Boosting the signal: Contextualization, extraction, and exploration of in-text patent citations[*]

Cyril Verluise[†] Gabriele Cristelli[‡] Kyle Higham[§] Gaétan de Rassenfosse[¶]

September 2022

Abstract

We apply modern machine learning methods to extract patent-to-patent citations from the text of USPTO patent documents. Overall, we are able to recover around 49 million "in-text" citations, 37 million of which are not found among traditional front-page ones. We show that in-text citations bring a different type of information compared to front-page citations. First, we observe weak relationships between bibliometric measures derived from each citation type, such as forward citation counts and various kinds of self-citation. Second, the positive relationship between inventions' reliance on science and their impact is stronger when measuring the latter using in-text citations. Lastly, in-text citation pairs exhibit higher textual similarity and are more geographically localized than front-page ones. The dataset is available at [weblink redacted to preserve anonymity] (CC-BY-4).

JEL classification: C81, O30
Keywords: Citation, Patent, Open data

[*] We are grateful to Philippe Aghion, Antonin Bergeaud, Francesco Lissoni, Bronwyn Hall, and Adam Jaffe for useful comments. We would like to thank Patrice Lopez and Ian Wetherbee for invaluable help. We are also indebted to Lucas Violon for outstanding research assistance. All errors are ours.

[†] QuantumBlack, Collège de France, and Paris School of Economics, cyril.verluise@gmail.com

[‡] Stanford University, gcristel@stanford.edu

[§] Hitotsubashi University, higham@iir.hit-u.ac.jp (corresponding author)

[¶] École Polytechnique Fédérale de Lausanne, gaetan.derassenfosse@epfl.ch

1 Introduction

Patent documents are an invaluable source of information about technological progress. They provide a detailed account of inventive activities, sometimes as early as the mid-nineteenth century (Sokoloff, 1988; Moser and Nicholas, 2004; Akcigit et al., 2017; Andrews, 2021). Researchers across all fields of sciences and engineering exploit them as a knowledge repository as well as for technology foresight and competitive intelligence analysis, among other applications (Porter et al., 2008; Benson and Magee, 2015; Candia et al., 2019). Researchers in the social sciences exploit them to study various facets of the innovation process.

Early work exploiting patent documents focused on easily accessible metadata; in particular, citations to patents and other public information have proved particularly attractive source of information over the past several decades (Jaffe and de Rassenfosse, 2017). Citations represent explicit relationships among technological and scientific developments, which can be used in their "raw" form for applications such as the measurement of invention "quality" (Carpenter et al., 1981; Trajtenberg, 1990; Higham et al., 2021), or the location of inventions within a broader technological network (Von Wartburg et al., 2005; Fontana et al., 2009; Higham et al., 2022). Further, in combination with other easily accessible metadata such as technology class and applicant location, researchers have used citations to track knowledge flows across spatial (Jaffe et al., 1993; Jaffe and Trajtenberg, 1999), technological (Verspagen and De Loo, 1999; Karvonen and Kässi, 2013), and temporal dimensions (Caballero and Jaffe, 1993; Higham et al., 2017). More recently, the field has been moving towards exploiting the full text of patent documents using a range of more modern computational techniques to better identify technological trajectories, measure invention similarity, and assess various dimensions of patent quality (Kaplan and Vakili, 2015; Younge and Kuhn, 2016; Arts et al., 2018; Kong et al., 2020; Kelly et al., 2021).

In this work, we focus on one aspect of full-text data that has eluded the attention of scholars, namely *in-text citations to patent documents*. Patent offices—and, therefore, the major patent datasets—provide structured data on so-called front-page citations. These citations are usually made for procedural reasons; they list prior art that is relevant for assessing the patentability of the claimed invention and can originate from applicants (or their attorneys and inventors), examiners, or third parties. They may appear on the patent application at the time of filing, be added during the substantive examination before grant, or after grant when a patent is opposed or re-examined. These generation mechanisms make front-page citations conceptually and functionally very different from citations typically found in scientific papers (Meyer, 2000).

In contrast, in-text patent citations appear in the patent text itself. They are made to fulfil enablement requirements; to make arguments for novelty and non-obviousness; or to make arguments for usefulness. As these justifications for adding in-text citations do not perfectly overlap with those that drive the generation of front-page citations, in-text citations likely contain truly unique information over and above that reflected in front-page citations.

Scholars have recently extracted in-text citations to the scientific literature, that is, patent-to-article citations (Bryan et al., 2020; Marx and Fuegi, 2020a; Verluise and de Rassenfosse, 2020). Given the intense interest in patent citation data in the past, the lack of treatment of in-text patent-to-patent citations is a surprising gap in the literature. Such data are likely to be particularly important for specific applications, such as the improved measurement of industrial knowledge flows. Indeed, inventors often contribute to the drafting of the text, and the references they mention are likely to better capture knowledge flows than front-page references. Despite this potential, scarce research exists that test this intuition due to the lack of easily accessible data.

To fill this gap, we have extracted patent citations from the full-text of 16,781,144 publications filed at the U.S. Patent and Trademark Office (USPTO) from 1790 to 2018. About 95 percent of these publications are granted patents or patent applications.[1] For the sake of simplicity, unless specified, we use the term "patent" to designate all publications in the dataset in the rest of the paper. We relied on "Grobid", an open-source machine learning

[1] The remaining 5 percent is composed of design patents, plant patents, reissued patents and statutory invention registration (SIR).

library leveraging Natural Language Processing (NLP) to extract and parse citations (Lopez, 2010), then performed extensive validation exercises, revealing high performance. Our extraction task in particular achieves a 97 percent precision and 82 percent recall (f1-score nearing 90 percent). Overall, we extracted 63,854,733 in-text patent citations, suggesting that in-text patent citations are by no means a marginal phenomenon. A total of 49,409,629 (77.5 %) of them were matched to a commonly-used (DOCDB) publication number, ensuring interoperability with other patent databases.

After extraction, we perform an in-depth quantitative analysis of these citations, with a focus on points of difference between in-text and front-page citations. We discovered four noteworthy elements. First, in-text citations weakly overlap with front-page citations. Overall, we are able to identify more than 37 million citations that are found among in-text and not on the patents' front-page. We observe concomitantly weak relationships between bibliometric measures derived from each type of citation, such as forward citation counts and various kinds of self-citation. Second, we find that the positive relationship between patented inventions' reliance on the scientific literature and their impact is much stronger when the latter is measured using in-text forward citation counts, compared to front-page ones. Third, relative to front-page citations, we find that in-text citations link patents that are more textually similar. This finding suggests that in-text citations generally reflect stronger technological linkages than front-page citations. Fourth, we find that in-text citations display much stronger geographical localisation than front-page citations, even after sample restrictions attempting to isolate “real” spillovers for both citation types.

We interpret this evidence as supporting the intuition that, due to a different generation process, in-text citations may contain unique information, over and beyond the signal provided by their front-page counterpart. In particular, our first two findings suggest that in-text patent citations may provide a valuable signal about patent impact and quality across dimensions that are reflected only in part in front-page citations. Our last two results suggest that in-text patent citations may be a less noisy measure of technical knowledge flows.

The full dataset is publicly available on Google Cloud Big Query and Zenodo. Additional technical documentation and usage guides are available on the project repository and the documentation website. [2] In addition to the final output, we have also released the validation data and the code to allows for replication and follow-on improvements. [3]

The remainder of the document is organized as follows. Section 2 discusses the nature of in-text citations in terms of their legal functions and generation mechanisms, with exemplars, and the practical implications of these attributes. Section 3 briefly describes the citation extraction process and subsequent validation exercises that are used to measure extraction performance, with a detailed technical description made available as an online appendix. Section 4 offers a quantitative overview of in-text citation data, directly comparing them with front-page citations through a battery of empirical tests assessing geographical, technological, and general bibliometric differences. Section 5 concludes and suggests future work that could shine additional light on the information context of in-text citations.

2 The epistemology of in-text citations

There are three patentability requirements enshrined in U.S. patent law that give rise to in-text citations to all types of prior art: to fulfil *enablement* requirements; to make arguments for *novelty and non-obviousness*; and to make arguments for *usefulness*. As these justifications for adding in-text citations do *not* perfectly overlap with those that generate front page citations, in-text citations likely contain truly novel information over and above that reflected in front-page citations. Further, we suggest that this novel information is likely to be associated with inventor input into the drafting process and, therefore, knowledge flows (Bryan et al., 2020). For a similar reason, we argue that in-text patent citations provide a valuable signal of patent importance.

[2] See [weblink redacted to preserve anonymity] for the project documentation.

[3] [Two weblinks redacted to preserve anonymity]

2.1 A legal perspective on in-text patent citations

The in-text citation justifications listed above relate to specific legal obligations that an applicant must fulfil in order for their application to be deemed patentable. While *novelty and non-obviousness* are usually judged by the examiner using direct comparison to the prior art, *enablement* and *usefulness* are also necessary for patentability and are primarily argued by the applicant in the detailed description of the patent application. To concretely demonstrate the use of in-text citations in context, Table 1 displays their different uses, together with their (often explicitly stated) legal purpose.

Enablement is necessary due to 35 U.S Code § 112, which explicitly states:

> *"The specification shall contain a written description of the invention, and of the manner and process of making and using it, in such full, clear, concise, and exact terms as to enable any person skilled in the art to which it pertains, or with which it is most nearly connected, to make and use the same, and shall set forth the best mode contemplated by the inventor or joint inventor of carrying out the invention."*

The enablement requirement is core to the modern conception of a government-issued patent. It ensures that when a patent falls into the public domain, others can (in theory) replicate and use the invention after reading the information in the patent description. Prior art citations may be incorporated by reference where appropriate and can make this description much more succinct; if the construction or use of an invention relies on previously patented or published information, the applicant may cite the information source in the text of the patent specification.[4] These kinds of citations are not necessarily material to the invention's patentability and, when this is the case, not required to be disclosed by the applicant via an information disclosure statement. As such, these "enablement" citations are not necessarily duplicated on the front page of the patent document. This is particularly

[4] 37 CFR 1.57.

true of citations accompanying specific examples that describe how the invention may be used in practice ("best modes"), which may be complementary (and not necessarily similar) to the invention described and may even be hypothetical (Freilich, 2019).

The *novelty and non-obviousness* requirements depend crucially on prior art.[5] For the most part, they are argued for implicitly through Information Disclosure Statements (IDS) submitted by the applicant throughout the application and patent prosecution processes—these references are the citations that appear on the front page of a patent.[6] However, the applicant can also make these arguments explicitly in the patent text by pointing out shortcomings of, or distinctions from, the most pertinent prior art, accompanied by citations to this art. As such, one may expect that citations intended to bolster an argument for novelty or non-obviousness would be duplicated on the front page.

