Scientometrics (2020) 122:895–913
https://doi.org/10.1007/s11192-019-03314-6

The impact of a paper's new combinations and new components on its citation

Yan Yan[1] · **Shanwu Tian**[1] · **Jingjing Zhang**[2]

Received: 20 March 2019 / Published online: 4 December 2019

Abstract
A paper's novelty enhances its impact and citation. In this paper, we examine two dimensions of a paper's novelty: new combinations and new components. We define new combinations as new pairs of knowledge elements in a related research area, and new components as new knowledge elements that have never appeared in a related research area previously. The importance of both dimensions is stressed, and we analyze the mechanisms that affect the frequency of a paper's citation; we believe that a paper's new combinations and new components both have an inverted U-shaped effect on its citation count. Utilizing a text-mining approach, we develop a novel method for constructing new combinations and new components using a paper's keywords. Using keywords from papers published in the wind energy field between 2002 and 2015 as our sample, we conduct an empirical analysis on the above-mentioned relationships. To do so, we use the negative binomial regression method and several robustness tests. The results provide support for our hypotheses that propose a paper's new combinations and new components significantly affect its impact. Specifically, new combinations and new components lead to more citation counts up to a specific threshold. When the number of new combinations and new components exceed the threshold, the paper is likely to be cited less frequently. Finally, we discuss the theoretical contributions, methodological contributions, and practical implications of these findings.

Keywords Novelty · New combinations · New components · Citation

✉ Jingjing Zhang
zhangjingjing@ucas.ac.cn

Yan Yan
yanyan@rmbs.ruc.edu.cn

Shanwu Tian
tianshanwu@ruc.edu.cn

[1] School of Business, Renmin University of China, Beijing 100872, People's Republic of China

[2] School of Public Policy and Management, University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China

Introduction

How to improve a paper's citation is a topic that is currently receiving extensive attention in academic research (Guan et al. 2017; Lee and Brusilovsky 2019; Leydesdorff et al. 2018, 2019; Verhoeven et al. 2016). The frequency of citation varies greatly from one paper to another. Specifically, some papers are cited hundreds of times, while some papers receive no citation. Citation count is seen as an important research impact indicator because it measures the influence of an individual or a group of scientists on others (Bornmann and Daniel 2006). A highly cited paper illustrates the usefulness of the research for a relatively large number of researchers (Garfield 1979).

There are many motivations for researchers to cite a paper (Bornmann and Daniel 2008). Different frameworks, methods, and networks have been used to study the factors affecting a paper's citation frequency (Abbasi and Jaafari 2013; Bornmann et al. 2008; Guan et al. 2017; Tahamtan et al. 2016). The citation process is complex and is influenced by many factors. Tahamtan and Bornmann (2018a) have pointed to the core elements that determine why scholars cite a paper. These factors include document, author, and journal features (Tahamtan et al. 2016), such as abstract length (Letchford et al. 2016), funding (Boyack and Börner 2003), acknowledgements (Wang and Shapira 2011), number of authors (Batista et al. 2006), journal impact factor (Bensman 2008), and WoS category (Skilton 2006).

Some researchers have verified that scientific works with significant novelty can inspire subsequent works, which can have a far-reaching impact (Min et al. 2018). One line of research stresses that a paper's novelty is an antecedent to its citation frequency (Agovino et al. 2017). However, there are two key challenges in assessing the effect of scientific novelty on a paper's citation. First, there is no consensus definition of scientific novelty, which makes the concept hard to quantify. Second, the mechanism between a paper's scientific novelty and the number of times it is cited is unclear.

Regarding the first challenge, prior studies on scientific novelty are relatively fragmented, and most of them consider the concept of novelty as a single dimension. There are some preconditions for novelty, such as the generation of ideas (Utterback 1971; Yan et al. 2019) and the need to establish original pathways instead of expanding existing pathways (Arthur 2009). Moreover, many researchers believe that a major source of novelty is generated by either a new recombination of previously uncombined elements or the combination of an established element with a new concept (Fleming 2001; Lee et al. 2015; Schumpeter 1934; Strumsky and Lobo 2015; Zhang et al. 2019). Therefore, we believe that the source of novelty is often reflected in the new combination of existing knowledge and generation of new ideas. These new combinations can result from sharing knowledge with colleagues, thorough interdisciplinary exchange, and solving practical problems (Strumsky and Lobo 2015; Uzzi et al. 2013; Tahamtan and Bornmann 2018b).

Meanwhile, Verhoeven et al. (2016) described the characteristics of innovation in terms of two dimensions of novelty: novelty in recombination of components and principles as well as novelty in knowledge origins. From this perspective, new combinations and new components are two significant dimensions of scientific novelty. For this study, we defined new combinations as new pairs of knowledge elements[1] in a related research area, and new

[1] A knowledge element refers to a socially defined category, including a set of tentative findings by the scientific or technological research community about facts, theories, operations, and procedures surrounding a subject (Wang et al. 2014). In this paper, we use keywords provided in scientific papers to indicate knowledge elements. Publication keywords are considered essential elements of identifying the primary focus of research and are often used to reveal knowledge structures in bibliometric research (Su and Lee

components as new knowledge elements that have never appeared in a related research area previously. Finding subdimensions of scientific novelty is both important and necessary. These subdimensions allow us to unpack, evaluate, and measure scientific novelty, which can help us better understand what novelty is and what drives it.

New combinations can be represented by the recombination of various final (or use) and technical features (Guan and Yan 2016; Gallouj and Weinstein 1997; Saviotti and Metcalfe 1984). The final features are the characteristics of the product seen from the perspective of the end user. The technical features describe the internal characteristics of the technology. Using the wind turbine as an example, the final features include the horizontal or vertical axis of the wind turbine. The technical features include related technologies on rotor, brake, energy storage and transfer, energy transformers, and charger controllers. In the wind turbine example, existing technology has been applied to new concepts, creating a new component. This new component generates new discoveries and introduces knowledge from another discipline (Guetzkow et al. 2004).

Regarding the second challenge, the relationship between a paper's scientific novelty and its citation count is complicated. In exploring the mechanism through which scientific novelty affects a paper's citation, one aim of this study is to reconcile two different views about that novelty. On one hand, some novelty is conducive to a paper's impact. The novelty and uniqueness of a paper indicate a significant combination of new and old ideas, or the generation of new ideas. High novelty has an increased probability of making a remarkable impact on other papers (Criscuolo et al. 2017). On the other hand, novelty does not always appear to positively correlate with the frequency of citation (Chandonia and Brenner 2006). A study that is too novel may not be cited frequently because it could be difficult to interpret or categorize. The resistance from incumbent scientific paradigms is also a significant concern (Wang et al. 2017).

