

Patent citations to scientific papers as early signs for predicting delayed recognition of scientific discoveries: a comparative study with instant recognition

Jian Du ¹, Peixin Li ², Robin Haunschild³, Yinan Sun ⁴, Xiaoli Tang*⁵

¹ du.jian@imicams.ac.cn ² li.peixin@imicams.ac.cn ⁴ sun.yinan@imicams.ac.cn ⁵ tang.xiaoli@imicams.ac.cn
Institute of Medical Information & Library, Chinese Academy of Medical Sciences, Beijing, 100005, Beijing
(China)

³ r.haunschild@fkf.mpg.de

Max Planck Institute for Solid State Research, Heisenbergstraße 1, 70569 Stuttgart (Germany)

Abstract

In this study, we investigate the extent to that patent citations to papers can serve as early signs for predicting delayed recognition using a comparative study with a control group, i.e., instant recognition papers. We identify the two opposite groups of papers by the Bcp measure, a parameter-free index for identifying papers which were recognized with delay (also called “sleeping beauties” in science). Combined with a typical case study, it appears that papers with delayed recognition show a stronger and longer technical impact than instant recognition papers. We provide indication that in the more recent years papers with delayed recognition are awakened more often and more earlier by a patent rather than by a scientific paper (also called “prince”). We also found that patent citations seem to play an important role to avoid instant recognition papers to level off or to become a so called “flash in the pan”. It also appears that the sleeping beauties may firstly encounter negative citations and then patent citations and finally get widely recognized. In contrast to the two focus fields (biology and chemistry) for instant recognition papers, delayed recognition papers are rather evenly distributed in biology, chemistry, psychology, geology, materials science, and physics. We discovered several pairs of “science sleeping”-“technology inducing”, such as biology-biotechnology/pharmaceuticals, chemistry-chemical engineering, psychology-computer/control technology, and physics-computer technology. We propose in further research to discover the potential ahead of time and transformative research by using citation delay analysis, patent & NPL analysis, and citation context analysis.

Keywords:

Delayed recognition papers; Citation delay analysis; Patent & NPL analysis; Scientific discoveries; Early signs

Introduction

According to our understanding of Kuhn’s paradigm on the *Structure of Scientific Revolutions*, scientific knowledge proceeds incrementally (incremental research), occasionally punctuated by paradigm-shifting discoveries (transformative research) (Kuhn & Hawkins, 1963). In contrast to incremental research, which moves forward through the continuous, incremental accumulation of knowledge, transformative research drives science forward by radically changing our understanding of a concept, causing a paradigm shift, or opening new frontiers (Trevors, Pollack, Saier, & Masson, 2012). Prioritization of transformative research has become pervasive among funding agencies (Sen, 2017). Such research brings great rewards, but also carries great risks for funding agencies because transformative research projects are very hard to identify in their early stages. An on-going challenge lies in identifying transformative research projects at the time they are proposed. Although it is rarely possible to predict the transformative nature of research during the proposal stage, yet it is more predictable during the research process or even for a long time after the discovery. Further, transformative research should not be understood as just the opposite of incremental research. Actually, most transformative research began with incremental goals, and the transformative potential was recognized later (Gravem et al., 2017).

Premature discoveries and transformative research are crucial for the development of science, but they are often initially neglected or resisted by the scientific community and thus are subject to delayed recognition (Figure 1). In a report by the National Academies of Sciences, Engineering, and Medicine in 2016, the committee reviewed five transformative areas of geographical research that have taken shape over the past 65 years to explore how transformative research has emerged. They found that transformative innovations can arise from older and long-ignored ideas (National Academies of Sciences & Medicine, 2016). Such ideas are often called “Sleeping Beauties” (SB) in science, one type of publications that goes unnoticed (or “sleeps”) for a long time and then, almost suddenly, attracts a lot of attention (or “is awakened by a Prince”) (A. F. J. Van Raan, 2004). This concept in terms of a citation curve is actually a quantitative description of “delayed recognition of scientific achievements”, a phenomenon widely discussed in sociology of science (Hook, 2002). To the best of our knowledge sociologist Stephen Cole was the first to propose the notion of measuring delayed recognition in science using citations (Cole, 1970).

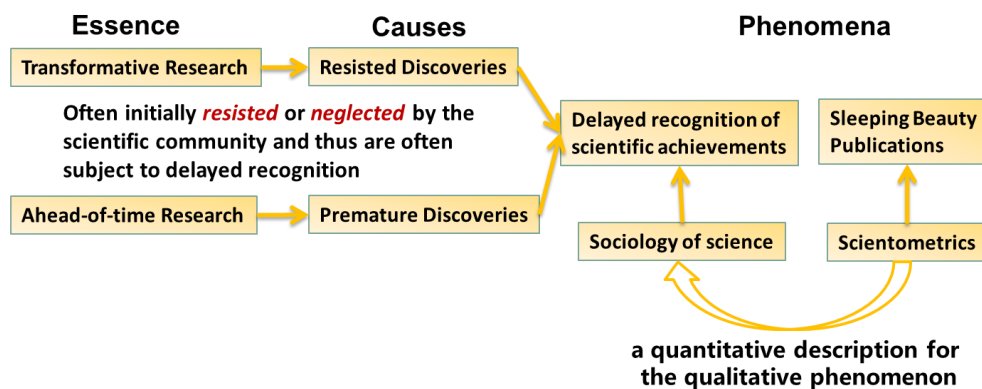


Figure 1. A schematic model of sleeping beauty and delayed recognition in science

