# Replicable Patent Indicators
# Using the Google Patents Public Datasets

George ABI YOUNES
Gaétan DE RASSENFOSSE

November 2023

Innovation and Intellectual Property Policy
Working Paper series no. 24

Available at: https://ideas.repec.org/p/iip/wpaper/24.html

EPFL Working Paper Series
STiP lab

# Replicable Patent Indicators Using the Google Patents Public Datasets

George Abi Younes[*] Gaétan de Rassenfosse[*†]

November 26, 2023

**Abstract**

Recognizing the increasing accessibility and importance of patent data, the paper underscores the need for standardized and transparent data analysis methods. By leveraging the BigQuery language, we illustrate the construction and relevance of commonly used patent indicators derived from Google Patents Public Datasets. The indicators range from citation counts to more advanced metrics like patent text similarity. The code is available in an open Kaggle notebook, explaining operational intricacies and potential data issues. By providing clear, adaptable queries and emphasizing transparent methodologies, this paper hopes to contribute to the standardization and accessibility of patent analysis, offering a valuable resource for researchers and practitioners alike.

**Keywords:** BigQuery language, data transparency, patent analytics, patent data.

[*]Ecole polytechnique fédérale de Lausanne, College of Management of Technology, CH-1015 Lausanne.

[†]Corresponding author: gaetan.derassenfosse@epfl.ch

# 1 Introduction

A variety of databases facilitate access to patent information. Besides paid services such as Derwent World Patent Index, PatentSight, and PatSnap, recent years have witnessed a democratization of patent data, with free services such as Espacenet and Google Patents or partly free such as The Lens. Accessing patent data at scale and computing one's patent indicators is now within everyone's reach.

With the democratization of patent data comes a crying need for standardized and replicable methods to analyze such data. For instance, Smith et al. (2017) point to insufficient quality of reporting in patent landscapes. Domain experts are well aware that "patents differ greatly in their technical and economic significance" (Griliches, 1990) and that "many patents are virtually worthless" (Lemley and Shapiro, 2005). Furthermore, the patent system is complex and generates data that can be challenging to process. Design choices may also prevent the comparability of indicators across studies. Finally, patent indicators are riddled with statistical pitfalls, calling for sound and transparent methods.

The objective of this paper is twofold. First, it explains commonly used patent indicators using insights from the patent analytics literature. Second, it provides computer code ('queries') to create such indicators, with a view of further facilitating access to patent data and ensuring replicability. The present paper substantially revises de Rassenfosse et al. (2014). It focuses on a different database that has been gaining popularity recently and exploits a query language different from that in the original 2014 article.

We use the BigQuery language to construct the indicators and source the raw data from the Google Patents Public Datasets.[1] We present and document the queries in an online, openly accessible Kaggle notebook.[2] The notebook explains in detail how each query works and what issues might arise with the data. The queries and the indicators are of varying complexity, ranging from a simple citation count to a metric of the 'originality' of an invention.

A couple of remarks are in order. The list of indicators presented herein is limited, and the companion notebook provides additional indicators. To facilitate

[1] However, we note that many 'raw data' patent databases, such as PATSTAT or PatentsView, are structured in similar ways. The queries can be adapted to these other databases. The discussion in this report abstracts away from technical details of the query language or the data structure.

[2] The notebook is available at `https://www.kaggle.com/code/georgeabiyounes/paper3/edit/run/105473614`.

the readability of the article, we do not delve into the technical details of the query, which we leave to the online companion. Instead, we explain the indicators' relevance and use in the patent literature. We have paid particular attention to making the queries as clear and flexible as possible. Finally, we illustrate the results of the queries using biotechnology patents as a use case, but we note that the indicators apply equally well to all fields of science.[3] Note that the queries best run on patents filed at the U.S. Patent and Trademark Office (USPTO). There are subtle differences in data acquisition and processing across jurisdictions, which makes international comparisons particularly tricky and beyond the scope of this paper.

# 2 The indicators

## 2.1 Count of citations

Patent citations are references to prior patents appearing in patent documents. According to the Manual of Patent Examining Procedure at the USPTO, inventors and patent agents have "a duty of candor and good faith" to disclose all information known to them to be material to the patentability of the invention.[4] They must list the prior art that is relevant in assessing the claimed invention's novelty, non-obviousness, and usefulness. Patent examiners then enrich the list of references during the search and examination phase.

As Jaffe and de Rassenfosse (2017, p. 1362) put it, the citations that appear in a patent (its 'backward' citations) inform us about the technological antecedents of the patented invention. Conversely, "the citations received by a patent from subsequent patents ('forward' citations) inform us about the technological descendants of the patented invention. A patent that is never cited was a technological dead end. A patent with many or technologically diverse forward citations corresponds to an invention that was followed by many or technologically diverse descendants."

[3] We identify biotechnology patents using the International Patent Classification (IPC), a hierarchical classification system that indicates the technology to which a patent pertains. The first letter of the IPC code represents the section (e.g., A), the first two digits represent the class (e.g., A61), and the subsequent letter represents the sub-class (e.g., A61K). To create our sample, we randomly selected 400 patents from each of the following IPC sub-classes: A61K, G01N, C12P, and C07K. According to OECD (2009), these sub-classes include biotechnology patents. The final sample includes 1600 U.S. patents granted between 2002 and 2015.