In trying to balance these two viewpoints, we proposed that an intermediate level of scientific novelty, both in terms of new combinations and new components, is likely to have a more significant impact, reflected as an inverted U-shaped effect on citation counts. That is, citation counts first increase with new combinations and new components at a decreasing rate to reach a maximum, after which citation counts decrease at an increasing rate (Haans et al. 2016). We speculated that scientific novelty has both positive and negative effects on citation counts. When the negative effect overwhelms the positive effect, the scientific novelty is negatively correlated with citation counts, which expands the previous linear studies of scientific novelty.

We proposed hypotheses that both a paper's new combinations and new components have inverted U-shaped effects on its citation count. To test our hypotheses, we used WoS and JCR databases to gather data on papers in the field of wind energy research published between 2002 and 2015. The wind energy industry is ideal for this research because this industry produces many papers that offer international, novel information (Guan et al. 2017), and this field is attracting more scholarly attention. Our research focused on the

Footnote 1 (continued)

2010; Zhang et al. 2015). Meanwhile, keywords are generally treated as the main method by which papers on related topics are identified and retrieved (McCain 1989). Therefore, we utilized a text mining approach to analyze knowledge structures and measured the new combinations and new components by using the keywords of a paper, which is different from prior studies using reference or citation approaches to measure scientific novelty (Uzzi et al. 2013). This approach is consistent with the viewpoint that innovation is an evolutionary search process which involves knowledge combination and generation (Lee et al. 2015).

paper level. We applied a negative binomial regression model with robust standard errors to test the above-mentioned relationships. The results of our empirical analysis provide support for our hypotheses.

Theory and hypotheses

New combinations and citation counts

Drawing on the combinatorial perspective of the scientific process, novelty may originate from the combination of knowledge elements (i.e., keywords) in a new fashion. Previous studies have provided similar arguments. For example, Nelson and Winter (1982, p. 130) claimed that "the creation of any sort of novelty in art, science, or practical life—consists to a substantial extent of a recombination of conceptual and physical materials that were previously in existence." Verhoeven et al. (2016) suggested that if the components of an invention differ from previous ones, they are considered to have novelty in terms of technical recombination. Following similar logic, new combinations in this study refer to new pairs of knowledge elements in a related research area. Based on the combinatorial perspective, new combinations in a paper might influence its citation count.

The new combinations, generated from combinations of existing scientific knowledge, affect a paper's citation count in at least two ways. First, from the perspective of recombination, scientific output pushes frontiers when it explores new ideas or pushes boundaries of knowledge (Phene et al. 2006). Breaking down barriers between different fields of research and then recombining this information is more likely to achieve higher research value (Ahuja and Lampert 2001). Thus, the research value of frontier-pushing science affects a paper's citation count (Criscuolo et al. 2017). Atypical combinations of knowledge can bring innovation while maintaining the advantages of traditional knowledge (Tahamtan and Bornmann 2018b; Uzzi et al. 2013).

Second, there is a need to expand scientific interaction among different fields to avoid siloing, which can limit innovation (Sternberg 1999). When a paper has a greater number of new combinations, it spans a larger number of knowledge fields. A paper that relies on literature (or references) from a variety of disciplines, is more likely to be accessed by readers from different disciplines. Due to new combinations of different references (or disciplines), such papers are more likely to be read by more scholars and consequently more likely to be cited in subsequent papers by scholars in those disciplines.

By contrast, when many new combinations are used in a paper, the paper is less likely to be cited. One possible reason is that new combinations dealing with very broad ideas may be too difficult to interpret (Sternberg 1999). By offering too wide of a perspective, authors may not have sufficient insight into foundational knowledge or understand the potential weaknesses of their research (Kaplan and Vakili 2015). Papers might also be cited less frequently because reviewers may fail to grasp the paper's main idea. A clear understanding from reviewers is crucial because peer review is generally regarded as an essential step in the path to publication (Mahoney 1977). The review process is typically organized within the most relevant discipline. Since new combinations involve multiple knowledge elements being combined in novel ways, reviewers might fail to recognize the full value of the research, which can hinder the work's publication process. Publication may be delayed by several years or the paper may be published in a lesser known journal, both of which can decrease its citation count.

Considering these insights, we expected that the relationship between new combinations in a paper and its citation count would be an inverted U-shaped relationship. Therefore, we propose:

Hypothesis 1 Apaper's new combinations have an inverted U-shaped effect on its citation count.

New components and citation counts

In this paper, new components refer to new knowledge elements (i.e., keywords) that have never appeared in a related research area previously. When a new component or idea is first studied, the research always involves new discoveries, experiments, methods, and theories (Dirk 1999; Guetzkow et al. 2004). When a paper explores new information components, new concepts and conclusions are offered through the research. Thus, a paper's new components and its novelty are highly associated.

New components offer several advantages to a paper's citation count in at least two ways. First, a paper with many new components indicates an exploratory R&D project. However, if it succeeds, it may have a profound impact, generating substantial new knowledge and winning broad acclaim (Foster et al. 2015). When the new components of a paper increase from a low to moderate level, the number of original knowledge elements in the paper increases (Verhoeven et al. 2016). More scholars cite such papers because they represent the novelty of the academic frontier. Some ideas that open a new field of scientific research may have dramatically more impact than others (Schilling and Green 2011). Innovation is integrated within the science network, which stems from the generation of new knowledge elements (Schumpeter 1934; Guan et al. 2015a). Second, scientific credit is awarded to researchers who publish first, which means subsequent related work is likely to cite their work or extrapolate from it (Wang et al. 2014). The creation of new knowledge, therefore, is more likely to receive greater attention. Thus, new components likely have a positive impact on a paper's citation count.

However, there are limits to these positive effects, which are mainly caused by ambiguity issues. Pursuing innovation is a gamble, and on average, there is not enough return to prove that the risk is reasonable. When adopting a risk-taking strategy in scientific activities, scientists fail more frequently and might be unproductive or have a lesser impact for longer periods (Foster et al. 2015). Because of the increase of a paper's new components, the paper may be cited less frequently. If a paper has too many original elements, it can be difficult for peers to understand (Dirk 1999). It has also been suggested that specific textual features (e.g., low-frequency words) can cause difficulties in understanding and impose a high cognitive burden (Lenzner et al. 2010). Jansen and Pooch (2001) claimed that more than half of researchers retrieve papers by querying keywords in online public access catalogue systems. Since people are querying keywords with which they are familiar, too many original keywords in papers will make it difficult for researchers to locate these papers if they are outside of specialized fields. This likely results in reduced citation counts in such circumstances.

