A Preliminary Study on the Difference between the Citation Counts of Issued Patents and Their Corresponding Pre-Grant Publications

Chung-Huei Kuan, Hsiang-Jui Cheng
Graduate Institute of Patent, National Taiwan University of Science and Technology, Taipei, Taiwan, ROC

Abstract--**This study tries to address a basic 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 issued subsequently. These so-called pre-grant publications 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 pre-grant publications. This omission may lead to erroneous analytic result as a pre-grant publication has its own citations in parallel with its corresponding patent. This study assesses the impact of such omission by using empirical data from United States patent database. The result shows that citations to the pre-grant publications can be significantly more than those to the patents, and an analyst should not ignore the citations to the pre-grant publications when evaluating patents or conducting patent citation analysis.**

I. INTRODUCTION

A patent for an invention is the grant of a property right to the applicants of the patent. A patent's citation count (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) has long been considered as an indication to a qualitative feature of the patent. This qualitative feature is given different names by various researchers, such as the patent's quality, value, impact, or importance of the patent (cf. [3][5][12]). The patent citation count is actually one of the earliest patent bibliometric indicators after Narin pointed out its significant similarity to paper citation count in his pioneering work [10]. A comparison of the two can be found in [8].

Other than simple 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) have become a valuable source of information for various bibliometric applications such as mapping technological trajectories (cf. [1][9]), detecting technological changes (cf. [6][7]), assessing knowledge spillover (cf. [4]), monitoring science-technology interaction (cf. [11]) etc., to name just a few.

However a major but often ignored difference between patents and papers is that an application for a patent (i.e., the patent application) will undergo an "early disclosure" process before it is granted by the authority. This early disclosure process is designed so that the public are able to access the content of an invention seeking patent protection earlier, and the early disclosed patent applications serve important notices to the public of the potential liability for infringing these inventions if patents are indeed issued in the future.

Taking United States of America as example, according to U.S. Patent Act (35 U.S.C.S § 122(b)), "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. " The same article also specifies that an application shall not be published if that application is (i) no longer pending; (ii) subject to a secrecy order; (iii) a provisional application; or (iv) an application for a design patent. In other words, a pending utility or plant patent application is published 18 months after filing unless it is subject to a secrecy order. There are some additional exceptions such as that, at the request of the applicant, an application may be published earlier than the end of such 18-month period.

These published patent applications, referred to as pre-grant publications (PGPubs) by United States Patent and Trademark Office (USPTO), are public documents and they can be cited by applicants or examiners of subsequent patent applications. Moreover, their citations may occur of course before they are granted, and even after they are granted. On the other hand, their corresponding patents are also citable public documents after they are issued. Therefore a patent application's PGPub and patent can be cited individually and in parallel by the subsequent patent applications.

A patents and its corresponding pre-grant publication disclose the same invention, but analysts often consider only the citations to the patents and ignore the citations to their pre-grant publications. The reason behind this omission is not clear. It is speculated that it is due to the PGPub, unlike the patent, is not granted a property right yet, and the bibliometric information of the PGPub is not as complete and stable as the patent (e.g., the PGPub does not have information about backward citations). However this omission may lead to erroneous analytical result. For example, if a patent is cited 5 times whereas its PGPub is cited 30 times, an analyst may seriously underestimate the patent's qualitative feature by only considering the 5 citations to the patent. If an analyst tries to determine how a technology evolves based on forward citations, he or she may very possibly derive incorrect result if some vital citations to the PGPubs are ignored.

Since omitting the citations to the PGPubs is a common practice, this study therefore tries to address a basic question: do we miscount patent citations by leaving out citations to the PGPubs? To answer this question, this study assesses the

impact of such omission using empirical data from United States patent database.

II. AN EXEMPLARY CASE

To gain some initial understanding of the individual and parallel accumulation of the citations to an application's PGPub and its subsequent issued patent, an exemplary case is provided as follows.

A patent application, titled "Sound tube tuned multi-driver earpiece," is randomly picked from USPTO public databases. The patent application was published on 2006/06/22 (i.e., the publication date) with publication number 2006/0133636 (hereinafter, the sample PGPub), and was granted on 2008/01/08 (i.e., the issued date) with patent number 7,317,806 (hereinafter, the sample patent) about one and half years later. It was found on January 15, 2014 that there are 8 patents citing the sample patent whereas there are 14 patents citing the sample PGPub. The information about these 22 citing patents, sorted ascendingly according to their filing dates, is summarized in Table 1.

