lengths chosen by the four information criteria (see Fig. 2). The distribution is substantially skewed toward shorter lags when using BIC and HQIC to select time delays. For BIC, about half of the time lags, it selects are 1, and nearly all fall within the 1–3 range. Most (89.23%) of HQIC's suggested time lag lengths fall between 1 and 5. In contrast, the time lag lengths selected by AIC and FPE have a more uniform distribution, with nearly equal chances of being picked.

Moreover, we investigated the similarities and differences in the time lag lengths chosen for the same topic under various combinations of the four information criteria (see Fig. 3). It was shown that for 40,370 topics, both AIC and FPE selected identical time lag lengths. Furthermore, considering that AIC is more commonly used than FPE, AIC is finally chosen as the basis to determine the optimal time lag in this article after much deliberation.

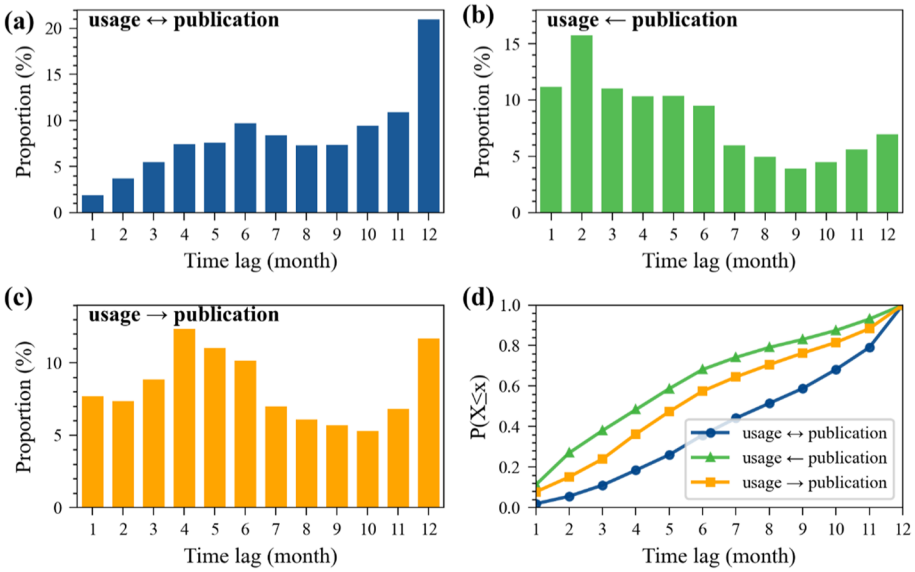

Fig. 5 The distribution of time lag lengths. **a** usage ↔ publication; **b** usage ← publication; **c** usage → publication; **d** cumulative distribution function of time lag lengths

Table 2 The descriptive statistics of time lag lengths in different Granger causality types

Granger causality types	Mean	SD	Min	Max	Skewness	Kurtosis
Usage ↔ publication	7.952	3.257	1	12	− 0.307	− 1.114
Usage ← publication	5.344	3.405	1	12	0.558	− 0.853
Usage → publication	6.373	3.450	1	12	0.226	− 1.125

Results

Granger causality types: usage & publication

Granger causality between article usage and publication counts was examined for all 56,343 topics. The results are shown in Fig. 4. Overall, 77.6% of the topics exhibit a statistically significant Granger causality effect between article usage and publication counts. Among these topics, 32% demonstrated bidirectional Granger causality, while 45.6% showed unidirectional Granger causality. Of the topics with unidirectional causality, 33.5% exhibited a causal relationship in which article usage influence publication counts, and 12.1% showed the opposite direction of causality. However, for 22.4% of the time series, there was no evidence of either bidirectional or unidirectional Granger causality.

In addition, the robustness of the Granger causality test findings is examined by varying the maximum lag: one is to reduce the lag, i.e., the maximum lag is set to 10; the other is to increase the lag, i.e., the maximum lag is set to 14. The results can be seen in Fig. 7 in the Appendix.