[4] See Manual of Patent Examining Procedure, Ninth Edition, Revision 07.2022, Published February 2023, available at `https://www.uspto.gov/web/offices/pac/mpep/s2001.html`.

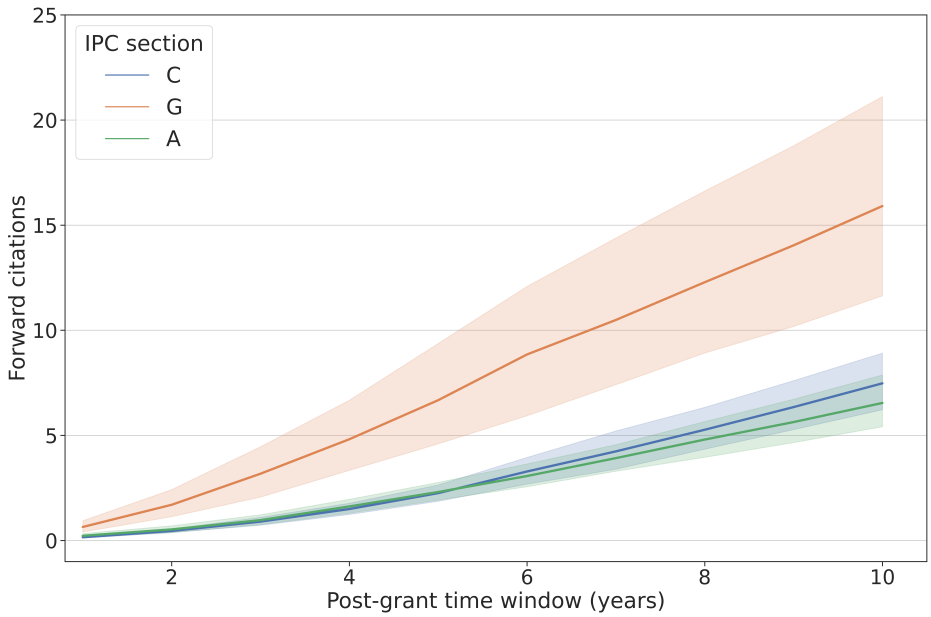
Carpenter et al. (1981), Trajtenberg (1990), and Hall et al. (2005), among others, have shown that the count of forward citations correlates with various measures of patent 'value.' Here, 'value' (or 'importance') is defined in a broad sense to capture an invention's technological merit and economic potential (see Higham et al., 2021, for an in-depth discussion). The count of forward citations is the most widely used measure of patent value in the literature. Since Jaffe et al. (1993), researchers have also used citation data to track knowledge spillovers and diffusion.

The number of forward citations a patent receives increases over time as new patent applications arrive. Thus, when dealing with patents of different cohorts, it is good practice to limit the count to a set time window (usually starting at the patent application date or grant date). Another solution to treat temporal issues involves weighting forward citations by some time factor that accounts for changes in the arrival rate of citations over time. Note that patent citation practices also differ across technological fields. Therefore, comparing citations across fields requires some field normalization. One possibility involves computing the within-field percentile rank of a patent's citations or applying some statistical correction in downstream analyses (such as using technology field fixed effects in regression models).

Query 1 in the online notebook calculates the number of forward citations received by the focal patents in the sample with a time window of $n$ years from the filing date. The number of years $n$ can be modified to suit the user's needs. Figure 1 depicts how the number of forward citations increases when we expand the time window $n$ from 1 to 10 years.[5]

[5]Since the most recent patents in the sample were granted in 2015, a 10-year time window implies that we should be counting citations accruing until 2025 for the youngest cohort. As the time of data collection is the end of 2021, only the time windows $n \leq 6$ are valid.

Figure 1: Forward citations count with different time windows




*Notes: 1. We computed the number of forward citations for discrete time windows from 1 to 10 years. 2. We group the patents at the IPC section level. 3. N = 1600.*