Given this context, we expected that the relationship between new components in a paper and its citation count would be an inverted U-shaped relationship. Therefore, we propose:

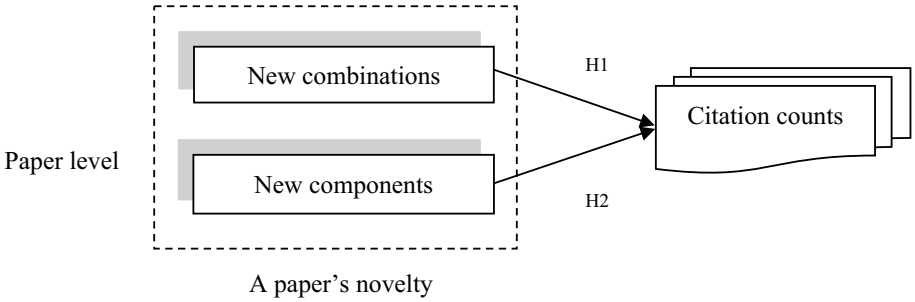

Fig. 1 Research framework

Hypothesis 2 Apaper's new components have an inverted U-shaped effect on its citation count.

Based on the above analysis, we propose the research framework shown in Fig. 1.

Data and methodology

Data and context

We used SCI and SSCI in WoS Core Collection to search papers. We retrieved wind energy paper records with "wind power*," "wind turbine*," "wind energy*," "wind farm*," "wind generation*," "wind system*" in the title, abstract, or keywords in WoS, and extracted the list of journals in which the papers were published. We downloaded these journals' information from JCR for each year from 2002 to 2015, and then integrated the impact factors (5 years) of all journals annually in our dataset (Guan et al. 2017; Guan et al. 2015b; Sanz-Casado et al. 2013). We used these data to study the worldwide scientific activity of wind energy research and investigate the new combinations and new components in scientific papers. The dataset includes papers' information, such as the type and language of the paper, author keywords, keywords plus, published journals, and forward and backward citations. Finally, we retrieved 15,224 records of publications, which were chosen as the data set for all steps of data processing.

Measure

Dependent variable

Citations A paper's citation count was used to evaluate the impact of that paper, given that papers with higher citation counts are often considered to be of higher quality and scientific influence than others (Tahamtan et al. 2016; Zhang and Guan 2016). In this study, we calculated the dependent variable by using the normalized citation count of each paper. Consistent with previous research (Cannella and McFadyen 2016; Guan et al. 2017), the citation count of each paper was standardized using the mean and standard deviation of citation counts of all papers published in the same year. This approach eliminates positive

citation bias found in previous studies (Cannella and McFadyen 2016; Guan et al. 2017). If a paper i is published in the year t, its normalized citations calculation method is:

$$\text{Normalized citations}_{it} = \frac{\text{Citations}_i - \text{Mean citations}_{\text{all papers}_t}}{\text{SD citations}_{\text{all papers}_t}}$$

where $\text{Mean citations}_{\text{all papers}_t}$ indicates the mean value of all papers' citations in the year t, and $\text{SD citations}_{\text{all papers}_t}$ indicates the standard deviation of all papers' citations in the year t.

Independent variables

This research has two objectives: (1) to propose a novel method to measure new combinations and new components, and (2) to test the inverted U-shaped relationship between new combinations and a paper's citation count as well as the inverted U-shaped relationship between new components and a paper's citation count. Therefore, the core independent variables in this study are new combinations and new components. We adopted a text-mining approach to handle keywords first. The analysis was conducted in R software and utilized several text-mining packages such as "tm" and "SnowballC." The package "tm" is a framework for text mining applications within R. "SnowballC" is an R interface to the C "libstemmer" library for collapsing words to a common root to aid the comparison of vocabulary. As a preprocessing step, we removed all non-alphabetic characters, punctuation, white space, and stop words. Then, all keywords were converted into lowercase and cleaned or stemmed, meaning that singular and plural forms of the keyword or keywords with the same root were counted together. After the preprocessing steps, we calculated new combinations and new components as follows.

New combinations New combinations in a paper are measured by new keywords pairs in related research area in the past 5 years. This is an indicator created by using a paper's keywords. A paper i's new combinations will be calculated as follows:

$$\text{new combinations}_i = \frac{\text{new combination pairs of keywords}_i}{\text{all potential combination pairs of keywords}_i}$$

Expressed more generally, the formula of new combinations is:

$$\text{new combinations}_i = \frac{\sum_{j=1}^{N_i} x_j}{C_{N_i}^2}$$

where x_j is a binomial variable. If the combination of keyword pairs has never appeared in papers in the past 5 years in the wind energy field, we define x_j as 1, otherwise x_j is 0. $C_{N_i}^2$ refers to all potential combination pairs of keywords in the paper. We defined the new combinations of a paper as the ratio of new combination pairs of keywords to all potential combination pairs of keywords. The pair of a previous element and a new element is considered to be a new combination. For example, as shown in Table 1, there are six keywords in Paper 1, which was published in 2010. The number of all potential combination pairs of keywords is 15. We found two pairs, keywords 1 and 4 and keywords 4 and 5, that have never appeared in wind energy field papers in the past 5 years (2004–2009), and thus there are two new combination pairs of keywords in Paper 1. According to the calculation formula, the new combinations in Paper 1 is $2/15 = 0.133$.