Among the many papers written by colleagues on delayed recognition or SBs which have been discussed in our foregoing papers (Du & Wu, 2016, 2018), we concentrate on a few important recent developments and focus on the early identification of such type of publications and/or transformative research. Can we know in early stages if a research project looks promising or might lead to transformative research? The studies mentioned below may provide some insight into early signs of the awakening of SB publications or the recognition of premature discoveries and transformative research. It has been shown by Marx (2014) using the example of the paper by Shockley and Queisser (1961) that delayed recognition papers often start getting cited when a highly cited paper or another prominent author has drawn attention to them.

Dey, Roy, Chakraborty, and Ghosh (2017) analyzed the features of a given paper which may become an SB, and were the first to investigate the early identification of SBs in computer science. They developed a methodology to predict early, i.e., as soon as possible after the paper is published, whether a paper is likely to become an SB. By distinguishing two classes of papers—SBs and non-SBs—based on a set of features derived from each paper, they observed that the entropy of the number of fields from where the target paper has received citations is the most important feature. The more the paper has potential to attract attention from multiple fields, the higher the probability that it qualifies as an interdisciplinary paper that can become popular eventually. This observation is corroborated by Ke, Ferrara, Radicchi, and Flammini (2015). They found that in many cases, the awakening of SBs occurs when an application in a field outside of the SB’s field is found, such as statistical methods that became useful in biology, chemistry, or physics.

Another important discovery is the correlation between the SBs and the scientific non-patent references in patents. One is perhaps more inclined to believe that SBs relate to more fundamental, basic, and less application-oriented work. But a surprising finding is that half of the SBs in physics, chemistry, engineering, and computer science are application-oriented (A. F. J. Van Raan, 2015) and significantly more cited in patents than in scientific papers (A. F. J. van Raan, 2017). By investigating the time lag between the publication year of the SBs as scientific non-patent references (SNPR) in patents and their first citation in a patent, he found evidences that this time lag was becoming shorter in recent years (A. F. J. van Raan, 2017). In a very recent study, (A. F. J. van Raan & Winnink, 2018) investigated this further by using as recent as possible SBs cited in patents. Their observations suggest that, on average SBs are awakened increasingly earlier by a patent (“technological prince”) rather than by a scholarly paper (“scientific prince”) in the more recent years. It may suggest that SBs with technological importance are “discovered” earlier in an application-oriented context. Then, papers may be also cited in a scientific context because of the earlier recognized technological relevance. Thus, early recognized technological relevance may “prevent” papers from becoming an SB. Very recently, the scientific and technological impact of sleeping beauties in medical research fields was analyzed by Anthony F. J. van Raan and Winnink (2019). Du and Wu (2018) also found that 60% of the extreme SBs published in *Science* and *Nature* have been cited by patents, and the SB’s first citation in terms of priority date (the earliest application date) in a patent usually appears to be earlier than the awakening year in the scientific context.

But studies by A. F. J. van Raan (2017) and A. F. J. van Raan and Winnink (2018) on patent citations to SBs are observational studies, not comparative studies. In their foregoing papers, they investigated a set of SBs with such thresholds as: (1) the average number of citations per year is at most one during 10 years after publication and, (2) the average number of citations per year is at least five during the next 10 years after 10 years of sleep. So, we can expect that the SBs have been cited at least 50 times. They identified 389 SBs for physics, 265 SBs for chemistry, and 367 SBs for engineering and computer science and found that 62 (16%), 92 (35%), and 108 (29%) of those SBs are also cited by patents. The possibility of SB papers being cited by patents is obviously higher than the proportion of all Web of Science (WoS) or MEDLINE covered publications cited by patents (about 4%). Ahmadpoor and Jones (2017) traced references from all 4.8 million patents issued by the U.S. Patent and Trademark Office (USPTO) from 1976 to 2015 to all 32 million journal articles published from 1945 to 2013 as indexed by the Web of Science. They found that 1.41 of 32 million (4.4%) WoS papers were cited by USPTO patents. This estimate is similar to a recent study where papers in MEDLINE rather than WoS were considered: a very small portion (4%) of MEDLINE papers published until 2013 are cited by USPTO patents (Ke, 2018). The percentage 4% is calculated based on all papers indexed by WoS/MEDLINE, including the papers which have never been cited. According to an investigation on the uncited papers in WoS (Van Noorden, 2017), the proportion of uncited papers levels off between five and ten years after publication in most disciplines, although the proportion is different in each discipline. Of all biomedical papers published in 2006, just 4% are uncited today; in chemistry, that number is about 8% and in physics it is close to 11%. In engineering and technology, the uncitedness rate is with 24% much higher than in the natural sciences. Thus, we can expect a dependence of the patent citation rate on scientific fields. Out of the 39 million research papers across all disciplines recorded in the Web of Science from 1900 to the end of 2015, about 21% haven’t been cited, yet (Van Noorden, 2017). Since these papers have no scientific impact, they are likely to have no technical impact and thus will probably not be cited by patents.

