Do We Miscount Patent Citations? An Empirical Study on the Impact of Overlooking the Citations to a Patent's Pre-grant Publication

C.-H. Kuan[1*], H.-J. Cheng[1]
[1]Graduate Institute of Patent, National Taiwan University of Science and Technology, Taipei, Taiwan ROC
([*]maxkuan@mail.ntust.edu.tw)

Abstract - **Utility patent applications are usually published 18 months after they are filed and before patents are actually issued. These so-called pre-grant publications and their corresponding issued patents are both cited individually and concurrently by the applicants or examiners of subsequent patent applications as relevant prior art. Most patent analysts however overlook the citations to the pre-grant publications and consider only those to the issued patents. This study assesses the impact of such omission by comparing the citations to about 140,000 US utility patents and their pre-grant publications. The statistics shows that 70% of the patents are underestimated by various degrees if the citations to their pre-grant publications are ignored, suggesting that analyst should combine the citations to the pre-grant publications and to the patents together when evaluating patents or conducting patent citation analysis.**

Keywords – **Patent, pre-grant publication, citation**

I. INTRODUCTION

Patents assigned to competitors are an important source of technological intelligence and patent bibliometrics has offered a wealth of investigation tools. Among them a patent's citation counts (i.e., the number of forward citations, or the number of times the patent is referenced or cited as relevant prior art by the applicants or examiners of subsequent patent applications) is one of the earliest patent bibliometric indicators after Narin pointed out its significant similarity to paper citation counts in his pioneering work [1]. A comparison of the two can be found in [2].

Patent citations are actually more objective than paper citations because citations to patents are produced not only by the applicants but also by the patent examiners who can be considered as objective third parties; and self-citation is rare since there is little benefit for an applicant to cite his or her prior applications or patents in establishing the application's patentability [3].

In addition to the usage as arithmetic counts, patent citations, including both forward citations and backward citations (i.e., the prior public documents referenced or cited by the applicants or examiners of the patent applications under examination) are also commonly utilized in various patent analysis works such as mapping technological trajectories (cf. [4][5]), detecting technological changes (cf. [6][7]), assessing knowledge spillover (cf. [8]), monitoring science-technology interaction (cf. [9]) etc.

However an often overlooked difference between patents and papers is that a patent application usually undergoes an *early publication process* before the patent is finally issued by the authority or before the patent application is given up by the applicant after failing to manifest the patent application's patentability to the authority.

The early publication process is a common practice for authorities across various nations and regions. For example, U.S. Patent Act (35 U.S.C.S § 122(b)) specifies that, "each application for a patent shall be published … promptly after the expiration of a period of 18 months from the earliest filing date for which a benefit is sought under this title." This early published patent application is referred to as *pre-grant publication* (PGPub) by United States Patent and Trademark Office (USPTO). There are indeed exceptions that an application is not early published if the application is (i) no longer pending; (ii) subject to a secrecy order; (iii) a provisional application; (iv) an application for a design patent; or (v) requested by the applicant. These exceptions are rare and for utility patent applications, which are the most common type of patent applications, if a patent is indeed issued, it is very possible that there is a corresponding PGPub. According to our empirical study detailed in a following section, there are 157,502 utility patents issued in 2007, 19,538 (about 12%) of them do not have corresponding PGPubs mostly due to the requests by the applicants (i.e., exception (iv)). Please note that, unlike the U.S. regulation, the early publication of an application cannot be avoided by applicant request in most regions or nations.

Patent analysts often consider only the citations to the issued patents and overlook the citations to their PGPubs. The reasons behind this omission are not clear and, to our best knowledge, no related discussion can be found in the literature. We speculate that it is due to the PGPubs, unlike the patents, are not granted a property right yet, and some of the bibliometric information of the PGPubs is not as complete and stable (e.g., the PGPubs do not have information about backward citations). Another possible reason is that, as described above, some patents do not have corresponding PGPubs. It is also possible that the citations to the patents and the PGPubs have to be queried separately and this adds to the analytic workload.

However, issued patents and their PGPubs are both public documents and they can be cited by applicants or examiners of subsequent patent applications individually and concurrently. This scenario can be observed from a sample case in the following section.

978-1-4799-6410-9/14/$31.00 ©2014 IEEE

II. A SAMPLE CASE

A patent application, titled "Sound tube tuned multi-driver earpiece," is randomly picked from USPTO database, which was published on 2006/06/22 (i.e., the publication date) with publication number 2006/0133636 (hereinafter, the sample PGPub), and was granted about one and half years later on 2008/01/08 (i.e., the issued date) with patent number 7,317,806 (hereinafter, the sample patent).

