Patent Wars: Attack of the IP5 Metrics[*]

Kyoungwon Kim[†] Hyunwoo Woo[‡]

September 27, 2024

Patent metrics, including the number of patents and citations, exhibit variability across time and technological domains. This variation can be attributed to shifts in technological significance or changes within the patent system (Bernstein, 2015). To compare innovation accurately, individual patent measures are adjusted by comparing them against the average metrics of comparable patents, identified based on their relevant dates (filing, publication, or grant) and technological classification. Otherwise, unadjusted patent metrics provide partial information, leading to biased or even erroneous implications in the worst case. However, it is a complex and time-consuming task to generate average patent metrics for all relevant dates and technological classifications. Researchers may face this challenge each time they need metrics related to their sample.

An important contribution to addressing this challenge is Hall et al. (2001). They describe the *NBER U.S. Patent Citations Data* (NBER) and introduce various patent measures across six technology fields, using the data from 1975 to 1999.[1] Their results not only provide a comprehensive view of U.S. patenting activity but also serve as a benchmark for scaling patent metrics in other studies across different time periods and technological fields. Many studies use these metrics to explore relationships between patent activity, economic growth, firm performance, and technological advancement. (e.g., Arora et al., 2021; Bernstein, 2015; Chatterji, 2009; Cohen et al., 2013; Flammer and Bansal, 2017; Li et al., 2014; Tian and Wang, 2011). However, the dataset has not been updated since 2006, limiting its applicability to recent trends in patenting and technological innovation. Consequently, while the NBER dataset is foundational for historical analysis, researchers studying recent trends need to seek alternative or complementary datasets. For example, Kogan et al. (2017) made patent-level panel data connected with firm data in *the Center for Research in*

[*] We thank Kwangwon Ahn, Jaewon Choi, Bias Daniel, Michael Ewens, Minmo Gahng, Jaehoon Hahn, Yewon Jang, Kee-Youn Kang, Daniel Sungyeon Kim, Edward Kim, Ji-hyun Kim, Josh Lerner, Harry Mamaysky, Taeyoung Park, Laura Veldkamp, Emmanuel Yimfor, and Eojin Yi.

[†] Warrington College of Business, University of Florida, kyoungwon.kim@warrington.ufl.edu

[‡] Columbia Business School, hw2980@columbia.edu

[1] The NBER is one of the most widely used datasets in innovation research to date, including over three million U.S. patents granted between 1976 and 2006. See `https://data.nber.org/patents/`.

Security Prices, LLC (CRSP) from 1926 to 2022, including the monetary value of innovation.[2] The *Private Capital Research Institute* (PCRI) offers a VC-backed patent database from 1970 to 2021, easily mergeable with the PatentsView database.[3] Ewens and Marx (2023) provide founding year data for U.S.-based assignees of USPTO patents granted from 1975 to 2021.[4]

Despite these efforts, several limitations still remain as follows. Firstly, all of the aforementioned data and studies focus solely on U.S. patents in the era of technological globalization. Although U.S. patents are often considered to have the highest technological value, they miss the opportunity to capture innovation in other countries. For instance, the relationship in which patents from a country are cited by those from other countries can be used to identify innovation networks between countries. Secondly, the expanded datasets, along with the original NBER, primarily focus on granted patents; and therefore do not fully account for ungranted patent applications as their data is derived from PatentsView, which is based on granted patents. However, ungranted patent applications can be effectively utilized in research such as evaluating the frequency of R&D efforts or assessing the innovation efficiency within firms or countries. Even rejected patents, which are excluded from the expanded data, may provide valuable information (Kline et al., 2019). Finally, the NBER and most of its related literature or subsequent datasets restrict forward citation timing to the time difference between the filing or grant years of the two patents. In addition to this classic approach, considering various timing points can provide new insights into studies on information flow and knowledge spillover. Since patent information is officially disclosed at the publication date, the time between the cited patent's publication and the citing patent's filing can be used to measure forward citation time. Apart from these limitations, expanded datasets generally do not calculate or disclose values such as originality or generality beyond forward citations.