In this study, we will answer two questions based on van Raan’s work: (1) what is the extent to that patent citations can serve as early signs of delayed recognition using a comparative

study with a control group, i.e., instant recognition papers? Delayed recognition in science is a phenomenon where papers went unnoticed until they are re-discovered some years after publication. By contrast, instant recognition (also called "flashes in the pan") in science is a phenomenon where papers received a lot of citations shortly after publication, but were ignored very quickly (Ye & Bornmann, 2018). (2) What is the pattern of the interaction in terms of citation relations between the sleeping science and the technology inducing its recognition? Based on our previous investigations on systematic identification of SBs and on their awaking mechanisms (Du & Wu, 2016, 2018), this contribution will further validate and detect early signs of the awakening of SBs. Our aim is to detect potential ahead-of-time discoveries or transformative research in order to shorten the time lag for original research to get recognized.

Data and Methods

Although being an SB sounds like a yes/no situation, it is clear that delayed recognition is not a clear-cut phenomenon (Rousseau, 2018). We are interested in detecting SBs in the hidden, under-cited publications with delayed impact. Note that our definition of "under-cited" is in contrast to "highly-cited". Currently, scholars in bibliometrics mostly focus on highly cited papers and ignore less highly cited or never cited ones. Research on SBs and flashes in the pan is valuable in turning scholars' attention to less highly or never cited papers which should have not been neglected. To characterize delayed recognition papers, it is necessary to compare them with instant recognition papers. We will turn an apparent yes/no question into a continuous phenomenon.

A parameter-free index for measuring the extent of delayed recognition

Based on the identification framework of the "beauty coefficient" (B) introduced by Ke et al. (2015) which takes the whole citation history of the publications concerned into account, Du and Wu (2018) substituted yearly citations in the beauty coefficient with yearly cumulative percentage of citations. The value of the modified beauty coefficient is denoted as Bcp. Bcp depends on the shape of the cumulative citation curve, especially when there is a cumulative citation burst in the whole life cycle, but not on the total number of citations of a given paper. Bcp works better than B in at least two aspects: (1) it "punishes" the situations when the SBs experienced early citations instead of continuous sleeping; (2) it allows comparisons of the extent of delayed citation impact of publications in different fields with different citation patterns.

In general, it is a sign of delayed recognition if a given paper's cumulative citation curve is concave ($Bcp > 0$). Early citations are indicated by a convex cumulative citation curve ($Bcp < 0$). The larger the Bcp value, the more delayed is the recognition of a paper in terms of the citation curve. The maximum value of the Bcp index is $(n-1)/2$, where n is the age of a given paper. This case occurs when the total number of citations is received in the last year. Hence, no citations are gathered in previous years. The smaller the Bcp's negative value, the more instant is the recognition of a paper. Just like the "top 1%" is usually used to select highly cited papers, we will also use "top 1%" versus "bottom 1%" for grouping delayed recognition and instant recognition papers.

In line with the earlier definition on the awakening year when the abrupt increase in the accumulation of citations of sleeping beauties occur, Du and Wu (2018) defined the "falling year" as the time when the abrupt decrease in the accumulation of citations of flash-in-the-pan papers occur. We will use this definition for "falling year" in our current study, too.

Delayed recognition (top 1%) versus instant recognition (bottom 1%) papers

The framework of Bcp was used to identify SBs published between 1970 and 2005 in Science and Nature. Citation data were included until the end of 2015. As we wish to have at least ten years of citation history after the latest publication year, 2005 is the most recent publication year included in our study. Articles with at least 200 citations, in total 20,000 publications were included in the following analysis. These 20,000 papers were ordered by their Bcp value. We selected the top 1% (N=200) as delayed recognition papers and the bottom 1% (N=200) as instant recognition papers.

Patent-citing related indicators

Patent documents provide citations to earlier patents issued (prior art) and to non-patent literature (NPL), which includes peer-reviewed research and other published documents. Earlier patents may be cited by the inventor to demonstrate their difference from prior art or added by the examiner to limit the scope of the patent. Patent backward citations to NPL are considered stronger indicators of the impact of scientific research on technical invention than citations to patents (Roach & Cohen, 2013). So, using the patent backward citations to NPL, one can measure the technological impact of the scientific knowledge. We compare the extent to which the delayed recognition papers and the instant recognition papers show up as NPL in patents. In this study, the linkage between patents and NPLs was gathered by searching the platform lens.org, created by Cambia, a non-profit organization in Australia dedicated to facilitating innovation, and Queensland University of Technology. The platform lens.org has the world's patent information to most of the scholarly literature via collaborations with CrossRef and National Library of Medicine (Jefferson et al., 2018). “The Nature Index 2017 Innovation” published the top 200 institutions ranked by the Lens score, shedding new light on the impact academic research has had on innovation by examining how research articles are cited in third party patents¹. We mainly focused on the following indicators.