Table 1 Illustration example for new combinations

Papers	Keywords	(Examples) combination pairs	Compare with before	Present before?	New combinations
Paper 1 (published in 2010)	Keyword 1	1–2	Comparing with 2004–2009	Yes	$2/C_6^2 = 0.133$
	Keyword 2	1–3		Yes	
	Keyword 3	1–4		No	
	Keyword 4	…		…(all yes)	
	Keyword 5	5–4		No	
	Keyword 6	5–6		Yes	
Paper 2 (published in 2007)	Keyword 7	7–8	Comparing with 2002–2006	No	$1/C_5^2 = 0.100$
	Keyword 8	7–9		Yes	
	Keyword 9	7–10		…(all yes)	
	Keyword 10	…		Yes	
	Keyword 11	10–11		Yes	

Table 2 Illustration example for new components

Papers	Keywords	Compare with before	Present before?	New components
Paper 3 (published in 2007)	Keyword 1	Comparing with 2002–2006	Yes	$^2/_5 = 0.400$
	Keyword 2		Yes	
	Keyword 3		No	
	Keyword 4		Yes	
	Keyword 5		No	
Paper 4 (published in 2008)	Keyword 6	Comparing with 2003–2007	Yes	$^1/_7 = 0.143$
	Keyword 7		Yes	
	Keyword 8		Yes	
	Keyword 9		Yes	
	Keyword 10		Yes	
	Keyword 11		No	
	Keyword 12		Yes	

New components New components in a paper are measured by keywords that have never appeared in a related research area in the past 5 years. That is, if a paper i was published in the year t, its new components will be calculated as follows:

$$\text{new components}_i = \frac{\text{the number of new appearing keywords}_i}{\text{all keywords}_i}$$

Expressed more generally, the formula of new components is as follows:

$$\text{new components}_i = \frac{\sum_{j=1}^{N_i} y_j}{N_i}$$

where y_j is a binomial variable. If the keyword j has never appeared in papers in the wind energy field in the past 5 years, we define y_j as 1; otherwise, y_j is 0. This is compared to all keywords in the dataset in the past 5 years. N_i refers to the number of all keywords in the paper. This definition is convenient for standardizing results, making the result of new components between 0 and 1, where 0 means no new keywords and 1 means all new keywords. We defined the new components in a paper as the ratio of new appearing keywords to all keywords in the paper. For example, as shown in Table 2, there are five keywords in Paper 3, which was published in 2007, with two keywords that have never appeared in papers within the wind energy field in the past 5 years (2002–2006). In this case, the number of new components in this paper is 2/5 = 0.400.

As another example, the keywords of Lutz et al. (2017) include "trailing-edge noise," "turbulent flows," "prediction," and "simulation." The new components of this paper are "trailing-edge noise," and thus the ratio of new components is 0.25. The combination of "prediction" and "simulation" has appeared in the previous 5 years, whereas the other five combinations have not, which means the ratio of new combinations is 5/6 = 0.833.

Control variables

Many factors have been shown to influence publication citations. In the process of empirically testing our two hypotheses, we controlled eight variables: the number of authors,

journal impact factor (journal IF), abstract length, WoS categories, funding, acknowledgements, title, and affiliation. Every publication in WoS is assigned to at least one subject category. In this paper, WoS categories indicate the number of categories under which the paper was published. Thus, the WoS categories variable is a continuous variable, as shown in Table 3.

Model

Given that our dependent variable is a count variable with overdispersion, we use the negative binomial regression method to examine the relationship between new combinations/new components and citation counts. Negative binomial regression is similar to conventional multiple regression except that the dependent (Y) variable is an observed count that follows a negative binomial distribution. A negative binomial (NB) regression model is suitable for asymmetric data sets because the number of citations is excessively discrete (Ajiferuke and Famoye 2015). In negative binomial regression, the exposure time t and a set of k regressor variables (the x's) determine the mean of y. The expression relating these quantities is

$$\mu_i = \exp(\ln(t_i) + \beta_i x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki})$$

According to the above principle, we tested the quantitative relationship between new combinations/new components and citation counts. In addition, this research conducted a sensitivity analysis using the Ordinary Least Squares (OLS) method to assess the robustness of our findings.

Regression analysis

The means, standard deviations, and correlations of key variables are listed in Table 4. The binary correlation values are quite small, indicating that our analysis is not affected by multicollinearity problems. In addition, we can see in Table 4 that the VIF values of all variables are less than 5, which indicates that there is no obvious multicollinearity problem in the data.

Because citation counts are excessively dispersed, the negative binomial regression model is suitable for asymmetric data sets (Ajiferuke and Famoye 2015). We used the stepwise regressions approach, entered all control variables in Model 1, entered independent variables separately in Models 2 and 3, and entered all variables in Model 4 to test the two hypotheses. Table 5 shows the results. We used r to indicate the coefficients below. As shown in Model 2, the coefficients of new combinations and new combinations squared are stably significant, while the coefficient of new combinations is positive ($r=1.3427$, $p<0.001$), and the coefficient of new combinations squared is negative ($r=-0.9455$, $p<0.01$). These results indicate that Hypothesis 1, which states that new combinations in a paper have an inverted U-shaped effect on its citation count, is fully supported. As shown in Model 3, the coefficients of new components and new components squared are stably significant, while the coefficient of new components is positive ($r=0.7843$, $p<0.01$), and the coefficient of new components squared is negative ($r=-1.1067$, $p<0.001$). These results indicate that Hypothesis 2, which states that new components in a paper have an inverted U-shaped effect on its citation count, is fully supported. The p value of ΔR^2 between Model 2 and Model 4 is 0.001 ($p<0.01$). Thus, adding the new components in the

Table 3 Definition and source of control variables

Control variables	Definition	Theoretical source
The number of authors	The number of authors of the paper	More authors might have more opportunities to be cited (Batista et al. 2006)
Journal IF	Journal impact factor (5 years)	Journals are an important factor in citation counts (Bensman 2008)
Abstract length	The number of sentences in the abstract	The characteristics of the abstract influence a paper's citations (Letchford et al. 2016)
WoS categories	The number of Web of Science categories under which the paper was published	Research areas or subject influence a paper's citations (Skilton 2006)
Funding	The account of funding records	Financial support impacts an article's quality (Boyack and Börner 2003)
Acknowledgements	The number of sentences in the acknowledgement	Peer reports influence an article's quality (Wang and Shapira 2011)
Title	The number of characters in the title	Titles and citations are positively correlated (Jamali and Nikzad 2011)
Affiliation	The number of authors' affiliations	Authors' affiliations are a powerful predictor of citations (Walters 2006)