This paper introduces a new patent dataset that addresses the limitations of the NBER and its existing expanded versions, providing more up-to-date and broader coverage. To support future research, our dataset spans a longer period and includes patents not only from the U.S. but also from other countries. Specifically, this data includes approximately 70 million patents filed from 1975 to 2022, of which 44 million were granted. It covers the intellectual property 5 (IP5) countries, also called Big 5, — *United States*, *Europe*, *China*, *South Korea*, and *Japan*.[5] Our dataset provides various patent metrics, including patent counts, generality, originality, the time required for the initial forward citation, and forward citations across different combinations of application and publication years. We categorize patent measures based on relevant years (filing, publication, or grant), technological classifications, and grant status. We collected patent information from the

[2] See `https://github.com/KPSS2017/Technological-Innovation-Resource-Allocation-and-Growth-Extended-Data`.

[3] See `http://www.privatecapitalresearchinstitute.org/pages/research/patent.html`.

[4] See `https://foundingpatents.com/data/`.

[5] These five national offices handle approximately 90 percent of the world's patent applications. See `https://www.fiveipoffices.org/about`.

Patent Statistical Database (PATSTAT), maintained by the *European Patent Office* (EPO), which is widely regarded as one of the most comprehensive sources of international patents covering over 100 countries.[6] We collected information on each patent application in the IP5. This includes the patent office location, title, abstract, application date, publication date, grant status, and the grant date, if applicable. We also collected forward and backward citation information and IPC codes to calculate citation-related measures for each patent, as well as aggregated measures by technology classification. We use the field classification provided by Goto and Motohashi (2007), which provides a crosswalk between the NBER's technology classifications and the IPC codes.

This extended abstract provides a portion of our dataset due to the page limits.[7] We introduce the number of patent filings and that of forward citations as they are widely used as a proxy for innovation (Bernstein, 2015; Kogan et al., 2017). Considering the quantitative aspects of innovation, the following observations can be made. Figure 1 shows the annual changes in the number of patent applications by country for each technology field. Generally, until the early 2000s, Japan had a relatively high number of patent applications. Then, the U.S. caught up or even surpassed Japan in some fields, and it is noted that China has filed the most patents by a significant margin since the 2010s. However, when it comes to the qualitative aspects of innovation, considering the time difference between the publication year of a cited patent and the filing year of a citing patent as a citation timing, the number of forward citations shows a different pattern from the trend in the number of patent applications. Figure 2 shows the forward citations of granted patents for three years from the publication year. Overall, the innovation quality of U.S. patents has been significantly higher across the entire period, while China has not shown comparable quality relative to the number of patent applications. Meanwhile, the recent decline in citations around 2020 seems to be due to a truncation issue, as insufficient time has passed since the patents were filed. These two figures are comparable to Figures 4 and 6 in Hall et al. (2001) to validate our dataset. As suggested by Lanjouw and Schankerman (2004), early forward citations are more closely linked to the economic value and significance of a patent compared to long-term citations. Based on this argument, we calculated the time taken until the first forward citation of a patent shown in Figure 3. The shortening of citation time over the years implies an increasing importance of patent-based innovation.

We continuously update the data to maintain precision and informativeness, leveraging ongoing efforts to integrate variables from other data sources. Beyond our advanced dataset for each technical field that can shorten the time spent in the pre-research stage, we believe our research can contribute to academia related to innovation, valuing data or information, and text analysis. Moreover, our data can facilitate the expansion of research topics, particularly in international studies.

[6] See `https://www.uspto.gov/web/offices/ac/ido/oeip/taf/data/other_sites.htm` and `https://inspire.wipo.int/patstat-online`.

[7] To check the whole results and codes, see `https://github.com/hw-woo/patent_ip5`.

