Journal of Informetrics 14 (2020) 100997

Contents lists available at ScienceDirect

Journal of Informetrics

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

The overlooked citations: Investigating the impact of ignoring citations to published patent applications

Chung-Huei Kuan [a], Dar-Zen Chen [b,d], Mu-Hsuan Huang [c,*]

[a] National Taiwan University of Science and Technology, Graduate Institute of Patent, No. 43, Sec. 4, Keelung Rd., Taipei, Taiwan, ROC
[b] National Taiwan University, Department of Mechanical Engineering, No. 1, Sec. 4, Roosevelt Rd., Taipei, Taiwan, ROC
[c] National Taiwan University, Department of Library and Information Science, No. 1, Sec. 4, Roosevelt Rd., Taipei, Taiwan, ROC
[d] National Taiwan University, Center for Research in Econometric Theory and Applications (CRETA), No. 1, Sec. 4, Roosevelt Rd., Taipei, Taiwan, ROC

ARTICLE INFO

Article history:
Received 23 August 2019
Received in revised form
27 November 2019
Accepted 27 November 2019
Available online 19 December 2019

Keywords:
Patent bibliometrics
Published patent application
Citation
Value capture rate
Main path analysis

ABSTRACT

A utility patent application may result in two citable documents: a published patent application (PPA) and a patent if the application is granted. Most analytic works consider only citations to the patent and ignore those to the PPA. This study gathers more than 270,000 U.S. utility patents granted in 2014 and their PPAs, and compares their citation counts up to 2018. Statistics show that citations to patents, on the average, account for less than 50 % of those to the patents and their PPAs combined together, indicating a significant underestimation to the value or impact of the patents. The degree of depreciation is worse when the time gaps between patents and their PPAs are longer, as the PPAs not only have accumulated citations for a longer period, but also individually, concurrently, and continuously receive citations after the patent is granted. This study further applies Main Path Analysis to a conventional citation network involving only citations to the patents and another network augmented with those to the PPAs, using empirical data from United States Patent and Trademark Office (USPTO) Cancer Moonshot Patent Data. The main path derived from the augmented network is almost entirely different from that of the conventional network.

1. Introduction

Citations between patents, just like citations between academic articles, are widely utilized in patent bibliometric works. In addition to the common practice of using citation count as an indication to quality and influence, patent citations are employed to detect relatedness and considered as a proxy to knowledge flow between citing and cited patents. Then, evolving technological trends may be discovered by observing patents along their citation chains; the cooperation/competition relationship and knowledge exchange between firms, institutions, counties, and industries may be examined and inferred based on the citations between their patent portfolios.

Due to the rich corpus of literature, a small number of studies applying patent citations to knowledge or technology spillover published after 2013 are sampled below. Karvonen and Kässi (2013) investigated the new industry segments emerging from the field of radio-frequency identification (RFID) using European Patent Office (EPO) patents and their cita-

* Corresponding author.
E-mail addresses: maxkuan@mail.ntust.edu.tw (C.-H. Kuan), dzchen@ntu.edu.tw (D.-Z. Chen), mhhuang@ntu.edu.tw (M.-H. Huang).

https://doi.org/10.1016/j.joi.2019.100997
1751-1577/

U.S. Patent Documents		
7906371	March 2011	Kim
9466545	October 2016	Scanlan
2004/0038510	February 2004	Munakata
2011/0127654	June 2011	Weng et al.
2012/0133032	May 2012	Tsai
2015/0296667	October 2015	Hirose
2016/0035680	February 2016	Wu
2016/0225733	August 2016	Wilcoxen
2016/0227680	August 2016	Hyun
2016/0268216	September 2016	Kim
2017/0347462	November 2017	Miwa

Fig. 1. Patents and PPAs cited by 10244670.

tions. Murata, Nakajima, Okamoto, and Tamura (2014) indicated that knowledge spillover are significantly localized using cited-citing relationships between U.S. patents. Li (2014) studied the effects of distance and subnational/national borders on international and intranational knowledge spillovers through patent citations across a number of most cited countries and U.S. metropolitan areas. Figueiredo, Guimarães, and Woodward (2015) found that industry localization may offset the adverse effect of distance, using a set of citing-cited pairs of U.S. patents. Kim, Lee, and Sohn (2016) studied the spillover of unmanned aerial vehicle (UAV) technology to different industries by U.S. patents citing the UAV patents. Noailly and Shestalova (2017) used citations of EPO patents to see what technologies are built on the knowledge developed in renewable energy. Ji, Barnett, and Chu (2019) investigated how genetically modified crop technology diffused and distributed globally over time using patent citation networks.

These works consider only citations to patents. There are, however, citations to *published patent applications* (PPAs) as illustrated in Fig. 1, which is a partial screen capture of a randomly picked, recently granted utility patent 10244670[1] from the United States Patent Trademark Office (USPTO) full-text database. As revealed, this patent's examiner and/or applicant considered that each of the two patents (i.e., the top two numbers) and the nine PPAs (i.e., those having 4-digit year prefixes) discloses materials related to at least some of the patent's technical content, and these eleven documents are therefore listed as references in the patent's publicized document.

An application for utility patent is usually *early published* as PPA a period of time after filing and before it is granted, a practice adopted by all patent offices. U.S. Patent Act, therefore, prescribes that "each application for a patent shall be published . . . promptly after the expiration of a period of 18 months from the earliest filing date" (35 U.S.C. § 122(b)(1)). The application is then publicized again when it is granted a patent. In other words, a utility patent application usually involves two public documents, and sometimes it is the PPA and sometimes it is the patent that is cited by a later patent, as illustrated in Fig. 1.

Since a patent and its PPA are both publications of the same patent application, the usual approach of considering just the citations to patents may miss a significant amount of information. Again using Fig. 1 as an example, the conventional method in tracing knowledge flow considers only citations from the two patents to 10244670, whereas those involving the more numerous PPAs are ignored.

As another example, 7657849[2] is Apple's first patent disclosing the slide-to-unlock function on all iPhone and iPad devices. The patent was cited 509 times up to 2018/12/31. Its PPA 2007/0150842, however, was even more frequently cited for 532 times. Among the 509 and 532 citations, 197 of them actually cited both. Then, in evaluating the impact or value of the slide-to-unlock patent, should the citation count be limited to those involving the patents only, or should those targeting the PPAs be included as well? The former (509) is clearly an underestimation compared to the latter (844 = 509 + 532-197).

Whether patent citations are used as proxies to their value or impact, or considered as channels for knowledge flow, overlooking citations to the patents' PPAs, if they are significant compared to their patent citations, may lead to distorted or erroneous analytic result. In addition, if citations to PPAs cannot be ignored, then, probably all citation-based patent bibliometric indicators should be generalized to include PPA citations. For example, the traditional patent indicators Originality and Generality proposed by Trajtenberg, Henderson, and Jaffe (1997)) respectively consider the degrees of concentration of the patent classification symbols from backward and forward citations. Then, to more accurately calculate the two indicators, not only patents' backward and forward citations, but also PPAs' backward and forward citations should be included. Furthermore, the backward and forward citations should not be restricted to those to and from other patents, and should include those to and from other PPAs as well.

