Information Processing and Management 58 (2021) 102429

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

# How wide is the citation impact of scientific publications? A cross-discipline and large-scale analysis

Yi Bu [a], Wei Lu [b], Yifei Wu [c], Hongkan Chen [d], Yong Huang [e,*]

[a] Department of Information Management, Peking University, Beijing 100871, China
[b] Information Retrieval and Knowledge Mining Laboratory, School of Information Management, Wuhan University, Wuhan, Hubei 430072, China
[c] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[d] Department of Information Management, Peking University, Beijing 100871, China
[e] Information Retrieval and Knowledge Mining Laboratory, School of Information Management, Wuhan University, Wuhan, Hubei 430072, China

ARTICLE INFO

*Keywords:*
Citation impact
Scientific impact
Citation analysis
Scientometrics
Science of science

ABSTRACT

Although scientometricians have focused on the strength of the citation impact of scientific publications, only a few have paid special attention to the width of the citation impact. In this article, we aim to understand the width by establishing our empirical study on a previously built structure, namely ego-centered citation networks (ECCNs). We remove the direct links to the focal publication and only retain the other edges in the network. Particularly, we examine the number and the size of the connected components of sub-networks of ECCNs in the whole Web of Science database. We find that the number and size of the connected components increase as the number of citations of publications rises, regardless of discipline, and that they are greater in recent publications than in older ones. We also observe that there are some differences in terms of the number/size of the connected components; particularly, the number/size of connected components for publications in Arts and Humanities is more "special" compared with other disciplines, and Natural Sciences tends to have greater values. These details are indicative in terms of the specializations of academia. Moreover, the findings have far-reaching implications on research assessment practices and literature searching.

## 1. Introduction

A recent economic study argued that scientific research might be one of the most valuable ways for a government to spend its money (Science Business Publishing, 2017). Indeed, funding for scientific research is a fundamental way of creating jobs, realizing economic growth, and maintaining the innovative edge for a country (Audretsch et al., 2002; Stephan, 1996). These reasons are why many countries spend billions of dollars in fostering scientific research and development each year. With the astronomical investment in and preferential policies on scientific research, governments, industries, and other funding providers are eager to know whether their investments pay off and, if so, to what extent. To this end, they need to systematically assess the outputs of the research they fund and expect that the outputs are impactful in scientific communities, industries, and/or society. As one of the most important and common outputs of research (King, 1987), scientific publications offer a convenient way to evaluate academic work, because they are a "countable end product of research in their final and public form" (Cano, 1989, p. 284). Thus, understanding the characteristics of

* Corresponding author.
*E-mail address:* yonghuang1991@whu.edu.cn (Y. Huang).

https://doi.org/10.1016/j.ipm.2020.102429
Received 13 July 2020; Received in revised form 29 September 2020; Accepted 27 October 2020
Available online 5 November 2020

scientific publications has received much attention from funding providers, policy makers, and research managers in the past several decades.

A scientific publication might have impacts on different aspects, such as economically and societally, among which its scientific community would particularly benefit from the publication because of, for example, its novel ideas, algorithms, findings, summarizations, and/or advancement of knowledge. The impact within the scientific system of publications is called scientific impact, i.e., the impact of the scientific publication on other studies. That is to say, scientific impact is the degree to which one's publications have contributed to knowledge and have impacted the thinking of others (Bollen, Sompel, Hagberg & Chute, 2009). This type of impact is the simplest and the most direct one among all. Among the many measurements of the scientific impact of a publication, citation-based indicators are very frequently used (Wang, Song & Barabási, 2013). The citation impact of a scientific publication is the scientific impact measured by its citations, which is narrower than the scientific impact itself.[1] The citation impact has long been regarded as an important characteristic of a scientific publication because it offers a useful way of examining the scientific impact of a publication (e.g., Bornmann & Daniel, 2008; Tahamtan & Bornmann, 2019) and assists us in understanding knowledge diffusion and the use of information (Cole & Cole, 1974).

Many previous researchers have studied the citation impact of scientific publications, paying particular attention to the strength of their citation impact (e.g., Ding, 2011), either the raw citation counts (Waltman, 2016), normalized citation counts (e.g., Zhang, Cheng & Liu, 2014), or content-based citation strength (e.g., Ding et al., 2014). Yet, only a few have focused on how wide the citation impact is for a certain publication. For instance, both publications A and B in Fig. 1 have received five citations. However, the citing publications of A (i.e., A1, A2, A3, A4, and A5) are strongly connected with each other, but those of B are "clustered" as two parts, namely {B1, B2} and {B3, B4, B5}, by recorded citing relationships. In this case, A has a widespread impact on more scientific communities while B has a narrow impact on a limited number of communities. Such a phenomenon should be distinguished in characterizing the citation impact of scientific publications because it offers more invaluable details on how a specific piece of research unfolds scientific advances and assists research assessment.

Width might be understood from two distinct but related perspectives. Let us consider the citation networks of a certain publication. Firstly, the number of connected components of the citation network, for instance, is a typically used metric in network science. In epidemiology, the number of connected components is used to quantitatively determine the expectations of the number of persons that need to be infected for the epidemic to "affect" the entire population in a certain geographical area (Audenaert, Hanson & Datta, 2019); its mathematical properties have also been widely discussed (e.g., Frank, 1978). In our research context, the number of connected components of a given citation network, for example, indicates the width of the impact of the focal publication, whereby more connected components in a network illustrate a widespread impact on more scientific communities whilst fewer connected components demonstrate a narrower impact on a limited number of scientific communities. Thus, we are interested in the distribution of the number of connected components, and how the distributions differ for publications with different characteristics (e.g., publications within various research domains, with a distinct number of citations, and in different years). These details help us understand one aspect of the width of the citation impact of scientific publications.

Secondly, sizes of connected components of a citation network indicate the largeness of the impacted scientific communities. Quantifying the distribution of sizes provides a clearer understanding of how and to what extent they differ. In network science, sizes of connected components have been widely employed for different research objectives, for instance, in social media, as a way of tracking political abuse (Ratkiewicz et al., 2011). In our research context, we are interested in exploring the distribution of the sizes of connected components, and how the distributions differ for publications with different characteristics (e.g., publications within various research domains, with a distinct number of citations, and in different years). These offer more in-depth details of the width of the citation impact.

The research objective of this article is to explore the width of the citation impact of scientific publications from two levels, namely the number of connected components and the size of the connected components of the citation network. Empirically, we use the whole Web of Science in our empirical study and examine all disciplines instead of just a part of science.

The rest of this article is outlined as follows. First, we discuss related work from the perspective of the width of the citation impact of scientific publications. Then, we describe the dataset and the methodology used in this article, which is followed by a discussion and comparison of our results with existing related studies. Finally, we present conclusions, implications, and suggestions for future work.

## 2. Related work

### *2.1. Citation impact indicators*

Previous studies proposed many citation impact indicators (CIIs) as ways of characterizing citation-based impact. CIIs are defined as indicators of scientific impact based on analyses of citations received by scientific publications (Waltman, 2016). They have been researched, developed, and extended in various contexts and play an important role in research evaluation nowadays (Mingers & Leydesdorff, 2015; Vinkler, 2010; Wildgaard, Schneider & Larsen, 2014). Waltman (2016) proposed five types of "basic CII," namely total citation count, average number of citations per publication, h-index, number of highly cited publications, and proportion of

[1] We should note that sometimes publications may also have citation impacts from patents. In these cases, citation impact may not be limited to a scientific impact—it also involves, for instance, a technological impact. However, in the current research, when mentioning "citation impact," we do not count any citations other than those from scientific publications.