Usefulness, perhaps the most subjective requirement, is described in 35 U.S. Code § 101. It requires the described invention to be "new and useful" to be patentable. The first part of this clause is covered by the novelty and non-obviousness requirements described above. However, the second (usually referred to as the "utility" requirement) requires the invention to be useful to the public as described and, as such, may overlap with *enablement* requirements. The word "useful" is particularly open to interpretation, but generally requires the patented invention to work, and is something that people may want or need (Machin, 1999). In the former case, while there is no burden on the applicant to prove that the invention works (Cotropia, 2009), the applicant may add citations to allay doubts that, for example, a claimed function of the invention is physically possible. An examiner is unlikely to question the latter (Machin, 1999).

[INSERT TABLE 1 HERE]

[5] 35 USC 102; 35 USC 103.

[6] 37 CFR 1.56.

2.2 In-text patent citations as valuable paper trails of knowledge flows

Applicants add in-text citations (to both patents and other bibliographic sources) on their patents for several reasons, necessitated by the above patentability requirements laid out in U.S. patent law. Some of these reasons overlap with those that require applicants to submit an IDS, the prior art listed on which often reach the front page of a granted patent. However, some prior art, and particularly those items deemed necessary to meet enablement or usefulness requirements, do not need to be submitted to the patent office in the form of an IDS because they do not directly limit the scope of the claims in the patent application. Further, examiners do not need specific pieces of the prior art to justify a rejection under the enablement or usefulness requirements.[7] Therefore, the front-page will not contain in-text citations added for these purposes (unless, of course, they are also relevant for the assessment of novelty and non-obviousness).[8]

Due to their resemblance to citations in academic articles, it is tempting to assume that in-text citations are more likely than front-page citations to have been added by the people directly involved in the discovery process, namely the inventors. We suggest that this is probably true, for two reasons. First, the in-text citations that are duplicated on the front page, as prior art material to patentability, are likely the most relevant pieces of prior art against which the invention needs to be judged as novel and non-obvious. The fact that these citations are also in the patent description would imply that they either fulfilled multiple requirements, or were so technologically close to the citing patent that applicants need to make explicit arguments for novelty in the description with reference to specific items in the prior art. In either case, the inventor was likely aware of this prior art during the invention process, or at least collaborated with the patent attorney drafting the application to make

[7] Manual of Patent Examining Procedure, Section 2107.02; Manual of Patent Examining Procedure, Section 2164.

[8] Manual of Patent Examining Procedure, Section 2120.

the appropriate technical distinctions.[9]

Second, those citations that are *not* duplicated on the front page are most likely included to address the enablement or usefulness requirements. While utility is often assumed, and rejections based on lack of utility are rare for most technology types (providing little incentive to add citations (Chien and Wu, 2018)), the enablement requirement states that a hypothetical "person skilled in the art" should be able to make and use the invention, and applicants add in-text citations to assist these hypothetical persons.[10] As such, this information was almost certainly necessary during the invention process, and the inventors were, therefore, aware of it. Believing otherwise would come with the implication that it is the *attorneys* who are writing instructions for those "skilled in the art" and, hence, are at least as skilled as these readers.

Both of the arguments above point towards inventors having more input into selecting in-text citations than they do for front-page citations. For these reasons, we suggest that in-text citations provide a promising measure of knowledge flow.

2.3 In-text patent citations as valuable signals of patent importance

In addition to their utility for capturing noisy signals of knowledge flows, researchers have also used front-page forward citations for decades as indicators of technological impact (Carpenter et al., 1981; Albert et al., 1991). Even if a particular cited patent was not a real knowledge input, the fact that it appears on the front page means that it is likely to be in the same technological space as the citing patent. As such, a patent receiving many front-page citations is either: useful and frequently reused information for the production of new inventions; in a dense technological space against which many new technologies happen to abut, or; a combination of these. This interpretation of front-page forward citation counts

[9] This latter point suggests that even if an inventor was not aware of a prior invention cited in the application's specification, it may nonetheless become part of that inventor's knowledge pool going forward.

[10] 35 USC 112.

is a consequence of the legal purpose of front-page citations; namely, to delineate the prior art material to the patentability of the citing patent. However, this is not the sole purpose of in-text citations.

In-text citation counts, as described above, also serve to fulfill enablement and utility requirements. Applicants sometimes do so by referring to their own patents; for example, firms producing consumer goods may have patents on multiple complementary inventions that, while not necessarily technologically similar, come together in the final product and are cited to demonstrate how the invention is used in practice. For these reasons, the interpretation of a patent accumulating a large number of in-text forward citations can be more complicated than for front-page citations.

On the one hand, the technologically similar inventions cited in-text are those from which the applicant of the citing patent or application has had to provide additional distinction, and, therefore, are likely to be those most likely to be justification for rejection. On the other hand, the technologically complementary inventions cited in-text are likely to be more generalizable technologies, as they are not technologically close enough to the citing patent to be considered material to patentability. Sometimes this relationship is made explicit, as indicated, for example, in U.S. patent 8,524,730 (emphasis added):

> *"More concretely, examples of the other active ingredients that can be combined with a compound of the invention as different or the same pharmaceutical compositions are shown below, which, however,* ***do not restrict the invention****."*
> *(a list of references follows)*

Patents cited in this fashion are not in the same technological space as the citing patent and are cited for their compatibility with other inventions. A large number of these kinds of citations may, therefore, indicate generality outside of the technical domain of the cited invention.

These generation mechanisms for in-text citations color our understanding of how exactly a large number of forward in-text citations relate to the intrinsic properties of the cited patent or invention. However, we strongly suspect that these citations are more likely to originate with the inventors themselves, rather than the attorneys or examiners. This scenario is an interesting one from the point of view of interpretation. The number of reasons for citing a patent in-text are more numerous than those made on the front page, but the resulting citations (often accompanied by context) are more thought-out and meaningful. As an analogy, if front-page citations were a single radio station plagued by significant and persistent static, in-text citations are the result of numerous stations broadcasting loud and clear the same frequency, to the point where it is difficult to make out what any individual station is saying. However, some may prefer this to static—the disentangling of these frequencies is undoubtedly possible to an extent. With both data and code publicly available, future research can build on this work to add further (textual) context to in-text citations and, ultimately, better understand what a highly-cited patent represents in this setting.

3 Citation extraction and validation

To construct the dataset we present in this work, it was necessary to accurately identify, extract and validate citations to other (potentially foreign) patent documents from the full text of over 16 million USPTO patent applications and grants. This task was not trivial and, as such, we provide an online appendix with complete details of each stage of this undertaking, including descriptions of the many challenges associated with the parsing of unstructured patent citations. The remainder of this section will provide an outline of the extraction and validation process.

First, we obtain the full-text specification (i.e., excluding claims and front-matter) of every patent document available in the Google Patents public data that was granted by the USPTO since its inception in 1790. In practice, these data are relatively complete from about 1836, but reliance on optical character recognition for early patents means that extracted citation numbers only reach sufficiently high reliability thresholds, for most use-cases, from about 1976.[11]

In order to extract the citations from the patent specifications, we use established statistical models and methods. In particular, we use the open-source GROBID library by Lopez (2010) to conduct machine-learning-based named-entity recognition, which relies on conditional random fields to extract and classify the patent citations.[12] Once we identify a potential citation, we parse its contents to obtain the issuing authority, the kind code, and the cited patent number itself. We then generate a list of potential synonyms for each citation (such as adding or removing zeroes that act as padding), and feed these through the Google Patents' Linking API to be matched to unique document IDs that can be linked to other well-known databases such as PATSTAT or the USPTO's own PatentsView database.[13] This process resulted in over 49 million in-text citations matched to almost 14 million unique patent documents.

Following the extraction of these citations, we carried out several validation exercises using a scriptable annotation tool, Prodigy.[14] The most notable impacts on the reliability of the extracted citations were due to format changes of USPTO documents (such as the locations of front-page citations). For example, between 1971 and 1975, two patent formats were in use at the USPTO, one more structured than the other. During this period, many of the "specifications" (from which we extract citations) mistakenly include what are, in fact, front-page citations. Noting these kinds of limitations, our human-annotated validation process allowed us to assess the performance of the various stages of the extraction process.

Full performance metrics for every aspect of the above tasks are available in the Online Appendix, (Section 2). In summary, we find that we are able to achieve 97 percent precision for the extraction task and around 95 percent accuracy for the parsing task. Notably, the recall of the extraction task was 82 percent, reflecting the difficulty of extracting citations from older patents. That is, our citation extraction process was conservative in nature—we

[11] Many patents were irretrievably lost in a fire at the USPTO in 1836 (Federico, 1937).

[12] https://github.com/kermitt2/grobid

[13] https://patents.google.com/api/match

[14] https://prodi.gy/

miss about 18 percent of potential citations, but we have high confidence that the citations we do extract are indeed citations listed in the patent specification. The matching process achieved 93 percent precision and 89 percent recall. In the Online Appendix, we conduct additional breakdowns, with specific examples, of the kinds of errors made by the extraction, parsing, and matching processes. Therefore, if needed, users of these data have a starting point for any appropriate pre-processing of the data before application to their use-case.

4 A first look into in-text citation data

Scholars have used front-page patent citations extensively over the past decades and have produced multiple studies assessing their validity as indicators and discussing their pitfalls. The purpose of this section is to provide an overview of the empirical characteristics of in-text citations as compared to front-page citations, focusing on the different bibliometric, semantic, and geographical patterns associated with each citation type. Unless specified, we consider all U.S. patents published from 1790 to 2018. As the results of most of these analyses are presented in graphical form, Table 2 presents a broad numerical summary of some key statistical properties for each citation type.

[INSERT TABLE 2 HERE]

4.1 Bibliometrics

Simple statistical measures. From the 16,781,144 U.S. patents in our dataset, we find that 9,453,181 patents cite at least one patent in the body of their description, corresponding to 56.3 percent of all U.S. patents, compared to 71.3 percent of the this set making at least one front-page patent citation. We also observe high variability in the former figure over time; the share of U.S. patents citing at least one patent in its description has increased from less than five percent in the second half of the nineteenth century to 70 percent in the 2010s.

In total, we extracted 63,854,733 in-text patent citations, about one third of the

203,557,201 front-page citations made by the same set of patents the during same period. While the distribution is very skewed, on average, the body of a patent contains 3.8 patent citations (unconditional average), increasing to 6.7 for the sub-sample of patents that make at least one citation. Once again, there is high variability in these values over time, from an (unconditional) average of less than one in-text patent citation until the early 1960s to more than five since the beginning of the twenty-first century.