[1] The content of the patent 10244670, titled "Electronic component, electric component manufacturing apparatus, and electronic component manufacturing method," can be accessed at: http://patft.uspto.gov/netacgi/nph-Parser?patentnumber=10244670.

[2] The content of the patent 7657849, titled "Unlocking a device by performing gestures on an unlock image," can be accessed at: http://patft.uspto.gov/netacgi/nph-Parser?patentnumber=7657849.

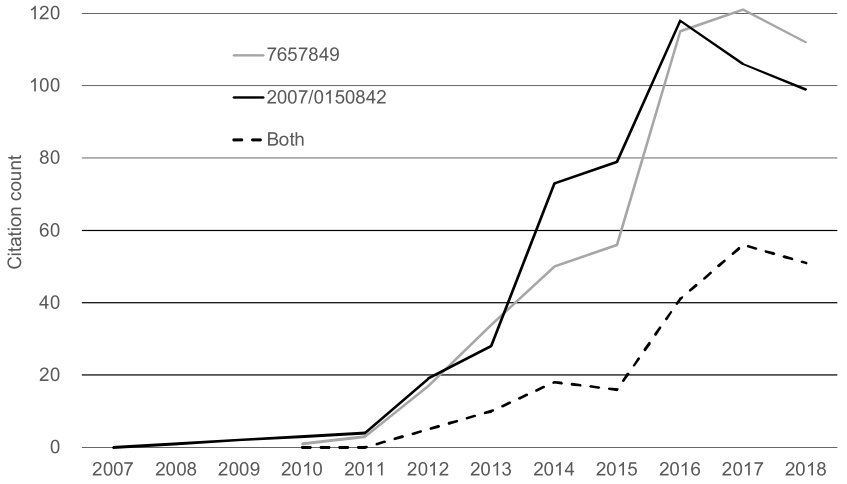

Fig. 2. Yearly citation counts of 7657849 and its PPA 2007/0150842 from 2007 to 2018.

2. Citing patent or published application

For patents or PPAs to be cited, they should be granted or published earlier so that they are "visible" to the examiners and/or applicants of later patents. Therefore, regarding the citations listed in Fig. 1, one may speculate that the two patents, but not their PPAs, were included because they were not early published, and the nine PPAs, but not their patents, were listed as they were not granted yet.

Most patent offices started early publication of patent applications after 2001 (e.g., USPTO published its first PPA on 2001/03/15). Therefore, utility patents filed before that time very possibly do not have a PPA. But for the patent 7906371[3] listed in Fig. 1, which was filed in 2008, it did have a PPA 2009/0294928. The examiner and/or applicant, however, chose to cite the patent, instead of its PPA, even though the PPA was published even earlier.

For patents after 2001, it is still possible that they are granted without being early published first. The other patent 9466545[4] is indeed such a case. There are a number of reasons for this exemption of early publication. Most patent offices are refrained from early publication when an application is subject to a secrecy order (cf. 35 U.S.C. § 122(b)(2)(A)). Some offices allow the omission of early publication if requested by the applicant (35 U.S.C. § 122(b)(2)(B)). These exceptions, however, are rare. According to an earlier study, there are 157,502 U.S. utility patents granted in 2007, 137,964 of them (about 88 %) do have corresponding PPAs (Kuan & Cheng, 2014). Therefore, for a patent filed within the last two decades, it is very possible that it has an associated PPA.

On the other hand, the PPA 2012/0133032 listed in Fig. 1 was later granted a patent 9111945[5], both of which were published several years before 10244670 was filed, and as such equally possible to be cited. However, it was the PPA, not the patent, that was cited by 10244670.

Some may argue that, once a patent is granted, examiners and/or applicants of later patents would stop citing the PPA and switch to cite the patent, implying that ignoring citations to the PPAs would not introduce significant bias. Again, there is no ground for this speculation. Fig. 2 shows the year-by-year numbers of citations to the slide-to-unlock patent granted in 2010 and to its PPA published in 2007. As illustrated, the citations to the PPA (black curve) actually continue to grow after the patent is granted, and there are even more citations to the PPA than to the patent (grey curve) between 2014 and 2016.

Fig. 2 also reveals that the simultaneous availability of the slid-to-unlock patent and its PPA for citation was indeed noted by some later examiners and/or applicants. They were, therefore, cited together by some patents granted after 2010 (dashed curve). There were, however, numerous other examiners and/or applicants chose to cite only one of them.

The strongest argument in leaving out citations to PPAs is that a PPA and its patent may contain different contents as the application may be amended to overcome the rejection by the patent office after the PPA is publicized. The PPA and its patent are as such distinct documents, justifying their separate consideration. Again using the slide-to-unlock patent and its PPA as examples, a word-for-word comparison indicates that their 13,500-word specifications are identical, except that the patent has an additional sentence of 40 words in the Summary section, and most of the patent's claims contain more details than those of the PPA.

[3] The content of the patent 7906371, titled "Semiconductor device and method of forming holes in substrate to interconnect top shield and ground shield," can be accessed at: http://patft.uspto.gov/netacgi/nph-Parser?patentnumber=7906371.

[4] The content of the patent 9466545, titled "Semiconductor package in package," can be accessed at: http://patft.uspto.gov/netacgi/nph-Parser?patentnumber=9466545.

[5] The content of the patent 9111945, titled "Package having ESD and EMI preventing functions and fabrication method thereof," can be accessed at: http://patft.uspto.gov/netacgi/nph-Parser?patentnumber=9111945.

Table 1
Summary citation statistics for empirical data.

	All	≤1 yr.	(1, 2] yrs.	(2, 3] yrs.	(3, 4] yrs.	>4 yrs.
# of Pats.	276,940	54,865	84,044	65,209	33,545	39,277
Avg. Pat. Cit.	2.34 (10.01)	2.58 (9.46)	2.25 (9.24)	2.24 (10.57)	2.22 (9.83)	2.47 (11.42)
Avg. PPA Cit.	6.60 (19.51)	2.90 (8.43)	4.36 (10.48)	6.17 (14.02)	8.66 (24.07)	15.54 (37.62)
Avg. Pat.∩PPA Cit.	0.49 (2.96)	0.51 (3.41)	0.46 (2.75)	0.46 (2.51)	0.51 (3.22)	0.56 (3.17)
Avg. Pat.∪PPA Cit.	8.45 (23.34)	4.97 (13.05)	6.14 (15.04)	7.95 (19.77)	10.37 (27.45)	17.46 (41.39)
Max. Pat. Cit.	434	359	356	434	359	431
Max. PPA Cit.	2852	414	401	477	2709	2852
Max. Pat.∩PPA Cit.	259	172	185	211	222	259
Max. Pat.∪PPA Cit.	2852	484	613	664	2709	2852
Spearman Coefficient	0.3830	0.3940	0.4130	0.4225	0.4232	0.4199
Pearson Coefficient	0.2982	0.3690	0.3630	0.4170	0.3070	0.3012
VCR	44.76 %	63.29 %	49.73 %	40.76 %	33.65 %	24.35 %
VCR (Pat. Cit. ≥ 1)	50.61 %	68.09 %	55.24 %	46.57 %	39.90 %	30.24 %

Could differences in the specification and/or the claims between a patent and its PPA lead people to cite only one of them? The authors believe that it is quite unlikely.