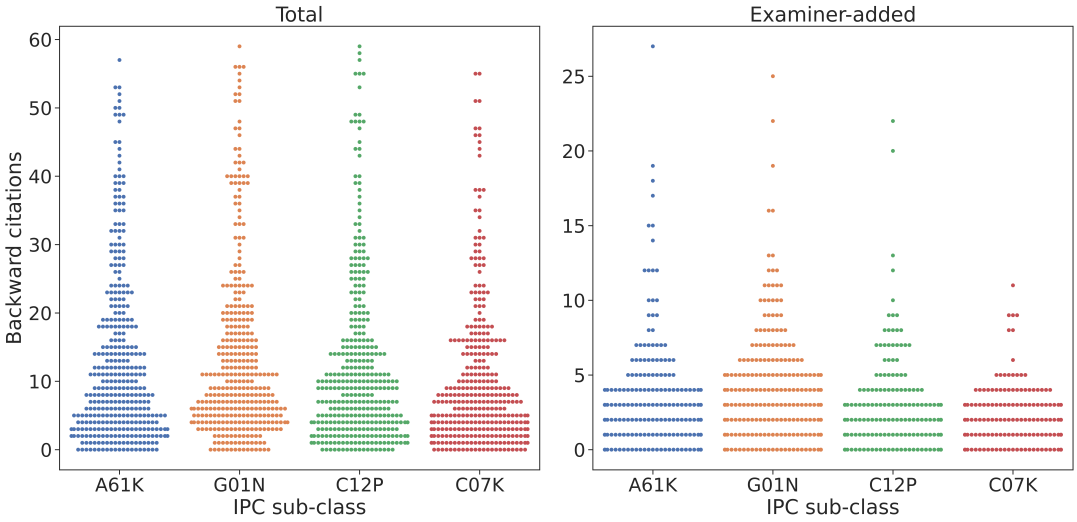
The Figure shows that patents in IPC section 'G' receive more forward citations over the different time windows than in sections 'C' and 'A.' This result does not necessarily imply that patents in section 'G' are more important or valuable than patents in the other sections. Indeed, the number and growth rate of patents filed in each section may differ, which affects the number of citations that antecedent patents will receive. Besides, citation practices may also differ across technology groups. Note that expanding the time window increases the risk of data truncation affecting the citation count.

Citations fall into different categories relating to their origin and their type. Regarding origin, citations predominantly originate from applicants (inventors or their patent attorneys) during the patent drafting stage and from examiners during the prosecution process. (In rarer cases, they can also originate from third parties.) Evidence shows that examiner citations are a stronger predictor of patent value than applicant citations (Hegde and Sampat, 2009). Regarding the type, not all citations refer to other patents. Patents can sometimes cite non-patent

literature (NPL), such as scientific articles, technical norms, or any other relevant publicly available prior art. The count of references to the scientific literature is an (imperfect) proxy of the scientific linkage of an invention (Narin et al., 1997; Meyer, 2000; Roach and Cohen, 2013).

Query 2 in the Kaggle notebook computes the number of backward citations for patents in the sample. The query restricts the count to the examiner-added references for illustration, as illustrated in Figure 2. The notebook further explains how to differentiate between the different types of references.

Figure 2: Swarm plots of backward citations




*Notes: 1. The Figure excludes patents with total backward citations* $> 60$ *and examiner-added citations* $> 30$. *2.* $N = 1542$.

## 2.2 Priority patents

Patent rights are jurisdictional, meaning they are valid only in the jurisdiction that grants them. To obtain international protection for an invention, inventors must file a patent in each country where they desire protection. The patent document first filed anywhere in the world is called the priority filing or the priority patent document. The 'extensions' in other countries are called second filings. Sometimes, priority filings and their second filings co-exist within the same country. Patent applications can claim priority to applications within the same country in case of

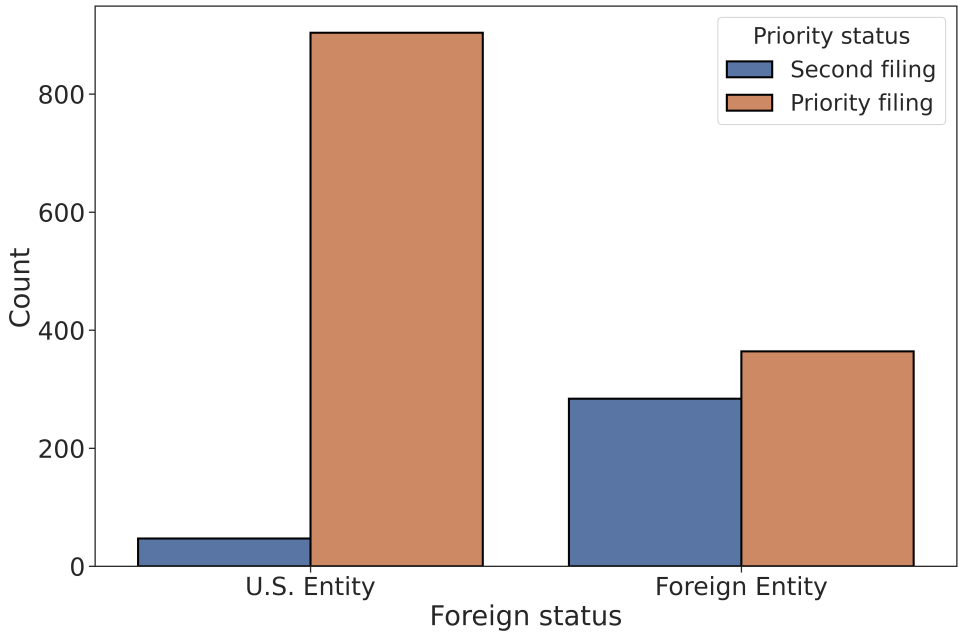
continuations or divisional patent applications, for instance.[6]

When it comes to counting patents, priority patent applications and second filings are like apples and pears. The former describes unique inventions, whereas the latter tells us something about an invention's value or market relevance (because only worthwhile inventions will be subject to international patent protection). It follows that counting U.S. patents by entities puts U.S. and foreign entities on unequal footing. The patent portfolio of U.S. entities will be composed primarily of priority filings ('unique inventions'). In contrast, the U.S. portfolio of foreign entities will be composed mainly of second filings—and thus will only form a subset of foreign entities' inventions. To better measure inventive activity, de Rassenfosse et al. (2013) propose an indicator that counts the number of priority applications filed worldwide.