Citation overlap. We now examine the citation-level overlap of in-text and front-page citations. Comparing unique citing-cited pairs of each citation type, we obtain three disjoint sets: citations appearing in the text only, citations observed on the front-page only, and citations recorded in both. Figure 1 depicts the number of patent citations appearing in each of these sets. There are 11,868,037 patent citations appearing both in the text and on the front-page, which represent only 5.79 percent of all front-page citations and 24.2 percent of all in-text citations.[15] Note also that, before 1947, front-page patent citations did not exist; before that date, all patent-to-patent citations were available only in-text. Over the whole period, we find 37,541,592 in-text citations that are not found among front-page citations, comprising 15.3 percent of all citations made by patents.

[INSERT FIGURE 1 HERE]

One concern we have about these ostensibly "non-front-page" citations is that they may actually be cited on the front page as non-patent-literature (NPL) if they refer to translations or abstracts rather than the original patent document. To check this possibility, we trained a text classifier to determine whether a front-page NPL citation actually refers to a patent document, achieving a 78.31 percent precision and 89.04 percent recall on the test set. We then applied the classifier to the universe of front-page NPL citations recorded in the DOCDB database and estimate that there are 1,714,260 such citations listed in the front-page NPL sections of U.S. patents since 1947. That is, we estimate the lower bound (when every one of these NPL citations are also made in-text) of the fraction of all citations that are made

[15] These figures include only in-text citations which were successfully matched with a standard publication number.

in-text to be about 14.63 percent.

Self-citations. We consider two forms of self-citation: family-level self-citation and applicant-level self-citation. The former provides very little information of any kind—the patent is effectively citing an earlier version of itself, usually in the "related applications" section of the patent description. The latter kind of self-citations are much more interesting. When studying knowledge spillovers, researchers generally remove applicant-level self-citations as they do not reflect knowledge "spilling over" into the inventive activities of other assignees. At the same time, however, they may contain information about continued development a firm's own inventions.

To assess family-level self-citations, we first map each citing and cited patent to its patent family and compute the share of citations citing a patent belonging to its own family, considering DOCDB simple families and INPADOC extended families separately (Martinez, 2010).[16] We find that the share of in-text citations belonging to the same DOCDB (INPADOC) family is 6.42 (10.65) percent. This is significantly higher than front-page citations' self-references figures, which are 0.69 (1.63) percent, reflecting the family linkages that are explicitly noted in the "related applications" section. From a practical perspective, these citations are very easy to omit and, in any case, the vast majority of in-text citations are not of this type.

Turning to applicant-level self-citation, we look at the share of citations having at least one common inventor or at least one common assignee. We rely on the harmonized names reported in the IFI CLAIMS dataset, labeling as same-patentee citations those where the name of at least one inventor (assignee) is the same for the citing and cited patent. This choice casts a wide net—inventors can move between firms, and harmonization is more likely to unintentionally merge two names that actually refer to different entities than split one into two entities (in cases of, e.g., typos). Both of these effects will lead to more "self-citations." We find that 17.43 (22.46) percent of in-text patent citations have at least one inventor

[16] INPADOC families are more permissive than DOCDB families as they group together all documents sharing directly or indirectly (e.g., via a third document) at least *one* priority filing.

(assignee) in common with their citing patent, compared to 5.98 (9.26) percent for front-page citations. This result confirms the relative importance of self-reliance in knowledge creation which appears to be even more visible through the lens of in-text citations, making these citations potentially a valuable new source of information for research on within-firm innovation or the activities of individual inventors.

4.2 Forward citations and dimensions of "quality"

Forward citations, those that are received from subsequent patents, have been extensively used by innovation scholars across a variety of contexts (Albert et al., 1991; Jaffe et al., 1993; Hall et al., 2001; Harhoff et al., 1999, 2003; Jaffe and de Rassenfosse, 2017). This section reports some statistical comparisons between forward citations received through in-text citations and those received through front-page citations.

First, as suggested above, we remove self-citations at the INPADOC family level so as to only include citations that are unlikely to be listed in a "related applications" section (i.e., that are not "true" in-text citations with respect to the information in Section 2). We then aggregate forward citations at the DOCDB family-level (rather than at the publication level), in order to avoid repeated counts of citations effectively relating to the same invention pairs. This choice is more conservative than INPADOC families, which may contain closely related inventions that are still distinct (Martinez, 2010).

First, we see that in-text and front-page citations are respectively directed to two partly disjoint sets of DOCDB patent families. In-text citations point to 5,506,374 distinct families, front-page citations point to 13,817,609 distinct families, and 4,262,548 of these families are in both sets. That is, for 23 percent of families that receive in-text citations, those citations are the only direct links to subsequent technological developments that we can observe using patent data. Second, forward citation counts based on in-text citations are only weakly related to the same metric obtained from front-page citations. Restricting to the set of DOCDB patent families with a positive count of forward citations both on the front-page and in the text, the raw correlation between the two measures is 0.23.

To add further color to the correlation analysis, we look at the correlation between in-text citations counts and the counts for applicant, examiner, and total front-page citations as a function of family priority date. For comparability across a broad time-span, we limit the citation lag to 10-years between the priority dates of the cited and citing patents; we only count citations from granted patents; and we only consider one citation per cited-citing family pair (i.e., we ignore duplicated citations on continuing filings). We calculate the correlations by priority date for the subset of patents that receive at least one in-text citation, and additionally run the same calculations for the logarithms of the citation counts to mitigate the impact of outliers. Lastly, we consider priority dates from 1980 to 2010 for the correlations with total front-page citation counts, and from 2000 to 2010 for correlations with applicant and examiner citation counts (which were only made accessible for patents granted after January 1st, 2001).

Figure 2 depicts the results of this analysis, allowing us to draw several conclusions. First, the correlation between in-text citations and total front-page citations is remarkably stable at around 0.3 once we consider the logarithms of the citation counts. Second, the raw correlation with front-page applicant citations dramatically increases for patent families filed from the mid-2000s, but the correlation between the logarithms of the same counts remains stable. These patterns suggest that the abrupt increase appears to be driven by a small set of outliers that receive many citations of both types. A manual check of some of these outliers suggests that this phenomenon is due to the copying of in-text and front-page citations across patents whereby firms or patent attorneys use boilerplate citation sets across related inventions. Lastly, the raw correlation between in-text citations and examiner citations has remained stable in recent years, but we note a dramatic fall once we use the logarithms of these counts.

These patterns suggest a real decoupling of in-text citation counts from examiner citation counts, perhaps masked by outliers on the citing side of the citation that make many in-text citations from which examiners draw (e.g., from the boilerplate sets described above). Indeed, we confirm that upon removal of the top 1 percent most-cited patents of each type, the correlation time series for raw and log-transformed citations counts (unreported) are almost identical and resemble the log-transformed series in Figure 2. All of these observations point to the conclusion that recent trends in applicants' front-page citation strategy (Kuhn et al., 2020) have likely affected in-text citations as well, though to a lesser extent and likely through different mechanisms. In any case, we suggest that researchers should apply correctional strategies to in-text citations where appropriate and in a similar manner to those applied to front-page citations. For example, when studying knowledge flows, one could consider only those in-text citations that originate from patents that make less than some threshold number of citations. Such a measure could be implemented with the assumption that any individual citation made by those patents containing hundreds or thousands of citations in their specifications is unlikely to reflect real knowledge flow (Kuhn et al., 2020).

[INSERT FIGURE 2 HERE]

The third set of observations relates to the distribution of forward citation counts. Figure 3 compares the empirical probability (panel 3a) and cumulative (panel 3b) distribution function of forward citations counts. It reveals two notable properties. First, the front-page distribution stochastically dominates the in-text distribution. Second, the tail of the front-page distribution is larger. This is perhaps unsurprising—on average, many more citations are made on the front page than in text, leading to more opportunity for a patent to be cited and build a cumulative advantage for future citations that may not be entirely reflective of quality (Higham et al., 2019). These observations can be partly explained by a fundamental difference in the citation generating process. Whereas in-text citations are mostly in the hand of the inventors, hence decentralized among many agents, front-page citations are determined by a finite number of examiners who, by nature, are likely to be aware of a limited number of patents on each subject. This leads to the emergence of highly cited patents, the so-called 'focal patent,' which participate in the larger tail observed in the distribution of front-page forward citations count. It is interesting to note that 'focal patents' might well be so partly independently on their intrinsic social or private value but because of examiners' biases.

[INSERT FIGURE 3 HERE]

Recent research has shown that patented inventions directly relying on scientific knowledge tend to be highly impactful and valuable (Ahmadpoor and Jones, 2017; Poege et al., 2019; Marx and Fuegi, 2020b; Watzinger et al., 2021). (Front-page) forward citations have been employed as a measure of patent impact and value. Our last test in this section focuses on the relationship between patents' reliance on scientific knowledge and forward citations, comparing front-page and in-text citations. We follow previous studies and measure patents' dependence on scientific knowledge using backward citations to articles published in academic journals, extracted from the patent text by Verluise and de Rassenfosse (2020).

We start by assessing the univariate relationship between backward citations to scientific articles and in-text and front page forward citations. Our sample includes USPTO patents granted between 1980–2010 with at least one inventor reporting a U.S. address, a total of 1,259,314 patents. Similarly to Kogan et al. (2017), we aggregate our data into 100 quantiles based on their forward citation counts (either in-text or front page) and plot the average number of citations in each quantile versus the average value of backward citations to science (both log-transformed). Before aggregation, we scale each patent's forward and backward citation count by their mean value in the same grant year cohort. Figure 4 shows a strong positive correlation between backward citations to the scientific literature and in-text forward citations. We notice also a positive correlation between backward citations to science and front-page forward citations, although decisively weaker than their in-text counterpart.

[INSERT FIGURE 4 HERE]

Next, we deepen our investigation using econometric tools. We adopt Ahmadpoor and Jones's (2017) strategy and regress an indicator for top-cited patents—either by front-page or in-text citations—on an indicator for patents citing at least one scientific article in their text. Top-cited patents are defined as those in the top 95th percentile of forward citations in a given grant year × NBER technology subcategory cohort. We classify 69,068 patents as front-page top-cited and 74,985 patents as in-text top-cited. Both groups include some patents classified as top-cited by both citation types, which in total are 27,676 patents. A total of 329,273 patents in our sample (around 26% of the total) cite at least on scientific article in their text. Our regressions include fixed effects for a patent's grant year, NBER technology subcategory, assignee type (*i.e.,* private firm, universities, or government laboratory), inventor team size, and total number of (front-page) backward citations.