All patent offices strictly prohibit amendments that extend the original application. U.S. Patent Act clearly specifies that "[n]o amendment shall introduce new matter into the disclosure of the invention" (35 U.S.C. § 132(a)). The PPA and the patent, no matter how their application is amended, are both bounded by the scope of the patent application originally filed. On the other hand, all patent offices also require that the specification must describe the invention outlined in the claims in sufficient details (cf. 35 U.S.C. § 112(a)), and each claim should be adequately supported by the specification (cf. Nemec and Zelenock (2007)).

Patent applicants as such often incorporate more specific details from the specifications into the claims in amending their applications to overcome rejections, as in the case of the slide-to-unlock patent, but the specifications undergo little change. A patent and its PPA may have different sets of claims, but whatever disclosed in their claims should all be described in their specifications, whose contents are confined by the same original application.

To demonstrate that whether to cite a patent or its PPA should have little to do with their difference, another example is as follows. A patent 7193232[6] granted on 2007/03/20 has been cited 141 times up to 2018/12/31, yet its PPA 2004/0136494 receives much greater 725 citations within the same time span between 2007/03/21 and 2018/12/31. Their hugely distant citation counts, however, cannot be ascribed to their being different documents, because text comparison reveals that both have almost identical specifications and claims.

As discussed above, it is not clear why it is the patents, not their PPAs, or the other way around, that are cited, when both are available to the examiners or applicants of later patents. Little discussion can be found in the literature regarding how and why examiners and applicants make their decisions.

Even though accurate explanation is hard to come by, ignoring citations to PPAs indeed risks missing a significant amount of information or some vital data. This study, therefore, conducts a comprehensive empirical study to assess the impact of such omission by observing the citation counts received by patents and their PPAs from a large dataset. Specifically, we would like to see (1) how much depreciation would be incurred if citations to PPAs are ignored, (2) whether the time gaps between the PPAs and patents may affect the degree of depreciation, and (3) when both PPAs and their patents are both available for citation, whether people tend to cite patents instead of their PPAs.

3. Empirical analysis

This study collects 300,677 utility patents granted in 2014 from USPTO database.[7] Hall, Jaffe, and Trajtenberg (2001) indicated that citations take time to ramp up, and similar phenomenon indeed can be seen from Fig. 2. The year 2014 is chosen so that these patents are not too old and they are given 4–5 years to accumulate their citations. Then, 23,737 (7.9 %) of them are removed as having no corresponding PPAs to compare with. The final dataset includes 276,940 patent-PPA pairs.

As outlined in the "All" column of Table 1, the patents on the average receive 2.34 ("Avg. Pat. Cit.") citations from their grant dates in 2014 up to 2018/12/31. Their earlier published PPAs have more time to accumulate citations and, as expected, achieve a greater average 6.60 ("Avg. PPA Cit."). Their standard deviations are provided within parentheses.

The dataset is separated into five groups of comparable sizes, based on the time difference between each patent's grant date and its PPA's publication date converted into 365-day years. Statistics for these five groups up to 2018/12/31 are summarized in separate columns of Table 1. For example, the "(1,2) yrs." column covers patents that are granted more than

[6] The content of the patent 7193232, titled "Lithographic apparatus and device manufacturing method with substrate measurement not through liquid," can be accessed at: http://patft.uspto.gov/netacgi/nph-Parser?patentnumber=7193232.

[7] We actually find 326,038 patents, but remove 23,657 design patent, 1,072 plant patents, 626 reissue patents, and 6 Statutory Invention Registration (SIR) documents.

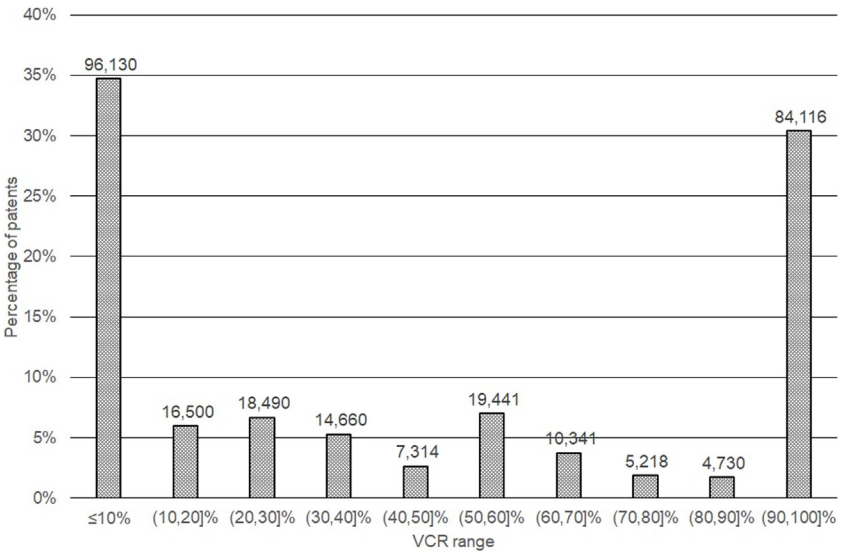

Fig. 3. VCR distribution of 276,940 patents.

365 days (1 yr.) later but no more than 730 days (2 yrs.) from when their PPAs are published. As illustrated, the average patent citations ("Avg. Pat. Cit.") ranges between 2.20 and 2.60 across these five groups. We speculate that the average patent citations do not vary significantly as the patents are still in the process of accumulating citations. As to the average PPA citations ("Avg. PPA Cit."), they increase from 2.90 to 15.54 as the PPAs are published further ahead in time (before 2014) and, therefore, are able to gather more citations.

3.1. Value capture rate

To investigate the impact of ignoring PPA citations to a patent, this study defines *Value Capture Rate* (VCR) as the ratio the patent citation count the patent's *combined citation* count:

$$VCR = \frac{\left|Patent\ Citations\right|}{\left|Combined\ Citatons\right|} = \frac{\left|Patent\ Citations\right|}{\left|Patent\ Citations \cup PPA\ Citations\right|}$$

$$= \frac{\left|Patent\ Citations\right|}{\left|Patent\ Citations\right| + \left|PPA\ Citations\right| - \left|Patent\ Citations \cap PPA\ Citations\right|}$$

A patent's combined citations is the union of its citations and the citations of its PPA, and this set should more realistically reflect the patent's impact or value. To calculate a patent's combined citation count, for example, if it is cited 10 times (i.e., |patent citations|), its PPA is cited 5 times (i.e., |PPA citations|), and 3 are common to the 10 and 5 citations (i.e., |patent citations∩PPA citations|), the combined citation count is 12 (= 10 + 5-3).