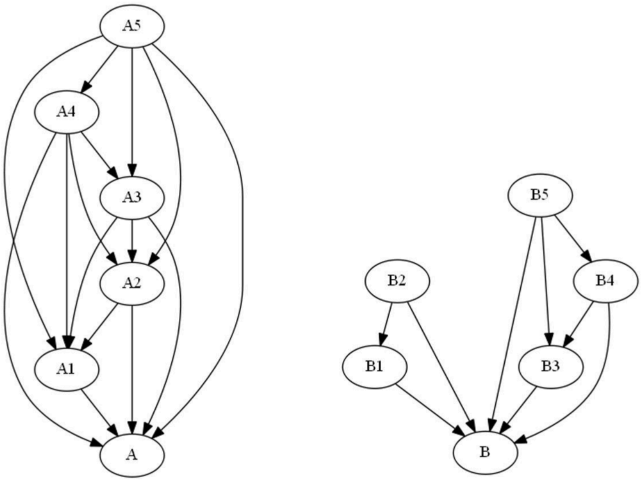


**Fig. 1.** A toy example for illustrating the width of the citation impact of scientific publications.

highly cited publications. Nonetheless, one of the deficiencies that basic CIIs have is that it is difficult to compare the impact of publications in different disciplines with different ages, and in different types. Therefore, scientometricians proposed several ways of normalization to improve basic CIIs, such as mean- (e.g., Albarrán, Crespo, Ortuño & Ruiz-Castillo, 2011; Moed, De Bruin & Van Leeuwen, 1995; Waltman et al., 2012) and highly cited publication-based strategies (e.g., Bornmann & Daniel, 2008; Van Raan, Visser, Van Leeuwen & Van Wijk, 2003).

Yet, CIIs listed above do not take into consideration citation networks when calculated. As a result, the calculation of these count-based CIIs cannot reflect any global patterns in a publication's domain. To solve this issue, there is an important thread of research focusing on the usage of citation networks on CIIs. However, the step from count- to network-based CIIs did not happen overnight. Instead, scientometricians first attempted small citation networks and only considered the citation count of citing publications; in a word, the network they used only contains the publication and its citing publications (a two-layer citation network), as well as their number of citations. For instance, Ding and Cronin (2011))) distinguished different citing publications based on whether they were highly cited or not, and found that, in the field of information retrieval, some authors were highly cited (high popularity) but not cited by highly cited authors (high prestige). These studies used the citation count of citing publications of a given article as a weight to normalize the citation count of this article; yet, its citing publication's citation count is not normalized. In a word, they simply painted a static picture on a small (local) part of a citation network. Additionally, they failed to consider any citing relationships between citing publications of a focal publication; thus, regretfully, though presenting local-level analyses, they did not show sufficient local details.

Much extant research in this track employed an iteration method to measure micro-level CIIs. On the publication level, the PageRank (Brin & Page, 1998) algorithm has been applied to analyze publications' citation networks. PageRank works by counting the number of citations of a research unit and the citation count of citing research units to determine the quality of this research unit (Waltman & Yan, 2014). Chen, Xie, Maslov and Redner (2007) proposed a standard PageRank algorithm for assessing the citation-based impact of publications; yet, their proposed approach has the drawback that older publications tend to have higher PageRank values than more recent publications. To solve this, a correction for the age of a publication was introduced in the approach proposed by Walker, Xie, Yan and Maslov (2007).

### 2.2. *Width of the citation impact*

The width of the citation impact of scientific publications has previously been discussed in scientometric literatures. For instance, Bu, Waltman and Huang (2019) proposed a multi-dimensional perspective and examined level, depth/breadth, and dependence/independence of the citation impact of scientific publications. Particularly, they used the number/proportion of citing publications that do not cite other citing publications of the focal publication as measurements. Mohapatra, Maiti, Bhatia and Chakraborty (2019) proposed a new indicator to quantify the impact of scientific publications and concluded that their indicator shows good performances on predicting highly influential publications. Similarly, della Briotta Parolo, Kujala, Kaski and Kivelä (2020) focused on Min, Chen, Yan, Bu and Sun (2020)) "citation cascades" and built up two new measurements (the persistent influence score and the diffusion model) to track the cumulative knowledge spreading from the perspective of multiple generations of a citation network. However, both measurements proposed by Mohapatra et al. (2019) and della Briotta Parolo et al. (2020) need great computational power in practice—for example, Mohapatra et al. (2019) removed a considerable number of edges in citation networks based on their rules, and this is time-consuming. Also, the assumption that the most influential article is that with equal values of width and depth seems problematic and lacks theoretical underpinning. Moreover, their empirical studies were limited to a single domain instead of all science. Hence, in this article, we expect to propose a more effective, easy-to-understand strategy for investigating the width of the citation impact, and implement a cross-discipline, large-scale comparison on all scientific disciplines.

## 3. Data

The dataset used in this research comes from Clarivate Analytics' Web of Science (WoS) XML raw data housed at the Indiana University Network Science Institute (IUNI). The WoS database covers bibliographic metadata of high-quality journal and book series (included in the Science Citation Index Expanded [SCI-E], the Social Sciences Citation Index [SSCI], and the Arts & Humanities Citation Index [A&HCI]), as well as conference proceedings and monographs (included in the Conference Proceedings Citation Index and the Book Citation Index). The Web of Science (WoS) dataset housed at the Indiana University Network Institute (IUNI) contains various types of publication such as research articles, review articles, conference papers, and books. These publications range from the years of 1900 through to 2016. The total number of publications covered is 63,590,916. Each publication in Web of Science (WoS) is labeled with one or more subjects (micro-level disciplines), such as "anthropology" and "business," and there are 254 different micro-level disciplines in total. All these 254 micro-level disciplines are categorized into six macro-level disciplines (in a tree structure), namely: (1) Arts and Humanities; (2) Clinical, Pre-clinical, and Health; (3) Engineering and Technology; (4) Life Sciences; (5) Physical Sciences; and (6) Social Sciences. The mapping relations between the macro-level and micro-level disciplines are shown in the link mentioned in the footnote.[2] Table 1 shows the number of publications categorized in each macro-level discipline, in which we can see that Life Sciences has the greatest number of publications (nearly 30% out of all publications). Physical Sciences, Engineering and Technology, and Clinical, Pre-clinical, and Health all have more than 20% of all publications. Social Sciences only has 14.18% of all publications, and Arts and Humanities has the fewest, only 7.75%.

Note that some publications have one label while some might have multiple labels for subjects, meaning that a specific publication might be categorized into a single or multiple micro-level disciplines in the database. Table 2 shows the number of micro-level discipline labels for each publication. From the table, we can see that about 75% of publications only have one label. About 22% of publications belong to two different micro-level disciplines.[3] There are only a limited number of publications with more than two labels for their subjects.

Furthermore, to see some potential interesting patterns in a small domain instead of macro-level disciplines such as Arts and Humanities, we also select "Scientometrics" as a specific micro-level discipline. The reasons for selecting Scientometrics are that: (1) Scientometrics is neither a too large (e.g., library and information science) nor a too small domain (e.g., topic modeling); and (2) we (the authors) are specialists in this domain, which means that we have more knowledge about it. To identify which publications in WoS "belong to" Scientometrics, we cannot use the WoS-generated labels because there are no labels named "Scientometrics" (perhaps the most similar label is "Information Science & Library Science," which is obviously not what we want). To this end, we employ a method introduced by Waltman and Van Eck (2012). Based on their method, all WoS-indexed publications[4] labelled with "article," "letter," and "review" were clustered based upon citing relationships. In total, 4047 clusters of publications were obtained. Clusters are non-overlapping, which means that each publication belongs to only one cluster. One of these 4047 clusters can be considered to represent the field of Scientometrics. We selected all 11,921 publications from 2000 to 2016 in this cluster as Scientometrics publications in this research. More details of the selection of Scientometrics publications can be found in our previous study (Bu et al., 2019).