Table 4 Means, standard deviations, and correlations

	1	2	3	4	5	6	7	8	9	10	11
1. Normalized citation	1										
2. The number of authors	0.09***	1									
3. Journal IF	0.39***	0.11***	1								
4. Abstract length	0.02	0.01	0.04*	1							
5. WoS categories	0.03^{+}	−0.06***	0.02	0.01	1						
6. Funding	0.01	0.05**	0.02	−0.04*	0.02	1					
7. Acknowledgements	−0.02	0.13***	0.04*	0.09***	0.02	−0.003	1				
8. Title	−0.02^{+}	0.06***	−0.02	0.02	0.06***	−0.01	0.03*	1			
9. Affiliation	0.07***	0.57***	0.11***	0.03*	−0.01	0.03	0.15***	0.03*	1		
10. New combinations	−0.01	0.03	−0.01	0.06***	0.02	−0.03^{+}	0.02	0.03^{+}	0.02	1	
11. New components	−0.05***	0.02	−0.04**	0.06***	0.04*	−0.03^{+}	−0.01	0.02	0.02	0.57***	1
M	0	3.02	2.11	8.12	1.76	0	0	86.32	1.78	0.78	0.35
SD	1	1.72	1.83	4.15	0.83	1	1	26.80	1.06	0.29	0.26
VIF	–	1.53	1.02	1.02	1.01	1.01	1.04	1.01	1.52	1.49	1.50

$^{+}p < 0.1$, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$

Table 5 Negative binomial regression results (with robust standard errors)

	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4
The number of authors	0.0279	0.0282	0.0283	0.0282
	(0.0189)	(0.0189)	(0.0187)	(0.0188)
Journal IF	0.3778***	0.3712***	0.3712***	0.3685***
	(0.0293)	(0.0299)	(0.0297)	(0.0300)
Abstract length	−0.0008	0.0001	0.0001	0.0004
	(0.0054)	(0.0055)	(0.0055)	(0.0055)
WoS categories	0.0620*	0.0617*	0.0643*	0.0639*
	(0.0283)	(0.0283)	(0.0281)	(0.0282)
Funding	−0.0112	−0.0108	−0.0112	−0.0107
	(0.0170)	(0.0171)	(0.0171)	(0.0171)
Acknowledgements	−0.0469*	−0.0497**	−0.0514**	−0.0526**
	(0.0186)	(0.0187)	(0.0185)	(0.0185)
Title	−0.0016	−0.0015	−0.0017*	−0.0016
	(0.0008)	(0.0008)	(0.0008)	(0.0008)
Affiliation	0.0546*	0.0580*	0.0569*	0.0588*
	(0.0262)	(0.0260)	(0.0260)	(0.0260)
New combinations		1.3427***		0.8435*
		(0.3547)		(0.3805)
New combinations squared		−0.9455**		−0.5406^{+}
		(0.2928)		(0.3275)
New components			0.7843**	0.5598
			(0.2596)	(0.3100)
New components squared			−1.1067***	−0.8655**
			(0.2932)	(0.3221)
Constant	2.3040***	1.8754***	2.2363***	1.9635***
	(0.1303)	(0.1607)	(0.1412)	(0.1639)
N	3407	3407	3407	3407
p	0.0000	0.0000	0.0000	0.0000

Standard errors in parentheses

$^{+}p<0.1$, $*p<0.05$, $**p<0.01$, $***p<0.001$

papers significantly improves the model. Similarly, the p value of ΔR^2 between Model 3 and Model 4 is 0.001 ($p<0.01$). Thus, adding the new combinations in the papers improves the model.

Based on the regression results in Table 5, we set all other variables to their average values and then showed the inverted U-shaped relationships between new combinations/new components and normalized citation counts in Figs. 2 and 3. The plots show that as the level of new combinations/new components increases from low to moderate, normalized citation counts increase. As the level of new combinations/new components increases from moderate to high, normalized citation counts decrease. The symmetry axis apex of new combinations is 0.7100; when new combinations are < 0.7100, new combinations are positively associated with citation counts. When new combinations are > 0.7100, new combinations are negatively associated with citation counts. The symmetry axis apex of new

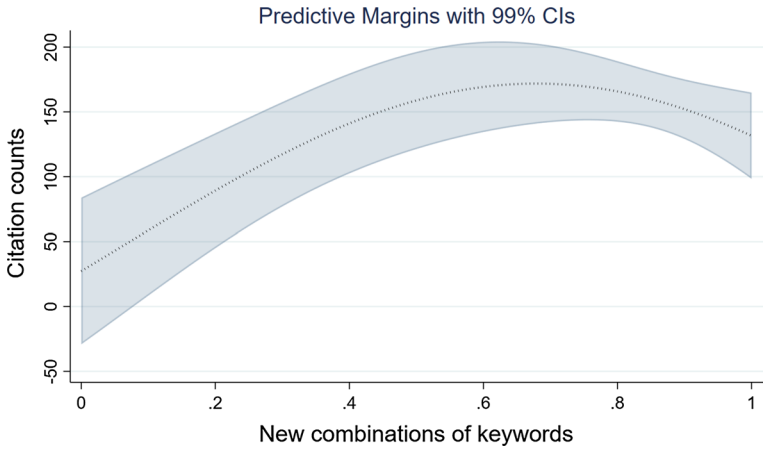

Fig. 2 The relationship between new combinations and citation counts

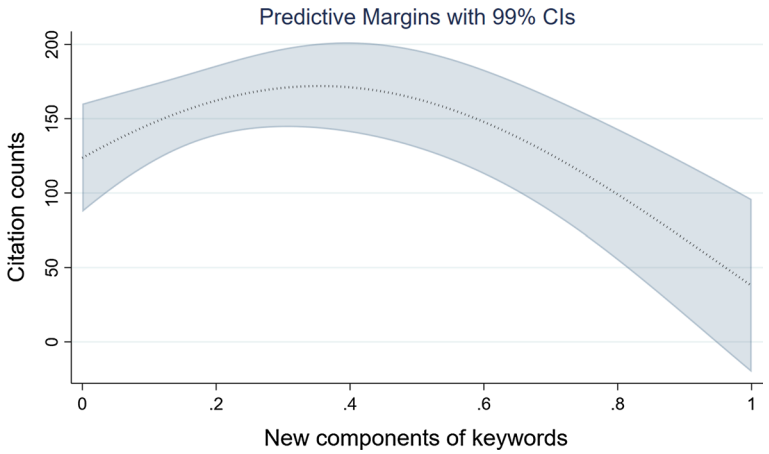

Fig. 3 The relationship between new components and citation counts

components is 0.3543; when new components are < 0.3543, new components and citation counts are positively associated. When new components are > 0.3543, new components and citation counts are negatively associated.