Table 3 reports estimation results from linear probability models. Column (1) shows that patents citing at least one scientific article are 3 percentage points more likely to be a front-page top-cited patent. In line with our descriptive results, column (2) shows that patents citing the scientific literature are 4.5 percentage points more likely to be an in-text top-cited patent. In column (3) we compare the two forward citation groups more explicitly, restricting the sample to top-cited patents. In this case, the dependent variable is a dummy equal to 1 for in-text top-cited patents and 0 for front-page top-cited ones. We estimate that patents citing scientific literature are around 8 percentage points more likely to be an in-text top-cited patent rather than a front-page top-cited one. In column (4), we replicate this last estimation excluding from the sample any patent classified as top-cited by both citation groups. The estimated coefficient only marginally shrinks, remaining statistically significant. In Appendix Table 12 we replicate our estimates using Marx and Fuegi's (2020) in-text backward citations to the scientific literature and obtain nearly identical results.[17]

[INSERT TABLE 3 HERE]

Taken together, our results indicate that the relationship between patented inventions' science dependence and in-text forward citations is similar in nature to that found for front-page citations by previous studies. Such a positive link, however, is stronger for in-text citations, driven especially by patents in top percentiles of the forward citation distribution.

[17] Compared to Ahmadpoor and Jones (2017), our estimates are slightly larger. We believe this result may be due to the use of in-text backward citations to scientific articles (rather than front-page ones) and to the different geographical and temporal scope of our sample.

Based on this evidence, it could be argued that in-text and front-page forward citation counts measure different dimensions of a patent's impact. Being more likely to come directly from inventors and resembling citations between scientific articles, in-text citations would measure the strict technological impact of a patent. Front-page citations, instead, define the set of prior art delimiting the scope of the intellectual property rights associated to the patent. As such, they would be more related to a patent's "legal" importance, its ability to restrict the breadth of the intellectual property rights assigned to subsequent patents and to limit competitors (potentially generating large private rents for the patent assignee), all characteristics only imperfectly correlated with technological impact. That would explain the positive albeit weak correlation between front-page and in-text citations described earlier in the section.

4.3 Textual similarity between citing and cited patents

In recent years, researchers have been increasingly interested in the link between technological or semantic similarity of patents and their value, novelty, or other quality dimension (Arts et al., 2021a,b; Kelly et al., 2021). It has also been noted that the similarity reflected by front-page citations has decreased over time (Whalen et al., 2020; Kuhn et al., 2020). Here, we contribute basic information about the relationship between in-text citations and semantic similarity.

We calculate semantic similarity for a given patent pair as the dot product of Google Patent's document embedding vectors, which were recently made available to researchers through Google Patents Public Datasets.[18] The embeddings are trained to predict CPC categories from each patent's full-text with a WSABIE algorithm (Weston et al., 2010). Figure 5 shows the pair-wise similarity distributions for in-text and front-page citations, alongside two reference distributions. The first reference distribution ("Within art unit")

[18] https://tinyurl.com/googlepatentdata. We note that a myriad of similarity measures exist (Younge and Kuhn, 2016; Arts et al., 2018, 2021c; Whalen et al., 2020); however, for the purposes of the current work, we required a low-dimensional vector form that could quickly and intuitively estimate the semantic distance between a large set of patents and choose Google Patents' embeddings for this reason.

is based on the similarity between randomly chosen pairs of patents examined by the same art unit. The second reference distribution ("Random") is based on the similarity between cited in-text patents matched to a random citing patent. Specifically, this set is constructed from our set of in-text citations, from which we randomly redirect each citation from their original destination to a different patent within the set, retaining the same set of citing and cited patents but with their citations reconfigured.

For ease of interpretation, we only use citing and cited patents that were granted by the USPTO in the years 2000–2009 to produce Figure 5. In Figure 5a we additionally exclude all within-INPADOC-family citations occurring for in-text and front-page citations (N=325,247). Pairs of patents used for the "Within art unit" and the "Random" distributions have been randomly omitted to match this new sample size. Citations between patents belonging to the same INPADOC family are much more common in the patent text than on the front-page, and removing them improves the comparability of the similarity distributions. In Figure 5b we report the same similarity distributions, excluding citations between patents belonging to the same DOCDB family. The in-text citation similarity distribution shown in Figure 5b clearly includes many near-identical patents, owing to the complexity of priority filing strategies. For this reason, we will focus on the distributions excluding within-INPADOC-family citations (Figure 5a), as they are more comparable to front-page citations.

[INSERT FIGURE 5 HERE]

We make several observations from this graphical comparison of similarities. First, in agreement with our validation measures, it is unlikely that a large portion of in-text citations are incorrectly matched, as these would be drawn from the random distribution. Indeed, because we cannot see a conspicuous lump in the in-text similarity distribution in the region where the random distribution peaks and because the shape is similar to that of the front-page citations, we may conclude that the error rates in these two sets of citations are similar. Second, the modal peak of the in-text citation distribution is shifted to slightly higher levels of similarity when compared to the distribution for front-page citations. This difference indicates that patents cited in-text are, on average, slightly more similar to the citing patent than patents cited on the front-page. Such a pattern would be expected when inventors need to make a particular effort to distinguish their application from the most similar prior art through in-text explanations (see, e.g., Table 1). Lastly, the in-text citation distribution displays a fatter tail at lower similarity levels, particularly around the similarity level expected from patents examined by the same art unit. This pattern is expected. Because patents cited in the patent text do not necessarily impact on patentability and do not have to be technologically similar to serve their purpose, they can be drawn from a wider (but still related) set of prior art. However, Figure 5a appears to indicate that these citations constitute a small minority of in-text citations, at least to the extent that semantic similarity can measure this kind of relationship.

This analysis reinforces our view of in-text citations as a promising indicator of knowledge flows, and are potentially a less "noisy" source of this information than front-page citations, in addition to being less affected by examiners' (Alcacer and Gittelman, 2006) or patent attorneys' (Jaffe et al., 2000) inputs into the patent prosecution process.

4.4 Geographic distribution and localization

Since the seminal study of Jaffe et al. (1993), a vast literature has examined the geographic spread of technological knowledge flows, finding them to be localized (*e.g.,* Almeida and Kogut, 1999; Peri, 2005; Singh, 2005; Thompson and Fox-Kean, 2005; Breschi and Lissoni, 2009; Singh and Marx, 2013). Patent (front-page) citations have been a crucial data source for these studies, employed as a proxy for the elusive "paper trail" of knowledge (Krugman, 1991) connecting patented inventions. However, despite their widespread use, front-page citations display non-negligible limitations. Scholars have argued that their generation process might be affected by applicants' strategic considerations (Sampat, 2010; Lampe, 2012, although see also Kuhn et al., 2021) or that they are sometimes added by patent attorneys instead of inventors (Jaffe et al., 2000), making them a noisy proxy of knowledge flows (Jaffe et al., 1998; Duguet and MacGarvie, 2005) or even a measure unlikely to capture them at all (Arora et al., 2018). In this section, we compare the geographic properties of in-text and front-page citations.

First, we take citing-cited inventor dyads in the two citation groups and calculate the distance, in kilometers, between the two inventors' geocoded addresses (de Rassenfosse et al., 2019). We consider citing patents granted between 1980–2010 and we exclude self-citations at the INPADOC family level. Figure 6 shows the probability distribution function of in-text and front-page citing-cited inventor dyads, for the entire sample (panel a) and for citation pairs within 200 kilometers (panel b). Both graphs portray in-text citations as slightly more localized than front-page ones. Panel (b) in particular, shows a higher share of in-text citations within a 50-kilometer distance.

Second, using the same sample, we look at the distribution of in-text and front-page citations across cited inventors' continents, considering the United States independently (Figure 7). In panel (a), we focus on citing patents with at least one inventor reporting a U.S. address. The majority of both in-text and front-page citations are directed to prior art from the United States, but in-text citations are directed to U.S.-based inventors at a slightly higher proportion than those on the front-page (respectively 89% and 86%). Front-page citations exhibit greater proportions of references to patents from outside the United States, particularly Europe. In panel (b), we exclude all citing patents listing one or more U.S.-based inventors. The majority of front-page citations are still directed to prior art from the United States (63%). The biggest group of in-text citations, however, is to prior art from Europe (49%), followed by prior art from the United States (43%).

The evidence from Figure 6 and 7 suggests that in-text citations may be more localized than front-page ones. We further investigate this possibility by comparing in-text citations with two subgroups of front-page citations, those added by the applicant and those added by the examiner, the data for which are available for patents granted from 2001 onwards.

Applicant front-page citations have been argued to be a less biased proxy of knowledge flows than examiner ones (Jaffe et al., 2000; Alcacer and Gittelman, 2006). As in Marx and Fuegi's study of patent-to-article citations, we adopt the econometric approach of Thompson (2006), regressing a measure of geographic distance between citing and cited patent on a citation category indicator and citing patent fixed effects. The sample includes patents granted between 2001–2010 and we continue to exclude self-citations at the INPADOC family level.

Table 4 reports our results. In line with Thompson (2006), we find that applicant front-page citations are more localized than those added by the examiner. This finding is true for both a coarse outcome, the probability of citing a patent originating from the same country (column 1, panel A), and a fine-grained outcome, the logarithmic transformation of distance between citing and cited patent (column 1, panel B). We find that in-text citations are also more localized than examiner front-page ones, but to a much greater extent than applicant front-page citations. Column (2) in panel A shows that in-text citations are, on average, 7 percentage points more likely to connect patents from the same country than examiner citations, while applicant citations only about 2 percentage points more likely to do so. Column (2) in panel B indicates that in-text citations are, on average, approximately 67 percent closer than examiner front-page citations, compared to a difference of about 14 percent between applicant and examiner front-page citations. Column (3) shows that in-text citations are also significantly more localized than applicant front-page citations, although the coefficients are slightly smaller than those estimated in comparisons with examiner citations.

The economically and statistically significant greater localization of in-text citations persists when we restrict our sample to citation pairs at up to 200 kilometers (panel C) or when we focus only on citation pairs within the United States (panel D).[19] Our results are also robust to the use of restricted samples that exclude citations older than ten years, patents with more than 100 front-page citations, and applicant self-citations (columns 4, 5, and

[19] Interestingly, when restricting the sample to citations at up to 200 kilometers, applicant front-page citations are *less* localized than examiner-added ones. However, the effect size is less than 2 percentage points.

6).[20] The persistence of in-text citations' greater localization when we exclude applicant self-citations is particularly interesting, as it implies that in-text citations may be used to better capture technological knowledge flows (or even spillovers) not only within, but also between organizations.