Web of Science (WoS) also records all citing relationships among these 63+ million publications (Type I citing relationships), as well as the citing relationships from any of these 63+ million publications to outside publications that are not covered by WoS (Type II citing relationships). In this research, we only consider Type I citing relationships, as WoS has no unique identifier for the cited publications outside WoS.[5] Also, WoS does not have the bibliographic metadata (e.g., journal, publication year, document type, title, author(s)) of these publications either. The total number of Type I citing relationships included is 1227,494,925. In this article, for simplicity, when we mention "citing relationships," it means Type I citing relationships. Fig. 2 indicates how the numbers of citations of publications in different macro-level disciplines are distributed (specifically, we use the complementary cumulative distribution functions [CCDFs] to show them), where we can see that citation distribution is highly skewed, which has been discussed quite frequently before (Waltman et al., 2012). Meanwhile, when comparing different macro-disciplines, we observe that publications in Arts and Humanities tend to have the fewest number of citations (Tahamtan, Afshar & Ahamdzadeh, 2016). For Scientometrics, we observe that the curve after 500 (the horizontal axis) does not follow the original pattern; this is because, in our selected Scientometrics publications, most have only received fewer than 500 citations with the exception of a limited number of publications, such as the h-index publications (with 2000+ citations).

More descriptive statistics of the dataset can be found in the Supplementary Information (i.e., Figure A1, Figure A2, and Figure A3).

[2] http://help.prod-incites.com/inCites2Live/indicatorsGroup/aboutHandbook/appendix/mappingTable.html

[3] Given the issue of duplication in categories, some scientometricians, such as Vincent Larivière and Cassidy R. Sugimoto, have started to use a National Science Foundation (NSF) subject classification strategy to determine the single subject of a certain publication (e.g., Lu et al., 2019).

[4] Note that this algorithm was run on a different version of WoS, which is housed at the Center for Science and Technology Studies (CWTS), Leiden University, the Netherlands. However, there are no significant differences between the two datasets.

[5] That is to say, if two publications in WoS (say A and B) both cited a common publication outside WoS (say C), WoS will assign two different IDs for C and we therefore cannot identify that they are the same publication.

**Table 1**
Number of publications in each macro-level discipline.

| Macro-level discipline | Number of publications (%) |
|---|---|
| Arts and Humanities | 5361,562 (7.75%) |
| Clinical, Pre-clinical, and Health | 16,493,933 (23.83%) |
| Engineering and Technology | 17,869,205 (25.82%) |
| Life Sciences | 20,368,543 (29.43%) |
| Physical Sciences | 18,053,623 (26.08%) |
| Social Sciences | 9814,173 (14.18%) |
| ALL (total) | 69,214,524 (100.00%) |

**Table 2**
Distribution of the number of micro-level discipline labels for each publication.

| Number of labels | Number of publications (%) |
|---|---|
| 1 | 52,189,631 (75.40%) |
| 2 | 15,409,995 (22.26%) |
| 3 | 1509,241 (2.18%) |
| 4 | 104,724 (0.15%) |

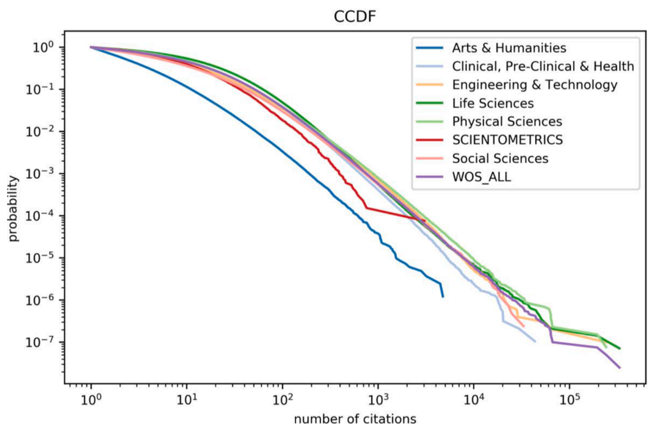


**Fig. 2.** Complementary cumulative distribution functions (CCDFs) of the number of citations for publications in all macro-level disciplines and Scientometrics.

## 4. Methods

### *4.1. Quantifying the citation impact of scientific publications*

When quantifying the citation impact of a certain publication, we should first focus on where the node representing this publication is located in the network, and then investigate which publication(s) cites the focal publication to examine how the focal publication impacts other publications. Thus, following Huang, Bu, Ding and Lu (2020), in this article, we need an ego-centered citation network (ECCN) to understand its citation impact. Intuitively, the network should at least involve the focal publication and its citing publications. Yet, these may not be enough—it may also be important to investigate the direct citations between citing publications (DCCPs). DCCPs are informative to the focal publication, because they show the paths of knowledge flow within the scope of the focal and citing publications (Huang et al., 2020). In their proposed ECCNs, there are two types of edge, direct citations and DCCPs[6]. The top graph of Fig. 3 shows an example of an ECCN (a.k.a., the ECCN of publication A). In this example, each node represents a publication, and an edge represents a citing relation. Note that citing relationships pointing to A are shown in solid lines, while those not pointing to A are in dotted lines. Meanwhile, the direction of the arrows shows the direction of the citing relationships, which is the reverse to the direction of knowledge diffusion.

[6] This networked structure was called a "citing cascade" in Huang, Bu, Ding, and Lu (2018).

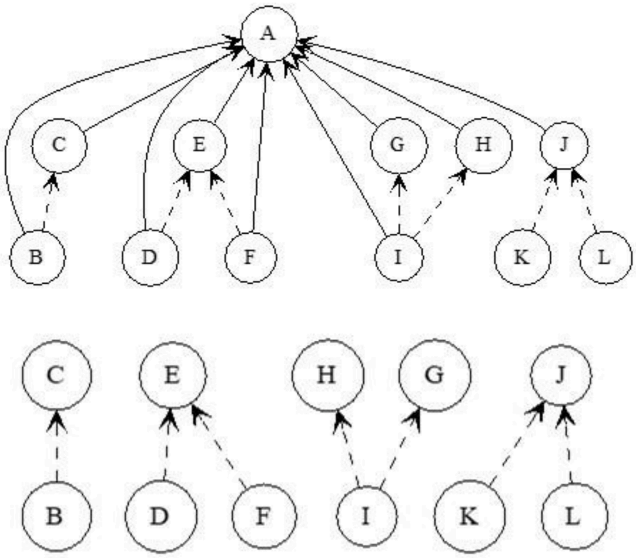


**Fig. 3.** A toy example of sub-networks of an ego-centered citation network (sub-ECCN, below) generated from an ego-centered citation network (ECCN, top).

### *4.2. Characterizing the width of the citation impact*

In many biology-related contexts, direct links (e.g., the citing relationship from C to A in the top sub-figure of Fig. 3) are often regarded as redundant (e.g., Bonneau et al., 2006), because there still remains another path in the network (e.g., the path C→B→A correspondingly). Therefore, many network scientists removed some of the existing edges in the network and examined the topological structures of the new "pruned" sub-networks. For instance, Clough, Gollings, Loach and Evans (2015), as mentioned before, implemented a transitive reduction of citation networks by removing direct citations between two publications given that they have other connected paths in the citation network. Inspired by the biological perspective, by the study of Clough et al. (2015), and by some other related studies (e.g., Vasiliauskaite & Evans, 2018), we plan to investigate the topological structures of ECCNs by removing all direct citations, as well as the focal publication; the remaining part in the ECCN will be the citing publications and the citing relationships among them. We call the remaining sub-network the sub-ECCN of the raw ECCN. Fig. 3 shows a toy example of an ECCN (the top sub-figure) and its sub-ECCN (the bottom sub-figure). For each ECCN, we calculate the number and the size of the connected components in our empirical study.