The priority status of patent applications is a useful indicator for constructing patent metrics. Query 3 in the Kaggle notebook returns a binary variable capturing the priority status of each patent application. The returned variable, priority, takes a value of 1 when the patent application is a priority filing and 0 otherwise.

[6] A continuation application allows the applicant to pursue additional claims for the invention that were not included in the original application. A divisional application is used to pursue claims for a distinct and independent invention that was disclosed, but not claimed, in the original application.

Figure 3: Count plot of priority filings for U.S. vs. foreign companies




*Notes: 1. We obtain the foreign entity status of the assignee from the USPTO's PatentsView database. An entity can be an individual, company, or government. 2. N = 1600.*

The Figure shows that U.S. entities overwhelmingly submit priority filings at the USPTO.[7] By contrast, foreign entities file a large proportion of second filings at the USPTO. de Rassenfosse et al. (2013) have shown that entities tend to file first in their home country and later extend patent protection to foreign countries for their most valuable inventions. This fact explains why the 'quality' of patents by foreign entities is, on average, higher than that of patents by local entities—the former set being a considerably selected set of patents.

[7]The origin of an entity is derived from the self-reported assignee's address on the patent document. Thus, a foreign-owned U.S. entity may appear as U.S. origin if the patent document was filed through the U.S. subsidiary.

## 2.3 Patent families

The discussion in the previous section leads us to the concept of the 'patent family.' In broad terms, a patent family is any group of patents sharing the same or similar technical content. Patent offices define two main types of patent families. The simple patent family, known as the DOCDB family, is a collection of patent documents that cover a single invention and, therefore, have the same technical content. Members of a DOCDB family share the same set of priorities. The extended patent family, or INPADOC, covers patents with similar, but not necessarily identical, technical content. Members of an INPADOC family share at least one priority filing in common.

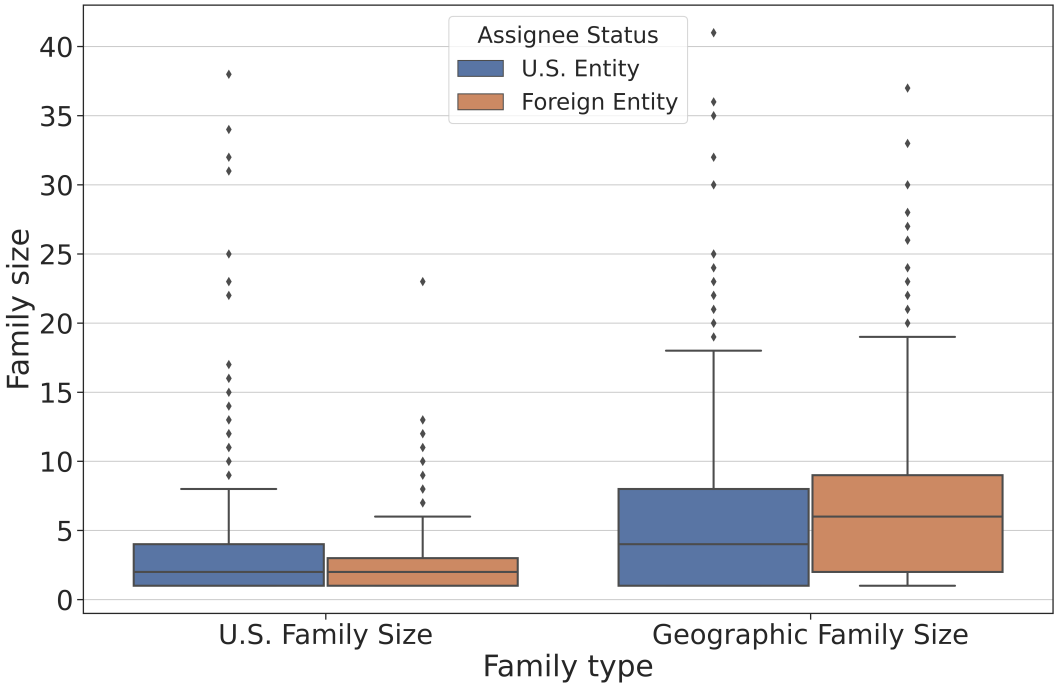
The size of the patent family correlates with the invention's value.[8] The patent application process is costly, and broader families usually cover more valuable inventions (Putnam, 1996; Harhoff et al., 2003). This argument holds for family size measured within a single office or across offices. The former metric indicates how many patents protect a unique invention in a given jurisdiction (using continuations or divisionals). The latter captures the market reach of the invention—its 'geographic' family size. Given that different definitions of a patent family can lead to different results (Martínez, 2011), one must use patent family-based measures with caution.

Query 4 in the Kaggle notebook computes the DOCDB family size for the focal patents in a given patent authority (i.e., country). For example, for a patent issued by the USPTO, the resulting count will include the number of unique U.S. patents within the same simple family as the focal patent.