Table 1 shows that the sample PGPub actually receives significantly more citations (14) than the sample patent does (8). The 14 patents citing the sample PGPub and the 8 patents citing the sample patent are arranged in the middle and rightmost two columns, respectively, where one column is for the patent numbers and the other column is for issued dates. As the amount of time spent in pendency varies significantly for different patents, the issued dates of these 22 citing patents are scrambled. For example, the No. 2 citing patent was filed 6 months before the sample patent was issued, but was still able to cite the sample patent.

One may speculate that, even though the sample PGPub receives more citations, perhaps most or these entire sample PGPub citations happened before the sample patent was issued and, once the sample patent was issued, people stopped citing the sample PGPub. Table 1 shows that the speculation is not true. A horizontal line in Table 1 between the Nos. 6 and 7 citing patents indicates the time when the sample patent was issued. As illustrated, there are 16 citing patents filed after the sample patent was issued (i.e., No.7 to No. 22), 9 of them chose to cite the sample PGPub and the remaining 7 chose to cite the sample patent.

Some may argue that there are of course more citations to the sample PGPub than to the sample patent since the former was published to the public about one and half years longer than the sample patent. To remove the bias that the sample PGPub has a longer period of time in the public, only the 16 citing patents (i.e., No. 7 to No. 22) filed after the sample patent are considered. For the 16 citing patents, their applicants or examiners are equally exposed to the sample PGPub and the sample patent when drafting or examining their respective patent applications. In other words, the sample PGPub and the sample patent should be equally possibly cited by the 16 citing patents. However more patents (9 vs. 7) still chose to cite the sample PGPub regardless of the presence of the sample patent within the same time window.

TABLE 1. 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

Table 1 indeed indicates a tendency that, as time advances, more recent citations are directed to the sample patent, rather than to the sample PGPub. For the first two years after the sample patent was issued (i.e., 2008 and 2009), significantly more citations were directed to the sample PGPub than to the sample patent (7 vs. 1). Then, from 2010 till January, 2014, significantly fewer citations were to the sample PGPub than to the sample patent (2 vs. 6). However there is no guarantee that new citations will only be limited to the sample patent. For example, for the latest 4 citing patents filed in 2011 and 2012, two still chose to cite the sample PGPub.

One of the strongest arguments to the study is that the content of an application's PGPub and patent can be different as the application may be amended after it is published so as to overcome the rejection or objection of the application's examiner. As such, an application's PGPub and patent should be considered as different documents and their citations should be counted separately. However, it is dubious that an application can be amended during the 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 from the other, especially considering that U.S. Patent Act (35 U.S.C.S. § 132(a)) clearly specifies that "No amendment shall introduce new matter into the disclosure of the invention." This so-called "new matter" refers to newly added material not supported by the disclosure at the time of filing. In other words, even though the presence of new matter is subjectively determined by the examiner, the content of the application's PGPub and patent is nevertheless bounded by what is disclosed at the time of filing. In this exemplary case, the sample PGPub and the sample patent are compared word for word, and they are completely identical in terms of their abstracts, specifications, and claims. This clearly demonstrates that the sample PGPub and the sample patent are not different documents and ignoring the citations to the sample PGPub is not a right move.

III. EMPIRICAL DATA

The exemplary case discussed in the previous section is surely not representative and a suitable set of patents should be collected and analyzed. Hall, Jaffe, and Trajtenberg [2] conducted an empirical study on the forward citation lag of patents issued in 1990. The forward citation lag of a target patent is the time difference between the time the target patent is issued and the time a subsequent patent citing the target patent is issued. It was found that the subsequent patents citing a target patent are mostly issued after the target patent was issued for 5, 6, and 7 years, and after 7 years, the number of citations gradually drops (specifically see Fig. 11 of Hall, Jaffe, and Trajtenberg [2]). This seems to suggest that, at least for this empirical study, a patent requires at least 7 years on the average so that its "potential" for attracting citations is given enough time to be appropriately developed.

We therefore choose the patents granted in the year 2007 so that these patents are given a suitable period of time to accumulate their citations, and that comparing the citation counts between them and their PGPubs is not seriously biased. However there are 182,928 patents issued in the year 2007 which are too many for a preliminary study like this paper. We therefore randomly choose the utility patents issued in the 26th week of the year 2007. The design patents are ignored as they are not subjected to the early disclosure process, and the plant patents are ignored as well since there is only a small portion of plant patents. Then there are total 3,087 utility patents issued in the 26th week of the year 2007.