Using the negative binomial regression results, we then determined the practical effects of new combinations and new components. New combinations have an inverted U-shaped effect on a paper's citation count, and thus the practical effects depend on the value of new combinations. Assuming that new combinations take its mean, a standard deviation increase of new combinations results in a 0.68 decrease in citation counts. Compared with the mean of citation counts, new combinations reduce citation counts by 2.38%. Similarly, assuming that new components take its mean, a standard deviation increase of new components results in a 0.55 decrease in citation counts. Compared with the mean of citation counts, new components reduce citation counts by 1.94%.

The distribution of typical citations is characterized by a tendency to be highly log-normal and skewed (Guan et al. 2017). In order to make the conclusion more stable, we

Table 6 Robustness checks (with robust standard errors)

	Normalized citation (OLS)			Ln (citation) (OLS)		
	(1)	(2)	(3)	(4)	(5)	(6)
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
The number of authors	0.0204	0.0206	0.0209	0.0055	0.0066	0.0068
	(0.0143)	(0.0144)	(0.0144)	(0.0160)	(0.0159)	(0.0159)
Journal IF	0.2163***	0.2183***	0.2166***	0.2567***	0.2598***	0.2572***
	(0.0349)	(0.0350)	(0.0351)	(0.0352)	(0.0341)	(0.0340)
Abstract length	0.0016	0.0014	0.0016	0.0077	0.0077	0.0081
	(0.0042)	(0.0042)	(0.0042)	(0.0047)	(0.0047)	(0.0048)
WoS categories	0.0201	0.0212	0.0202	0.0949***	0.0985***	0.0971***
	(0.0222)	(0.0220)	(0.0221)	(0.0226)	(0.0224)	(0.0224)
Funding	0.0195	0.0191	0.0191	−0.0083	−0.0101	−0.0099
	(0.0191)	(0.0191)	(0.0191)	(0.0158)	(0.0160)	(0.0159)
Acknowledgements	0.0063	0.0052	0.0050	−0.0162	−0.0198	−0.0203
	(0.0193)	(0.0192)	(0.0192)	(0.0174)	(0.0173)	(0.0173)
Title	−0.0008	−0.0009	−0.0009	−0.0003	−0.0004	−0.0004
	(0.0005)	(0.0005)	(0.0005)	(0.0007)	(0.0007)	(0.0007)
Affiliation	0.0263	0.0232	0.0248	0.0318	0.0252	0.0279
	(0.0244)	(0.0244)	(0.0243)	(0.0239)	(0.0238)	(0.0237)
New combinations	0.9156***		0.6987***	1.8477***		1.1229**
	(0.2113)		(0.2091)	(0.3282)		(0.3458)
New combinations squared	−0.7539***		−0.5922**	−1.5856***		−0.9128**
	(0.1945)		(0.1953)	(0.2635)		(0.2870)
New components		0.2995	0.3041		0.5686*	0.5197^{+}
		(0.1737)	(0.1959)		(0.2227)	(0.2676)
New components squared		−0.4983*	−0.4127*		−1.1779***	−1.0062***
		(0.1963)	(0.2107)		(0.2556)	(0.2805)
Constant	−0.7350***	−0.5324***	−0.7005***	1.5536***	1.9584***	1.6746***
	(0.1016)	(0.1108)	(0.1035)	(0.1417)	(0.1194)	(0.1450)
N	3407	3407	3407	3407	3407	3407
r2	0.1582	0.1579	0.1592	0.1693	0.1743	0.1767
F	12.7888	14.0800	13.7143	17.7432	20.9938	19.2824
p	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Standard errors in parentheses

$^{+}p<0.1$, $*p<0.05$, $**p<0.01$, $***p<0.001$

conducted some robustness checks, as shown in Table 6. First, we replaced the dependent variable (normalized citation counts) with the original number of citation counts. Second, we used the natural logarithm and then the OLS model, which has robust standard errors for regression (Thelwall and Wilson 2014). The results of multiple sensitivity analyses also support our findings.

Conclusion and discussion

Some relevant research found mixed relationships between scientific novelty and a paper's citation count (Leydesdorff et al. 2019; Verhoeven et al. 2016). We examined two dimensions of a paper's novelty and studied how new combinations and new components are related to a paper's citation count. We reconciled previous views by testing inverted U-shaped relationships between new combinations/new components and citation counts, respectively. New combinations lead to more citations up to a specific threshold, after which citation counts decrease rapidly. We believe this is due to additional difficulty for researchers to interpret and understand the paper's information. New components show the same relationship.

Citation counts are influenced by many factors besides scientific quality (Bornmann and Daniel 2006). We defined a paper's novelty in two dimensions—new combinations and new components—which affect the number of citations it receives. This study began with the concepts of new combinations and new components and defined formulas to measure these two constructs using a paper's keywords (Guan et al. 2017; Su and Lee 2010; McCain 1989; Zhang et al. 2015). Furthermore, we used data on wind energy papers from the WoS and JCR databases (2002—2015) to examine the relationships between new combinations/new components and a paper's citation count. The main finding of this research is that both the new combinations and new components in a paper have an inverted U-shaped effect on its citation count. When operating within a certain degree, higher new combinations and new components in a paper will cause more researchers to cite the paper, making the citation count relatively high. However, the new combinations/new components have a negative effect on citation counts beyond a certain degree, which was shown as a decrease in citation counts when new combinations/new components increase.

This study presents several theoretical contributions. First, we defined a paper's novelty as the sum of new combinations and new components. Scientific novelty involves the generation of new information and combinations among existing information (Schilling and Green 2011). Thus, in this study, we highlighted the two-dimensional nature of scientific novelty: new combinations and new components. The innovation process cannot be fully understood if we take a disjointed viewpoint and exclusively focus on a single dimension of scientific novelty. This is a step forward based on the research of Verhoeven et al. (2016), which described the characteristics of innovation in terms of two dimensions of novelty: novelty in recombination of components and principles as well as novelty in knowledge origins. Second, a large body of research has investigated the potential driving factors of a paper's citations. We advanced this work by stressing the importance of new knowledge combinations and new knowledge components in citation counts due to novelty. We complemented and advanced the research on the antecedents of paper citations by proposing two potential but overlooked influential factors: new combinations and new components of knowledge elements (i.e., keywords). Finally, to the best of our knowledge, there is no empirical research analyzing the impact of new combinations and new components on a paper's citations. Therefore, there is no clear indication of the quantitative relationship between them.