The distribution of time lag lengths in different Granger causality types

Table 2 shows that the average time lag for bidirectional Granger causality is 7.952 months, which is greater than the average time lag for unidirectional Granger causality. Specifically, for the unidirectional Granger causality, the average time lag is 5.344 months for the type "usage ← publication," which is one month shorter than the average time lag for the type "usage → publication."

Figure 5 presents the distribution of time lags for different Granger causalities. As a whole, bidirectional Granger causality is more likely to occur with longer time delays, and its proportion increases as the lag time increases. On the other hand, unidirectional Granger causality tends to favor shorter time lags, but there are differences between the two types, "usage ← publication" and "usage → publication." Specifically, "usage → publication" has a higher proportion of intermediate time lags, with the proportion increasing and then decreasing as the time lag length increases.

For "usage ↔ publication," the highest proportion of time lags occurs between the 10th and 12th months, accounting for 41.23%, with a peak at the 12th month (20.96%) followed by the 11th month (10.88%). For "usage ← publication," the highest proportion of time lags occurs between 1st and 5th months, with a peak at the 2nd month (15.77%). For "usage → publication," the highest proportion of time lags occurs between the 4th and 6th months, with a peak at the 4th month (12.35%). The accumulation rate of time lags, as shown in subfigure d of Fig. 4, is highest for "usage ← publication," followed by "usage → publication" and "usage ↔ publication."

The results demonstrate that, for the majority of research topics, there exists a circular pattern of "publish-use-publish" in the interaction between article usage and publication counts. This pattern elucidates the process of research topics' emergence, development, and evolution. Initially, a new research topic is typically prompted by a handful of research publications, often originating from pioneering researchers or teams. During this phase, the topic is utilized infrequently owing to a lack of widespread interest in the research area. Subsequently, as more researchers begin to focus on and study the topic, there is a concomitant increase in the publication of related research papers. These publications further contribute to the growth and evolution of the topic, leading to its more frequent usage. Ultimately, the rise in usage corresponds to an upsurge in the popularity of the topic, which in turn spurs an increase in the number of publications on the topic, thereby perpetuating the cycle. This ongoing cycle eventually results in the topic gaining widespread recognition as a "hot" research area. However, if the cycle stagnates, the topic may gradually become obsolete or transform into a new research topic.

Effect of topic importance on Granger causal direction

To investigate the effect of topic importance on Granger causality direction, we constructed a topic co-occurrence network and calculated four centrality indicators. In undirected complex networks, centrality measures are often used to determine the importance or influence of nodes, including degree centrality, betweenness centrality, closeness centrality, and

eigenvector centrality[2]. First, we used the *NetworkX* library[3] in Python to construct a co-occurrence network for all 56,343 topics, which took approximately a week due to the large number of nodes and complex relationship connections. The resulting network had over 6 million edges, with a theoretical potential of over 3 billion connections. Next, we extracted the degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality for each topic. Finally, we sorted and grouped the topics based on their centrality, as shown in Fig. 6. For instance, we categorized the top 10% of topics with the highest degree centrality into the first group, those with a degree centrality of 10–20% into the second group, and so on, until those with a degree centrality of 90–100% were categorized into the tenth group.

It is evident that the influence of topic importance on the direction of causal relationships between article usage and publication counts exhibited the same pattern across all four centrality indicators. As the value of these indicators changes, the proportion of bidirectional Granger causality and no Granger causality shifts less, while the proportion of the two unidirectional causal linkages exhibits a reversible trend. Specifically, the more influential the topic, the greater the tendency for article usage to be the Granger cause of the publication counts, with the proportion varying from the first to the last group by approximately 15%.

One possible explanation is that when a research topic is located at the periphery of the network, it receives less attention from scholars, resulting in an increase in article usage that tends to depend on an increase in the scientific literature related to the topic. On the other hand, as the topic moves towards the core of the network, it receives a growing amount of attention, and the usage statistics will contribute to the rise in publication counts.