### *4.3. Null model (baseline)*

To reveal "real" patterns, we follow Fontana, Iori, Montobbio and Sinatra (2020) and establish a null model by randomly reshuffling the whole Web of Science (WoS) citation edge list (see details in Fig. 3 of Fontana et al. (2020)). The edge of implementing a null model in scientometric studies has been adopted by many studies (AlShebli, Rahwan & Woon, 2018). Particularly, given the edge list containing all Type I citing relationships of WoS (detailed in the "Data" sub-section), we record the publication year and the labeled subject(s) of the target node (i.e., the cited publication). For any two real citing relationships, if they have the same source nodes (i.e., the citing publications), and the publication year and labeled subject of the target nodes are the same, we reshuffle the two edges (i.e., citing relationships). To do this, in practice, we build up a list of edges for each publication year and each subject. We then randomize the target nodes with random.choice() function in the Python random package for each list. We duplicate this process ten times to obtain more stable randomization results. We actually also tried another randomization strategy by only keeping the publication year of two target nodes the same. However, we do not think this is a reasonable way of performing randomization as too many characteristics are discarded, although we find that the curves representing two randomization strategies do not vary obviously.

## 5. Results and discussion

### *5.1. Number of connected components*

In this section, we show some interesting patterns regarding the number of connected components of sub-ECCNs for different publications. Essentially, the number of connected components indicates into how many "groups" (small communities) the citing publications of a focal publication are clustered. Recall that the number of connected components of a given sub-ECCN indicates the width of the impact of the focal publication (a widespread impact of more scientific communities vs. a narrow impact on a limited number of communities).

Fig. 4 manifests the distribution of the number of connected components of sub-ECCNs for publications. Note that in figures related to this article, all six macro-level disciplines are denoted using lower-case letters, whilst Scientometrics is denoted using upper-case letters (i.e., "SCIENTOMETRICS"); this is for a better comparison in the figures. Additionally, in these figures, "ALL" indicates results relating to all publications, regardless of their respective disciplines. "RANDOMIZE" indicates the results from the null model detailed in the previous sub-section.

Typically, we can see that most macro-level disciplines, as well as the whole WoS database, show a typical power-law distribution, as their CCDFs are straight lines in the log-log coordinate. This indicates that most sub-ECCNs of publications only have a few components, while there are only a limited number of sub-ECCNs that have many components. Among all eight real-data curves in Fig. 4, there are only two curves that are not quite "similar" to the others, i.e., the curves representing Arts and Humanities and Scientometrics. Specifically, we can see that, generally, Arts and Humanities publications have fewer components in their sub-ECCNs compared with other disciplines. For example, there are about 1% of publications in other disciplines that have ten or more components, but this is equal to 0.3% for Arts and Humanities. As for Scientometrics, those with eight or fewer components have similar patterns of the complement cumulative distribution function compared with other macro-level disciplines, but one can observe that there are obviously fewer Scientometrics publications with more than eight components in their sub-ECCNs, compared with other disciplines (including Social Sciences). This is quite likely because of the fact that we only have a limited number of Scientometrics publications that have many citations (Fig. 2). Moreover, we find that disciplines might have different maximum values of the number of connected components. For example, Life Sciences has around 700 as its maximum value while Arts and Humanities has approximately forty.

The relatively lower values of the number of connected components for publications might be interpreted from three different perspectives: First, Arts and Humanities is found to be reflexive and non-cumulative compared with Natural Sciences (Reale et al., 2018; Weingart & Schwechheimer, 2007). This reveals potential different citing behavior (e.g., citing habits and "culture") between Arts and Humanities and other domains, e.g., self-citations (Lariviere & Sugimoto, 2019). Second, as concluded by Bod (2013) and Molas-Gallart (2015), Arts and Humanities is more interested in novel approaches instead of new theories and findings. The relative preference for methods compared with other domains indicates different patterns of citation networks of publications in Arts and Humanities. This consequently influences the patterns of the number of connected components and the number of direct citations between citing publications, as observed by Huang et al. (2020). Third, given that we use WoS as the main source, there are many uncaptured records for Arts and Humanities (also for Social Science). This is due to the following reasons: (1) Arts and Humanities is found to be more national- or regional-oriented compared with Natural Sciences (Frandsen & Nicolaisen, 2008); (2) many Arts and Humanities publications are not shown as research articles; instead, they could be monographs, regional or national serials, and/or reports (Broadus, 1971); (3) Fewer Arts and Humanities publications are written for English outlets compared to other disciplines (Engels, Ossenblok & Spruyt, 2012; Sivertsen, 2014)—of course, this is highly related to (1) as mentioned. We also find that Scientometrics is not similar to other disciplines, partly because of the limited number of articles in this field compared with other macro-level disciplines.

Fig. 5 presents the maximum, mean, and minimum values of the number of connected components in sub-ECCNs of highly (top 10% cited publications in a certain macro-level discipline), medium (top 10–50% cited publications in a macro-level discipline), and lowly cited (bottom 50% cited publications in a macro-level discipline) publication groups. Note that we have removed all publications without any direct citations between citing publications (i.e., the dotted arrows in Fig. 3) because these publications will actually have the same number of connected components as the number of citing publications, which would never shed light on our data analysis. From Fig. 5, we can see that most macro-level disciplines have similar patterns—lowly cited publications have the lowest number of connected components in their sub-ECCNs, while highly cited publications have the greatest number of connected components. Yet, we also find that the range of the number of connected components for highly cited publications tends to be greater than that of lowly cited publications, which hints that there are diverse patterns among the ECCNs of highly cited publications.

As a robustness check, Fig. 6 presents how the number of connected components relates to a publication's number of citations. In the figure, we observe that different macro-level disciplines have similar patterns but there are also some minor differences. The similar pattern is that the number of connected components increases as the number of citations rises, which is consistent to what we have found in Fig. 5. Specifically, when taking into consideration all disciplines' publications, there are on average 2.6 components for publications with 100 citations; this number equals approximately 8.0 for publications with 1000 citations. Meanwhile, we find some discipline-level differences, especially when the number of citations is greater than 1000. Note that the curve indicating publications from all disciplines outperforms others on the top right of the figure, which hints that many publications have more than one subject label (see Table 2) and that many multiple-label publications are from Life Sciences. It is reasonable to have a positive relation between the number of citations of publications and their numbers of components because, based upon the definition of components in network science, a small network cannot have too many components.

We also explore how the number of connected components changes over the years, as shown in Fig. 7 and Figure A4 (in the Supplementary Information), in which we observe that the number of connected components increases as time goes by for publications before 2000. In 1980, for instance, on average there were about 8.5 components in sub-ECCNs for publications, and this number equaled approximately 7.7 in 1960. After 1960, the speed of increase for Arts and Humanities is transparently slower than that of other disciplines. As a supplement, Figure A4 in the appendix shows how the number of connected components in sub-ECCNs is differently distributed in publications in different time periods (1910–1920, 1930–1940, 1950–1960, 1970–1980, 1990–2000, and 2010–2016[7]).

[7] Since Scientometrics publications are only from 2000 to 2016, in this figure, we do not consider them.