Among the 3,087 utility patents, 7 patents were withdrawn (e.g., due to the applicant's petition) and 392 patents were issued directly without being published first (e.g., due to the applicant's request). These patents are removed as they don't have any publication information and finally there are total 2,688 patents left.

For the 2,688 patents, the following pieces of information are gathered from USPTO: (1) their patent numbers, (2) the publication numbers of their corresponding PGPubs, (3) their citation counts, (4) the citation counts of their corresponding PGPubs, (5) their citation counts from those filed after the 2,688 patents were issued, and (6) the citation counts of their corresponding PGPubs from those filed after the 2,688 patents were issued, by running programs to simulate manual interaction with the USPTO advanced search interface (http://patft.uspto.gov/netahtml/PTO/search-adv.htm). The citation counts of (3) to (6) are limited to those occurring before or on January 6, 2014.

The purpose of collecting the citation counts of (5) and (6) is that these citations occurred over the same period of time (i.e., all after the patents were issued) and that they are equally possible to cite the patents or the PGPubs. As such whether citations to the PGPubs will cease to happen after their corresponding patents are issued can be observed, similar to what is done to the exemplary case in the previous section.

IV. RESEARCH RESULT

Tables 2 provides statistics to the various citation counts of the 2,688 patents and their PGPubs. Table 2 is arranged so that the citation statistics from all citing patents and those from the citing patents filed after the 2,688 patents were issued are in the middle and the rightmost two columns, respectively, where one column is for those citing the 2,688 patents and the other column is for those citing the 2,688 PGPubs.

According to Table 2, the 2,688 patents and their PGPubs receive roughly the same numbers of citations and therefore the same average numbers of citations. For citing patents filed after the 2,688 patents were issued, the number of citations to the PGPubs is indeed smaller as we speculated in the previous section, but still is up to two third of the number of citations to their patents. The significant amount of citations received by the PGPubs again suggests that the citations to the PGPubs should not be ignored easily. Also according to Table 2, the citations to the 2,688 patents and their PGPubs widely varies, as suggested by the significant standard deviations.

TABLE 2. STATISTICS OF PATENTS CITING THE 2,688 PGPUBS AND PATENTS.

	From all citing patents		From citing patents filed after the patents were issued	
	To patents	To PGPubs	To patents	To PGPubs
Total citation count	13,441	13,422	8,846	6,102
Max. citation count	146 (38)	228 (46)	112 (9)	119 (2)
Avg. citation count	5.0	5.0	3.3	2.3
Standard deviation	9.9	11.6	6.7	5.9

For all citing patents, the PGPub 2002/0026394 actually receives the greatest 228 citations among all 2,688 PGPubs whereas its corresponding patent 7,236,950 receives only 46 citations (as shown in the parentheses to the right). On the other hand, the patent 7,236,687 receives the greatest 146 citations among all 2,688 patents where its corresponding PGPub 2002/0164152 receives only 38 citations (as shown in the parentheses to the right). Similarly, for citing patents filed after the 2,688 patents were issued, the patent 7,234,624 receives the greatest 112 citations whereas its PGPub 2006/0043148 receives only less than one tenth of the citations. The PGPub 2005/0256627 receives the greatest 119 citations whereas its patent 7,236,871 receives only 2 citations, which is less than 2%.

Table 3 provides some additional statistics about the absolute differences between the citations to the 2,688 PGPubs and their patents. As illustrated, the above-mentioned PGPub 2002/0026394 achieves the greatest difference in terms of citation counts (228-46=182), and the above-mentioned PGPub 2005/0256627 achieves the greatest difference in terms of citation counts (119-2=117) from all citing patents filed after the patents were issued. One can imagine that, if the two patents are considered in the conventional manner to have citation counts 46 and 2, respectively, the two patents are seriously underestimated. Again, the significant standard deviations suggest that the distribution of the citation count difference varies widely.

In the following an analyst's point of view is assumed so as to investigate the impact if the analyst ignores the citations to the PGPubs. To achieve this objective, a PGPub Citation Ratio of a patent is defined as follows: (citation count of the patent's PGPub)/(citation count of the patent's PGPub + citation count of the patent).