- 1) Number of citing patent families: we group patent publications describing the same invention in “patent families” to prevent double counting when counting the number of patent citations to a given paper.
- 2) Interval of priority year between the earliest and the latest citing patent: by this measure, we can figure out the durability of patent citations to a given paper.

Fields of study

In order to compare the field of technology of papers with fields of study, we use the hierarchical fields of study from Microsoft Academic which are provided by a semantic algorithm on the paper basis. We appended the field of study from a local in-house database of Microsoft Academic to the top 1% and bottom 1% papers via the DOI and from Lens.org via PMID. Starting in August 2018, all scholarly papers cited by patents will have the information of field of study thanks to a partnership with Microsoft Academic². Not all papers in Microsoft Academic database have a field of study attached to them but some papers have multiple fields of study at different levels. We found at least one field of study for 198 top 1% papers and for 196 bottom 1% papers with DOIs and/or PMIDs. For the rest of papers, we give the top level field of study based on their research areas reflected by title and/or abstract.

¹ https://www.nature.com/press_releases/nature-index-2017-innovation-supplement.html

² <https://www.microsoft.com/en-us/research/project/academic/articles/sharpening-insights-into-the-innovation-landscape-with-a-new-approach-to-patents/>

Results of a comparative study

Identifying the two opposite groups of papers by Bcp measure

Figure 2 shows citation curves of the first and the last paper ranked by Bcp and the distribution of citation percentiles for the two groups of papers. The awakening year for the most delayed recognition paper is 2004 (until the 33rd year after publication) and the falling year for the most instant recognition paper is 1977 (just in the 7th year after publication). Many of the most delayed recognition papers are lowly cited, whereas many of the most instant recognition papers are highly cited. We can see that Bcp is not very dependent on the total number of citations of a given paper. It is appropriate for distinguishing those publications with delayed recognition from those with instant recognition although they are not so highly cited.

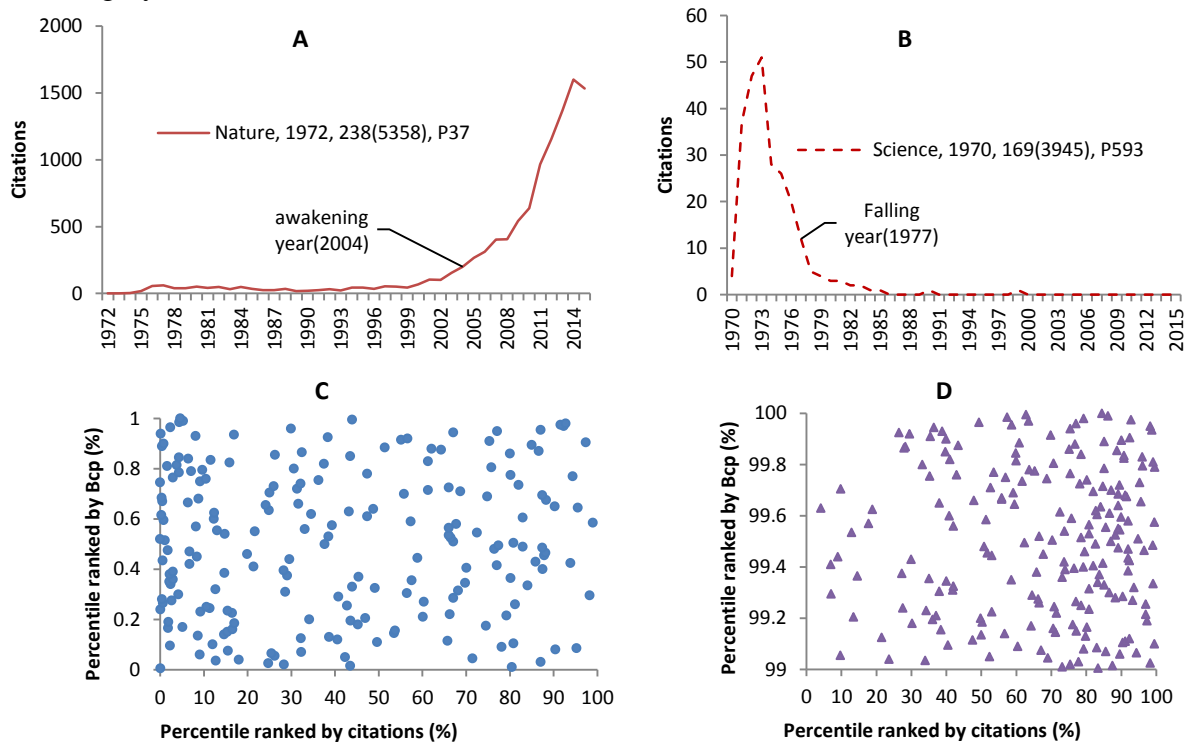


Figure 2. Citation curves of the first (A) and the last (B) paper ranked by Bcp and distribution of citation percentiles for top 1% (C) and bottom 1% (D) papers