As summarized in Table I, there are 14 patents citing the sample PGPub and there are only 8 patents citing the sample patent (data collected on 2014/01/15). Both groups of patents are sorted according to their filing dates and arranged in the middle and rightmost columns, respectively.

A horizontal line in Table I between the Nos. 6 and 7 citing patents indicates the issued date (i.e., 2008/01/08) of the sample patent. We can see that, even after the sample patent is issued, people continue to cite the sample PGPub concurrently despite the presence of the sample patent, and there are actually more citations (9) to the sample PGPub than those (7) to the sample patent. The No. 2 patent is the only one that cites the sample patent but is filed before the sample patent is issued. This is because the No. 2 patent's lengthy examination process has spanned across the sample patent's issued date, and its examiner has the chance of locating the sample patent during the examination process.

Using this sample case as example, the sample patent may be significantly underestimated if only its 8 citations are considered and the 14 citations to the sample PGPub are ignored. Similarly, if an analyst tries to determine how a technology evolves based on patents and their forward citations, the analyst may very possibly derive an incorrect trajectory if some crucial citations to the PGPubs are not included and considered.

TABLE I
PATENTS CITING THE SAMPLE PGPUB AND THE SAMPLE PATENT

		Patents citing the sample PGPub		Patents citing the sample patent	
No.	Filing Date	Patent No.	Issued. Date	Patent No.	Issued Date
1	2007/03/27	8,194,911	2012/06/05		
2	2007/06/13			8,170,249	2012/05/01
3	2007/08/28	8,098,854	2012/01/17		
4	2007/08/30	8,135,163	2012/03/13		
5	2007/09/28	8,290,187	2012/10/16		
6	2007/11/05	8,300,871	2012/10/30		
7	2008/10/31	8,447,059	2013/05/21		
8	2008/12/10	8,238,596	2012/08/07		
9	2008/12/17	8,189,804	2012/05/29		
10	2009/01/11	8,509,468	2013/08/13		
11	2009/03/11	8,311,259	2012/11/13		
12	2009/03/27	8,213,645	2012/07/03		
13	2009/11/17			8,116,502	2012/02/14
14	2009/12/17	8,116,502	2012/02/14		
15	2010/07/09			8,538,061	2013/09/17
16	2010/07/09			8,548,186	2013/10/01
17	2010/07/09			8,549,733	2013/10/08
18	2010/10/25			8,437,489	2013/05/07
19	2011/08/04	8,611,969	2013/12/17		
20	2011/08/04	8,625,834	2014/01/07		
21	2011/12/09			8,567,555	2013/10/29
22	2012/01/03			8,488,831	2013/07/16

One may argue that the content of a PGPub may be different from that of its corresponding patent, as the applicant may apply amendment during the patent examination process so as to overcome the rejection by the authority and, as such, the PGPub and its patent are not a same document, thereby justifying their citations being counted separately. However, we have compared the sample PGPub and the sample patent word for word, and they are completely identical. Additionally, U.S. Patent Act (35 U.S.C.S. § 132(a)) clearly specifies that "[n]o amendment shall introduce new matter into the disclosure of the invention." The so-called *new matter* refers to newly added material not supported by the patent application at the time of filing. In other words, the contents of the PGPub and the patent should both be bounded by what is disclosed at the time of filing. Therefore, it is dubious that a patent application can be amended during the patent examination process to such an extent that some piece of information can only be found and cited in one of its PGPub and patent, but not in the other. However, we have to admit that at the moment we cannot prove this speculation.

III. EMPIRICAL DATA

In order to assess the impact of overlooking the citations to the PGPubs, we decide to use the U.S. utility patents issued in the year 2007. The year 2007 is chosen because, according to an empirical study by Hall, Jaffe, and Trajtenberg [10], a target patent are most frequently cited by subsequent patents issued after the target patent was issued for 5 to 7 years, and after 7 years, the number of citations gradually drops (specifically see Fig. 11 of Hall, Jaffe, and Trajtenberg [10]). In other words, the year 2007 seems to be a reasonable choice as the year's utility patents are given a fair period of time to accumulate their citations, and comparing their citation counts with those of their PGPubs, which are out in the public for a longer period of time, is expected not to be seriously biased.

We have written a program to find 157,502 utility patents issued in 2007 from USPTO databases, and 137,964 patents of them have corresponding PGPubs. We then gather the citation counts to the 137,964 patents and their PGPubs, respectively, by running another program to query the USPTO databases. The citations are limited to those occurring before and on 2013/12/31.