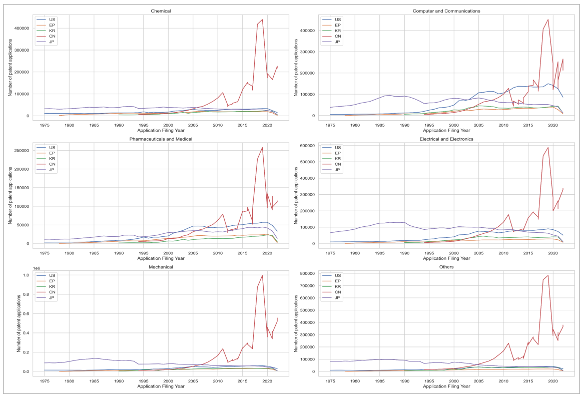

Figure 1. Number of Patent Filings by Year

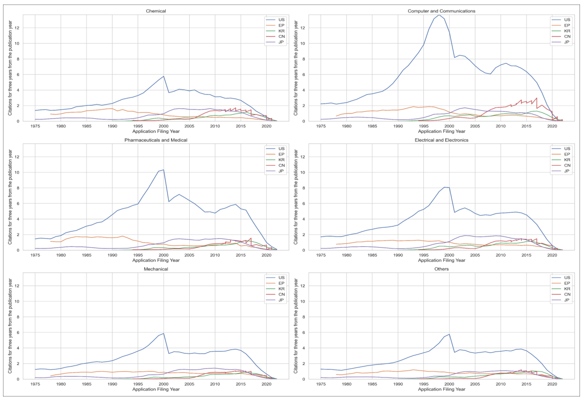

Figure 2. Forward Citations of Granted Patents for three Years from the Publication Year

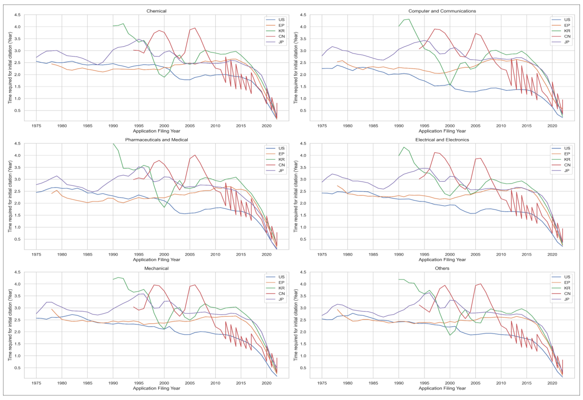

Figure 3. Time Required for Filing Patents' Initial Forward Citation

References

Arora, A., Belenzon, S., and Sheer, L. (2021). Knowledge spillovers and corporate investment in scientific research. *American Economic Review*, 111(3):871–98.

Bernstein, S. (2015). Does going public affect innovation? *Journal of Finance*, 70(4):1365–1403.

Chatterji, A. K. (2009). Spawned with a silver spoon? entrepreneurial performance and innovation in the medical device industry. *Strategic Management Journal*, 30(2):185–206.

Cohen, L., Diether, K., and Malloy, C. (2013). Misvaluing Innovation. *The Review of Financial Studies*, 26(3):635–666.

Ewens, M. and Marx, M. (2023). Firm Age and Invention: An Open Access Dataset. *Working Paper*.

Flammer, C. and Bansal, P. (2017). Does a long-term orientation create value? evidence from a regression discontinuity. *Strategic Management Journal*, 38(9):1827–1847.

Goto, A. and Motohashi, K. (2007). Construction of a japanese patent database and a first look at japanese patenting activities. *Research Policy*, 36(9):1431–1442.

Hall, B. H., Jaffe, A. B., and Trajtenberg, M. (2001). The nber patent citation data file: Lessons, insights and methodological tools. NBER Working Paper.

Kline, P., Petkova, N., Williams, H., and Zidar, O. (2019). Who profits from patents? rent-sharing at innovative firms. *Quarterly Journal of Economics*, 134(3):1343–1404.

Kogan, L., Papanikolaou, D., Seru, A., and Stoffman, N. (2017). Technological Innovation, Resource Allocation, and Growth*. *The Quarterly Journal of Economics*, 132(2):665–712.

Lanjouw, J. O. and Schankerman, M. (2004). Patent Quality and Research Productivity: Measuring Innovation with Multiple Indicators. *The Economic Journal*, 114(495):441–465.

Li, G.-C., Lai, R., D'Amour, A., Doolin, D. M., Sun, Y., Torvik, V. I., Yu, A. Z., and Fleming, L. (2014). Disambiguation and co-authorship networks of the u.s. patent inventor database (1975–2010). *Research Policy*, 43(6):941–955.

Tian, X. and Wang, T. Y. (2011). Tolerance for Failure and Corporate Innovation. *The Review of Financial Studies*, 27(1):211–255.