A patent having the PGPub Citation Ratio 0% means that its PGPub is not cited and the patent is not underestimated if considering only the patent's citation count. On the other hand, a patent having the PGPub Citation Ratio 100% means that only its PGPub is cited and the patent is completely underestimated if considering only the patent's citation count (which is 0 in this case). In other words, if a patent's PGPub Citation Ratio is closer to 100% or 0%, the patent is more or less underestimated. A special condition is that the citation counts of the patent and its PGPub are both zero, and the PGPub Citation Ratio is considered as 0%. If the citation count of a patent is 0, its PGPub Citation Ratio is always 100% as long as the citation count of the PGPub is greater than 0. Similarly, if the citation count of a PGPub is 0, the corresponding patent's PGPub Citation Ratio is always 0% as long as the citation count of the patent is greater than 0.

Fig. 1 is a histogram showing the distributions of the PGPub Citation Ratios of all 2,688 patents where all citing patents filed before and after the 2,688 patents were issued are considered. As illustrated in FIG. 1, there are 819 patents out of the 2,688 patents having PGPub Citation Ratios equal to 0%, 38 patents having PGPub Citation Ratios greater than 0% but less than and equal to 10%, 148 patents having PGPub Citation Ratios greater than 10% but less than and equal to 20%, and so on.

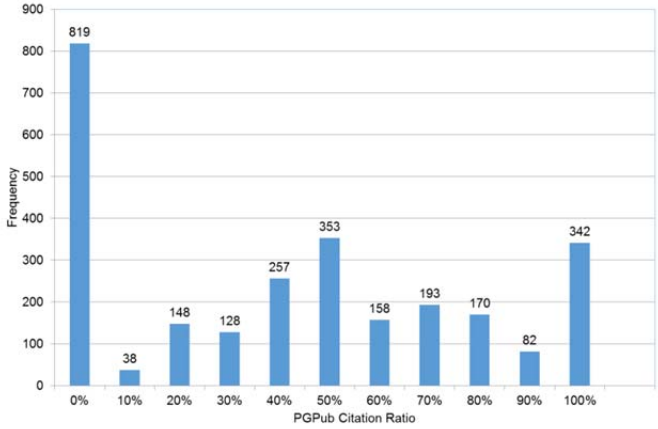

Fig. 1. The histogram of PGPub Citation Ratios of all 2,688 patents from all citing patents.

TABLE 3. STATISTICS OF CITATION COUNT DIFFERENCE BETWEEN THE 2,688 PATENTS AND THEIR PGPUBS.

	From all citing patents	From citing patents filed after the patents were issued
Max. citation count difference	182	117
Avg. citation count difference	4.4	2.8
Standard deviation	9.5	6.2

From Fig. 1, it is found that only about 30% (=819/2,688) of the patents are not underestimated at all if considering only the citation counts of the patents. For the remaining 70%, they are underestimated to various degrees. For example, about 13% (=342/2,688) of the patents have PGPub Citation Ratios above 90% and therefore are completely or nearly completely underestimated. On the other hand, there are about 35% (=(158+193+170+82+342)/2,688) of the patents have PGPub Citation Ratios above 50%, implying their patent citation counts only account for at most 50% of the total citations to both the patents and their PGPubs.

Fig. 2 is another histogram showing the distribution of the PGPub Citation Ratios of all 2,688 patents where only the citing patents filed after the 2,688 patents were issued are considered. Under this scenario, as described earlier, the patents and their PGPubs are equally citable over a same period of time. As illustrated, the citation counts of the PGPubs are significantly reduced compared what is shown in Fig. 1, and it is found that about 47% (=1,260/2,688) of the patents are not underestimated at all if considering only the citation counts of the patents. In other words, there are still about 53% of patents that are underestimated to various degrees. Fig. 2 also shows that, even after the patents were issued, their PGPubs are still cited and, for about 11% (=310/2,688) of the patents, this still leads to their complete or nearly complete underestimation.

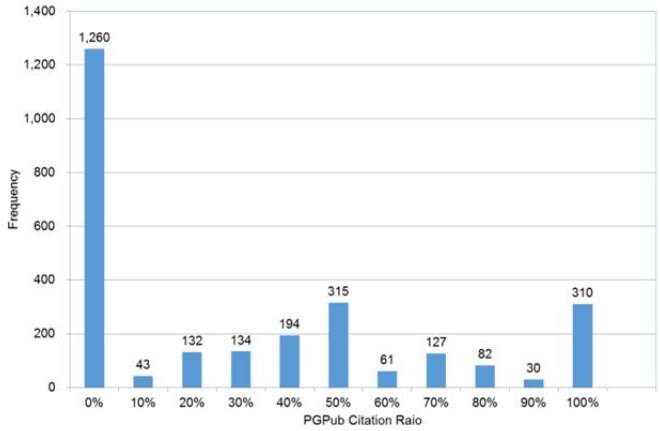

Fig. 2. The histogram of PGPub Citation Ratios of all 2,688 patents from citing patents filed after the 2,688 patents were issued.