Conclusions

We analyzed usage data for various topics (keywords) from IEEE Xplore to examine the Granger causality between article usage and publication counts. While previous studies have suggested that papers' usage data can indicate research tendencies, it remains unclear whether there is a statistically causal relationship between usage and publications and, if so, what that relationship is. Our study finds that nearly 80% of the topics exhibit significant usage-publication interactions from a time-series perspective.

[2] Degree centrality refers to the number of direct connections a node has with other nodes (in a topic co-occurrence network, the nodes represent topics). Betweenness centrality measures the frequency of a node's appearance on all shortest paths within the network, quantifying its role and importance as a mediator or "bridge" within the network. Closeness centrality is defined by calculating the reciprocal of the shortest path lengths from a certain node to all other nodes in the network, which is used to measure the average proximity of a node to all other nodes within the network. Eigenvector centrality involves the adjacency matrix of the network and the eigenvector corresponding to its largest eigenvalue, with the underlying idea that one's own importance depends on the importance of the nodes to which one is connected.

[3] https://networkx.org/

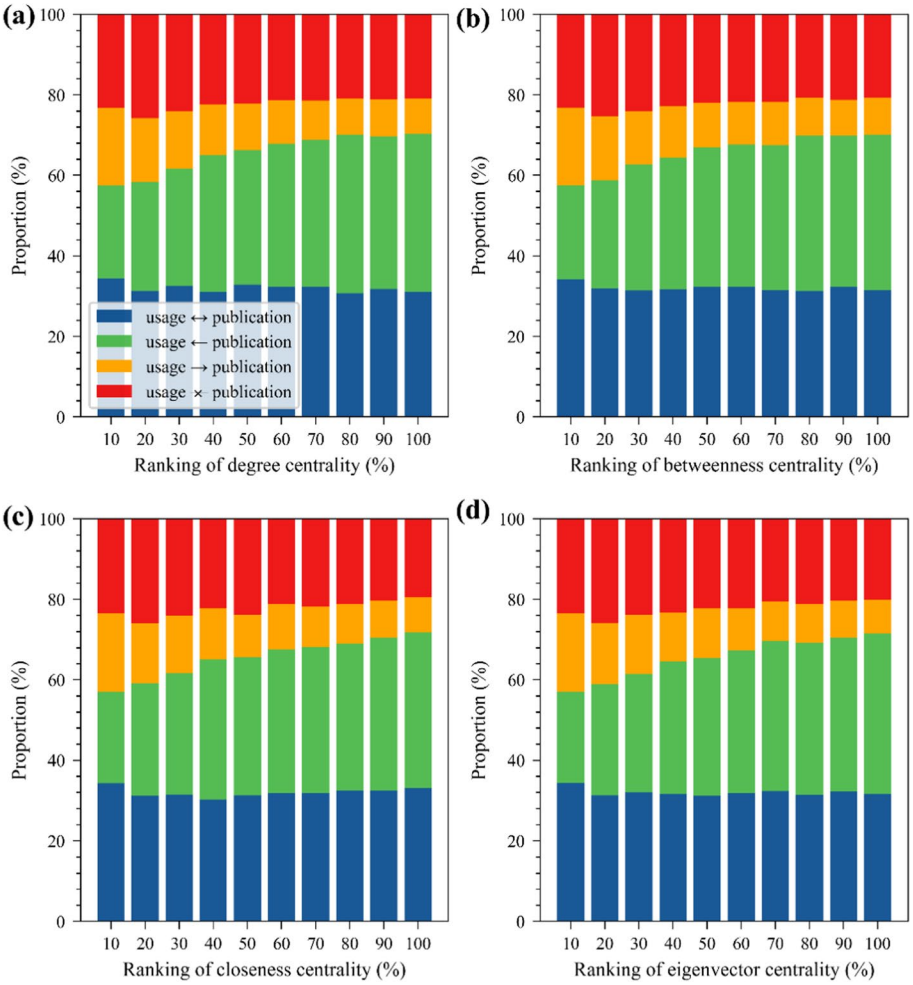

Fig. 6 Distribution of the four Granger causality types under different centrality indicators. The x-axis of subfigures a to d represents the grouping of different centrality indicators, e.g., "10" on the x-axis of subfigure a represents topics with the top 10% of degree centrality, and "20" represents topics with the top 10%–20% of degree centrality