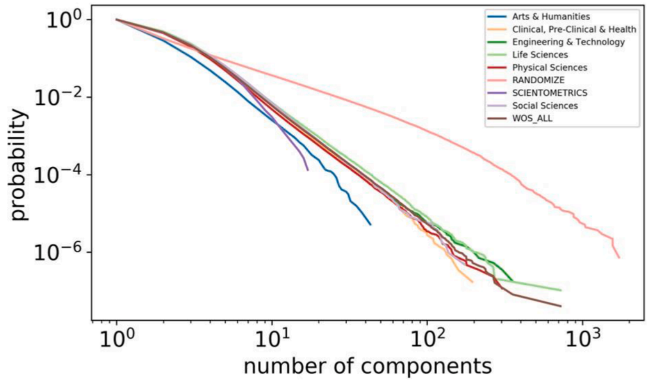


**Fig. 4.** Complementary cumulative distribution functions (CCDFs) of the number of connected components in sub-ECCNs of publications in different disciplines.

Similar to Fig. 7, we can see that the number of connected components manifests an increasing trend over the years. For instance, in 1910–1920, there are 0.03 publications (when considering publications from all disciplines) with five or more components, but this number increases to 0.10 for 1990–2000. A robustness test with 95% confident interval is displayed in Figure A5 in the Supplementary Information.

The fact that there are more DCCPs in recent publications than in older ones could be interpreted from the following two perspectives:

On the one hand, publications are found to contain more references than before (Dong, Ma, Shen & Wang, 2017; Petersen, Pan, Pammolli & Fortunato, 2019; Sinatra, Deville, Szell, Wang & Barabási, 2015). This is also empirically reflected in Figure A1. This fact could be attributed to the following three reasons:

(1) The rapid development of research leads to the accumulation of more literature and makes better science;
(2) Having multiple ways of searching and acquiring literature raises the availability of previous publications; and
(3) The "more complex, interdisciplinary nature of modern science that has developed over time" (Dong et al., 2017, p. 1442).

On the other hand, publications tend to include more information than before. For instance, Vale (2015) compared all publications in the first half of 1984 and 2014 in Cell, Nature, and Journal of Cell Biology, and observed that the number of panels in the experimental figures of these publications increased by two to four times, partly due to demands from reviewers.

These two perspectives perhaps explain why more recent publications tend to be cited more than older publications. This also increases the possibilities of forming new links in citation networks, which consequently raises the number of connected components.

The comparison between document type and the number of connected components can be found in Figure A6 in the Supplementary Information.

### *5.2. Size of the connected components*

The size of the connected components indicates how large scientific communities endorsing the focal publication are. Fig. 8 shows the CCDF of size of the connected components in sub-ECCNs of scientific publications, where we find a similar pattern among the number of connected components: There are many components that are quite small while there are only a few components that are large. Different macro-level disciplines have similar distributions, except Arts and Humanities.

Fig. 9 shows the relationship between the size of the connected components in sub-ECCNs of scientific publications and their numbers of citations in different disciplines. One can see a clear positive relation between the two variables. Meanwhile, such a finding is domain-unspecific. As for domain comparison, Physical Sciences has the largest number of connected components while Arts and Humanities has the lowest. The special features revealed by Arts and Humanities, again, might be attributable to its nature and/or the coverage of WoS (Reale et al., 2018).

Again, we are also interested in how size of the connected components differs in sub-ECCNs for publications in different years. From Fig. 10, we can see that the size of the connected components increases steadily over the years before 2000. Clinical, Pre-Clinical, and Health, and Life Sciences, have the greater values for size of the connected components. Arts and Humanities and Social Sciences have the lowest values in terms of the size of the connected components. A robustness check has also been implemented in Figure A7 and it shows consistent results.

The comparison between document type and the size of the connected components can be found in Figure A8 in the Supplementary Information.

In the above analyses, we showcase how the size of the connected components is differently distributed in various disciplines, and how it differs in publications with different numbers of citations and in different years. Below, in Table 3, we extend our analyses on the maximum, mean, median, and diversity of the size of the connected components in sub-ECCNs for publications. In the table, we

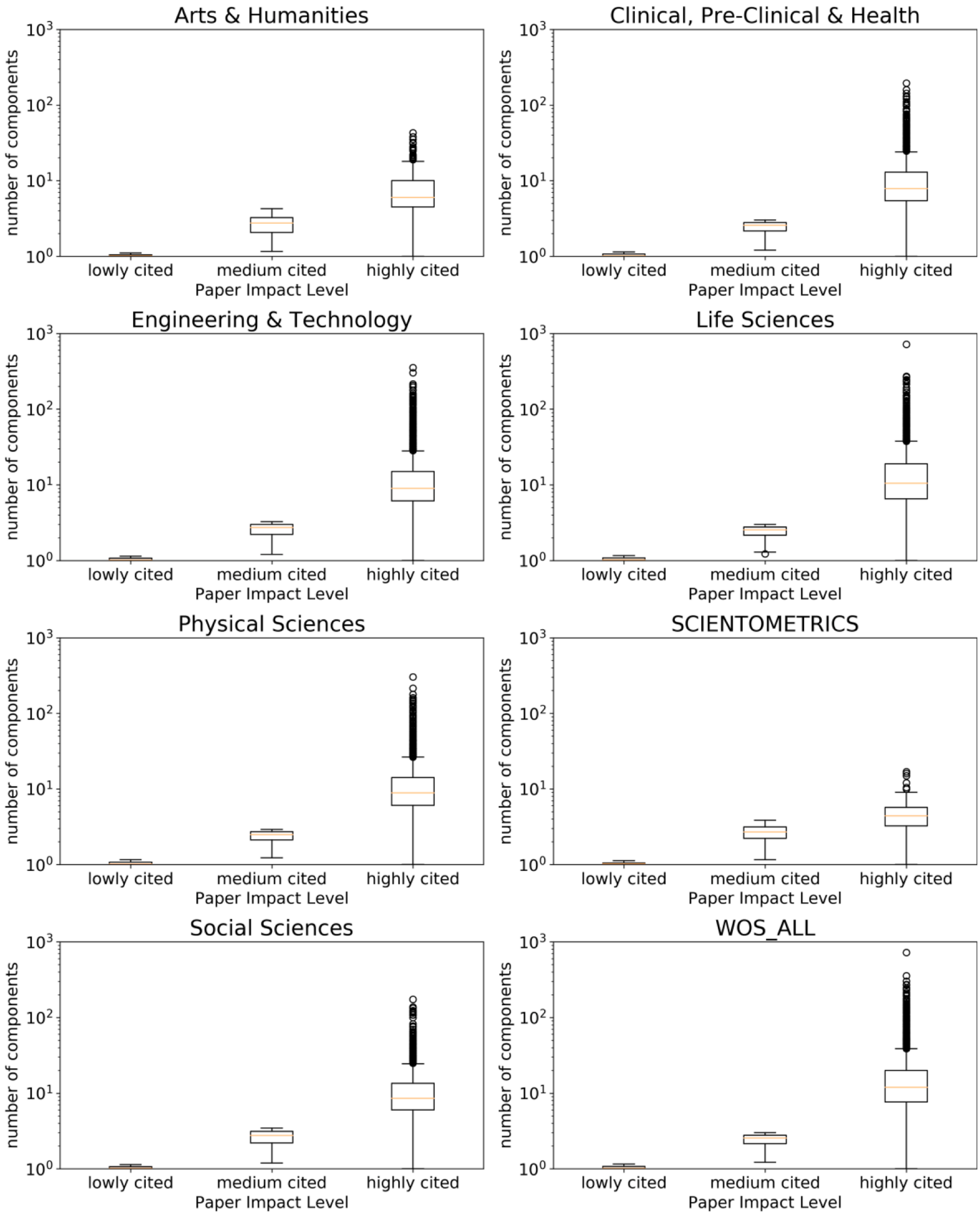


**Fig. 5.** Maximum, mean, and minimum values of the number of connected components in sub-ECCNs for lowly, medium, and highly cited publication groups. All scientific publications that have no direct citations between citing publications are removed.

examine highly (top 10% cited publications in a certain discipline), medium (top 10–50% cited publications), and lowly cited (bottom 50% cited publications) publication groups, respectively.