Taken together, our results indicate that patent-to-patent in-text citations are decisively more localized than front-page ones. Is this evidence that in-text citations better capture inventors' knowledge? A precise answer would require comparing in-text and front-page citations with a "ground-truth" measure of scientists' and engineers' knowledge inputs to the patented invention, as provided by R&D laboratories' or inventor surveys (Cohen et al., 2000; Jaffe et al., 2000). Nevertheless, our results might still suggest that in-text citations are a less noisy proxy of inventors' knowledge than front-page ones, under the assumption that knowledge flows are anchored to geography through inventors' locations. Just as greater localization of applicant front-page citations relative to examiner ones might reflect the presence of fewer references unknown to the inventors among the former, in-text citations' strong bias towards the locations where inventors work and live (and possibly source a large portion of the their knowledge) could be the result of even fewer "noisy" references being made in patent specifications.

5 Discussion and conclusions

This paper puts the focus on a neglected component of patent-to-patent citations. It does so by introducing a novel dataset on patent citations that provides 63,854,733 million citations identified in the full-text of 16,781,144 million U.S. patent documents from 1790 to 2018. To the best of our knowledge, it is the first openly-released and extensively validated dataset of the sort. Given the importance of citation data across the social sciences, we expect these

[20] The results in columns (4)–(6) indicate a smaller difference between in-text and front-page citations than our baseline estimates, while a larger one between applicant and examiner front-page ones (except in panel A). Nevertheless, in-text citations are still shown to be substantially more localized than both groups of front-page ones. Appendix Table 13 reports three sets of results for the exclusion of each category of citations or patents.

data to be of considerable interest to the scientific community.

Several significant observations can be made, and conclusions drawn, from the extraction and exploration of these new data. First, we find little overlap between commonly-used front-page citations and the in-text citations we extract. The inclusion of in-text citations adds approximately 15 percent more patent citations compared to using front-page citations alone.

Second, in addition to adding *more citations*, the inclusion of in-text citations also adds information of *a different nature* due to a different data generation process compared with front-page citations. In particular, we have argued that, conceptually, these generation processes are more likely to lead to a less noisy signal of knowledge flows compared to front-page citations. We provide preliminary empirical evidence supporting this claim through analyses of the relationship between in-text citations and their associated semantic and geographic properties. We also suggest that in-text citation generation mechanisms are likely to provide valuable signals about patent quality across dimensions that are not reflected in front-page citations. Capturing knowledge flow and measuring patent quality are two of the most popular uses of patent citations and, as such, we expect these new data to be of great utility to the research community.

In conclusion, we hope that the public release of the dataset, alongside our contextualization and initial empirical explorations, will enable further study within the innovation studies field and beyond.

5.1 Future work

In this work, we present in-text citations having mostly distinct origins to front-page citations, while at the same time advocating for their use in place of, or in addition to, front-page citations for certain applications. Indeed, while front-page patent citations have served as indicators for a wide variety of phenomena in the past (Jaffe and de Rassenfosse, 2017), we suggest that lack of alternatives have led to an over-dependence on these data. In particular, their actual information content regarding patent quality and knowledge spillovers have faced criticism in recent years while front-page citation generation mechanisms appear to be systematically decoupling citation-based measurement from these phenomena (Arora et al., 2018; Kuhn et al., 2020; Higham et al., 2021). In-text citations appear to be a promising alternative, particularly with respect to the measurement of knowledge flows, but more work is needed to measure the signal present in these citations.

We assert that in-text citations are the result of very different data-generation processes to front-page citations, and that these processes are shaped by very different incentives and parties during drafting and application. We also find, empirically and consistent with the above assertion, that there is little overlap between these two sets of citations and that the similarity profiles of linked patent documents are distinct. Of course, this does not rule out the possibility that both citation types could be noisy indicators of the same phenomena, but it is unclear whether in-text citations should be thought of as a substitute or complement to forward citations at present.

In terms of striking a balance between the importance of a phenomenon and our ability to reliably measure it, continued validation of the information that in-text citations may hold about knowledge flows appear to be an ideal candidate for future work. To this end, we suggest that a large scale survey of inventors regarding the relationship between patent citations and knowledge flow is well overdue, the last being over 20 years ago (Jaffe et al., 2000).[21] The survey route is clearly very resource intensive, but with the use of “new” tools that may be utilized in such a survey, including internet-based professional networks (formal and informal), such a survey could provide a large ground-truth data set for future work on technical knowledge flows of many kinds.

There are also proxies for testing whether knowledge flow is reflected in patent citations. The most common of which are processes of elimination—by controlling for as many confounding factors as possible that might be mistaken for knowledge flow (but are not), we can

[21] We also note a survey of inventors in Japan (Nagaoka and Yamauchi, 2015); however, this survey focused on their use of science and the associated citations to non-patent literature.

try to make the case that whatever is left is real 'knowledge flow.' This general philosophy was first adopted by Jaffe et al. (1993), and has been relied upon by many others since to provide evidence that front-page citations contain information about interpersonal knowledge flow. Our preliminary geographical analyses of in-text citations in this work demonstrate great potential for a more comprehensive empirical approach. Such an analysis could, for example, jointly assess the roles of both physical and social distances (Breschi and Lissoni, 2009; Diemer and Regan, 2022).

A second avenue for future research could infer the sources of different kinds of citations with respect to the knowledge pools of the parties involved in the drafting and examination process. Most examiners, inventors, and attorneys have a history consisting of patented inventions they have examined, invented, or drafted, and the citations on these patents in each parties history makes up a knowledge pool of potential future citations. As such, it is possible to check whether different kinds of citations are more likely to be drawn from different sources and make inferences about who knew of a particular invention of the past at the time that their new patent application was filed. This kind of analysis may highlight differences between citation types in terms of the likely citation origin, which will in turn provide information about whether in-text citations are good at capturing inputs into the invention process. Because in-text citations are unlikely to have been added by examiners, for reasons discussed in section 2, this undertaking would aim to compare front-page applicant citations and in-text citations as coming from either an inventor or an attorney/agent.

In-text citations will undoubtedly be applied and stress-tested in a wide variety of ways in the near future. To this end, a full complement of empirical tools will be required to establish a complete picture of the information content of this novel data source. Alongside these efforts we see great potential for more comprehensive, nuanced, or alternative analyses of established phenomena of interest that would benefit from additional citation context such as strategic disclosure (Lampe, 2012; Kuhn et al., 2021), knowledge sourcing and technological search (Stuart and Podolny, 1996; Almeida, 1996; Wagner et al., 2014; Corsino et al., 2019), and patent generality (Hall and Trajtenberg, 2004; Raiteri, 2018; Higham and Yoshioka-Kobayashi, 2022).

References

Ahmadpoor, M. and Jones, B. F. (2017). The dual frontier: Patented inventions and prior scientific advance. *Science*, 357(6351):583–587.

Akcigit, U., Grigsby, J., and Nicholas, T. (2017). Immigration and the rise of american ingenuity. *American Economic Review*, 107(5):327–31.

Albert, M. B., Avery, D., Narin, F., and McAllister, P. (1991). Direct validation of citation counts as indicators of industrially important patents. *Research Policy*, 20(3):251–259.

Alcacer, J. and Gittelman, M. (2006). Patent citations as a measure of knowledge flows: The influence of examiner citations. *The Review of Economics and Statistics*, 88(4):774–779.

Almeida, P. (1996). Knowledge sourcing by foreign multinationals: Patent citation analysis in the us semiconductor industry. *Strategic Management Journal*, 17(S2):155–165.

Almeida, P. and Kogut, B. (1999). Localization of knowledge and the mobility of engineers in regional networks. *Management Science*, 45(7):905–917.

Andrews, M. J. (2021). Historical patent data: A practitioner's guide. *Journal of Economics & Management Strategy*, 30(2):368–397.

Arora, A., Belenzon, S., and Lee, H. (2018). Reversed citations and the localization of knowledge spillovers. *Journal of Economic Geography*, 18(3):495–521.

Arts, S., Cassiman, B., and Gomez, J. C. (2018). Text matching to measure patent similarity. *Strategic Management Journal*, 39(1):62–84.

Arts, S., Cassiman, B., and Hou, J. (2021a). Technology differentiation and firm performance. *Harvard Business School Strategy Unit Working Paper*, (22-040).

Arts, S., Hou, J., and Gomez, J. C. (2021b). Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. *Research Policy*, 50(2):104144.

Arts, S., Hou, J., and Gomez, J. C. (2021c). Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. *Research Policy*, 50(2):104144.

Benson, C. L. and Magee, C. L. (2015). Quantitative determination of technological improvement from patent data. *PLOS ONE*, 10(4):e0121635.

Breschi, S. and Lissoni, F. (2009). Mobility of skilled workers and co-invention networks: an anatomy of localized knowledge flows. *Journal of Economic Geography*, 9(4):439–468.

Bryan, K. A. and Ozcan, Y. (2021). The impact of open access mandates on invention. *Review of Economics and Statistics*, 103(5):954–967.

Bryan, K. A., Ozcan, Y., and Sampat, B. (2020). In-text patent citations: A user's guide. *Research Policy*, 49(4):103946.

Caballero, R. J. and Jaffe, A. B. (1993). How high are the giants' shoulders: An empirical assessment of knowledge spillovers and creative destruction in a model of economic growth. *NBER Macroeconomics Annual*, 8:15–74.

Candia, C., Jara-Figueroa, C., Rodriguez-Sickert, C., Barabási, A.-L., and Hidalgo, C. A. (2019). The universal decay of collective memory and attention. *Nature Human Behaviour*, 3(1):82–91.

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

Chien, C. V. and Wu, J. Y. (2018). Decoding patentable subject matter. *Patently-O Patent*

Law Journal 1, Santa Clara University Legal Studies Research Paper, 1.

Cohen, W. M., Nelson, R. R., and Walsh, J. P. (2000). Protecting their intellectual assets: Appropriability conditions and why US manufacturing firms patent (or not). *NBER Working Paper*, (w7552).

Corsino, M., Mariani, M., and Torrisi, S. (2019). Firm strategic behavior and the measurement of knowledge flows with patent citations. *Strategic Management Journal*, 40(7):1040–1069.

Cotropia, C. A. (2009). The folly of early filing in patent law. *Hastings Law Journal*, 61:65–129.

de Rassenfosse, G., Kozak, J., and Seliger, F. (2019). Geocoding of worldwide patent data. *Scientific Data*, 6(1):1–15.

Diemer, A. and Regan, T. (2022). No inventor is an island: Social connectedness and the geography of knowledge flows in the US. *Research Policy*, 51(2):104416.

Du Plessis, M., Looy, B. V., Song, X., and Magerman, T. (2009). Data production methods for harmonized patent indicators: Assignee sector allocation. *EUROSTAT Working Paper and Studies*.

Duguet, E. and MacGarvie, M. (2005). How well do patent citations measure flows of technology? Evidence from french innovation surveys. *Economics of Innovation and New Technology*, 14(5):375–393.