VCR indicates that how much of a patent's value is captured by the conventional approach that considers only patent citations. A 100 % VCR means that leaving out PPA citations does not lead to any depreciation. This condition may occur when the patent's PPA is not cited at all, or whatever cites the PPA also cites the patent. A 0 % VCR indicates the opposite and may occur when the patent has zero citation. The closer the VCR is to 0 % or 100 %, the more or less underrated the patent is. A special case occurs when neither a patent nor its PPA is cited, and there are 53,340 such patents. Their VCR is assumed to be 100 %, as no underestimation is involved. One may also define *Value Depreciation Rate* (VDR) as 1-VCR to show much a patent's value is overlooked by considering only patent citations.

For the 276,940 patent-PPA pairs, there are on the average 0.49 citations citing both a patent and its PPA ("Avg. Pat.∩PPA Cit.", hereinafter, *common citations*), 8.45 combined citations ("Avg. Pat.∪PPA Cit."), and 44.76 % VCR, meaning that, using only patent citations accounts for, on the average, merely about 45 % of the patent value. The average patent citations ("Avg. Pat. Cit.") do not vary much across the five groups of Table 1, the common citations, therefore, reveal no significant difference. The combined citations ("Avg. Pat.∪PPA Cit."), influenced by the average PPA citations ("Avg. PPA Cit."), increase from 4.97 to 17.546.

As real-life applications often concern only patents that are cited one or more times, this study simulates this condition by considering only those 139,518 of the 276,940 patents that are cited at least once. Their average VCR ("VCR (Pat. Cit. $\geq$ 1)") is only slightly improved to 50.61 %, suggesting that still about 50 % of patent value evades detection.

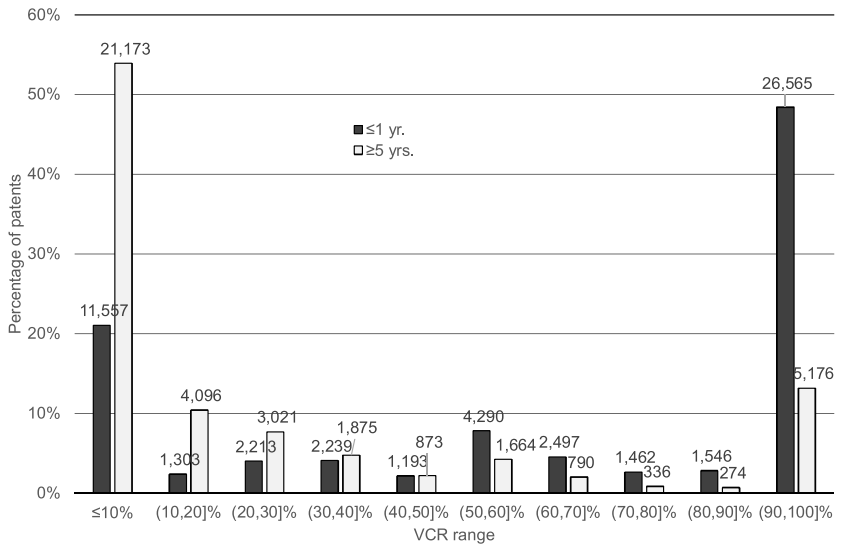

Fig. 4. VCR distributions for groups "≤1 yr." and ">4 yrs".

The degree of value depreciation actually may be worse. Fig. 3 provides the distribution of the 276,940 patent-PPA pairs divided into 10 groups according to their VCRs. As illustrated, there are two major groups of patents respectively located around 0 % and 100 %. For patent-PPA pairs within the 90∼100 % VCR interval (the rightmost group), a significant portion of them (53,340 out of 84,116, about 63 %) includes the no-citation-all-all patent-PPA pairs (i.e., patents and their PPAs have not been cited) whose VCR is assumed to be 100 %. If these pairs are excluded, the average VCR would certainly be even lower. On the other hand, the majority of the pairs having 0∼10 % VCRs (84,082 out of 96,130, about 87 %) is those having no patent citation but whose PPAs are cited at least once. They have 0 % VCR regardless of how many times their PPAs are cited. Further calculation shows that these PPAs actually have an average citation count 5.5 (maximum 563). If these PPA citations are not ignored, the deprecation should be even worse.

3.2. Time gap between patents and their PPAs

The above observation indicates that PPAs do accumulate more citations than their later publicized patents. If a PPA is published farther ahead in time, it would have a greater number of citations, and the patent would be more depreciated.

As illustrated in Table 1, the ">4 yrs." group, whose patents and PPAs having the greatest time gaps ranging from 5 to 14 years, indeed has the largest citation difference (15.54 vs. 2.47) and, therefore, the worst VCR 24.35 %. In contrast, for patents and PPAs in the "≤1 yr." group that are granted and published within 365 days (1 yr.), they reveal the lowest citation difference (2.90 vs. 2.58) and the best VCR 63.29 %.

This observation confirms the effect of publication time difference on the VCR. An interesting by-product is, as all five groups have comparable average citation counts (between 2.22 and 2.58) and common citation counts (between 0.46 and 0.56), it seems patents' citations are not affected by how long ago their PPAs are published. Having its PPA published farther in the past does not necessarily make the patent more probable for citation by later applicants and examiners. Spearman and Pearson's correlation coefficients between patent and PPA citation counts are calculated for these groups and included in Table 1 (all have p-values less than 2.2e-16). The groups of longer gaps do not reveal a correlation stronger than those of shorter gaps.

To see why there is such a huge difference between their VCRs, the VCR distributions of the two groups having the shortest and longest time gaps are depicted in Fig. 4. As illustrated, when patents and their PPAs are both granted and published within a short window, they have the highest average VCR 63.29 % because close to half of them achieves 100 % value capture. As to the group whose patents and PPAs are at least five years apart, more than 50 % has no more than 10 % capture rates because their PPAs very possibly have already accumulated sizable citations.

Table 1 also provides a number of maximum citation counts for these groups, and two extreme cases are manifested. One is the patent 8679650[8] and its PPA 2010/0092800, where the patent has just one citation whereas the PPA is cited 2709

[8] The patent 8679650, titled "Substrate for growing wurtzite type crystal and method for manufacturing the same and semiconductor device," was filed by the Japanese company Canon Kabushiki Kaisha, and respectively published and granted on 2010/04/15 and 2014/03/25, about 4 years apart. The content of the patent can be accessed at: http://patft.uspto.gov/netacgi/nph-Parser?patentnumber=8679650.