Here, the Gini index (a.k.a., Gini coefficient) is employed as a proxy of diversity measurement on the size of the connected components. The Gini coefficient measures diversities by quantifying the inequality among values of a frequency distribution. It has been widely used in various fields, such as economics (Choi, 2006), political science (Dye, 1969), anthropology (Hammel, 2005), public health (Althoff, Hicks, King, Delp & Leskovec, 2017; Navarro et al., 2006), and network science (Pan, Petersen, Pammolli &

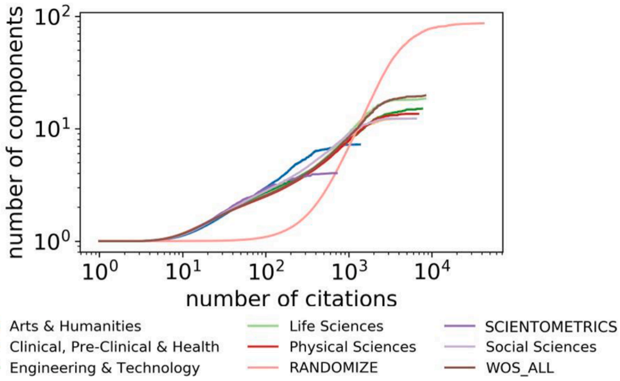


**Fig. 6.** Relationships between the number of connected components in sub-ECCNs and the different numbers of citations in different disciplines.

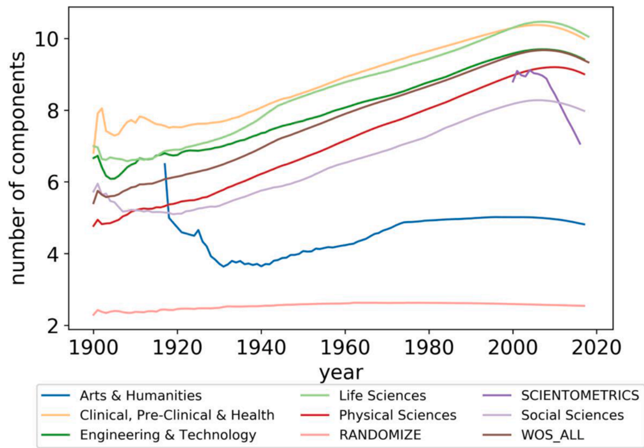


**Fig. 7.** Relationships between the number of connected components in sub-ECCNs of publications and their publication years in different disciplines.

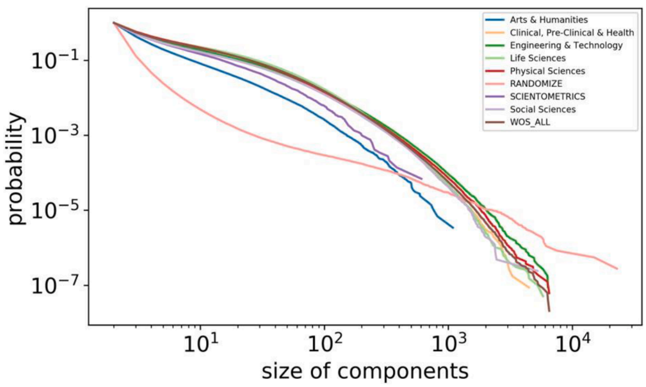


**Fig. 8.** Complementary cumulative distribution functions (CCDFs) of the size of the connected components in sub-ECCNs of publications in different disciplines.

Fortunato, 2018). A famous example using the Gini index is to measure the inequality of people's income. The Gini coefficient is mathematically defined based on the Lorenz curve (Egghe, 2005; Nijssen, Rousseau & Van Hecke, 1998) and the line of equality. The Lorenz curve measures as the vertical axis the percentage of people's total income and as the horizontal axis the cumulative bottom n% of the people. LoE, on the other hand, is shown as a 45-degree line in the same diagram. The Gini coefficient is calculated as the ratio of the area lying between the line of equality and Lorenz curve over the total area under the line of equality. If the Gini coefficient is zero,

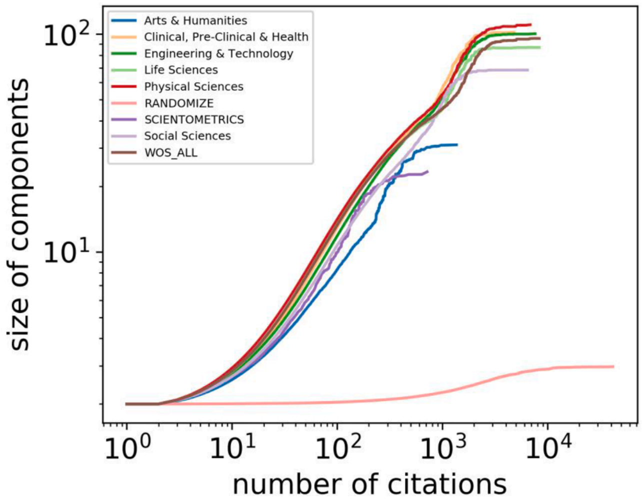


**Fig. 9.** Relationships between the size of the connected components in sub-ECCNs of publications with different numbers of citations in different disciplines.

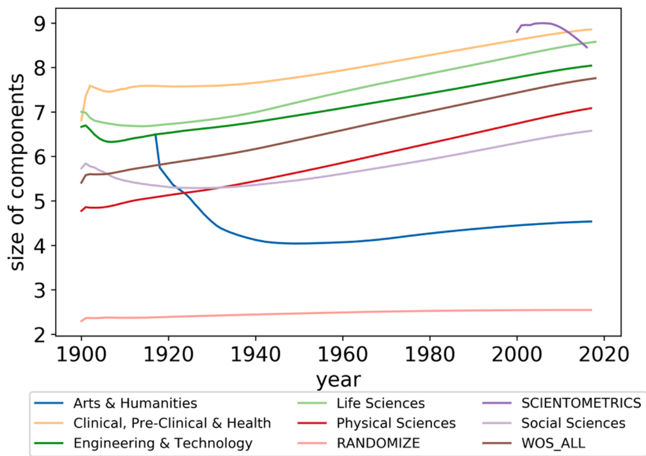


**Fig. 10.** Relationships between the size of the connected components in sub-ECCNs of publications in different years in different disciplines.

all values in the group are exactly the same, indicating perfect equality. A Gini coefficient of one expresses the maximal inequality among all values. The range of the Gini coefficient is therefore between zero and one. In practice, we calculate $G_{size\_of\_components}$ by using the Python Gini index calculator.[8]

From Table 3, we observe that, in each discipline, the maximum and average (mean) values of the size of the connected components in sub-ECCNs for highly cited publications are greater than those of medium cited publications, which are greater than lowly cited publications. These are consistent with the results in Fig. 9. As for the median, we can find that all groups, regardless of discipline and citation impact, have values of two or three. In terms of diversity (Gini index), we see that lowly and medium cited publications have a lower Gini index ($<$ 0.50 including Scientometrics) while highly cited publications obviously have a higher Gini index ($>$ 0.75 including Scientometrics). We know that a publication with a greater value of Gini index for its size of the connected components indicates that both huge and tiny components exist in the sub-ECCN. This is quite reasonable because, in ECCNs of lowly and medium cited publications, there are fewer citing publications and therefore the possibilities of forming huge components are quite limited compared to highly cited publications.