Delayed recognition papers showing a stronger and longer technical impact than instant recognition papers

We find that about half of the 200 delayed recognition papers (DR-NPL) are cited by patents and about one-third of the 200 instant recognition papers (IR-NPL) are cited by patents. Similar to citations given by publications, also the number of citations by patents is characterized by a skewed distribution. For instance, about one third of the DR-NPLs are cited by only 1 or 2 patents, and six are cited by more than 300 patents. In addition, about half of the IR-NPLs are cited by only 1 or 2 patents. One is cited by 120 patents, but the rest is cited by no more than 40 patents. In total, the 99 DR-NPLs are cited by 3988 patents, and the 70 IR-NPLs are cited by 543 patents. Delayed recognition articles were 5 times more cited by patents than instant recognition papers (on average 40.3 patent families versus 7.8 patent families), showing a stronger technical impact (Table 1). Next, we determined for the DR-NPLs and IR-NPLs the filing years of the earliest and of the latest citing patent until Dec 23, 2018. The difference between the filing years of the latest and the earliest patent indicates the

durability of patent citations. The average durability per paper is 15 years for DR-NPLs and 10 years for IR-NPLs, showing a longer technical impact for delayed recognition papers

Table 1. Extent to that the delayed recognition papers versus the instant recognition papers show up as NPLs in patents.

	<i>top1%</i>	<i>bottom 1%</i>
Total number of papers	200	200
Number of papers cited by patents	99	70
% of papers cited by patents	49.5	35.0
Total number of citing patent families	3988	543
Number of citing patent families per paper	40.3	7.8
Average durability of patent citations per paper	15.2	10.1

Note: Patent citations to non-patent literature accessed via lens.org by Dec 23, 2018.

Patent citations: earlier awakening, delayed recognition, and avoiding “flashes in the pan”

First, we compare the earliest patent citing year and the awakening year for the 99 delayed recognition NPLs, and find that for 70% (n=69) of the papers the first patent citing year is earlier than the awakening year; for 5% (n=5) of the papers the first patent citing year is the same as the awakening year; only 25% (n=25) of the papers are cited by patents after awakening. The difference between the filing year of the earliest citing patent and the awakening year, i.e., when the citations of the DR-NPL begin to abruptly increase define the time lag between the first citation by a patent and its “reviving”. This time lag ranges from -19 to 29 years (average 6.7, SD=10.1). For example, for a Science article published in 1976 (10.1126/science.996549), the awakening year is 1994, and the year of the first patent citation is 2013. So, time lag between the year of the first patent citation and awakening year is -19. The average value relates to a long measuring period. In order to find out if there is a trend over time, we calculate the averages for successive, partly overlapping 5-year periods (Figure 3). In the case of the 99 DR-NPLs, these periods are 1970-1974, 1971-1975,..., and 1990-1994. Remarkably, the time lag is fluctuating in the earlier years but becomes rapidly shorter in the recent years. In other words, for the more recent DR-NPL, once it is cited by a patent, it will be awakened more quickly.

Afterwards, we analyze the time lag between the first patent citation and the falling year for instant recognition papers. The difference between the filing year of the earliest citing patent and the falling year, i.e., when the citations of the IR-NPL begin to abruptly decrease define the time lag between the first citation by a patent and its “perishing”. Obviously, the time lag becomes rapidly longer in the measured period 1970-1983 (Figure 3). In the more recent years, even the instant recognition papers will be more likely to exhibit a long time window of citations once they are cited by a patent. Both observations suggest that, on average, in the more recent years, the delayed recognition papers are awakened increasingly earlier by a patent (“technological prince”) rather than by a scholarly paper (“scientific prince”). Patent citations seem to play a more important role to avoid instant recognition papers to level off or become “flashes in the pan”.