Federico, B. M. (1937). The patent office fire of 1836. *J. Pat. Off. Soc'y*, 19:804.

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

Freilich, J. (2019). Prophetic patents. *UC Davis Law Review*, 53:663–731.

Hall, B. H., Jaffe, A. B., and Trajtenberg, M. (2001). The NBER patent citation data file: Lessons, insights and methodological tools. Working Paper 8498, National Bureau of Economic Research.

Hall, B. H. and Trajtenberg, M. (2004). Uncovering GPTs with patent data. *NBER Working Paper*, (w10901).

Harhoff, D., Narin, F., Scherer, F. M., and Vopel, K. (1999). Citation frequency and the value of patented inventions. *Review of Economics and Statistics*, 81(3):511–515.

Harhoff, D., Scherer, F. M., and Vopel, K. (2003). Citations, family size, opposition and the value of patent rights. *Research policy*, 32(8):1343–1363.

Higham, K., Contisciani, M., and De Bacco, C. (2022). Multilayer patent citation networks: A comprehensive analytical framework for studying explicit technological relationships. *Technological Forecasting and Social Change*, 179:121628.

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

Higham, K. and Yoshioka-Kobayashi, T. (2022). Patent citation generation at the triadic offices: Mechanisms and implications for analysis. *Available at SSRN 4022851*.

Higham, K. W., Governale, M., Jaffe, A., and Zülicke, U. (2017). Fame and obsolescence: Disentangling growth and aging dynamics of patent citations. *Physical Review E*, 95(4):042309.

Higham, K. W., Governale, M., Jaffe, A., and Zülicke, U. (2019). Ex-ante measure of patent quality reveals intrinsic fitness for citation-network growth. *Physical Review E*, 99(6):060301.

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

Jaffe, A. B., Fogarty, M. S., and Banks, B. A. (1998). Evidence from patents and patent citations on the impact of NASA and other federal labs on commercial innovation. *The Journal of Industrial Economics*, 46(2):183–205.

Jaffe, A. B. and Trajtenberg, M. (1999). International knowledge flows: Evidence from patent citations. *Economics of Innovation and New Technology*, 8(1-2):105–136.

Jaffe, A. B., Trajtenberg, M., and Fogarty, M. S. (2000). Knowledge spillovers and patent citations: Evidence from a survey of inventors. *American Economic Review*, 90(2):215–218.

Jaffe, A. B., Trajtenberg, M., and Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *The Quarterly Journal of Economics*, 108(3):577–598.

Kaplan, S. and Vakili, K. (2015). The double-edged sword of recombination in breakthrough innovation. *Strategic Management Journal*, 36(10):1435–1457.

Karvonen, M. and Kässi, T. (2013). Patent citations as a tool for analysing the early stages of convergence. *Technological Forecasting and Social Change*, 80(6):1094–1107.

Kelly, B., Papanikolaou, D., Seru, A., and Taddy, M. (2021). Measuring technological innovation over the long run. *American Economic Review: Insights*, 3(3):303–20.

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.

Kong, N., Dulleck, U., Jaffe, A. B., Sun, S., and Vajjala, S. (2020). Linguistic metrics for patent disclosure: Evidence from university versus corporate patents. Technical report, National Bureau of Economic Research.

Krugman, P. R. (1991). *Geography and trade*. MIT press.

Kuhn, J., Younge, K., and Marco, A. (2020). Patent citations reexamined. *The RAND Journal of Economics*, 51(1):109–132.

Kuhn, J., Younge, K., and Marco, A. (2021). Strategic citation: A reassessment. *The Review of Economics and Statistics*, pages 1–24.

Lampe, R. (2012). Strategic citation. *Review of Economics and Statistics*, 94(1):320–333.

Lopez, P. (2010). Automatic extraction and resolution of bibliographical references in patent documents. In *Information Retrieval Facility Conference*, pages 120–135. Springer.

Machin, N. (1999). Prospective utility: A new interpretation of the utility requirement of Section 101 of the Patent Act. *California Law Review*, 87:421–456.

Martinez, C. (2010). Insight into different types of patent families. *OECD Science, Technology and Industry Working Papers*, 2010(2):1.

Marx, M. and Fuegi, A. (2020a). Reliance on science by inventors: Hybrid extraction of in-text patent-to-article citations. *Journal of Economics & Management Strategy*.

Marx, M. and Fuegi, A. (2020b). Reliance on science: Worldwide front-page patent citations to scientific articles. *Strategic Management Journal*, 41(9):1572–1594.

Meyer, M. (2000). What is special about patent citations? Differences between scientific and patent citations. *Scientometrics*, 49(1):93–123.

Moser, P. and Nicholas, T. (2004). Was electricity a general purpose technology? Evidence from historical patent citations. *American Economic Review*, 94(2):388–394.

Nagaoka, S. and Yamauchi, I. (2015). The use of science for inventions and its identification: Patent level evidence matched with survey. *Research Institute of Economy, Trade and Industry (RIETI)*.

Peri, G. (2005). Determinants of knowledge flows and their effect on innovation. *Review of Economics and Statistics*, 87(2):308–322.

Poege, F., Harhoff, D., Gaessler, F., and Baruffaldi, S. (2019). Science quality and the value of inventions. *Science advances*, 5(12):eaay7323.

Porter, A. L., Youtie, J., Shapira, P., and Schoeneck, D. J. (2008). Refining search terms for nanotechnology. *Journal of Nanoparticle Research*, 10(5):715–728.

Raiteri, E. (2018). A time to nourish? Evaluating the impact of public procurement on technological generality through patent data. *Research Policy*, 47(5):936–952.

Sampat, B. N. (2010). When do applicants search for prior art? *The Journal of Law and Economics*, 53(2):399–416.

Singh, J. (2005). Collaborative networks as determinants of knowledge diffusion patterns. *Management Science*, 51(5):756–770.

Singh, J. and Marx, M. (2013). Geographic constraints on knowledge spillovers: Political borders vs. spatial proximity. *Management Science*, 59(9):2056–2078.

Sokoloff, K. L. (1988). Inventive activity in early industrial America: Evidence from patent records, 1790–1846. *The Journal of Economic History*, 48(4):813–850.

Stuart, T. E. and Podolny, J. M. (1996). Local search and the evolution of technological capabilities. *Strategic Management Journal*, 17(S1):21–38.

Thompson, P. (2006). Patent citations and the geography of knowledge spillovers: Evidence from inventor-and examiner-added citations. *The Review of Economics and Statistics*, 88(2):383–388.

Thompson, P. and Fox-Kean, M. (2005). Patent citations and the geography of knowledge spillovers: A reassessment. *American Economic Review*, 95(1):450–460.

Trajtenberg, M. (1990). A penny for your quotes: Patent citations and the value of innovations. *The RAND Journal of Economics*, pages 172–187.

Verluise, C. and de Rassenfosse, G. (2020). Patcit: A comprehensive dataset of patent citations (version 0.15) [data set]. *Zenodo. http://doi.org/10.5281/zenodo.3710994*.

Verspagen, B. and De Loo, I. (1999). Technology spillovers between sectors. *Technological Forecasting and Social Change*, 60(3):215–235.

Von Wartburg, I., Teichert, T., and Rost, K. (2005). Inventive progress measured by multi-stage patent citation analysis. *Research Policy*, 34(10):1591–1607.

Wagner, S., Hoisl, K., and Thoma, G. (2014). Overcoming localization of knowledge? The role of professional service firms. *Strategic Management Journal*, 35(11):1671–1688.

Watzinger, M., Krieger, J. L., and Schnitzer, M. (2021). Standing on the shoulders of science. *Harvard Business School Working Paper 21-128*.

Weston, J., Bengio, S., and Usunier, N. (2010). WSABIE: Scaling up to large vocabulary image annotation. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Three*, pages 2764–2770.

Whalen, R., Lungeanu, A., DeChurch, L., and Contractor, N. (2020). Patent similarity data and innovation metrics. *Journal of Empirical Legal Studies*, 17(3):615–639.

Younge, K. A. and Kuhn, J. M. (2016). Patent-to-patent similarity: A vector space model. *Available at SSRN: https://ssrn.com/abstract=2709238*.