Query 5 computes the geographic family size of the focal patents. It corresponds to the number of unique countries covered by patents belonging to one patent family. As discussed in the notebook, patents from regional offices (such as the European Patent Office) or the World Intellectual Property Organization may require special treatment depending on user needs.

[8]Scholars have also used information on DOCDB 'equivalent' patent applications across jurisdictions as an identification strategy, see Webster et al. (2014); de Rassenfosse et al. (2019).

Figure 4: Box plot of family sizes for U.S. vs. foreign entities




*Notes: 1. The Figure excludes patents with a family size* $> 45$. *2.* $N = 1571$.

Figure 4 shows that U.S. entities have larger U.S. family sizes than foreign entities. U.S. entities tend to protect their inventions with more patents and adopt more sophisticated filing strategies in the United States since it is their home market. The geographic family size of inventions by foreign entities is larger than that of U.S. entities. This is a direct consequence of the fact that many patent applications by foreign entities are second filings—and thus have a family member elsewhere.

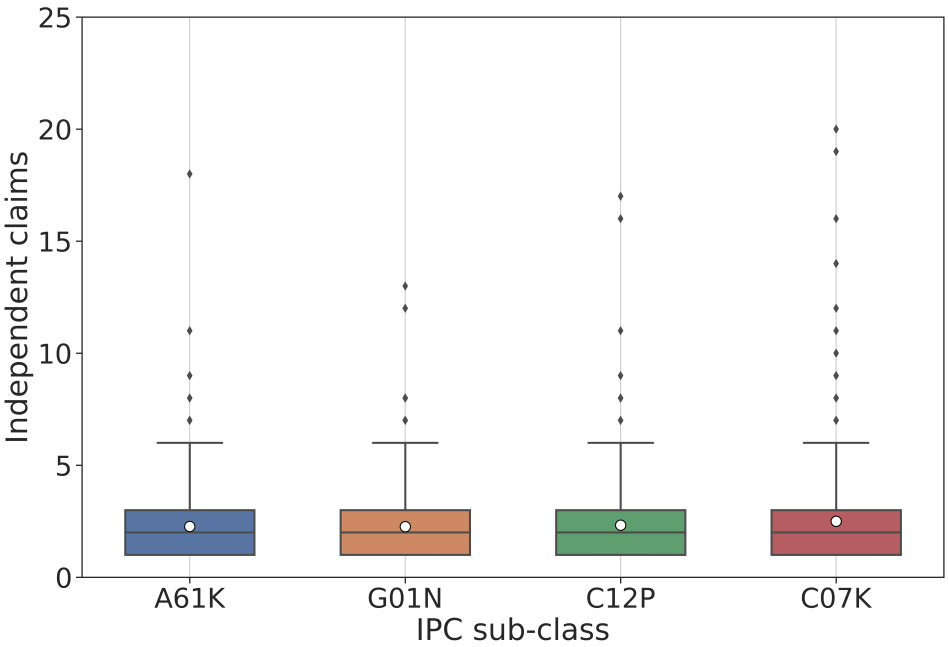
## 2.4 Claims

Patent claims are the building blocks of a patent application. They describe what the invention consists of and, consequently, delineate the scope of the protection that the patent applicant seeks. An independent claim is a claim that stands on its own and does not depend on any other claim for its meaning or scope. The dependent claims depend on one or more other independent or dependent claims for their meaning or scope. They are used to provide more specific definitions or examples of the invention that are covered by the independent claims.

The count of independent patent claims is a proxy for the scope of the invention, with more claims or shorter claims indicating a broader scope (Marco et al., 2019). Along this line, Kuhn and Thompson (2019) use the number of words added by the examiner to the first claim and compare it to the average of the examiner's art unit to derive an indicator of a patent examiner's toughness—more words added to the claim implying a restriction of the scope of the invention. Other researchers have proposed counting the number of independent claims instead of the number of patents to measure technological performance (Tong and Frame, 1994).

The Kaggle notebook presents two queries that deal with patent claims. Since the Google publications table reports the raw text of the claims, one first needs to parse the text of the claims in Query 6.1 in order to identify individual claims. Query 6.2 counts the number of independent claims in a patent document. The notebook explains the details of text parsing.

Figure 5: Number of independent claims per IPC sub-class.




*Notes: 1. The plot excludes patents with more than 25 independent claims. 2. N = 1597.*

Figure 5 reports the distribution of the number of independent claims per sub-class. The median number of independent claims is two for all four sub-classes. The mean varies slightly, with sub-class C07K having the highest. Within each sub-class, patents exhibit heterogeneity in the number of independent claims, although the vast majority have fewer than five independent claims.

## 2.5 Text similarity

Most patent metrics exploit the structural metadata of patent documents, such as citations or technology classes. However, patent documents also contain rich textual data. In addition to the claims, which constitute the core of the document, other text sections, such as the title, abstract, and descriptions, provide additional information about the invention. With the increasing processing power and the improvement of statistical and natural language processing (NLP) models, the full-text analysis of patent documents has become within reach of non-experts.