Furthermore, we observe that the time lag lengths differ depending on the direction of the Granger causality. Bidirectional Granger causality between article usage and publication counts is associated with a longer time lag (7.952) compared to "usage←publication" (5.344) or "usage→publication" (6.373). Finally, we find that article usage is more likely to be the Granger cause of publication counts for important topics with large network centralities.

The results of this study have important implications for predicting research trends. While previous studies have utilized text mining and knowledge networks for this purpose (Liang et al., 2021; Masoumi & Khajavi, 2023), our findings suggest that dynamic usage data can serve as a valuable supplement for predicting research trends. Additionally, our

results may encourage data platforms to provide more finely-grained dynamic usage data, enabling further exploration of the value of usage data and expanding our findings' applicability to other fields.

However, there are limitations to our study that future research could address. Granger causality measures the statistical correlation between time series data, and while it can reveal the direction of causality, it cannot estimate the magnitude of the causal effects. Future research could use machine learning-based causal inference methods to more depth estimate the causal effects between article usage and publication counts. Second, since the formation of research hotspots is the result of multiple factors, the causal relationship between usage counts and research hotspots still requires more rigorous econometric methods for inference based on controlling other variables. In the future, time series data could be transformed into panel data, and additional control variables could be added to utilize fixed effects models and instrumental variable methods for a more precise discrimination of their logical relationship. Finally, the empirical analysis in this study is facilitated by the IEEE Xplore database in the field of electronic communications, and whether its conclusions are applicable to other academic fields still needs further verification.

Appendix

See Fig. 7.

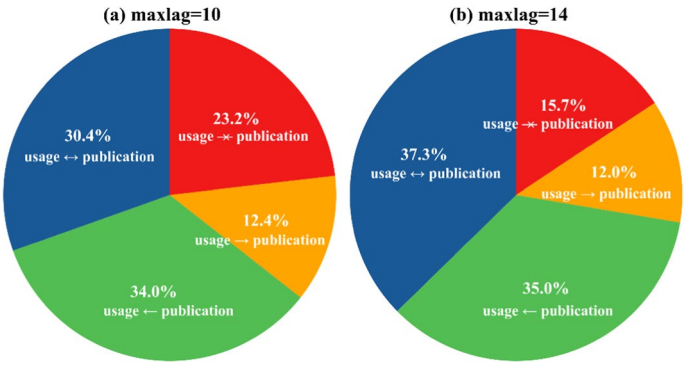

Fig. 7 Results of robustness test. Both reducing and raising the maximum lag time demonstrate a strong statistically significant causal association between article usage and publication counts, demonstrating that the results from this study are robust. Specifically, when the maximum lag was set to 10, 76.8% of the topics exhibited an inherent logical link between article usage and publication counts; when the maximum lag was set to 14, this proportion was 84.3%

Acknowledgements The present study is an extended version of a paper presented at the 19th International Conference on Scientometrics and Informetrics 2023 (ISSI 2023), Bloomington, Indiana (USA), 2-5 July 2023 (Tian et al., 2023). This study is partially supported by the National Natural Science Foundation of China (71974029, 71974030) and LiaoNing Revitalization Talents Program (XLYC2007149). Wencan Tian is financially supported by the China Scholarship Council (202106060134). The authors are grateful to the anonymous reviewers for their helpful comments and suggestions.

Declarations

Conflict of interest The authors declare that they have no conflicts of interest.