This study presents several methodological contributions. This research takes a step forward in measuring a paper's novelty. Utilizing a text-mining approach, we measured new combinations and new components using a paper's keywords at the paper level and used this measure as a proxy of the novelty of the paper (Guan et al. 2017; Su and Lee 2010; McCain 1989; Zhang et al. 2015). We measured new combinations in a paper as the

percentage of new combinations of knowledge elements and new components in a paper as the new knowledge elements. Our approach is similar to some scholars' viewpoints that the recurrence of knowledge elements can indicate the recombination process of knowledge (Guan et al. 2017; Verhoeven et al. 2016; Carayol et al. 2019). At the same time, we further improved the indicator construction of scientific novelty of prior studies that used reference or citation-based indicators (Uzzi et al. 2013). Focusing on the knowledge elements of a paper can help us capture the fundamental information it is conveying and, therefore better understand its novelty.

This study has several practical implications. The findings suggest that a paper's citation count is affected by its new combinations and new components of keywords. Researchers are better equipped to understand interdisciplinary research if they can learn from both their own disciplines and other relevant disciplines. Furthermore, researchers should look for new ways to recombine knowledge elements, but avoid indulging in excessive new recombination, which can make it difficult for other scholars to interpret and understand the technical and practical value of their research.

This study also has limitations. In the process of cleaning the keywords, we used the SCI2 to stem and lemmatize the keywords, which is the best method available; however, it did not perfectly clean the keywords. Future work can develop new ways to stem and lemmatize a paper's keywords. Additionally, the citing process is more complex than what we have discussed in this paper. As Tahamtan and Bornmann (2018a) suggested, future studies can examine how other features such as the context of the cited and citing documents or processes from selection to citation, affect the citing process.

Acknowledgements This study is supported by National Natural Science Foundation of China (Grant No. 71904191), and by University of Chinese Academy of Sciences (Grant No. Y95402JXX2). This study is supported by the joint PhD programme scholarship from Business School, Renmin University of China. The authors are very grateful for the valuable comments and suggestions from Prof. Editor Wolfgang Glänzel and two anonymous reviewers.

References

Abbasi, A., & Jaafari, A. (2013). Research impact and scholars' geographical diversity. *Journal of Informetrics, 7*(3), 683–692.

Agovino, M., Aldieri, L., Garofalo, A., & Vinci, C. P. (2017). Quality and quantity in the innovation process of firms: A statistical approach. *Quality & Quantity, 51*(4), 1579–1591.

Ahuja, G., & Lampert, C. M. (2001). Entrepreneurship in the large corporation: A longitudinal study of how established firms create breakthrough inventions. *Strategic Management Journal, 22*(6/7), 521–543.

Ajiferuke, I., & Famoye, F. (2015). Modelling count response variables in informetric studies: Comparison among count, linear, and lognormal regression models. *Journal of Informetrics, 9*(3), 499–513.

Arthur, W. B. (2009). *The nature of technology: What it is and how it evolves*. New York: Free Press.

Batista, P. D., Campiteli, M. G., & Kinouchi, O. (2006). Is it possible to compare researchers with different scientific interests? *Scientometrics, 68*(1), 179–189.

Bensman, S. J. (2008). Distributional differences of the impact factor in the sciences versus the social sciences: An analysis of the probabilistic structure of the 2005 journal citation reports. *Journal of the American Society for Information Science and Technology, 59*(9), 1366–1382.

Bornmann, L., & Daniel, H. (2006). Selecting scientific excellence through committee peer review—a citation analysis of publications previously published to approval or rejection of post-doctoral research fellowship applicants. *Scientometrics, 68*(3), 427–440.

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

Bornmann, L., Mutz, R., Neuhaus, C., & Daniel, H. D. (2008). Citation counts for research evaluation: Standards of good practice for analyzing bibliometric data and presenting and interpreting results. *Ethics in Science and Environmental Politics, 8*(1), 93–102.

Boyack, K. W., & Börner, K. (2003). Indicator-assisted evaluation and funding of research: Visualizing the influence of grants on the number and citation counts of research papers. *Journal of the American Society for Information Science and Technology, 54*(5), 447–461.
Cannella, A. A., & McFadyen, M. A. (2016). Changing the exchange: The dynamics of knowledge worker ego networks. *Journal of Management, 42*(4), 1005–1029.
Carayol, N., Lahatte, A., & Llopis, O. (2019). *The right job and the job right: Novelty, impact and journal stratification in science*. *Cahiers du GREThA*, n°2019-05.
Chandonia, J., & Brenner, S. E. (2006). The impact of structural genomics: Expectations and outcomes. *Science, 311*(5759), 347–351.
Criscuolo, P., Dahlander, L., Grohsjean, T., & Salter, A. (2017). Evaluating novelty: The role of panels in the selection of R&D projects. *Academy of Management Journal, 60*(2), 2014–2861.
Dirk, L. (1999). A measure of originality: The elements of science. *Social Studies of Science, 29*(5), 765–776.
Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science, 47*(1), 117–132.
Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and innovation in scientists' research strategies. *American Sociological Review, 80*(5), 875–908.
Gallouj, F., & Weinstein, O. (1997). Innovation in services. *Research Policy, 26*(4–5), 537–556.
Garfield, E. (1979). Is citation analysis a legitimate evaluation tool? *Scientometrics, 1*(4), 359–375.
Guan, J. C., & Yan, Y. (2016). Technological proximity and recombinative innovation in the alternative energy field. *Research Policy, 45*(7), 1460–1473.
Guan, J. C., Yan, Y., & Zhang, J. J. (2015a). How do collaborative features affect scientific output? *Evidences from wind power field. Scientometrics, 102*(1), 333–355.
Guan, J. C., Yan, Y., & Zhang, J. J. (2017). The impact of collaboration and knowledge networks on citations. *Journal of Informetrics, 11*(2), 407–422.
Guan, J. C., Zhang, J. J., & Yan, Y. (2015b). The impact of multilevel networks on innovation. *Research Policy, 44*(3), 545–559.
Guetzkow, J., Lamont, M., & Mallard, G. (2004). What is originality in the humanities and the social sciences. *American Sociological Review, 69*(2), 190–212.
Haans, R. F. J., Pieters, C., & He, Z. L. (2016). Thinking about u: Theorizing and testing u- and inverted u-shaped relationships in strategy research. *Strategic Management Journal, 37*(7), 1177–1195.
Jamali, H. R., & Nikzad, M. (2011). Article title type and its relation with the number of downloads and citations. *Scientometrics, 88*(2), 653–661.
Jansen, B. J., & Pooch, U. (2001). A review of Web searching studies and a framework for future research. *Journal of the American Society for Information Science and Technology, 52*(3), 235–246.
Kaplan, S., & Vakili, K. (2015). The double-edged sword of recombination in breakthrough innovation. *Strategic Management Journal, 36*(10), 1435–1457.
Lee, D. H., & Brusilovsky, P. (2019). The first impression of conference papers: Does it matter in predicting future citations? *Journal of the Association for Information Science and Technology, 70*(1), 83–95.
Lee, Y., Walsh, J. P., & Wang, J. (2015). Creativity in scientific teams: Unpacking novelty and impact. *Research Policy, 44*(3), 684–697.
Lenzner, T., Kaczmirek, L., & Lenzner, A. (2010). Cognitive burden of survey questions and response times: A psycholinguistic experiment. *Applied Cognitive Psychology, 24*(7), 1003–1020.
Letchford, A., Preis, T., & Moat, H. S. (2016). The advantage of simple paper abstracts. *Journal of Informetrics, 10*(1), 1–8.
Leydesdorff, L., Bornmann, L., & Wagner, C. S. (2019). The relative influences of government funding and international collaboration on citation impact. *Journal of the Association for Information Science and Technology, 70*(2), 198–201.
Leydesdorff, L., Wagner, C. S., & Bornmann, L. (2018). Betweenness and diversity in journal citation networks as measures of interdisciplinarity—A tribute to Eugene Garfield. *Scientometrics, 114*(2), 567–592.
Lutz, T., Herrig, A., Würz, W., Kamruzzaman, M., & Krämer, E. (2017). Design and wind-tunnel verification of low-noise airfoils for wind turbines. *AIAA Journal, 45*(4), 779–785.
Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research, 1*(2), 161–175.
McCain, K. W. (1989). Descriptor and citation retrieval in the medical behavioral sciences literature: Retrieval overlaps and novelty distribution. *Journal of the American Society for Information Science, 40*(2), 110–114.
Min, C., Bu, Y., Sun, J., & Ding, Y. (2018). Is scientific novelty reflected in citation patterns? *Proceedings of the Association for Information Science and Technology, 55*(1), 875–876.