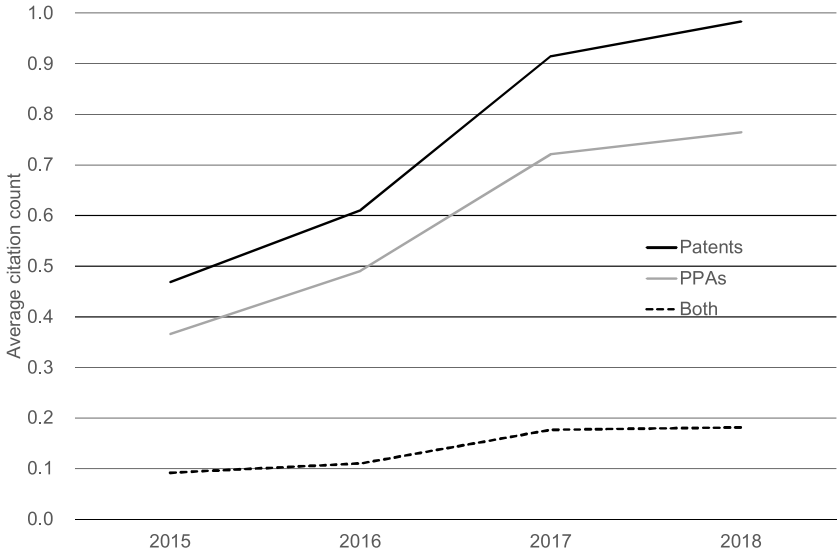

Fig. 5. Average citation counts of patents and PPAs issued and published within 2014/1∼2014/3 from 2015 to 2018.

times (including one also citing the patent). The other one is patent 8681077[9], which has mere two citations, whereas its PPA 2006/0208977 receives 2852 citations (including two also citing the patent). These two patents are all related to light emitting semiconductor devices. Text comparison again show that, despite having quite different sets of claims, both patents have almost identical specifications as their PPAs.

3.3. Patent and PPA citation growth

It is mentioned above that earlier PPAs do not make their patents more probable for citation, as revealed by the correlation between their citation counts. This seems to suggest that, when both a PPA and its patent are available for citation, they are cited individually and concurrently.

On one hand, for patents and their PPAs separated by less than a year and, therefore, they are at more equal footings, Table 1 indicates that both receive comparable citations (2.58 vs. 2.90) up to 2018/12/31.

To further investigate this phenomenon, the study selects 6,366 patents both published and granted in the first three months of 2014, so that the influence from PPAs' early publication is reduced as much as possible, while a reasonable sample size is attained. Their respective citations each year from 2015 to 2018 are depicted in Fig. 5. It is interesting to see that, for these four years, the patents actually receive a greater number of citations (black curve) each year than their PPAs (grey curve) do. Therefore, it is not necessarily true that PPAs are always cited more frequently than their patents; in this specific case, it is the patent citations that outnumber the PPA citations.

Another interesting phenomenon shown in Fig. 5 is that the patent and PPA citations grow in comparable manners. Even though PPAs do accumulate more citations, once the patents are issued, they and their PPAs seem to receive citations individually and concurrently.

4. Case study using main path analysis

The above empirical analysis reveals that, on the average, (1) citations to patents account for less than 50 % of all those to the patents and their PPAs combined together, indicating a significant underestimation to the value or impact of the patents, (2) the degree of depreciation is worse when the time gaps between patents and their PPAs are longer, and (3) people do not stop citing PPAs even after their patents are granted; PPAs individually and concurrently receive citations along with their patents.

Based on the above observations, one may immediately speculate that various patent bibliometric measures and analytic methods would be affected, if they involve the use of patent citations. This study, therefore, chooses to observe how overlooking citations to PPAs would affect the derivation of a representative trajectory from a patent citation network using Main Path Analysis (MPA).

[9] The patent 8681077, titled "Semiconductor device, and display device, driving method and electronic apparatus thereof," was filed by the Japanese company Semiconductor Energy Laboratory Co., Ltd., and respectively published and granted on 2006/09/21 and 2014/03/25, about 7.5 years apart. The content of the patent can be accessed at: http://patft.uspto.gov/netacgi/nph-Parser?patentnumber=8681077.

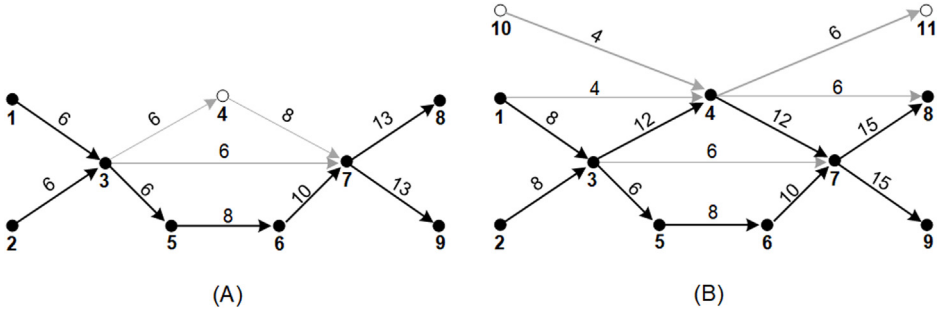

Fig. 6. Two fictitious citation networks.

4.1. Main path analysis

MPA is first used to derive a representative development trajectory for a scientific field using a citation network constructed from the field's research articles (Hummon & Dereian, 1989). This method is well-received and adopted in a wide range of bibliometric and technological management applications. For a few examples, MPA has been employed to detect technological changes and knowledge transformation (Lucio-Arias & Leydesdorff, 2008; Martinelli, 2012; Mina, Ramlogan, Tampubolon, & Metcalfe, 2007), to review a field's literature (Bhupatiraju, Nomaler, Triulzi, & Verspagen, 2012; Calero-Medina & Noyons, 2008; Colicchia & Strozzi, 2012; Harris, Beatty, Lecy, Cyr, & Shapiro, 2011; Liu, Lu, Lu, & Lin, 2013; Lu, Hsieh, & Liu, 2016), and to map technological development (Fontana, Nuvolari, & Verspagen, 2009; Park & Magee, 2017; Verspagen, 2007). The popular social network analysis software Pajek (Batagelj & Mrvar, 1998; De Nooy, Mrvar, & Batagelj, 2011) has built-in MPA functions.

MPA involves two key ingredients. Each arc of the network has to be assigned a weight related to its traversal count within the network. Then, a series of connected arcs is determined as the main path of the network. There are different weight assignment and path determination algorithms. This study chooses to use the search path link count (SPLC) algorithm (Hummon & Dereian, 1989), as it is suggested to be "closest to the knowledge diffusion scenario in science and technology development" (Liu, Lu, & Ho, 2019), and the global search method (Liu & Lu, 2012), as it often provides a single main path by selecting the chain of arcs having the highest sum of weights, which makes the subsequent comparison easier. Fig. 6 provides two fictitious citation networks (A) and (B) whose arc weights assigned using SPLC are shown besides the arcs, and main paths determined by the global search method include those black arcs.