References

Akaike, H. (1970). Statistical predictor identification. *Annals of the Institute of Statistical Mathematics, 22*(2), 203. https://doi.org/10.1007/BF02506337

Bai, R., Liu, B., & Leng, F. (2020). Frontier identification of emerging scientific research based on multi-indicators. *Journal of the China Society for Scientific and Technical Information, 39*(7), 747–760. https://doi.org/10.3772/j.issn.1000-0135.2020.07.007

Baker, K. S., & Mayernik, M. S. (2020). Disentangling knowledge production and data production. *Ecosphere, 11*(7), e03191. https://doi.org/10.1002/ecs2.3191

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*(4–5), 993–1022.

Bollen, J., & Van De Sompel, H. (2006). Mapping the structure of science through usage. *Scientometrics, 69*(2), 227–258. https://doi.org/10.1007/s11192-006-0151-8

Bollen, J., & Van De Sompel, H. (2008). Usage impact factor: The effects of sample characteristics on usage-based impact metrics. *Journal of the American Society for Information Science and Technology, 59*(1), 136–149. https://doi.org/10.1002/asi.20746

Borner, K., Penumarthy, S., Meiss, M., & Ke, W. (2006). Mapping the diffusion of scholarly knowledge among major US research institutions. *Scientometrics, 68*(3), 415–426. https://doi.org/10.1007/s11192-006-0120-2

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology, 61*(12), 2389–2404. https://doi.org/10.1002/asi.21419

Breitzman, A. (2021). The relationship between web usage and citation statistics for electronics and information technology articles. *Scientometrics, 126*(3), 2085–2105. https://doi.org/10.1007/s11192-020-03851-5

Brody, T., Harnad, S., & Carr, L. (2006). Earlier web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology, 57*(8), 1060–1072. https://doi.org/10.1002/asi.20373

Chen, C. M. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology, 57*(3), 359–377. https://doi.org/10.1002/asi.20317

Chen, W. M. Y., Bukhari, M., Cockshull, F., & Galloway, J. (2020). The relationship between citations, downloads and alternative metrics in rheumatology publications: A bibliometric study. *Rheumatology, 59*(2), 277–280. https://doi.org/10.1093/rheumatology/kez163

Chen, W., & Chen, W. (2022). Predicting popularity of emerging topics with multivariable LSTM and bibliometric indicators. *Data Analysis and Knowledge Discovery, 6*(10), 35–45. https://doi.org/10.11925/infotech.2096-3467.2022.0075

Chi, P.-S. (2020). The field-specific citation and usage patterns of book literature in the book citation index. *Research Evaluation, 29*(2), 203–214. https://doi.org/10.1093/reseval/rvz037

Chi, P.-S., & Glänzel, W. (2018). Comparison of citation and usage indicators in research assessment in scientific disciplines and journals. *Scientometrics, 116*(1), 537–554. https://doi.org/10.1007/s11192-018-2708-8

Chi, P.-S., Gorraiz, J., & Glanzel, W. (2019). Comparing capture, usage and citation indicators: An altmetric analysis of journal papers in chemistry disciplines. *Scientometrics, 120*(3), 1461–1473. https://doi.org/10.1007/s11192-019-03168-y

Clarkson, J. J., Janiszewski, C., & Cinelli, M. D. (2013). The desire for consumption knowledge. *Journal of Consumer Research, 39*(6), 1313–1329. https://doi.org/10.1086/668535

Dickey, D., & Fuller, W. (1979). Distribution of the estimators for autoregressive time-series with a unit root. *Journal of the American Statistical Association, 74*(366), 427–431. https://doi.org/10.2307/2286348

Ding, W., & Chen, C. (2014). Dynamic topic detection and tracking: A comparison of HDP, C-word, and cocitation methods. *Journal of the Association for Information Science and Technology, 65*(10), 2084–2097. https://doi.org/10.1002/asi.23134