## 6. Conclusions and future work

This article investigates the width of the citation impact of scientific publications. We establish our empirical study on a previously

[8] https://github.com/oliviaguest/gini

**Table 3**
Descriptive statistics of the maximum, mean, median, and Gini index (diversity) of the size of the connected components in sub-ECCNs for publications in different macro-level disciplines. H: Highly cited publications, M: Medium cited publications, L: Lowly cited publications.

| Disciplines | Citation impact | Max. | Avg. | Median | Gini |
|---|---|---|---|---|---|
| **Arts and Humanities** | L | 35 | 3.54 | 2 | 0.34 |
| | M | 114 | 9.67 | 3 | 0.64 |
| | H | 1084 | 19.49 | 2 | 0.81 |
| **Clinical, Pre-clinical, and Health** | L | 35 | 5.18 | 3 | 0.45 |
| | M | 119 | 15.82 | 3 | 0.65 |
| | H | 4458 | 38.10 | 2 | 0.80 |
| **Engineering and Technology** | L | 35 | 4.86 | 3 | 0.44 |
| | M | 118 | 14.52 | 3 | 0.65 |
| | H | 6337 | 37.76 | 2 | 0.82 |
| **Life Sciences** | L | 35 | 5.55 | 3 | 0.46 |
| | M | 119 | 16.56 | 3 | 0.65 |
| | H | 5860 | 36.44 | 2 | 0.81 |
| **Physical Sciences** | L | 35 | 5.54 | 3 | 0.46 |
| | M | 119 | 17.16 | 3 | 0.64 |
| | H | 6499 | 40.56 | 2 | 0.81 |
| **SCIENTOMETRICS** | L | 34 | 4.31 | 2 | 0.41 |
| | M | 107 | 11.36 | 3 | 0.64 |
| | H | 605 | 27.64 | 2 | 0.79 |
| **Social Sciences** | L | 35 | 4.55 | 3 | 0.43 |
| | M | 119 | 13.10 | 3 | 0.66 |
| | H | 5221 | 30.59 | 2 | 0.82 |
| **ALL** | L | 35 | 5.33 | 3 | 0.46 |
| | M | 119 | 16.23 | 3 | 0.65 |
| | H | 6499 | 37.25 | 2 | 0.81 |

built structure, namely ego-centered citation networks (ECCNs). Particularly, we examine the number and the size of the connected components of sub-ECCNs on the whole WoS database. We find that the number and size of the connected components increase as the number of citations of publications rises, regardless of discipline, and that they are greater in recent publications than in older ones. We also observe that there are some differences in terms of the number/size of the connected components; particularly, the number/size of the connected components for publications in Arts and Humanities is more "special" compared with other disciplines, and Natural Sciences tends to have greater values.

In this article, one of our findings is that more recent publications have more connected component counts than older ones, which is indicative in terms of specializations of science. As objectives of science become "deeper" and more complex, a new and effective way of doing science, namely the division of labor, has been proposed. This makes science more specialized (Casadevall & Fang, 2014; Crane, 1972; Foster, Rzhetsky & Evans, 2015). The observation that the sizes of connected components are greater than before demonstrates the proliferation of these specializations, which, in turn, advances science as we observe an exponential increasing of the number of publications over time.

Quite a few factors might affect the patterns of the number/size of the connected components, such as publications' disciplines, published year, and author-related factors, etc. However, the current study separately examines some of these factors instead of combining them, although it is quite likely that these factors are interconnected and might influence each other. Quantifying how each of these functions would be crucial to understand patterns of width of publications' citation impact. Thus, future studies should employ more sophisticated approaches, such as econometrics methods, to investigate how these factors interact with each other. Particularly, some regression analysis should be implemented, and endogeneity issues need to be carefully addressed with multiple strategies. Along these lines, there are many extant works that have used these methods to detect complicated relationships among variables (e.g., J. Y. Wang, Jones & Wang, 2019, 2017).

## CRediT authorship contribution statement

**Yi Bu:** Conceptualization, Validation, Investigation, Writing - original draft. **Wei Lu:** Investigation, Conceptualization. **Yifei Wu:** Investigation. **Hongkan Chen:** Investigation, Writing - original draft. **Yong Huang:** Investigation, Conceptualization, Writing - review & editing.

## Acknowledgments

The authors thank two anonymous reviewers for their constructive comments. Yi Bu is grateful to Ying Ding, Yong-Yeol Ahn, Johan Bollen, Stasa Milojević, Cassidy R. Sugimoto, Ludo Waltman, and Dashun Wang for invaluable suggestions to improve this research. Financial support from the Chinese National Natural Science Foundation (Number: 72004168) and the fellowship of China Post-doctoral Science Foundation (Number: 2019M662729) is gratefully acknowledged.

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ipm.2020.102429.

## References


Albarrán, P., Crespo, J. A., Ortuño, I., & Ruiz-Castillo, J. (2011). The skewness of science in 219 sub-fields and a number of aggregates. *Scientometrics, 88*(2), 385–397.

AlShebli, B. K., Rahwan, T., & Woon, W. L. (2018). The preeminence of ethnic diversity in scientific collaboration. *Nature Communications, 9*(1), 1–10.

Audenaert, K., Hanson, E.P., .& Datta, N. (2019). A continuity bound for the expected number of connected components of a random graph: A model for epidemics. ArXiv Preprint ArXiv:1912.05599.

Althoff, T., Hicks, J. L., King, A. C., Delp, S. L., & Leskovec, J. (2017). Large-scale physical activity data reveal worldwide activity inequality. *Nature, 547*(7663), 336–339.

Audretsch, D. B., Bozeman, B., Combs, K. L., Feldman, M., Link, A. N., & Siegel, D. S. (2002). The economics of science and technology. *The Journal of Technology Transfer, 27*(2), 155–203.

Bod, R. (2013). *A new history of the humanities: The search for principles and patterns from antiquity to the present*. Oxford University Press.

Bollen, J., Sompel, H. V.de, Hagberg, A., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PloS one, 4*(6), e6022. https://doi.org/10.1371/journal.pone.0006022.

Bonneau, R., Reiss, D. J., Shannon, P., Facciotti, M., Hood, L., & Baliga, N. S. (2006). The Inferelator: An algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biology, 7*(5). https://doi.org/10.1186/gb-2006-7-5-r36. R36.

Bornmann, L., & Daniel, H.-. D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation, 64*(1), 45–80.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*(1–7), 107–117.

Broadus, R. N. (1971). Literature of social sciences-survey of citation studies. *International Social Science Journal, 23*(2), 236–243.

Bu, Y., Waltman, L., & Huang, Y. (2019). A multidimensional perspective on the citation impact of scientific publications. ArXiv Preprint ArXiv:1901.09663.

Cano, V. (1989). Citation behavior: Classification, utility, and location. *Journal of the American Society for Information Science, 40*(4), 284–290.

Casadevall, A., & Fang, F. C. (2014). Specialized science. *American Society of Microbiology*.

Chen, P., Xie, H., Maslov, S., & Redner, S. (2007). Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics, 1*(1), 8–15.

Choi, C. (2006). Does foreign direct investment affect domestic income inequality. *Applied Economics Letters, 13*(12), 811–814.

Clough, J. R., Gollings, J., Loach, T. V., & Evans, T. S. (2015). Transitive reduction of citation networks. *Journal of Complex Networks, 3*(2), 189–203.

Cole, J. R., & Cole, S. (1974). *Social stratification in science*.

Crane, D. (1972). Invisible colleges; diffusion of knowledge in scientific communities.

della Briotta Parolo, P., Kujala, R., Kaski, K., & Kivelä, M. (2020). Tracking the cumulative knowledge spreading in a comprehensive citation network. *Physical Review Research, 2*(1), Article 013181.

Ding, Y. (2011). Applying weighted PageRank to author citation networks. *Journal of the American Society for Information Science and Technology, 62*(2), 236–245.