The weight assignment algorithms available from Pajek all determine an arc's weight by its structural connectivity within the network (Hummon & Dereian, 1989), i.e., how many source or preceding nodes reaching and how many sink or succeeding nodes reached through the arc. For example, an arc's SPLC weight is its traversal count from all preceding nodes to all sink nodes. Therefore, the arc 3→4 of network (A) has a weight 6, as it may be traversed from the three preceding nodes (1–3) to the two sink nodes (8, 9). Similarly, the arc 3→4 of network (B), has a weight 12 as the same preceding nodes may traverse the arc three times to reach node 11, six times to node 8, and three times to node 9. By comparing networks (A) and (B), it can be see that, by increasing the connectivity of arcs 3→4 and 4→7 in network (B), their increased weights have promote them into an additional main path (1,2) →3→4→7→(8, 9) in (B), in addition to the original one (1,2) →3→5→6→7→(8, 9).

4.2. Empirical data

The Cancer Moonshot Patent Data published by USPTO (Cancer Moonshot Patent Data, 2016) for the promotion of cancer-related research and development is used to construct the citation networks due to its official status. This dataset includes 269,353 U.S. cancer-related PPAs and patents filed between 1963 to the first half of 2016, and each is categorized into one or more of eight fields: Drugs & Chemistry, Diagnostic & Surgical Devices, Radiation Measurement, Data Science, Food & Nutrition, Model Systems & Animals, Cells & Enzymes, and Other & Pre-classification. This study further chooses the field Cells & Enzymes, which focus on cellular and molecular biology, and gathers the field's patents that were granted after 2006/01/01. The year 2006 is picked because USPTO instituted the early publication of patent applications since 2001 and, for patents granted after 2006, they most likely have corresponding PPAs.

There are 31,071 patents that either themselves or their PPAs are cited at least once by other patents of the field, or that they cite one or more of the patents or their PPAs. Among them, there are 45,990 patent citations, and 93,315 patent and PPA citations. Again, the patent citations account for 49 % of all citations, conforming to what is observed in the previous section. The former is then used to construct a conventional patent citation network (PCN), and both are processed into a patent and PPA citation network (PPCN) where each patent and its PPA are represented by a single node.

This case study also reveals that combining the citations to patents and their PPAs together may not be trivial. In setting up the PPCN, the combination of patents and their PPAs may lead to cycles in the PPCN. As shown in Fig. 7(A), one scenario is that a patent E is granted earlier and is cited by a later-granted patent L ($E \rightarrow L$) while the later patent L's PPA, which is published even earlier than E, is cited by E (L's PPA$\rightarrow E$). As patent L and its PPA are merged in the PPCN, as shown in Fig. 7(B),

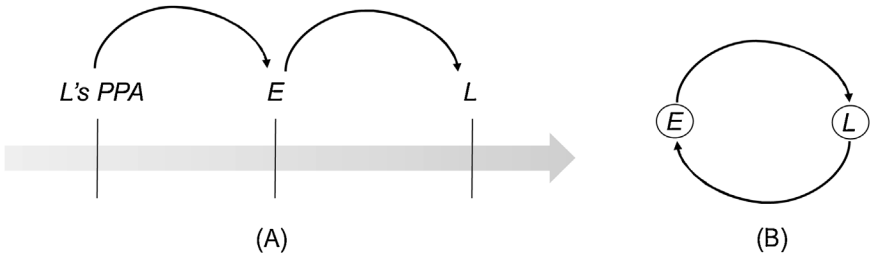

Fig. 7. A scenario causing cycle in PPCN.

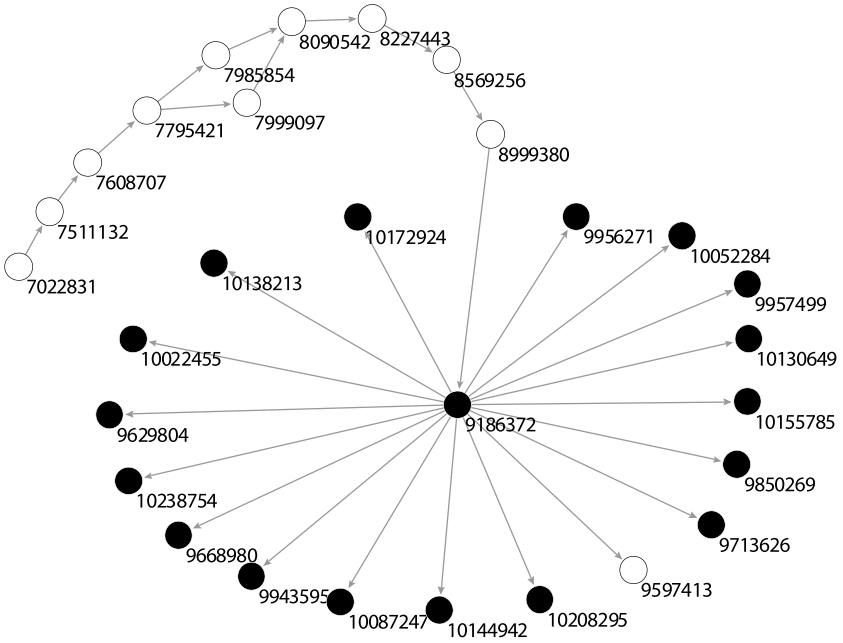

Fig. 8. Main path from the conventional PCN.

a cycle between nodes *E* and *L* is formed. To achieve an acyclic PPCN so that MPA may be conducted, every citation from a PPA to a patent causing a cycle is removed. This resolution is for simplicity's sake. Recently, Jiang, Zhu, and Chen (2019) have discussed alternative data structures and algorithms for MPA on cyclic citation networks. Finally, the PPCN includes 85,413 citations.

4.3. Main path comparison

For both PCN and PPCN, their main paths, derived using SPLC algorithm and global search method, are illustrated respectively in Figs. 8 and 9. Each node in the PCN is a patent whereas a PPCN node may a patent (if its PPA is not cited in the dataset), a PPA (if its patent is not cited), or a patent/PPA combination (if they both are cited), but they are all represented by their patent numbers in Fig. 9. The main path from the PCN depicts 29 nodes, and the one form the PPCN includes more numerous 71 nodes. Patents identified by both main paths are represented as black nodes.

There are 18 patents common to both main paths. 17 of them occur within the fan-out section near the end of each main path. They are not the representative ones as they are usually sink nodes and they are included in the main path simply because MPA fails to differentiate them due to their lack of citations. If these fanned-out patents are ignored, the two main paths share only one common patent, 9186372[10] .

The significant difference between the two main paths, both in terms of their length and content, is resulted from that the PPCN has become a structurally different network from the PCN, as the PCN includes 18,915 nodes and 45,990 arcs, whereas the PPCN involves 29,880 nodes and 85,413 arcs, even though they are derived from the same dataset and PPCN

[10] The content of the patent 9186372, titled "Split dose administration," can be accessed at: http://patft.uspto.gov/netacgi/nph-Parser?patentnumber=9186372.