Ding, Y., Dong, X., Bu, Y., Zhang, B., Lin, K., & Hu, B. (2021). Revisiting the relationship between downloads and citations: A perspective from papers with different citation patterns in the case of the Lancet. *Scientometrics, 126*(9), 7609–7621. https://doi.org/10.1007/s11192-021-04099-3

Dorta-González, P., & Dorta-González, M. I. (2023). The funding effect on citation and social attention: The UN sustainable development goals (SDGs) as a case study. *Online Information Review, 47*(7), 1358–1376. https://doi.org/10.1108/OIR-05-2022-0300

Fang, Z., Costas, R., Tian, W., Wang, X., & Wouters, P. (2020). An extensive analysis of the presence of altmetric data for web of science publications across subject fields and research topics. *Scientometrics, 124*(3), 2519–2549. https://doi.org/10.1007/s11192-020-03564-9

Glaenzel, W., & Gorraiz, J. (2015). Usage metrics versus altmetrics: Confusing terminology? *Scientometrics, 102*(3), 2161–2164. https://doi.org/10.1007/s11192-014-1472-7

Glänzel, W., & Thijs, B. (2012). Using' core documents' for detecting and labelling new emerging topics. *Scientometrics, 91*(2), 399–416. https://doi.org/10.1007/s11192-011-0591-7

Gorraiz, J., Gumpenberger, C., & Schloegl, C. (2014). Usage versus citation behaviours in four subject areas. *Scientometrics, 101*(2), 1077–1095. https://doi.org/10.1007/s11192-014-1271-1

Granger, C. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica, 37*(3), 424–438. https://doi.org/10.2307/1912791

Guerrero-Bote, V. P., & Moya-Anegón, F. (2014). Relationship between downloads and citations at journal and paper levels, and the influence of language. *Scientometrics, 101*(2), 1043–1065. https://doi.org/10.1007/s11192-014-1243-5

Hu, B., Ding, Y., Dong, X., Bu, Y., & Ding, Y. (2021). On the relationship between download and citation counts: An introduction of granger-causality inference. *Journal of Informetrics, 15*(2), 101125. https://doi.org/10.1016/j.joi.2020.101125

Jeong, D. H., & Song, M. (2014). Time gap analysis by the topic model-based temporal technique. *Journal of Informetrics, 8*(3), 776–790. https://doi.org/10.1016/j.joi.2014.07.005

Khan, M. S., & Younas, M. (2017). Analyzing readers behavior in downloading articles from IEEE digital library: A study of two selected journals in the field of education. *Scientometrics, 110*(3), 1523–1537. https://doi.org/10.1007/s11192-016-2232-7

Kurtz, M. J., & Henneken, E. A. (2017). Measuring metrics—a 40-year longitudinal cross-validation of citations, downloads, and peer review in astrophysics. *Journal of the Association for Information Science and Technology, 68*(3), 695–708. https://doi.org/10.1002/asi.23689

Lanham, R. A. (2007). The economics of attention: Style and substance in the age of information. University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/E/bo3680280.html

Lee, L. C., Lin, P. H., Chuang, Y. W., & Lee, Y. Y. (2011). Research output and economic productivity: A Granger causality test. *Scientometrics, 89*(2), 465. https://doi.org/10.1007/s11192-011-0476-9

Lee, W. H. (2008). How to identify emerging research fields using scientometrics: An example in the field of information security. *Scientometrics, 76*(3), 503–525. https://doi.org/10.1007/s11192-007-1898-2

Liang, Z., Mao, J., Lu, K., Ba, Z., & Li, G. (2021). Combining deep neural network and bibliometric indicator for emerging research topic prediction. *Information Processing & Management, 58*(5), 102611. https://doi.org/10.1016/j.ipm.2021.102611

Lippi, G., & Favaloro, E. J. (2013). Article downloads and citations: Is there any relationship? *Clinica Chimica Acta, 415*, 195–195. https://doi.org/10.1016/j.cca.2012.10.037