Ding, Y., & Cronin, B. (2011). Popular and/or prestigious? Measures of scholarly esteem. *Information Processing & Management, 47*(1), 80–96.

Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., & Zhai, C. (2014). Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information Science and Technology, 65*(9), 1820–1833.

Dong, Y., Ma, H., Shen, Z., & Wang, K. (2017). A century of science: Globalization of scientific collaborations, citations, and innovations. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1437–1446).

Dye, T. R. (1969). Income inequality and American state politics. *American Political Science Review, 63*(1), 157–162.

Egghe, L. (2005). *Power laws in the information production process: Lotkaian informetrics*. Emerald Group Publishing Limited.

Engels, T. C., Ossenblok, T. L., & Spruyt, E. H. (2012). Changing publication patterns in the social sciences and humanities, 2000–2009. *Scientometrics, 93*(2), 373–390.

Fontana, M., Iori, M., Montobbio, F., & Sinatra, R. (2020). New and atypical combinations: An assessment of novelty and interdisciplinarity. *Research Policy, 49*(7), Article 104063. https://doi.org/10.1016/j.respol.2020.104063.

Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and innovation in scientists' research strategies. *American Sociological Review, 80*(5), 875–908. https://doi.org/10.1177/0003122415601618.

Frandsen, T. F., & Nicolaisen, J. (2008). Intradisciplinary differences in database coverage and the consequences for bibliometric research. *Journal of the American Society for Information Science and Technology, 59*(10), 1570–1581.

Frank, O. (1978). Estimation of the number of connected components in a graph by using a sampled subgraph. *Scandinavian Journal of Statistics*, 177–188.

Hammel, E. A. (2005). Demographic dynamics and kinship in anthropological populations. *Proceedings of the National Academy of Sciences, 102*(6), 2248–2253.

Huang, Y., Bu, Y., Ding, Y., & Lu, W. (2018). Number versus structure: Towards citing cascades. *Scientometrics, 117*(3), 2177–2193.

Huang, Y., Bu, Y., Ding, Y., & Lu, W. (2020). Exploring direct citations between citing publications. *ArXiv Preprint ArXiv:1811.01120*.

King, J. (1987). A review of bibliometric and other science indicators and their role in research evaluation. *Journal of Information Science, 13*(5), 261–276.

Lariviere, V., & Sugimoto, C. R. (2019). The journal impact factor: A brief history, critique, and discussion of adverse effects. *Springer handbook of science and technology indicators (pp. 3–24)*. Springer.

Lu, C., Bu, Y., Dong, X., Wang, J., Larivière, V., & Sugimoto, C. R. (2019). Analyzing linguistic complexity and scientific impact. *Journal of Informetrics, 13*.

Min, C., Chen, Q., Yan, E., Bu, Y., & Sun, J. (2020). Citation cascade and the evolution of topic relevance. *Journal of the Association for Information Science and Technology*. https://doi.org/10.1002/asi.24370.

Mingers, J., & Leydesdorff, L. (2015). A review of theory and practice in scientometrics. *European Journal of Operational Research, 246*(1), 1–19.

Moed, H., De Bruin, R., & Van Leeuwen, T. H. (1995). New bibliometric tools for the assessment of national research performance: Database description, overview of indicators and first applications. *Scientometrics, 33*(3), 381–422.

Mohapatra, D., Maiti, A., Bhatia, S., & Chakraborty, T. (2019). Go wide, go deep: Quantifying the impact of scientific papers through influence dispersion trees. *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 305–314.

Molas-Gallart, J. (2015). Research evaluation and the assessment of public value. *Arts and Humanities in Higher Education, 14*(1), 111–126.

Navarro, V., Muntaner, C., Borrell, C., Benach, J., Quiroga, Á., & Rodríguez-Sanz, M. (2006). Politics and health outcomes. *The Lancet, 368*(9540), 1033–1037.

Nijssen, D., Rousseau, R., & Van Hecke, P. (1998). The Lorenz curve: A graphical representation of evenness. *Coenoses*, 33–38.

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.

Petersen, A. M., Pan, R. K., Pammolli, F., & Fortunato, S. (2019). Methods to account for citation inflation in research evaluation. *Research Policy, 48*(7), 1855–1865.

Ratkiewicz, J., Conover, M. D., Meiss, M., Gonçalves, B., Flammini, A., & Menczer, F. M. (2011). Detecting and tracking political abuse in social media. *Fifth International AAAI Conference on Weblogs and Social Media*.

Reale, E., Avramov, D., Canhial, K., Donovan, C., Flecha, R., & Holm, P. (2018). A review of literature on evaluating the scientific, social and political impact of social sciences and humanities research. *Research Evaluation, 27*(4), 298–308.

Science Business Publishing. (2017). Why fund research? A guide to why EU-funded research and innovation matters. https://sciencebusiness.net/report/why-fund-research-guide-why-eu-funded-research-and-innovation-matters.

Sinatra, R., Deville, P., Szell, M., Wang, D., & Barabási, A.-. L. (2015). A century of physics. *Nature Physics, 11*(10), 791–796. https://doi.org/10.1038/nphys3494.

Sivertsen, G. (2014). Scholarly publication patterns in the social sciences and humanities and their coverage in Scopus and Web of Science. *Proceedings of the Science and Technology Indicators Conference*, 598–604.
Stephan, P. E. (1996). The economics of science. *Journal of Economic Literature, 34*(3), 1199–1235.
Tahamtan, I., Afshar, A. S., & Ahamdzadeh, K. (2016). Factors affecting number of citations: A comprehensive review of the literature. *Scientometrics, 107*(3), 1195–1225.
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.
Vale, R. D. (2015). Accelerating scientific publication in biology. *Proceedings of the National Academy of Sciences, 112*(44), 13439–13446. https://doi.org/10.1073/pnas.1511912112.
Van Raan, A. F., Visser, M. S., Van Leeuwen, T. N., & Van Wijk, E. (2003). Bibliometric analysis of psychotherapy research: Performance assessment and position in the journal landscape. *Psychotherapy Research, 13*(4), 511–528.
Vasiliauskaite, V., & Evans, T. S. (2018). Diversity from the topology of citation networks. *ArXiv Preprint ArXiv:1802.06015*.
Vinkler, P. (2010). *The evaluation of research by scientometric indicators*. Elsevier.
Walker, D., Xie, H., Yan, K.-. K., & Maslov, S. (2007). Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment, 2007*(06), P06010.
Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics, 10*(2), 365–391.
Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C., Tijssen, R. J., & van Eck, N. J., van Leeuwen, T. N., van Raan, A. F., Visser, M. S., & (2012). The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology, 63*(12), 2419–2432.
Waltman, L., & Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology, 63*(12), 2378–2392.
Waltman, L., & Yan, E. (2014). PageRank-related methods for analyzing citation networks. *Measuring scholarly impact (pp. 83–100)*. Springer.
Wang, D., Song, C., & Barabási, A.-. L. (2013). Quantifying long-term scientific impact. *Science (New York, N.Y.), 342*(6154), 127–132.
Wang, J., Veugelers, R., & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy, 46*(8), 1416–1436.
Wang, Y., Jones, B. F., & Wang, D. (2019). Early-career setback and future career impact. *Nature Communications, 10*(1), 1–10.
Weingart, P., & Schwechheimer, H. (2007). Conceptualizing and measuring excellence in the social sciences and humanities.
Wildgaard, L., Schneider, J. W., & Larsen, B. (2014). A review of the characteristics of 108 author-level bibliometric indicators. *Scientometrics, 101*(1), 125–158.
Zhang, Z., Cheng, Y., & Liu, N. C. (2014). Comparison of the effect of mean-based method and z-score for field normalization of citations at the level of Web of Science subject categories. *Scientometrics, 101*(3), 1679–1693.