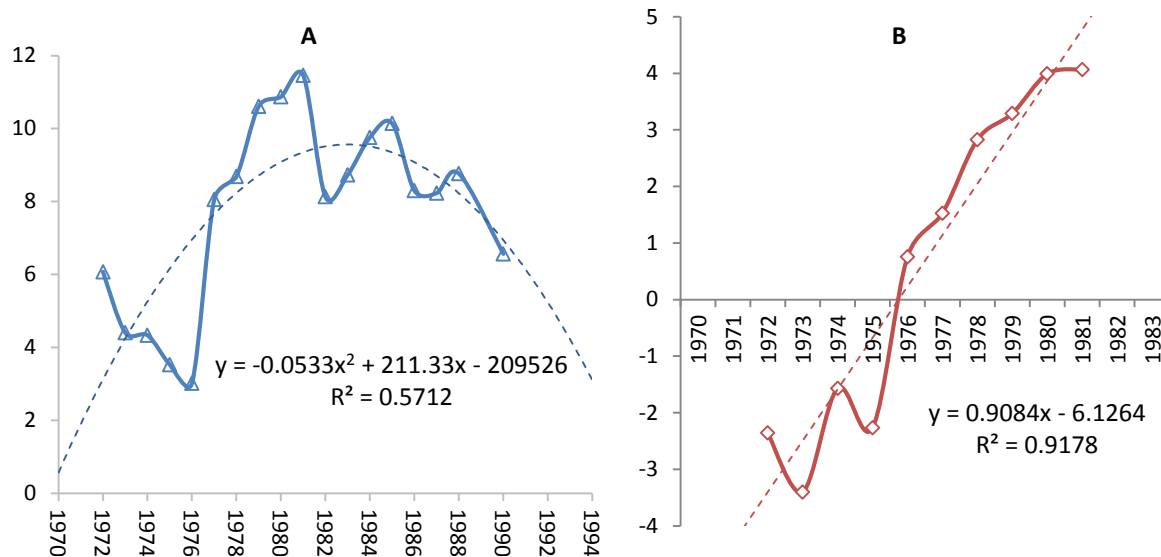


Figure 3. Time lag between the year of the first patent citation and the awakening year for DR-NPLs (A) and time lag between the first patent citation and the falling year for IR-NPLs (B).

Comparing the difference of science-technology interactions between DR and IR papers

The impact of discoveries may extend beyond the domain of science and may be crucial steps towards technological applications. It has been argued that technology-driven, or more specific, patent citations to papers, might be one of the awakening mechanisms for delayed recognition papers (Du, Sun, Zhang, & Tang, 2019). To reveal the whole picture of research fields for the scientific papers and the interactions with the technical focus of the citing patent families, we firstly match the field of study for each of the 200 delayed recognition and the other 200 instant recognition papers from Microsoft Academic, which determines the field of study based on machine learning parsing of all accessible text in a record. Microsoft Academic increases the power of semantic search by adding more fields of study³ (from February 15, 2018). There are now 19 top-level fields of study, including Biology, Medicine, Geology, Chemistry, Psychology, Philosophy, Sociology, Engineering, Economics, Computer Science, Art, Physics, History, Political Science, Materials Science, Mathematics, Geography, Business, and Environmental Science. The Microsoft Academic data contain fields of study with a six-level hierarchy. Using the technology classification groups or WIPO concordance table⁴, which links IPC symbols with 35 fields of technology we identified the fields of technology for each of the earliest citing patents in our two datasets. Afterwards, we map the interactions (by means of direct citations) between fields of study and fields of technology to figure out the different patterns for the two groups of papers.

There are 952 fields of study for the 200 delayed recognition papers, of which 55 (27.5%) are biology, followed by chemistry, psychology (n=25, 12.5%), geology (n=17, 8.5%), materials science (n=17, 8.5%), physics (n=16, 8%), and so on. However, there are 618 fields of study for the 200 instant recognition papers, of which almost 90% (n=180) are biology and nearly 10% (n=19) are chemistry. Figure 4 shows that delayed recognition papers in biology are mainly cited by patents in biotechnology and pharmaceuticals. Delayed recognition papers in chemistry are often cited by chemical engineering technology, biotechnology, and

³ <https://www.microsoft.com/en-us/research/lab/microsoft-research-asia/articles/microsoft-academic-increases-power-semantic-search-adding-fields-study/>

⁴ https://www.wipo.int/export/sites/www/ipstats/en/statistics/patents/pdf/wipo_ipc_technology.pdf

pharmaceuticals. Delayed recognition papers in materials science are mainly cited by patents in biotechnology and metallurgy materials. Delayed recognition papers in psychology are mainly cited by computer technology and control technology. Delayed recognition papers in physics are mainly cited by computer technology.



Figure 4. Interactions between fields of study (red color) of papers and their earliest citing patent families' fields of technology (blue color) for delayed recognition papers (A) and Instant recognition papers (B).

Discussion and Conclusion

Rousseau (2018) recently made an important observation that using citations to study delayed recognition is just a—convenient—operationalization of the concept, but that experts may agree on delayed recognition long before this is shown evidently by citations. He discovered a case in which an expert found delayed recognition several years before citation analysis could discover this phenomenon. This leads to the question: How good (or adequate) is citation analysis for detecting premature discoveries? To answer this question, we propose to prioritize the investigation into the lowly-cited papers instead of the most highly cited papers. Combining the frameworks of citation delay and beauty coefficient, we have proposed a parameter-free index known as Bcp index, for identifying under-cited SBs in science, which may indicate possible breakthroughs in an early stage (Du & Wu, 2018). Note that our

definition of “under-cited” is in contrast to “highly-cited”, from the perspective of the first generation of citations. By Bcp measure, we can distinguish delayed recognition from instant recognition papers.