Figures

Figure 1: Patent citations by origin

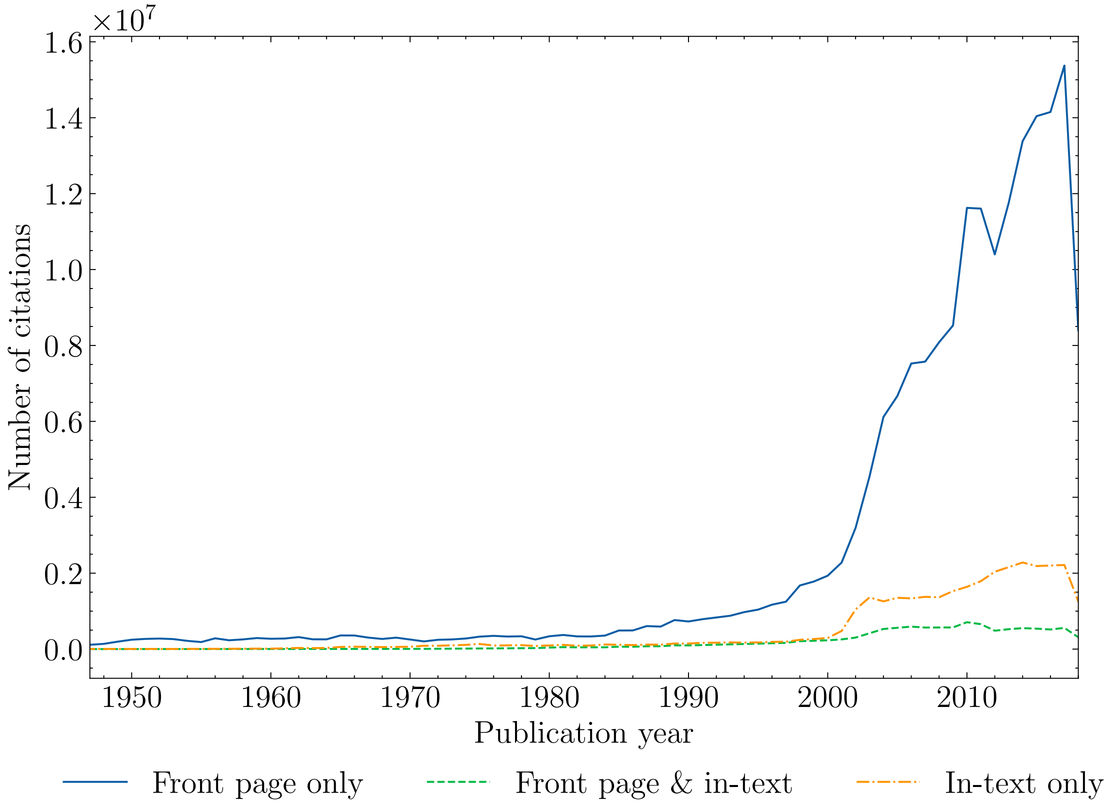
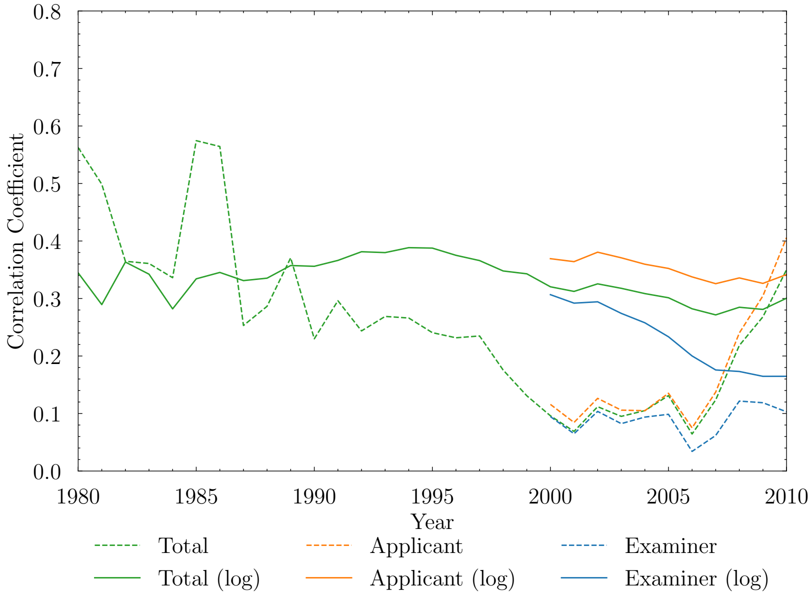
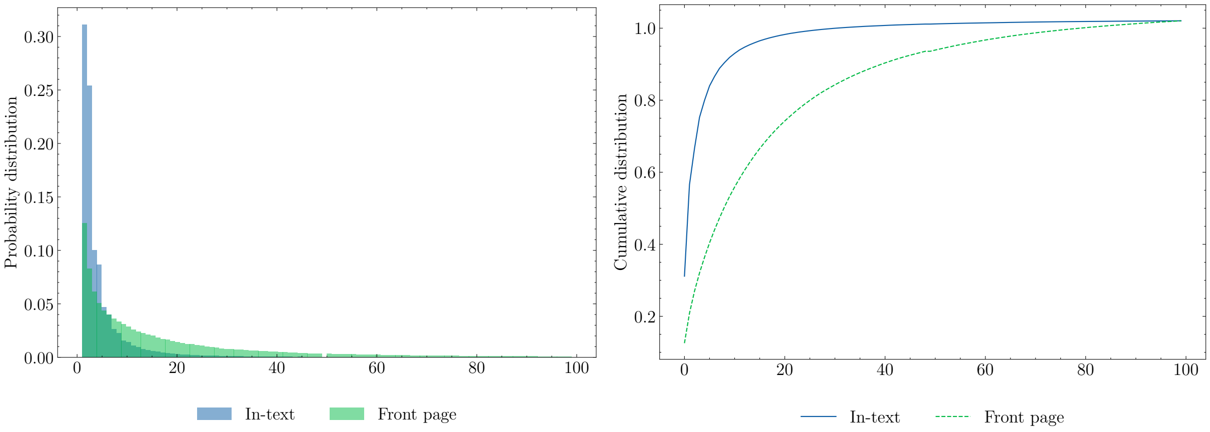


Figure 2: In-text and front-page citation counts: correlation over time



Notes:

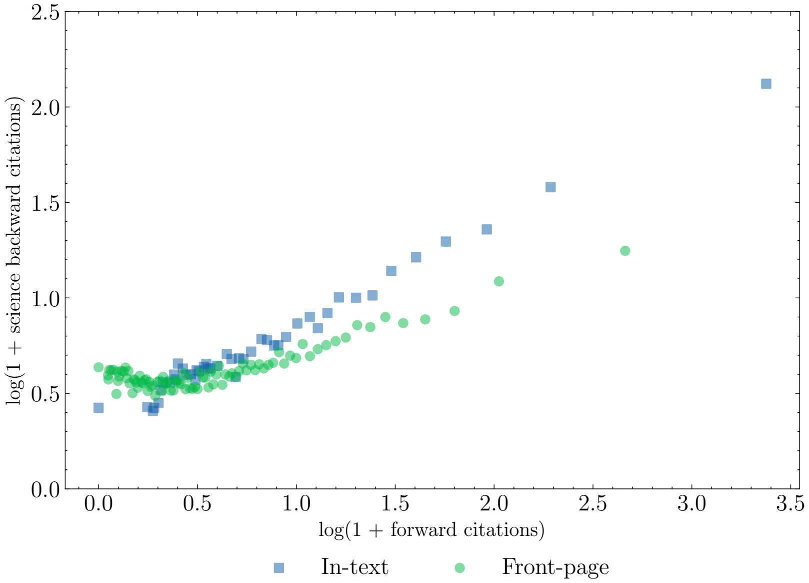
Figure 3: Empirical distribution of forward citations count



(a) Empirical probability distribution function (b) Empirical cumulative distribution function

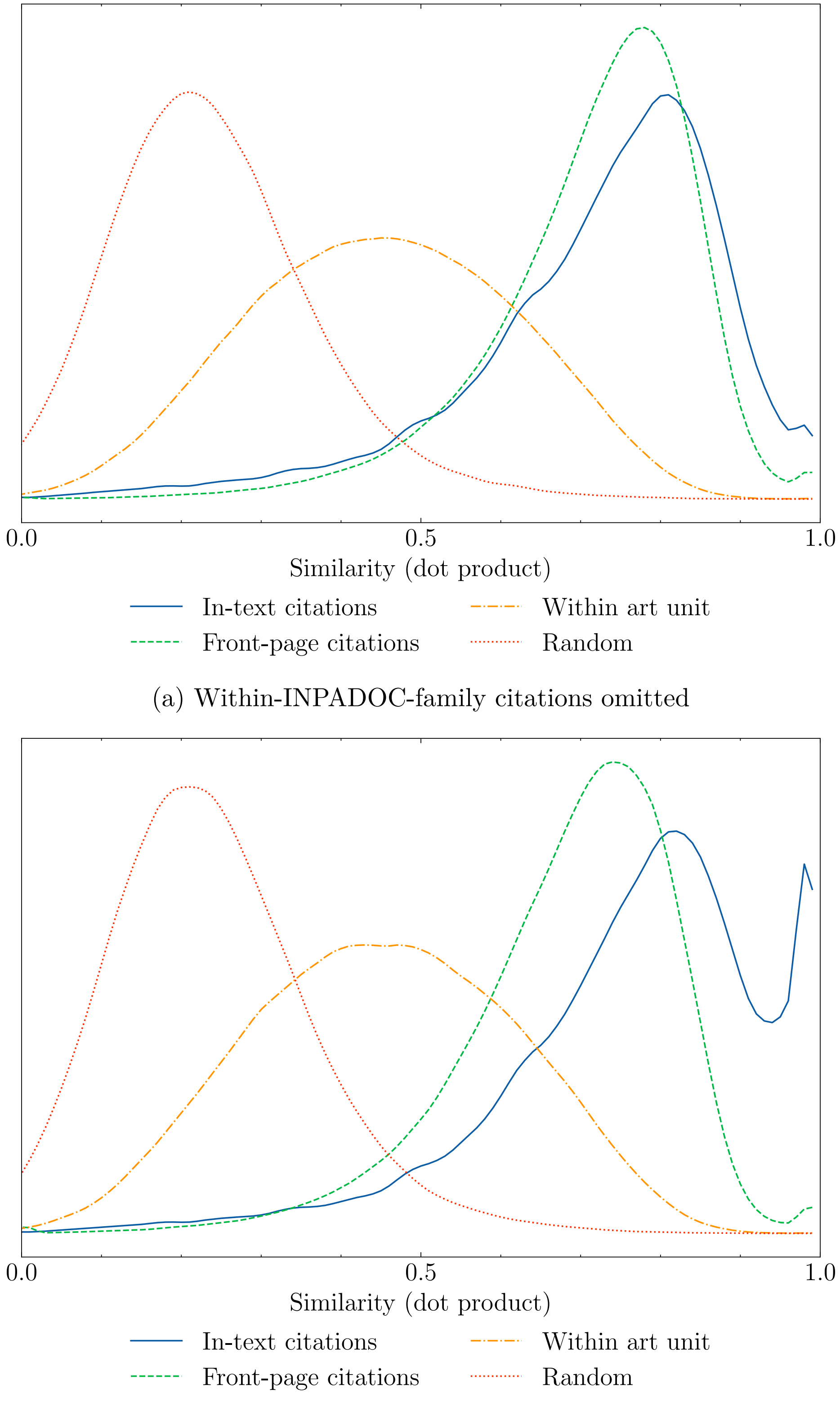
Notes: We use a 10 percent random sample of all DOCDB patent families with a positive front page and in-text forward citations count.

Figure 4: Backward citations to scientific articles and forward citations (in-text vs. front page)



Notes: In-text and front page forward citations are grouped into 100 quantiles based on their cohort adjusted citations (*i.e.,* for patent i granted in year t: $1 + fcites_{i,t}/\overline{fcites_t}$). The x axis plots the logarithmic transformation of the average cohort adjusted patent citation in each quantile. The y axis plots the logarithmic transformation of the average number of backward citations to scientific articles in each quantile (scaled by the average number of backward citations to scientific articles of patents granted in the same year). Backward citations to scientific articles are extracted from patents' text by Verluise and de Rassenfosse (2020).