IV. STATISTICS

To gain more insights to the impact of overlooking the citations to the PGPubs, Tables II provides statistics for the citation counts to the 137,904 patents and their PGPubs, respectively. As illustrated, the 137,904 patents

TABLE II
STATISTICS OF CITATIONS TO 137,964 PGPUBS AND PATENTS

	PGPubs	Patents
Total citation count	705,714	696,405
Avg. citation count*	5.12	5.05
Max. citation count**	405	392
Max. difference	360	363

*The standard deviations are 11.92 and 10.72, respectively.
**Their corresponding patent and PGPub receive 45 and 29 citations only.

have received fewer overall and average citations than their PGPubs. This rough data have already suggested that an analyst should be cautious about ignoring PGPub citations.

In an extreme case, a PGPub (No. 2004/0136494) actually receives 8 times more citations than its patent, as shown in Table II. The PGPub receives as high as 405 citations whereas its corresponding Patent (No. 7,193,232) receives only 45 citations. These numbers are obtained for citations occurring up to 2013/12/31. As of today (2014/05/31), PGPub No. 2004/0136494 receives 446 citations and Patent No. 7,193,232 receives 60 citations, indicating that the PGPub accumulates citations also at a faster rate (41 vs. 15). Interestingly, similar extreme case can also be found for patents. US Patent No. 7,297,977 receives 392 citations whereas its PGPub No. 2005/0199960 receives only 29 citations. In this case, the patent is also cited at a faster rate (from 392 on 2013/12/31 to 547 on 2014/05/31) than that of the PGPub (from 29 to 35). These two pairs of patents and PGPubs are also the ones producing the greatest citation differences 360 (=405-45) and 363 (=292-29).

We now assume an analyst's point of view so as to further investigate the impact of ignoring the PGPub citations. To do this, we define a patent's *PGPub Citation Share* as a percentage ratio of its PGPub citation count to the sum of citation counts of the patent and its PGPub. A 100% PGPub Citation Share indicates that the patent has no citation at all whereas its PGPub has at least one citation. In this case, the patent is completely underestimated. On the other hand, a 0% PGPub Citation Share indicates that the PGPub has no citation whereas its patent is cited at least once. In other words, the closer the PGPub Citation Share is to 100% or 0%, the more or less underestimated the patent is, if its PGPub citations are ignored. When the patent and PGPub citation counts are both zero, the PGPub Citation Share is assumed to be 0% since, in this special case, ignoring PGPub citation does not cause underestimation to the patent.

Fig. 1 is a histogram showing the distribution of the PGPub Citation Shares of all 137,964 patents by dividing them into 10 intervals. As illustrated, the leftmost bar represents those patents having PGPub Citation Shares less or equal to 10%. Its blue section represents the 41,087 patents having PGPub Citation Shares equal to 0%. For these 30% (=41,087/137,964) of all patents, ignoring PGPub citations does not lead to any underestimation. On the other hand, the remaining 70% of the patents are underestimated by various degrees if their corresponding

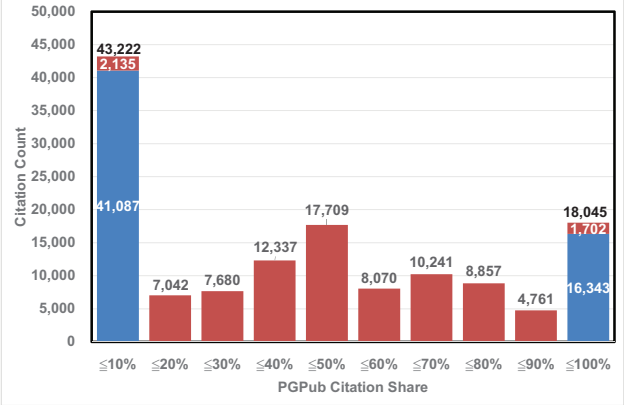

Fig. 1. Distribution of PGPub Citation shares of all 137,964 patents.

PGPub citations are ignored. This observation again strongly suggests that an analyst should exercise care when evaluating patents using their citations.

We can also see from Fig. 1 that there are 49,974 (=8,070+10,241+8,857+4,761+18,045) patents having PGPub Citation Shares greater than 50%. In other words, for these 36% (=49,974/137,964) of all patens, their PGPub citation counts account for more than 50% of the total citations to both the patents and the PGPubs, and therefore they would be evaluated to only one half of their real values by ignoring their PGPub citations. In the worst case, the rightmost bar represents those patents having PGPub Citation Shares above 90% and its blue section represents the 16,343 patents having the PGPub Citation Shares equal to 100%. For these 12% (=16,343/137,964) of the patents, they are completely underestimated if the PGPub citations are ignored.