Using articles published between 1970 and 2005 in the journals *Science* and *Nature*, we conducted a comparative study on delayed recognition with instant recognition papers. Combined with a case study, we found that delayed recognition papers show a stronger and longer technical impact than instant recognition papers. On average, in the more recent years the delayed recognition papers are cited increasingly earlier by a patent. Patent citations seem to play an important role to avoid instant recognition papers to level off or to become “flashes in the pan”. We provided further evidence to support the observation made by A. F. J. van Raan and Winnink (2018) that in the more recent years SBs are awakened increasingly earlier by a patent rather than by a scientific paper. This may suggest that early recognized technological relevance may “prevent” papers from becoming delayed recognition papers. It also appears that the sleeping beauties may firstly encounter negative citations, then patent citations, and finally get widely recognized.

We found that in contrast to the two focus fields (biology and chemistry) for instant recognition papers, delayed recognition papers are rather evenly distributed across biology, chemistry, psychology, geology, materials science, and physics. We also discovered several pairs of “science sleeping”-“technology inducing”, such as biology-biotechnology/pharmaceuticals, chemistry-chemical engineering, materials science-biotechnology/metallurgy materials, psychology-computer/control technology, and physics-computer technology.

The non-patent literature (NPL) cited by patents may provide insight into the awakening of delayed recognition publications, which may mean that the ahead-of-time discoveries get understood or the transformative potential of research is recognized. In a previous study, Du and Wu (2018) have discovered using citation context analysis that the extreme delayed recognition papers were all landmark publications of a specific research field, such as “the first report on ...” or “the classic theory about ...”. It appears that high quality publications tend to encounter delayed recognition and thus show delayed citation impact. One could identify transformative research through some text terms (such as “disagree”, “contradict”, “controversial”, “inconsistent”, “dispute”, ...). In order to discern such potential transformative research, one could observe whether the relevant documents get early citation from patents or not. Much transformative research has influence on technology and invention and thus SBs tend to be cited by patents. In such cases, SBs with technological importance tend to be ‘discovered’ and ‘awakened’ earlier in an application-oriented context. Therefore, we propose to discover the potential ahead of time and transformative research by a combination of citation delay analysis, patent & NPL direct citation analysis, and citation context analysis.

Future research

Inspired by our investigations in this study, we propose to combine citation delay analysis with patent & NPL direct citation analysis to identify potential ahead of time and transformative research. The Bcp index proposed by Du and Wu (2018) can be used to identify those under-cited papers that are now happening to be at the sleeping-awakening interface. Afterwards, one could further identify those delayed recognition papers which are also cited by patents. Finally, one could map the structure of the older and long-ignored ideas at both the sleeping-awakening interface and science-technology interface. These ideas and research topics may be the potential origin of transformative research.

Further, inspired by National Institute of Health (NIH)’s Translational Science Search (<http://tscience.nlm.nih.gov>) and SciTech Strategies Inc.’s procedure for identifying

discoveries in the biomedical sciences (Small, Tseng, & Patek, 2017), we argue that combining text mining based on authors' claims with citation context analysis from citers' comments, one may also discover potential transformative research. It may be possible to use text mining for identifying articles that are regarded by their authors as controversial (they challenge established dogma) or refutation (they disprove previously published data or hypotheses). The author's view can be compared with the citer's view by searching for specific terms (such as "disagree", "contradict", "contrast", "inconsistent", "dispute", ...) in the citation context. After a manual screening process to remove non-transformative research discoveries, it might be possible to provide a list of transformative research discoveries in the recent ten years from the perspectives of both author's claims and the community's comments.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (Grant No. 71603280) and the Young Elite Scientists Sponsorship Program by China Association for Science and Technology (Grant No. 2017QNRC001) and CAMS Initiative for Innovative Medicine (CAMS-I2M-3-018). Data from Microsoft Academic (Sinha et al., 2015) (see also <https://aka.ms/msracad>) were shared with one of us (RH).