V. CONCLUSION AND FUTURE WORK

In this study a limited set of 2,688 patents is used to gain some preliminary insight into the individual and parallel accumulation of citations to these patents and their PGPubs. It is found that, if the citations to the PGPubs are ignored, about 70% of the patents are underestimated in terms of their value, quality, or importance to various extents, and there are about 13% of the patents that are completely or nearly completely underestimated.

Ignoring the citations to the PGPubs therefore should be a dangerous choice by an analyst. On the other hand, this preliminary insight also suggests that people should be more cautious about the analytic result from patent citation analysis or patent evaluation based on citation counts, if the common practice of ignoring the citations to the PGPubs is adopted.

It should be pointed out that this study didn't compare the 2,688 patents word for word against their PGPubs, which is a daunting task. However, this study doubts that, even though there are indeed some differences between some of the patents and their PGPubs arising out of amendments in the examination process, these differences are really the reason why there is such diverse discrepancy between the citations to the patents and PGPubs. The study further speculates that the citation by an applicant or examiner towards a patent or its PGPub is actually accidental. For example, an examiner cites a PGPub because he or she searches PGPub database first and finds the PGPub, even though the corresponding patent is already issued. Similarly, an applicant cites a PGPub simply because he or she happens to be aware of this document and does not bother to check if the corresponding patent is issued or not.

Figs. 1 and 2 in the previous section seem to suggest that, as time advances, the citations to the PGPubs would gradually drop. Therefore, for aged patents, the impact of ignoring citations to their PGPubs may not be so significant. On the other hand, for young patents just issued recently, citations to their PGPubs should definitely be taken into consideration.

The collection of a more thorough set of empirical data is currently underway, which involves all utility patents issued in the years 2007, 2009, and 2011, and their corresponding pre-grant publications by USPTO. This new set of empirical data, which contains more than half of million patents, will be used to expand the observation conducted in this study, and will be used to compare the impact of ignoring PGPub citations for patents of different seniority.

ACKNOWLEDGMENTS

This study was supported by the National Science Council, Taiwan, ROC, under Grant No. NSC 102-2221-E-011-051.

REFERENCES

[1] Fontana, R., A. Nuvolari and B. Verspagen, "Mapping technological trajectories as patent citation networks. An application to data communication standards," *Economics of Innovation and New Technology*, vol. 18, no. 4, pp. 311-336, 2009.

[2] Hall, B. H., 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.

[3] Hall, B. H., A. B. Jaffe and M. Trajtenberg, "Market Value and Patent Citations," *RAND Journal of Economics*. vol. 36, no. 1, pp. 16-38, 2005.

[4] Jaffe, A. B., M. Trajtenberg and R. Henderson, "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations," *Quarterly Journal of Economics*, vol. 108, pp. 577-598, 1993.

[5] Jaffe, A. B., M. S. Fogarty and B. A. Banks, "Evidence from Patents and Patent Citations on the Impact of NASA and Other Federal Labs on

Commercial Innovation," *Journal of Industrial Economics*, vol. 46, no. 2, pp. 183-206, 1998.

[6] Lucio-Arias, D., and L. Leydesdorff, "Main-path analysis and path-dependent transitions in HistCite (TM)-based historiograms," *Journal of the American Society for Information Science and Technology,* vol. 59, no. 12, pp. 1948-1962, 2008.

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

[8] Meyer, M.; "What is special about patent citations? Differences between scientific and patent citations," *Scientometrics*, vol. 49, no. 1, pp. 93-123, 2000.

[9] Verspagen, B.; "Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research," *Advances in Complex Systems,* vol. 10, no. 1, pp. 93-115, 2007.

[10] Narin, F.; "Patent Bibliometrics," *Scientometrics*, vol. 30, no. 1, pp. 147-155, 1994.

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

[12] Trajtenberg, M.; "A Penny for your Quotes: Patent Citations and the Value of Innovations," *Rand Journal of Economics*, vol. 21, pp. 172-187, 1990.