Nelson, R. R., & Winter, S. G. (1982). *An evolutionary theory of economic change*. Cambridge, MA: The Belknap Press of Harvard University Press.
Phene, A., Fladmoe-Lindquist, K., & Marsh, L. (2006). Breakthrough innovations in the US biotechnology industry: The effects of technological space and geographic origin. *Strategic Management Journal, 27*(4), 369–388.
Sanz-Casado, E., Garcia-Zorita, J. C., Serrano-López, A. E., Larsen, B., & Ingwersen, P. (2013). Renewable energy research 1995–2009: A case study of wind power research in EU, Spain, Germany and Denmark. *Scientometrics, 95*(1), 197–224.
Saviotti, P. P., & Metcalfe, J. S. (1984). A theoretical approach to the construction of technological output indicators. *Research Policy, 13*(3), 141–151.
Schilling, M. A., & Green, E. (2011). Recombinant search and breakthrough idea generation: An analysis of high impact papers in the social sciences. *Research Policy, 40*(10), 1321–1331.
Schumpeter, J. A. (1934). *The theory of economic development*. Cambridge, MA: Harvard University Press.
Skilton, P. F. (2006). A comparative study of communal practice: Assessing the effects of taken-for-grantedness on citation practice in scientific communities. *Scientometrics, 68*(1), 73–96.
Sternberg, R. J. (1999). *Handbook of creativity*. Cambridge, UK: Cambridge University Press.
Strumsky, D., & Lobo, J. (2015). Identifying the sources of technological novelty in the process of invention. *Research Policy, 44*(8), 1445–1461.
Su, H., & Lee, P. (2010). Mapping knowledge structure by keyword co-occurrence: A first look at journal papers in technology foresight. *Scientometrics, 85*(1), 65–79.
Tahamtan, I., & Bornmann, L. (2018a). Core elements in the process of citing publications: Conceptual overview of the literature. *Journal of Informetrics, 12*(1), 203–216.
Tahamtan, I., & Bornmann, L. (2018b). Creativity in science and the link to cited references: Is the creative potential of papers reflected in their cited references? *Journal of Informetrics, 12*(3), 906–930.
Tahamtan, I., Safipour, A. A., & Ahamdzadeh, K. (2016). Factors affecting number of citations: A comprehensive review of the literature. *Scientometrics, 107*(3), 1195–1225.
Thelwall, M., & Wilson, P. (2014). Regression for citation data: An evaluation of different methods. *Journal of Informetrics, 8*(4), 963–971.
Utterback, J. M. (1971). The process of innovation: A study of the origination and development of ideas for new scientific instruments. *IEEE Transactions on Engineering Management, 18*(4), 124–131.
Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science, 342*(6157), 468–472.
Verhoeven, D., Bakker, J., & Veugelers, R. (2016). Measuring technological novelty with patent-based indicators. *Research Policy, 45*(3), 707–723.
Walters, G. D. (2006). Predicting subsequent citations to articles published in twelve crime-psychology journals: Author impact versus journal impact. *Scientometrics, 69*(3), 499–510.
Wang, C., Rodan, S., Fruin, M., & Xu, X. (2014). Knowledge networks, collaboration networks, and exploratory innovation. *Academy of Management Journal, 57*(2), 484–514.
Wang, J., & Shapira, P. (2011). Funding acknowledgement analysis: An enhanced tool to investigate research sponsorship impacts: The case of nanotechnology. *Scientometrics, 87*(3), 563–586.
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.
Yan, Y., Dong, J. Q., & Faems, D. (2019). Not every coopetitor is the same: The impact of technological, market and geographical overlap with coopetitors on firms' breakthrough inventions. *Long Range Planning*. https://doi.org/10.1016/j.lrp.2019.02.006.
Zhang, J. J., & Guan, J. C. (2016). Scientific relatedness and intellectual base: A citation analysis of un-cited and highly-cited papers in the solar energy field. *Scientometrics, 110*(1), 1–22.
Zhang, J. J., Yan, Y., & Guan, J. C. (2015). Scientific relatedness in solar energy: A comparative study between the USA and China. *Scientometrics, 102*(2), 1595–1613.
Zhang, J. J., Yan, Y., & Guan, J. C. (2019). Recombinant distance, network governance and recombinant innovation. *Technological Forecasting and Social Change, 143,* 260–272.