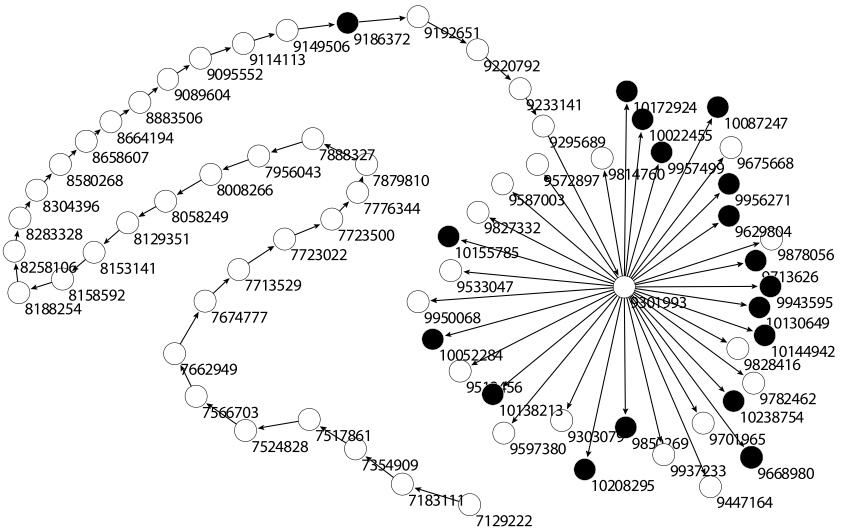

Fig. 9. Main path from the augmented PPCN.

is an extension of the PCN. This huge discrepancy suggests that analysts may actually obtain inaccurate analytic result by ignoring citations to the PPAs.

5. Summary and discussion

This study tries to address a question: do we miscount the patent citations? The citation count (i.e., the number of forward citations) of a patent is often considered an indication to the value or quality of the patent. However patents, specifically utility patents, are usually published 18 months after their applications are filed and before they are granted subsequently. These published patent applications (PPAs) and the corresponding patents disclose the same inventions, and are both citable as relevant prior art by the applicants or examiners of subsequent patent applications.

Most patent analysts, however, consider only the citations to the patents and ignore those to their PPAs. This omission may lead to erroneous analytic result, and this study assesses the impact of such omission by gathering more than 270,000 U.S. utility patents granted in 2014 and their PPAs, and compares their citation counts up to 2018. It is found that citations to patents, on the average, account for less than 50 % of those to the patents and their PPAs combined together, indicating a significant underestimation to the value or impact of the patents. The degree of depreciation would get worse when the time gaps between patents and their PPAs are longer, as the PPAs not only have accumulated citations for a longer period, but also individually, concurrently, and continuously receive citations after the patent is granted.

As to why later examiners/applicants do not show preference for issued patents while they are available for citations, the authors believe that the decision by the later examiners/applicants involves their experience, knowledge of the prior art, and examination/application practice. For example, an examiner may habitually cite a particular patent or PPA for applications related to a specific technological content. Randomness may also play a role; an applicant or examiner probably cites whatever come across first during his/her prior art search. One possible approach to investigate the motivation behind examiner/applicant decision is to separate pairs of patents and the PPAs into different categories and to explore what characteristics are jointly revealed by these categories. A feasible categorization is to divide patent-PPA pairs into those (1) whose patent gets most citations; (2) whose PPA gets most citations; and (3) whose patent and PPA both receive comparable numbers of citations.

This study further applies Main Path Analysis to a conventional citation network involving only citations to the patents and a second network supplemented with those to the PPAs, using USPTO Cancer Moonshot Patent Data. The main path derived from the augmented network is almost entirely different from that of the conventional network, suggesting a significantly undesirable impact if citations to PPAs are ignored.

There are a number of limitations and reminders to the findings of this study. Firstly, as the number of patent applications continues to grow worldwide, there are more citable prior documents, thereby causing an increasing trend of citations. This citation inflation phenomenon (Marco, 2007) suggests that earlier patents would suffer less depreciation than what is revealed here. This is confirmed by what is reported in Kuan and Cheng (2014) that used earlier patents for empirical analysis. On the other hand, future patents may subject to even greater degree of underestimation as more citations to the PPAs are overlooked, unless this citation inflation problem is addressed properly.

Secondly, since only utility patents have PPAs, the finding of this work is not applicable to other types of patents such as utility model patents and design patents. In addition, the degree of depreciation to utility patents may vary for non-U.S. patents. USPTO imposes an obligation on patent applicants to fully disclose all relevant prior art known to the applicants

(cf. Kuhn (2010)). Patent offices other than USPTO do not have such requirement and non-U.S. patents and PPAs, therefore, have relatively fewer citations. The depreciation caused by ignoring PPA citations is expected to be less severe for non-U.S. patents.

Another limitation is that, in order to observe more recent patents, this study chooses those granted in 2014 and calculates their citation counts up to 2018. This study, therefore, fails to investigate how patents and their PPAs accumulate their citations after five years. Observations for longer periods of time may be conducted in a future study.

Finally, despite the authors have provided a number of regulatory evidences, suggesting that patents and their PPAs are from the same inventions/applications, and their citation differences should have little to do with patents' being amended from the PPAs, some may still be unconvinced. The authors, however, believe that a concrete proof may probably never be attained as there is no information about which specific part of a patent or PPA a citation is made against.

The greatest contribution of this study lies in that it serves as a reminder to researchers about the vast amount of PPA citations. Ignoring them seems to be an awful waste, and means to harness this set of overlook data should be devised.

Additionally, the relation between a working paper (or, similarly, a conference paper) and its version later published in a journal seems to be similar to that between a PPA and its corresponding patent. Therefore, what is revealed in this work suggests that there may be some degree of depreciation to a journal paper if some part or all of it is first published as a working paper or in a conference proceeding. However, there are two major differences. Firstly, patent authorities stipulate that a patent application cannot be amended beyond what is disclosed in the original application, and patent examiners are tasked to strictly enforce this regulation. The working paper (or similar early publication) and its journal version are not governed by such requirement, even though they should be similar at least to some extent. The other difference is that PPA and its patent are explicitly linked as they are from the same patent application. The linkage between the working paper and its journal version, in contrast, is not so clear cut, and identification of such linkage is difficult.

This study may be extended in a number of directions. In addition to what is mentioned above, another possibility is to investigate, as many patent analytic works use patent families as their subjects, how families including and excluding PPAs as family members differ in their integrally counted citations.

Author contributions

Chung-Huei Kuan: Conceived and designed the analysis, Collected the data, Contributed data or analysis tools, Performed the analysis, Wrote the paper.

Dar-Zen Chen: Conceived and designed the analysis, Collected the data, Contributed data or analysis tools, Wrote the paper.

Mu-Hsuan Huang: Conceived and designed the analysis, Contributed data or analysis tools, Performed the analysis, Wrote the paper.

Acknowledgments

The authors express their appreciation to the anonymous reviewers for their careful reading of the manuscript and their helpful comments and suggestions.

This work was financially supported by the Center for Research in Econometric Theory and Applications (Grant no. 108L900204) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan, and by the Ministry of Science and Technology (MOST), Taiwan, under Grant Nos. MOST 102-2221-E-011-051-, MOST 108-3017-F-002-003-.

References

Batagelj, V., & Mrvar, A. (1998). Pajek - program for large network analysis. *Connections*, *21*(2), 47–57.