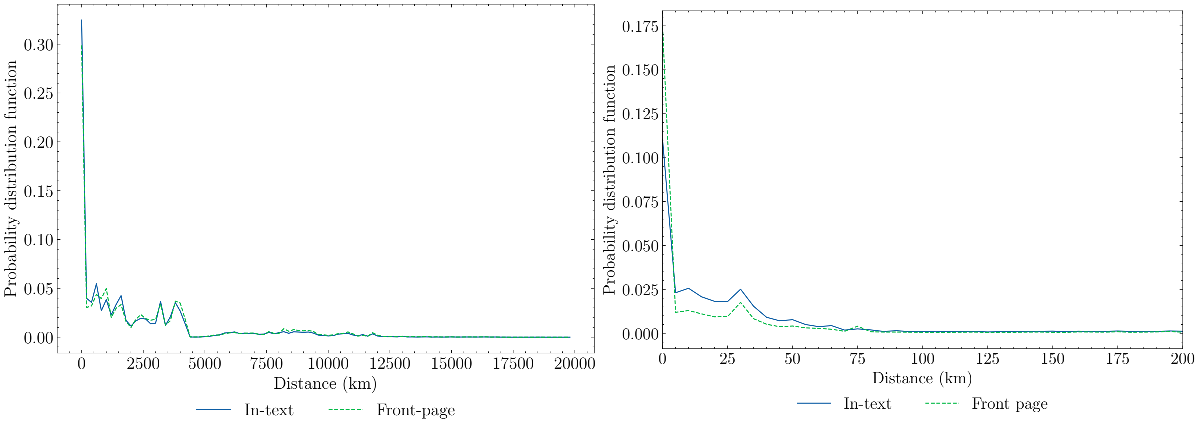
Figure 5: Citing-cited patent pair-wise similarity distribution



(a) Within-INPADOC-family citations omitted

(b) Within-DOCDB-family citations omitted

Figure 6: Citations distribution across cited inventors' location

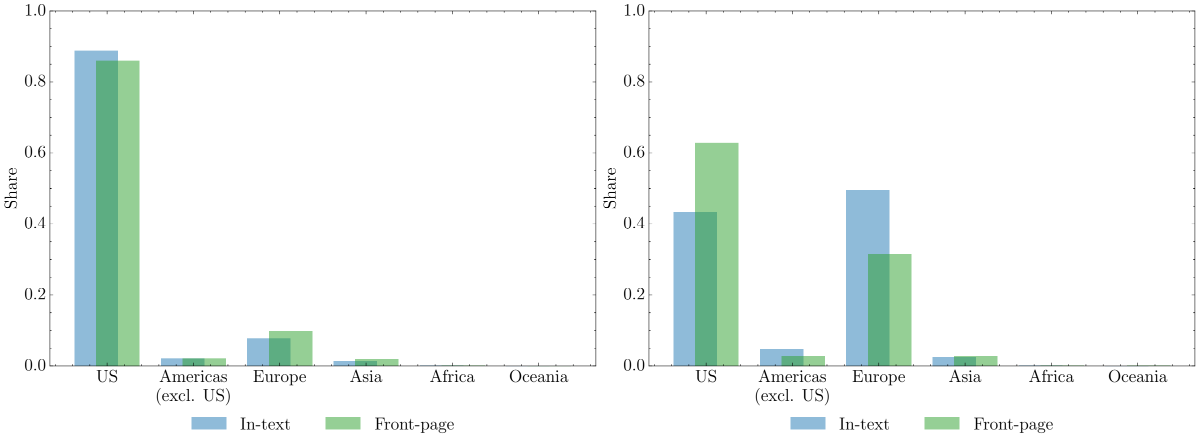


(a) All

(b) Distance ≤ 200km

Notes: Distance in kilometers is calculated from the latitude-longitude coordinates of the citing inventor's address to the latitude-longitude coordinates of cited inventor's address. The sample includes USPTO citing patents granted between 1980 and 2010. We exclude self citations at the INPADOC family level. In panel (a) we group observations by 200 km bins. In panel (b) we use 5 km bins.

Figure 7: Citations distribution by continent



(a) Citing patents listing US inventors

(b) Excl. citing patents listing US inventors

Notes: The sample includes USPTO citing patents granted between 1980 and 2010. We exclude self citations at the INPADOC family level.

Tables

Table 1: Typology of in-text citations

Citation Reason	**Example Patent**	**Citation and Context**
Enablement	9,607,299 (*Transactional security over a network*)	“Techniques for data encryption are disclosed in, for example, U.S. Pat. Nos. 7,257,225 and 7,251,326 (incorporated herein by reference) and the details of such processes are not provided herein to maintain focus on the disclosed embodiments.”
	9,606,907 (*Memory module with distributed data buffers and method of operation*)	“Examples of circuits which can serve as the control circuit ... are described in more detail by U.S. Pat. Nos. 7,289,386 and 7,532,537, each of which is incorporated in its entirety by reference herein.”
Novelty and non-obviousness	8,100,652 (*Ceiling fan complete cover*)	“U.S. Pat. No. 5,281,093, issued to Sedlak, et al., discloses a fan blade cover with a zipper. Sedlak, however, does not protect the fan’s housing and motor, nor does it prevent blades from spinning.”
	9,607,328 (*Electronic content distribution and exchange system*)	“One skilled in the art will readily appreciate that there is a great deal of prior art centered on methods for selecting programming for a viewer based on previous viewing history and explicit preferences, e.g., U.S. Pat. No. 5,758,257. The methods described in this application are unique and novel over these techniques as they suggest...”
Usefulness	9,607,730 (*Non-oleic triglyceride based, low viscosity, high flash point dielectric fluids*)	Applicant directly compares empirical results for the invention at hand with similar, previously granted patents.
	9,911,050 (*Driver active safety control system for vehicle*)	“For example, the interior rearview mirror assembly may comprise a prismatic mirror assembly, such as the types described in U.S. Pat. Nos. 7,249,860; 6,318,870;..., which are hereby incorporated herein by reference in their entireties.”

Table 2: In-text and front page citations summary statistics

	Front page	**In-text**
Number of patents	16,781,144	16,781,144
Number of patents with at least one citation	11,965,720	9,453,181
Share of patents with at least one citation	71.30%	56.33%
Number of citations	203,557,205	63,854,733
Number of citations^a	203,557,011	46,115,608
Average number of citations per patent	12.13	3.81
Average number of citations per patent - conditional on citing at least one patent	17.01	6.75
Number of US patent citations^a	181,162,466	32,827,382
Share of non U.S. citations^a	11%	28.82%
Median pairwise similarity (dot product) between citing and cited patent [lower quartile, upper quartile]^{a,c}	0.71 [0.62, 0.78]	0.80 [0.68, 0.88]
Share of citations in the same DOCDB family^b	0.69%	6.27%
Share of cited patents in the same INPADOC family^b	1.63%	10.51%
Share of cited patents with at least one shared inventor^b	5.98%	17.43%
Share of cited patents with at least one shared assignee^b	9.26%	22.46%

Notes: ^{a}: After 1947 only. ^{b}: Matched in-text only. ^{c}: After removing within-DOCDB family citations.

Table 3: Backward citations to scientific articles and top-cited patent status

	Top-cited front-page (dummy) (1)	Top-cited in-text (dummy) (2)	Top-cited in-text (dummy) (3)	Top-cited in-text (dummy) (4)
Cites-scientific-article	0.030*** (0.001)	0.045*** (0.001)	0.076*** (0.003)	0.067*** (0.004)
R^2	0.041	0.032	0.013	0.023
Observations	1,259,108	1,259,108	116,174	88,535
Fixed Effects				
Grant year	✓	✓	✓	✓
Tech. subcategory	✓	✓	✓	✓
Assignee type	✓	✓	✓	✓
Team size	✓	✓	✓	✓
No. of backward citations	✓	✓	✓	✓
Sample	Full	Full	Only top-cited front-page & in-text	Only top-cited front-page & in-text (strict)

Notes: * p<0.1, ** p<0.05, *** p<0.01. Robust standard errors in parentheses. Estimations by OLS. Top-cited patents are defined as those in the top 95th percentile of forward citations in a given grant year × NBER technology subcategory cohort. In column 1 the dependent variable is an indicator for top-cited patents according to front-page citations. In columns 2, 3, and 4 the dependent variable is an indicator for top-cited patents using in-text citations. In column 3 and 4 the sample is restricted to top-cited patents. In column 4 the sample further excludes any top-cited patent according to both front-page and in-text citations. Cites-scientific-article is an indicator for patents citing at least on scientific article in their text, using data from Verluise and de Rassenfosse (2020). Assignee type fixed effects distinguish patents filed by firms, universities, and government laboratories. Universities and government laboratories are identified following Bryan and Ozcan (2021) and Ahmadpoor and Jones (2017).

Table 4: Citations' geographic localization: In-text vs. applicant- & examiner-front-page

	Full sample			**Restricted sample**		
	(1)	(2)	(3)	(4)	(5)	(6)
A. Same country						
Front-page App.	0.019***			0.014***		
	(0.000)			(0.001)		
In-text		0.073***	0.042***		0.050***	0.023***
		(0.001)	(0.000)		(0.001)	(0.001)
R^2	0.481	0.564	0.464	0.550	0.639	0.538
Observations	8,371,251	4,253,715	6,714,744	4,484,126	2,409,491	3,012,787
B. log Distance						
Front-page App.	-0.156***			-0.124***		
	(0.002)			(0.002)		
In-text		-1.102***	-0.707***		-0.602***	-0.315***
		(0.004)	(0.003)		(0.005)	(0.004)
R^2	0.295	0.412	0.356	0.458	0.405	0.711
Observations	8,371,251	4,253,715	6,714,744	4,484,126	2,409,491	3,012,787
C. log Distance ($\leq$ 200 km)						
Front-page App.	0.017***			-0.035***		
	(0.004)			(0.007)		
In-text		-0.265***	-0.225***		-0.224***	-0.134***
		(0.006)	(0.004)		(0.014)	(0.007)
R^2	0.614	0.681	0.612	0.711	0.799	0.722
Observations	1,348,479	838,494	1,325,279	507,982	281,708	400,276
D. log Distance (within U.S.)						
Front-page App.	-0.087***			-0.079***		
	(0.002)			(0.003)		
In-text		-1.040***	-0.715***		-0.547***	-0.318***
		(0.005)	(0.003)		(0.006)	(0.005)
R^2	0.286	0.413	0.352	0.335	0.448	0.398
Observations	6,218,864	3,051,426	5,147,672	3,257,424	1,696,826	2,259,050
Citing patent FE	✓	✓	✓	✓	✓	✓
Reference group	Front-page Exa.	Front-page Exa.	Front-page App.	Front-page Exa.	Front-page Exa.	Front-page App.

Notes: * p<0.1, ** p<0.05, *** p<0.01. Robust standard errors in parentheses. Estimations by OLS. Same country is a dummy variable equal to 1 for citing-cited pairs where the inventor countries coincide. Distance measures the kilometers separating the latitude-longitude coordinates of the citing and cited inventor. In the regressions, we employ its logarithmic transformation $log(1+distance)$. The sample includes USPTO citing patents granted between 2001 and 2010. We exclude self citations at the INPADOC family level. In columns 4, 5, and 6 we restrict the sample by (i) considering only citations between patents filed at up to ten years of distance, (ii) excluding any patent with more than 100 front-page citations, (iii) excluding any self-citation at the patent applicant level. To identify unique patent applicants we use Du Plessis et al.'s (2009) identifiers.