V. CONCLUSION AND FUTURE WORK

In this study we have compared the citation counts to 137,964 U.S. utility patents with those to their PGPubs. For these patents, we find that, if the citations to their PGPubs are ignored, about 70% of the patents would be underestimated to various extents, about 36% of the patents would be significantly underestimated, and about 12% of the patents would be completely underestimated.

Ignoring PGPub citations therefore would be a risky choice by an analyst. The safest approach would be to consider a patent and its PGPub as a single entity and to combine their citations altogether. On the other hand, our observation also suggests that people should be cautious about a patent analysis work if the analyst adopts the common practice of ignoring the PGPub citations without assessing the impact of such omission in advance.

What we fail to investigate in this paper is that it seems, as time advances, the citations to the PGPubs would gradually drop, and the longer a patent has been issued,

the smaller its PGPub Citation Share is, and the weaker the impact of ignoring PGPub citations is. This speculation is implicitly revealed in Table I. Even though there are still more citations to the sample PGPub after the sample patent is issued, we can see that most citations to the sample patent occur recently whereas most PGPub citations occur earlier. A future extension of this study therefore is to collect patents issued in earlier and more recent years so as to verify our speculation. If our speculation is confirmed, the impact of ignoring PGPub citations would be more limited for senior patents. On the other hand, for young patents issued recently, citations to their PGPubs should not be ignored easily.

Additionally, even though we claim that it is unlikely a citation to a patent would reference some matter not contained in its PGPub, we have not compared the 137,964 patents word for word against their PGPubs. In addition to the regulation against incorporating new matter during the patent examination process mentioned in a previous section, we further speculate that the citation by an applicant or examiner towards a patent or its PGPub is probably accidental. For example, an examiner cites a PGPub probably because he or she happens to find the PGPub first, and does not bother to check whether the PGPub is granted or not. Similarly, an applicant cites a patent probably because he or she is aware of this document and whether there is a corresponding PGPub is irrelevant to the applicant. An indirect evidence to this speculation is that we find it is very rare that a patent and its PGPub are both cited by a same subsequent patent application or patent. However, this speculation still requires further investigation.

ACKNOWLEDGMENT

This study was supported by the Ministry of Science and Technology, Taiwan, ROC, under Grant No. NSC 102-2221-E-011-051.

REFERENCES

[1] F. Narin, “Patent Bibliometrics,” *Scientometrics*, vol. 30, no. 1, pp. 147–155, 1994.

[2] M. Meyer, “What is special about patent citations? Differences between scientific and patent citations,” *Scientometrics*, vol. 49, no. 1, pp. 93–123, 2000.

[3] C.-H. Kuan, M.-H. Huang, and D.-Z. Chen, “Locating the Key Competitors: A New Tool for Technology Manager,” in *Proc. of the 2012 IEEE International Conference on Industrial Engineering and Engineering Management*, *IEEM´12*, Hong Kong, China, pp. 671–676.

[4] R. Fontana, A. Nuvolari, and B. Verspagen, “Mapping technological trajectories as patent citation networks. An application to data communication standards,” *Econ. Innov. New Tech.*, vol. 18, no. 4, pp. 311–336, 2009.

[5] B. Verspagen, “Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research,” *Adv. Complex Syst.,* vol. 10, no. 1, pp. 93–115, 2007.

[6] D. Lucio-Arias, and L. Leydesdorff, “Main-path analysis and path-dependent transitions in HistCite (TM)-based historiograms,” *J. Am. Soc. Inf. Sci. Tec.,* vol. 59, no. 12, pp. 1948–1962, 2008.

[7] A. Martinelli, “An emerging paradigm or just another trajectory? Understanding the nature of technological changes using engineering heuristics in the telecommunications switching industry,” *Res. Policy,* vol. 41, no. 2, pp. 414–429, 2012.

[8] A. B. Jaffe, M. Trajtenberg, and R. Henderson, “Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations,” *Q. J. Econ.*, vol. 108, pp. 577–598, 1993.

[9] R. J. Tijssen, “Global and domestic utilization of industrial relevant science: patent citation analysis of science-technology interactions and knowledge flows,” *Res. Policy*, vol. 30, no. 1, pp. 35–54, 2001.

[10] B. H. Hall, A. B. Jaffe, and M. Trajtenberg, “The NBER patent citation data file: Lessons, insights and methodological tools,” No. w8498, National Bureau of Economic Research, 2001.