Luan, C., Deng, S., & Allison, J. R. (2022). Mutual granger "causality" between scientific instruments and scientific publications. *Scientometrics, 127*(11), 6209–6229. https://doi.org/10.1007/s11192-022-04516-1

Markusova, V., Bogorov, V., & Libkind, A. (2018). Usage metrics vs classical metrics: Analysis of Russia's research output. *Scientometrics, 114*(2), 593–603. https://doi.org/10.1007/s11192-017-2597-2

Masoumi, N., & Khajavi, R. (2023). A fuzzy classifier for evaluation of research topics by using keyword co-occurrence network and sponsors information. *Scientometrics, 128*(3), 1485–1512. https://doi.org/10.1007/s11192-022-04618-w

McGillivray, B., & Astell, M. (2019). The relationship between usage and citations in an open access mega-journal. *Scientometrics, 121*(2), 817–838. https://doi.org/10.1007/s11192-019-03228-3

Miao, Z., Du, J., Dong, F., Liu, Y., & Wang, X. (2020). Identifying technology evolution pathways using topic variation detection based on patent data: A case study of 3D printing. *Futures, 118*, 102530. https://doi.org/10.1016/j.futures.2020.102530

Park, I., Lee, K., & Yoon, B. (2015). Exploring promising research frontiers based on knowledge maps in the solar cell technology field. *Sustainability, 7*(10), 13660–13689. https://doi.org/10.3390/su71013660

Porter, A. L., Garner, J., Carley, S. F., & Newman, N. C. (2019). Emergence scoring to identify frontier R&D topics and key players. *Technological Forecasting and Social Change, 146*, 628–643. https://doi.org/10.1016/j.techfore.2018.04.016

Rowlands, I., & Nicholas, D. (2007). The missing link: Journal usage metrics. *Aslib Proceedings, 59*(3), 222–228. https://doi.org/10.1108/00012530710752025

Schloegl, C., & Gorraiz, J. (2010). Comparison of citation and usage indicators: The case of oncology journals. *Scientometrics, 82*(3), 567–580. https://doi.org/10.1007/s11192-010-0172-1

Schloegl, C., Gorraiz, J., Gumpenberger, C., Jack, K., & Kraker, P. (2014). Comparison of downloads, citations and readership data for two information systems journals. *Scientometrics, 101*(2), 1113–1128. https://doi.org/10.1007/s11192-014-1365-9

Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. *Research Policy, 43*(8), 1450–1467. https://doi.org/10.1016/j.respol.2014.02.005

Tahamtan, I., & Bornmann, L. (2019). What do citation counts measure? An updated review of studies on citations in scientific documents published between 2006 and 2018. *Scientometrics, 121*(3), 1635–1684. https://doi.org/10.1007/s11192-019-03243-4

Thelwall, M., & Maflahi, N. (2015). Are scholarly articles disproportionately read in their own country? An analysis of mendeley readers. *Journal of the Association for Information Science and Technology, 66*(6), 1124–1135. https://doi.org/10.1002/asi.23252

Tian, W., Fang, Z., Wang, X., & Costas, R. (2024). A multi-dimensional analysis of usage counts, mendeley readership, and citations for journal and conference papers. *Scientometrics, 129*(2), 985–1013. https://doi.org/10.1007/s11192-023-04909-w

Tian, W., Wang, Y., & Wang, X. (2023). Granger causality between usage counts and publication numbers. In Proceedings of the 19th international conference on scientometrics and informetrics - (ISSI 2023) 2-5 July 2023, Bloomington, Indiana, USA.