Bhupatiraju, S., Nomaler, O., Triulzi, G., & Verspagen, B. (2012). Knowledge flows—Analyzing the core literature of innovation, entrepreneurship and science and technology studies. *Research Policy*, *41*(7), 1205–1218.

Calero-Medina, C., & Noyons, E. C. M. (2008). Combining mapping and citation network analysis for a better understanding of the scientific development: The case of the absorptive capacity field. *Journal of Informetrics*, *2*(4), 272–279.

Colicchia, C., & Strozzi, F. (2012). Supply chain risk management: A new methodology for a systematic literature review. *Supply Chain Management: An International Journal*, *17*(4), 403–418.

(2016). *Cancer moonshot patent data*.. Retrieved from. https://www.uspto.gov/learning-and-resources/electronic-data-products/cancer-moonshot-patent-data

De Nooy, W., Mrvar, A., & Batagelj, V. (2011). . *Exploratory social network analysis with Pajek* (Vol. 27) Cambridge University Press.

Figueiredo, O., Guimarães, P., & Woodward, D. (2015). Industry localization, distance decay, and knowledge spillovers: Following the patent paper trail. *Journal of Urban Economics*, *89*, 21–31.

Fontana, R., Nuvolari, A., & Verspagen, B. (2009). Mapping technological trajectories as patent citation networks. An application to data communication standards. *Economics of Innovation and New Technology*, *18*(4), 311–336.

Hall, B. H., Jaffe, A. B., & Trajtenberg, M. (2001). *The NBER patent citation data file: Lessons, insights and methodological tools (No. w8498)*. National Bureau of Economic Research.

Harris, J. K., Beatty, K. E., Lecy, J. D., Cyr, J. M., & Shapiro, R. M. (2011). Mapping the multidisciplinary field of public health services and systems research. *American Journal of Preventive Medicine*, *41*(1), 105–111.

Hummon, N. P., & Dereian, P. (1989). Connectivity in a citation network: The development of DNA theory. *Social Networks*, *11*(1), 39–63.

Ji, J., Barnett, G. A., & Chu, J. (2019). Global networks of genetically modified crops technology: A patent citation network analysis. *Scientometrics*, 1–26.

Jiang, X., Zhu, X., & Chen, J. (2019). Main path analysis on cyclic citation networks. *Journal of the Association for Information Science and Technology*, http://dx.doi.org/10.1002/asi.24258
Karvonen, M., & Kässi, T. (2013). Patent citations as a tool for analysing the early stages of convergence. *Technological Forecasting and Social Change*, *80*(6), 1094–1107.
Kim, D. H., Lee, B. K., & Sohn, S. Y. (2016). Quantifying technology-industry spillover effects based on patent citation network analysis of unmanned aerial vehicle (UAV). *Technological Forecasting and Social Change*, *105*, 140–157.
Kuan, C. H., & Cheng, H. J. (2014). Do we miscount patent citations? An empirical study on the impact of overlooking the citations to a patent's pre-grant publication. In *2014 IEEE International Conference on Industrial Engineering and Engineering Management* (pp. 1034–1037).
Kuhn, J. M. (2010). Information overload at the US Patent and Trademark Office: Reframing the duty of disclosure in patent law as a search and filter problem. *Yale Journal of Law & Technology*, *13*, 89–141.
Li, Y. A. (2014). Borders and distance in knowledge spillovers: Dying over time or dying with age?—Evidence from patent citations. *European Economic Review*, *71*, 152–172.
Liu, J. S., & Lu, L. Y. (2012). An integrated approach for main path analysis: Development of the Hirsch index as an example. *Journal of the American Society for Information Science and Technology*, *63*(3), 528–542.
Liu, J. S., Lu, L. Y., & Ho, M. H. C. (2019). A few notes on main path analysis. *Scientometrics*, *119*(1), 379–391.
Liu, J. S., Lu, L. Y., Lu, W. M., & Lin, B. J. Y. (2013). Data envelopment analysis 1978-2010: A citation-based literature survey. *OMEGA: The International Journal of Management Science*, *41*(1), 3–15.
Lu, L. Y., Hsieh, C. H., & Liu, J. S. (2016). Development trajectory and research themes of foresight. *Technological Forecasting and Social Change*, *112*, 347–356.
Lucio-Arias, D., & Leydesdorff, L. (2008). Main-path analysis and path-dependent transitions in HistCite (TM)-based histograms. *Journal of the American Society for Information Science and Technology*, *59*(12), 1948–1962.
Marco, A. C. (2007). The dynamics of patent citations. *Economics Letters*, *94*(2), 290–296.
Martinelli, A. (2012). An emerging paradigm or just another trajectory? Understanding the nature of technological changes using engineering heuristics in the telecommunications switching industry. *Research Policy*, *41*(2), 414–429.
Mina, A., Ramlogan, R., Tampubolon, G., & Metcalfe, J. S. (2007). Mapping evolutionary trajectories: Applications to the growth and transformation of medical knowledge. *Research Policy*, *36*(5), 789–806.
Murata, Y., Nakajima, R., Okamoto, R., & Tamura, R. (2014). Localized knowledge spillovers and patent citations: A distance-based approach. *The Review of Economics and Statistics*, *96*(5), 967–985.
Nemec, D. R., & Zelenock, E. J. (2007). Rethinking the role of the written description requirement in claim construction: Whatever happened to possession is nine-tenths of the law. *Minnesota Journal of Law, Science & Technology*, *8*, 357–408.
Noailly, J., & Shestalova, V. (2017). Knowledge spillovers from renewable energy technologies: Lessons from patent citations. *Environmental Innovation and Societal Transitions*, *22*, 1–14.
Park, H., & Magee, C. L. (2017). Tracing technological development trajectories: A genetic knowledge persistence-based main path approach. *PloS One*, *12*(1), e0170895.
Trajtenberg, M., Henderson, R., & Jaffe, A. (1997). University versus corporate patents: A window on the basicness of invention. *Economics of Innovation and New Technology*, *5*(1), 19–50.
Verspagen, B. (2007). Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research. *Advances in Complex Systems*, *10*(1), 93–115.

Chung-Huei Kuan is an assistant professor of the Graduate Institute of Patent at National Taiwan University of Science and Technology, Taipei, Taiwan. His research interests include patent bibliometrics, patent information mining and analysis, and practices in patent prosecution, patent specification drafting, and patent/technology transfer and licensing.

Dar-Zen Chen is a professor of the Department of Mechanical Engineering and Institute of Industrial Engineering at National Taiwan University, Taipei, Taiwan. His research interests include intellectual property management, patentometrics, competitive analysis, robotics, automation, kinematics, and mechanism design. He also leads the Intellectual Property Analysis & Innovative Design Laboratory (IAID).

Mu-Hsuan Huang is a chair professor of the Department of Library and Information Science at National Taiwan University, Taipei, Taiwan. Her early research focused on information retrieval and information behavior, and turned to bibliometrics, science and technology policy, intellectual property, and patent information for late years. She is also the project investigator of Performance Ranking of Scientific Papers for World Universities (NTU Ranking).