Abbas et al. (2014) offer an early review of the literature on the use of textual data of patents. The uses include but are not limited to the identification of technology trends (Kim et al., 2009; Yoon and Kim, 2012), the detection of patent infringement (Park et al., 2012), and the study of patent quality (Trappey et al., 2013). Scholars have also used the patent text to compare the technological contents of patents. In broad terms, one can use patents' textual data to create similarity indices between patent pairs. The methods vary in complexity, but they all rely on converting textual data into numerical data through vector embeddings. Vector embeddings are numerical vectors that situate patents in an abstract 'knowledge space.' The dot product of two vector embeddings measures the distance between two patents in the knowledge space, i.e., their similarity.

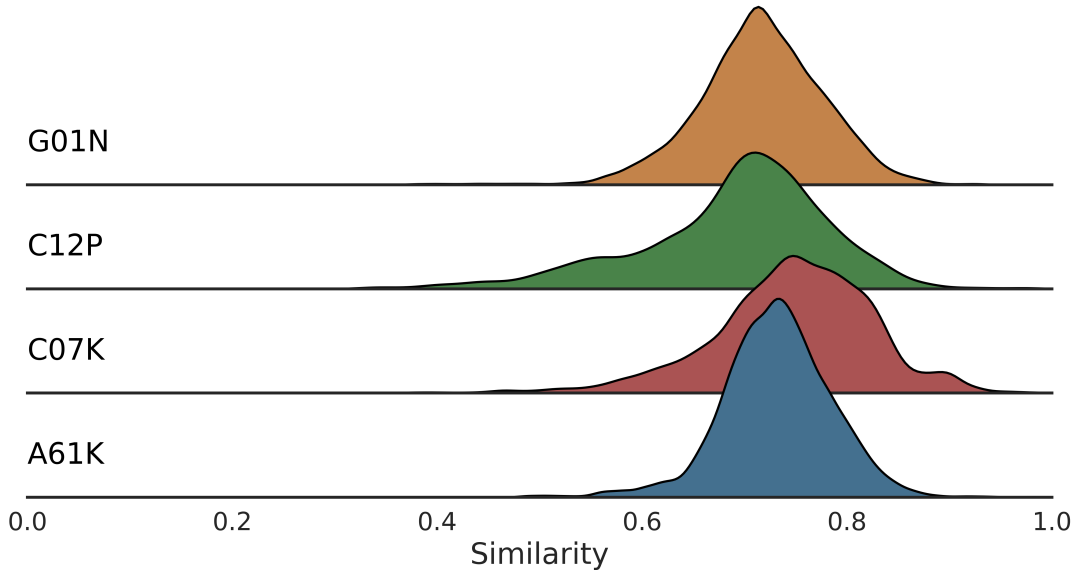
Query 8 calculates the similarity score between the focal patents and the set of patents sharing the same IPC sub-class and filing year. We calculate the dot products using the embeddings provided by Google. The Google similarity measure comes from a model that has learned a set-of-words embedding of the patent text to the technology grouping (CPCs) of that patent using the WSABIE embedding algorithm (Weston et al., 2011).[9] From a technical viewpoint, this method is superior to other approaches that have used unsupervised embeddings relying either on a single-layered neural network such as Word2Vec (Whalen et al., 2020) or more basic methods such as one-hot encoding (Arts et al., 2018) and TF-IDF (Younge and Kuhn, 2016). It has already been used in academic research, for instance, in de Rassenfosse and Raiteri (2022).

Figure 6 reports the distribution of similarity scores by IPC sub-class. We computed the similarity score for each sampled patent, limiting the analysis to the most similar 100 patents. The distributions are roughly similar across the subclasses. The distribution of C07K is flatter and has a thicker right tail compared to

[9]More details on the similarity algorithm are available at `https://media.epo.org/play/gsgoogle2017`, last accessed November 26, 2023.

other sub-classes, which may indicate a cluster of similar patents in that sub-class.

Figure 6: Similarity scores density per IPC sub-class.




*Notes: 1. We keep the top 100 most similar patents per focal patent, when applicable. 2. To avoid potential language bias, we only consider similarity scores with other U.S. patents. N = 74,781.*