Tian, W., Hu, Z., & Wang, X. (2019). Upgrading from 3G to 5G: Topic evolution and persistence among scientists. In *Proceedings of the 17th international conference on scientometrics and informetrics* (pp. 1156–1165)

Uddin, S., & Khan, A. (2016). The impact of author-selected keywords on citation counts. *Journal of Informetrics, 10*(4), 1166–1177. https://doi.org/10.1016/j.joi.2016.10.004

Vaughan, L., Tang, J., & Yang, R. (2017). Investigating disciplinary differences in the relationships between citations and downloads. *Scientometrics, 111*(3), 1533–1545. https://doi.org/10.1007/s11192-017-2308-z

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics, 10*(2), 365–391. https://doi.org/10.1016/j.joi.2016.02.007

Wan, J., Hua, P., Rousseau, R., & Sun, X. (2010). The journal download immediacy index (DII): Experiences using a Chinese full-text database. *Scientometrics, 82*(3), 555–566. https://doi.org/10.1007/s11192-010-0171-2

Wang, X., & Fang, Z. (2016). Detecting and tracking the real-time hot topics: A study on computational neuroscience. arXiv: 1608.05517

Wang, X., Liu, C., Mao, W., & Fang, Z. (2015). The open access advantage considering citation, article usage and social media attention. *Scientometrics, 103*(3), 1149–1149. https://doi.org/10.1007/s11192-015-1589-3

Wang, X., Mao, W., Xu, S., & Zhang, C. (2014). Usage history of scientific literature: Nature metrics and metrics of nature publications. *Scientometrics, 98*(3), 1923–1933. https://doi.org/10.1007/s11192-013-1167-5

Wang, X., Wang, Z., & Xu, S. (2013). Tracing scientist's research trends realtimely. *Scientometrics, 95*(2), 717–729. https://doi.org/10.1007/s11192-012-0884-5

Wu, H., Yi, H., & Li, C. (2021). An integrated approach for detecting and quantifying the topic evolutions of patent technology: A case study on graphene field. *Scientometrics, 126*(8), 6301–6321. https://doi.org/10.1007/s11192-021-04000-2

Xu, H., Winnink, J., Yue, Z., Zhang, H., & Pang, H. (2021). Multidimensional scientometric indicators for the detection of emerging research topics. *Technological Forecasting and Social Change, 163*, 120490. https://doi.org/10.1016/j.techfore.2020.120490

Ye, G., Wang, C., Wu, C., Peng, Z., Wei, J., Song, X., Tan, Q., & Wu, L. (2023). Research frontier detection and analysis based on research grants information: A case study on health informatics in the US. *Journal of Informetrics, 17*(3), 101421. https://doi.org/10.1016/j.joi.2023.101421

Zahedi, Z., & Haustein, S. (2018). On the relationships between bibliographic characteristics of scientific documents and citation and mendeley readership counts: A large-scale analysis of web of science publications. *Journal of Informetrics, 12*(1), 191–202. https://doi.org/10.1016/j.joi.2017.12.005

Zhang, C., Bu, Y., Ding, Y., & Xu, J. (2018). Understanding scientific collaboration: Homophily, transitivity, and preferential attachment. *Journal of the Association for Information Science and Technology, 69*(1), 72–86. https://doi.org/10.1002/asi.23916

Zhang, G., Shang, F., Wang, L., Xie, W., Jia, P., Jiang, C., & Wang, X. (2023). Is peer review duration shorter for attractive manuscripts? *Journal of Information Science*. https://doi.org/10.1177/01655515231174382

Zhang, G., Wang, Y., Xie, W., Du, H., Jiang, C., & Wang, X. (2021). The open access usage advantage: A temporal and spatial analysis. *Scientometrics, 126*(7), 6187–6199. https://doi.org/10.1007/s11192-020-03836-4

Zhao, S. X., Lou, W., Tan, A. M., & Yu, S. (2018). Do funded papers attract more usage? *Scientometrics, 115*(1), 153–168. https://doi.org/10.1007/s11192-018-2662-5

Zong, Q., Fan, L., Xie, Y., & Huang, J. (2020). The relationship of polarity of post-publication peer review to citation count evidence from publons. *Online Information Review, 44*(3), 583–602. https://doi.org/10.1108/OIR-01-2019-0027