References

- Ahmadpoor, M., & Jones, B. F. (2017). The dual frontier: Patented inventions and prior scientific advance. *Science*, 357(6351), 583-587. doi:10.1126/science.aam9527
- Cole, S. (1970). Professional Standing and the Reception of Scientific Discoveries. *American Journal of Sociology*, 76(2), 286-306.
- Dey, R., Roy, A., Chakraborty, T., & Ghosh, S. (2017). Sleeping beauties in Computer Science: characterization and early identification. *Scientometrics*, 113(3), 1645-1663. doi:10.1007/s11192-017-2543-3
- Du, J., Sun, Y. N., Zhang, Y., & Tang, X. L. (2019). Characterizing and Detecting the early scientometric signs of the potential transformative research. *Bulletin of National Natural Science Foundation of China*, 33(150), 1-11.
- Du, J., & Wu, Y. S. (2016). A Bibliometric Framework for Identifying "Princes" Who Wake up the "Sleeping Beauty"; in Challenge-type Scientific Discoveries. *Journal of Data and Information Science*, 1(1), 50-68. doi:10.20309/jdis.201605
- Du, J., & Wu, Y. S. (2018). A parameter-free index for identifying under-cited sleeping beauties in science. *Scientometrics*, 116(2), 959-971. doi:10.1007/s11192-018-2780-0
- Gravem, S. A., Bachhuber, S. M., Fulton-Bennett, H. K., Randell, Z. H., Rickborn, A. J., Sullivan, J. M., & Menge, B. A. (2017). Transformative Research Is Not Easily Predicted. *Trends in Ecology and Evolution*, 32(11), 825-834. doi:10.1016/j.tree.2017.08.012
- Hook, E. B. (2002). *Prematurity in Scientific Discovery: On Resistance and Neglect*: University of California Press.
- Jefferson, O. A., Jaffe, A., Ashton, D., Warren, B., Koellhofer, D., Dulleck, U., . . . Jefferson, R. A. (2018). Mapping the global influence of published research on industry and innovation. *Nature Biotechnology*, 36(1), 31-39. doi:10.1038/nbt.4049
- Ke, Q. (2018). Comparing scientific and technological impact of biomedical research. *Journal of Informetrics*, 12(3), 706-717. doi:10.1016/j.joi.2018.06.010
- Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and identifying Sleeping Beauties in science. *Proceedings of the National Academy of Sciences of the United States of America*, 112(24), 7426-7431. doi:10.1073/pnas.1424329112
- Kuhn, T. S., & Hawkins, D. (1963). The Structure of Scientific Revolutions. *American Journal of Physics*, 31(7), 554-555. doi:10.1119/1.1969660
- Marx, W. (2014). The Shockley-Queisser paper - A notable example of a scientific sleeping beauty. *Annalen der Physik*, 526(5-6), A41-A45. doi:10.1002/andp.201400806
- National Academies of Sciences, E., & Medicine. (2016). *Fostering Transformative Research in the Geographical Sciences*. Washington, DC: The National Academies Press.

- Roach, M., & Cohen, W. M. (2013). Lens or Prism? Patent citations as a measure of knowledge flows from public research. *Management Science*, 59(2), 504-525. doi:10.1287/mnsc.1120.1644
- Rousseau, R. (2018). Delayed recognition: recent developments and a proposal to study this phenomenon as a fuzzy concept. *Journal of Data and Information Science*, 3(3), 1-13.
- Sen, A. (2017). Island + Bridge: How transformative innovation is organized in the federal government. *Science and Public Policy*, 44(5), 707-721. doi:10.1093/scipol/scx007
- Shockley, W., & Queisser, H. J. (1961). DETAILED BALANCE LIMIT OF EFFICIENCY OF P-N JUNCTION SOLAR CELLS. *Journal of Applied Physics*, 32(3), 510-&. doi:10.1063/1.1736034
- Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J. P., & Wang, K. (2015). *An Overview of Microsoft Academic Service (MAS) and Applications*. Paper presented at the 24th International Conference on World Wide Web (WWW '15 Companion), Florence, Italy.
- Small, H., Tseng, H., & Patek, M. (2017). Discovering discoveries: Identifying biomedical discoveries using citation contexts. *Journal of Informetrics*, 11(1), 46-62. doi:10.1016/j.joi.2016.11.001
- Trevors, J. T., Pollack, G. H., Saier, M. H., & Masson, L. (2012). Transformative research: Definitions, approaches and consequences. *Theory in Biosciences*, 131(2), 117-123. doi:10.1007/s12064-012-0154-3
- Van Noorden, R. (2017). The science that's never been cited. *Nature*, 552(7684), 162-164. doi:10.1038/d41586-017-08404-0
- Van Raan, A. F. J. (2004). Sleeping Beauties in science. *Scientometrics*, 59(3), 467-472. doi:10.1023/B:SCIE.0000018543.82441.f1
- Van Raan, A. F. J. (2015). Dormitory of physical and engineering sciences: Sleeping beauties may be sleeping innovations. *PLoS ONE*, 10(10). doi:10.1371/journal.pone.0139786
- van Raan, A. F. J. (2017). Sleeping beauties cited in patents: Is there also a dormitory of inventions? *Scientometrics*, 110(3), 1123-1156. doi:10.1007/s11192-016-2215-8
- van Raan, A. F. J., & Winnink, J. J. (2018). Do younger Sleeping Beauties prefer a technological prince? *Scientometrics*, 114(2), 701-717. doi:10.1007/s11192-017-2603-8
- van Raan, A. F. J., & Winnink, J. J. (2019). Sleeping Beauties in Medical Research: Technological Relevance, High Scientific Impact. Retrieved from <https://arxiv.org/abs/1904.07658>
- Ye, F. Y., & Bornmann, L. (2018). "Smart girls" versus "sleeping beauties" in the sciences: The identification of instant and delayed recognition by using the citation angle. *Journal of the Association for Information Science and Technology*, 69(3), 359-367. doi:10.1002/asi.23846