## 2.6 Patent originality

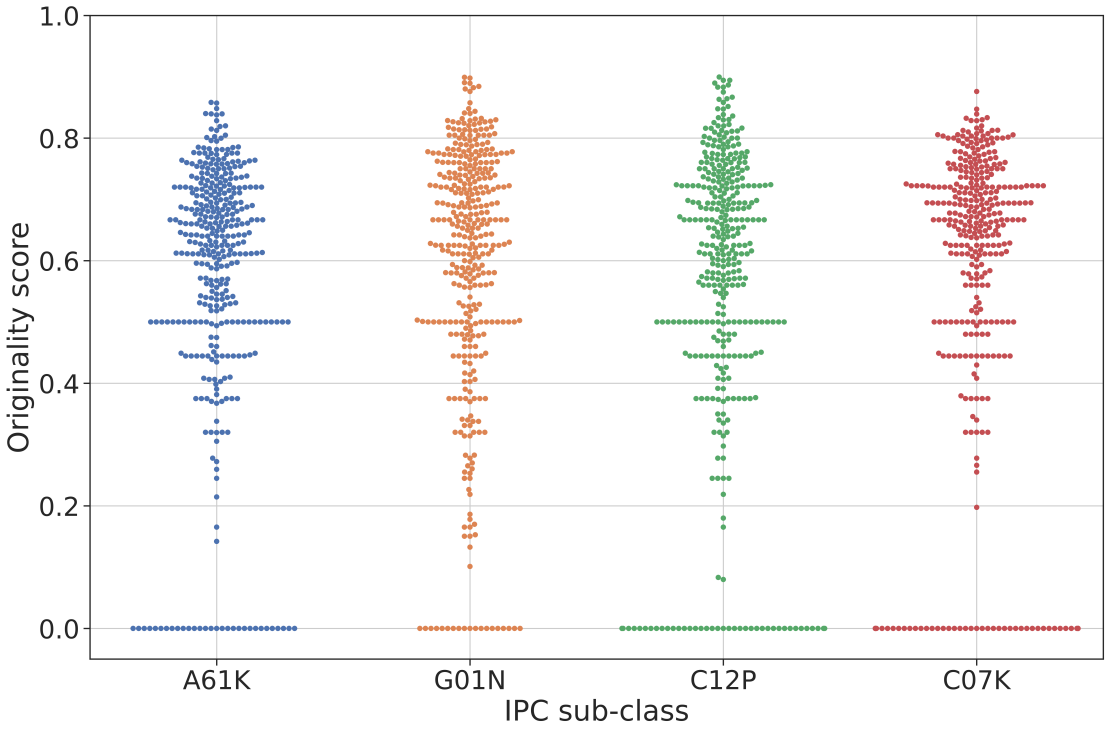
For an invention to be granted a patent, it must be non-obvious to a skilled practitioner of the relevant technology. Patent examiners ensure that all patented inventions exceed a certain inventive step threshold. Notwithstanding this common quality threshold, granted patents exhibit significant heterogeneity in their technical merits. Inventions can differ, e.g., in their originality, closeness to science, and scope.

Trajtenberg et al. (1997) propose a set of standard measures that capture key aspects of the technical merit of the invention. Two noteworthy measures are the originality and the generality of a patent. The originality $O$ of patent $i$ is defined as:

$$O_i = 1 - \sum_{k=1}^{N_i} \left( \frac{NCITED_{ik}}{NCITED_i} \right)^2 \tag{1}$$

where $NCITED$ is the number of patents cited by the focal patent, and $k$ is an index of the technology group to which the cited patents belong. By construction, the metric varies between 0 and 1. The more a patent builds on patents that belong to different technology groups, the higher the originality score. The generality score is similarly defined, except that it focuses on citing patents. A patent that is cited in a large number of technology groups is more general. Query 9 computes the originality score of the focal patents.

Figure 7: Swarm plot of the originality scores in different IPC sub-classes




*Notes: 1. We discard patents without references to other patents. 2. N = 1523.*

Figure 7 depicts the distribution of the originality scores. A large number of patents have originality scores of 0 and 0.5. By construction, a patent with only one backward citation (or all citing patents belonging to the same IPC sub-class) will obtain an originality score of 0. Patents with an originality score of 0.5 usually cite only two patents belonging to different IPC sub-classes.

# 3 Discussion

This paper has discussed patent metrics commonly used in the literature. The code for producing these metrics is openly available in the companion Kaggle notebook. In the interest of space, we have discussed only a few metrics. However, the notebook provides additional metrics and detailed information on how to build them. Other users have also shared pieces of code to produce other indicators.

It is clear by now that many subtle design choices affect the construction of patent metrics, hampering their comparability across studies. For instance, the computation of the originality score exploits the distribution of the technology groups of the cited patents. These technology groups can be computed at different levels of granularity (e.g., IPC section, class, or sub-class), which will affect the score. Similarly, a cited patent may belong to more than one technology class, and the treatment of these multi-class patents will also affect the final score. This case exemplifies the need for replicable patent indicators. The present paper makes a step in this direction, adding to other such initiatives.[10]

[10]See, for instance, `https://github.com/google/patents-public-data`.

# References


Assad Abbas, Limin Zhang, and Samee U Khan. A literature review on the state-of-the-art in patent analysis. *World Patent Information*, 37:3–13, 2014.

Sam Arts, Bruno Cassiman, and Juan Carlos Gomez. Text matching to measure patent similarity. *Strategic Management Journal*, 39(1):62–84, 2018.

Mark P Carpenter, Francis Narin, and Patricia Woolf. Citation rates to technologically important patents. *World Patent Information*, 3(4):160–163, 1981.

Gaétan de Rassenfosse and Emilio Raiteri. Technology protectionism and the patent system: strategic technologies in china. *Journal of Industrial Economics*, 70(1):1–43, 2022.

Gaètan de Rassenfosse, Helene Dernis, Dominique Guellec, Lucio Picci, and Bruno van Pottelsberghe de la Potterie. The worldwide count of priority patents: A new indicator of inventive activity. *Research Policy*, 42(3):720–737, 2013.

Gaétan de Rassenfosse, Hélène Dernis, and Geert Boedt. An introduction to the patstat database with example queries. *Australian Economic Review*, 47(3): 395–408, 2014.

Gaétan de Rassenfosse, Paul H Jensen, T'Mir Julius, Alfons Palangkaraya, and Elizabeth Webster. Are foreigners treated equally under the trade-related aspects of intellectual property rights agreement? *The Journal of Law and Economics*, 62(4):663–685, 2019.

Zvi Griliches. Patent statistics as economic indicators: A survey. *Journal of Economic Literature*, 28(4):1661–1707, 1990.

Bronwyn H Hall, Adam Jaffe, and Manuel Trajtenberg. Market value and patent citations. *RAND Journal of economics*, pages 16–38, 2005.

Dietmar Harhoff, Frederic M Scherer, and Katrin Vopel. Citations, family size, opposition and the value of patent rights. *Research Policy*, 32(8):1343–1363, 2003.

Deepak Hegde and Bhaven Sampat. Examiner citations, applicant citations, and the private value of patents. *Economics Letters*, 105(3):287–289, 2009.

Kyle Higham, Gaétan De Rassenfosse, and Adam B Jaffe. Patent quality: Towards a systematic framework for analysis and measurement. *Research Policy*, 50(4): 104215, 2021.

Adam B Jaffe and Gaétan de Rassenfosse. Patent citation data in social science research: Overview and best practices. *Journal of the Association for Information Science and Technology*, 68(6):1360–1374, 2017.

Adam B Jaffe, Manuel Trajtenberg, and Rebecca Henderson. Geographic localization of knowledge spillovers as evidenced by patent citations. *the Quarterly Journal of Economics*, 108(3):577–598, 1993.

Youngho Kim, Yingshi Tian, Yoonjae Jeong, Ryu Jihee, and Sung-Hyon Myaeng. Automatic discovery of technology trends from patent text. In *Proceedings of the 2009 ACM symposium on Applied Computing*, pages 1480–1487, 2009.

Jeffrey M Kuhn and Neil C Thompson. How to measure and draw causal inferences with patent scope. *International Journal of the Economics of Business*, 26(1): 5–38, 2019.

Mark A Lemley and Carl Shapiro. Probabilistic patents. *Journal of Economic Perspectives*, 19(2):75–98, 2005.

Alan C Marco, Joshua D Sarnoff, and AW Charles. Patent claims and patent scope. *Research Policy*, 48(9):103790, 2019.

Catalina Martínez. Patent families: When do different definitions really matter? *Scientometrics*, 86(1):39–63, 2011.

Martin Meyer. Does science push technology? patents citing scientific literature. *Research Policy*, 29(3):409–434, 2000.

Francis Narin, Kimberly S Hamilton, and Dominic Olivastro. The increasing linkage between us technology and public science. *Research Policy*, 26(3):317–330, 1997.

OECD. Biotechnology patents. *In: OECD Science, Technology and Industry Scoreboard 2009*, pages 66–67, 2009.

Hyunseok Park, Janghyeok Yoon, and Kwangsoo Kim. Identifying patent infringement using sao based semantic technological similarities. *Scientometrics*, 90(2): 515–529, 2012.

Jonathon Douglas Putnam. *The value of international patent rights*. PhD thesis, Yale University, 1996.

Michael Roach and Wesley M Cohen. Lens or prism? patent citations as a measure of knowledge flows from public research. *Management Science*, 59(2):504–525, 2013.

James A Smith, Zeeshaan Arshad, Hannah Thomas, Andrew J Carr, and David A Brindley. Evidence of insufficient quality of reporting in patent landscapes in the life sciences. *Nature Biotechnology*, 35(3):210–214, 2017.

Xuesong Tong and J Davidson Frame. Measuring national technological performance with patent claims data. *Research Policy*, 23(2):133–141, 1994.

Manuel Trajtenberg. A penny for your quotes: patent citations and the value of innovations. *RAND Journal of Economics*, pages 172–187, 1990.

Manuel Trajtenberg, Rebecca Henderson, and Adam Jaffe. University versus corporate patents: A window on the basicness of invention. *Economics of Innovation and new technology*, 5(1):19–50, 1997.

Amy JC Trappey, Charles V Trappey, Chun-Yi Wu, Chin Yuan Fan, and Yi-Liang Lin. Intelligent patent recommendation system for innovative design collaboration. *Journal of Network and Computer Applications*, 36(6):1441–1450, 2013.

Elizabeth Webster, Paul H Jensen, and Alfons Palangkaraya. Patent examination outcomes and the national treatment principle. *The RAND Journal of Economics*, 45(2):449–469, 2014.

Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

Ryan Whalen, Alina Lungeanu, Leslie DeChurch, and Noshir Contractor. Patent similarity data and innovation metrics. *Journal of Empirical Legal Studies*, 17 (3):615–639, 2020.

Janghyeok Yoon and Kwangsoo Kim. Trendperceptor: A property–function based technology intelligence system for identifying technology trends from patents. *Expert Systems with Applications*, 39(3):2927–2938, 2012.

Kenneth A Younge and Jeffrey M Kuhn. Patent-to-patent similarity: a vector space model. *Available at SSRN 2709238*, 2016.