Working Paper Series

The Missing 15 Percent of Patent Citations

Cyril Verluise
Gabriele Cristelli
Kyle Higham
Gaétan de Rassenfosse

December 2020

Innovation and Intellectual Property Policy
Working Paper series no. 13

Available at: https://ideas.repec.org/p/iip/wpaper/13.html

The Missing 15 Percent of Patent Citations[*]

Cyril Verluise
Collège de France & Paris School of Economics

Gabriele Cristelli
École Polytechnique Fédérale de Lausanne

Kyle Higham
Hitotsubashi University

Gaétan de Rassenfosse
École Polytechnique Fédérale de Lausanne

This version: December 2020

Abstract

Patent citations are one of the most commonly-used metrics in the innovation literature. Leading uses of patent-to-patent citations are associated with the quantification of inventions' quality and the measurement of knowledge flows. Due to their widespread availability, scholars have exploited citations listed on the front-page of patent documents. Citations appearing in the full-text of patent documents have been neglected. We apply modern machine learning methods to extract these citations from the text of USPTO patent documents. Overall, we are able to recover an additional 15 percent of patent citations that could not be found using only front-page data. We show that "in-text" citations bring a different type of information compared to front-page citations. They exhibit higher text-similarity to the citing patents and alter the ranking of patent importance. The dataset is available at patcit.io (CC-BY-4).

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

[*] Corresponding author: Cyril Verluise, cyril.verluise@college-de-france.com. We are grateful to Philippe Aghion and Antonin Bergeaud 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.

1 Introduction

Patent documents represent 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, 2020). 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 (Jaffe and de Rassenfosse, 2017).

Early work exploiting patent documents focused on easily accessible metadata, including citations and technology classes. Citation data are a particularly popular object of study; a Google Scholar search with the keyword "patent citation" returns about 15,000 results. Use cases are too numerous to list but cover the measurement of invention 'quality,' the placement of inventions in the broader invention network, and the tracking of knowledge flows. More recently, the field has been moving towards exploiting the full text of patent documents. Applications cover, e.g., keyword extraction, topic identification, and invention similarity (Kaplan and Vakili, 2015; Younge and Kuhn, 2016; Arts et al., 2018; Righi and Simcoe, 2019)

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 made for procedural reasons; they list prior art that is relevant for assessing the patentability of the claimed invention. They originate from applicants (or their attorneys and inventors), examiners, and third parties.[1] They may originate directly at the time of filing, during the substantive examination before grant as well as after grant in case of opposition, re-examination, revocation, etc. By their nature, front-page citations are thus conceptually different from citations typically found in scientific papers (Meyer, 2000).

By contrast, in-text patent citations appear in the patent text itself. They are

[1] An example of citations by third parties is Section 801 of the Patent Cooperation Treaty (Administrative Instructions), which allows third parties to make observations referring to relevant prior art.

made 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 drive the generation of front-page citations, in-text citations contain truly novel 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, 2020; Verluise and de Rassenfosse, 2020). Given the importance of citation data, the lack of treatment of in-text patent-to-patent citations is an obvious gap. Such data are likely to be particularly important for specific applications, such as for the measurement of knowledge flows. Indeed, inventors often contribute to the drafting of the text, and the references they mention are likely to be a better way of capturing knowledge flows than front-page references. Despite our strong suspicion that these data might be relevant for some applications, little research exists to confirm it—precisely because data were not readily available until now. It is thus critical to process these data and make them widely accessible.

We have extracted patent citations from the full-text of 16,781,144 publications filed in the U.S. Patent and Trademark Office (USPTO) from 1790 to 2018. About 95 percent of these publications are granted patents or patent applications.[2] 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 library leveraging Natural Language Processing (NLP) to extract and parse citations.[3] We performed an extensive validation exercises, revealing high performance: our extraction task in particular achieves a satisfying 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 standard publication number ensuring interoperability with other patent datasets. The data collection effort is part of PatCit, an open source project that aims at building a

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

[3] Grobid (2008-2020) https://github.com/kermitt2/grobid

comprehensive patent citation dataset.

We have performed an in-depth quantitative analysis of the difference between in-text and front-page citations. We discovered three noteworthy elements. First, by-and-large, in-text citations do not overlap with front-page citations. Overall, we are able to identify an additional 15 percent more citations than one would get using front-page data alone (that is, these citations are not listed on the front page). This figure jumps to 100 percent before 1947, meaning that our data will be an invaluable help to researchers interested in the pre-WWII period. Second, the data generation process of in-text citations intrinsically differs from that of front-page citations and, we believe, is particularly suited to capture knowledge flows. This intuition is reinforced by measures of textual similarity; we find that in-text citations are more similar to the focal citing patent than front-page citations. Third, we find surprisingly low correlation between the front-page forward citation count and the in-text forward citation count. Scholars have used such counts to measure invention importance (Trajtenberg, 1990; Lanjouw and Schankerman, 2004; Hall et al., 2005). The low correlation suggests that in-text citations provide valuable information to assess invention importance.

The 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.[4] In addition to the final output, we also release the validation data and the code with a view of ensuring replicability and follow-on improvements by the community.[5]

The remainder of the document is organized as follows. Section 2 discusses the nature of in-text citations. Section 3 sets forth the processing pipeline and provides technical details about the methods. Section 4 describes our validation procedure and reports performance measures for various critical steps of the data pipeline. Section 5 offers a quantitative overview of in-text citation data and compares them with front-page citations. Section 6 concludes.

[4] See http://patcit.io for the project documentation.

[5] https://github.com/cverluise/PatCit/tree/0.3.0

2 The epistemology of in-text citations

This section describes the characteristics of in-text patent citations, with a particular focus on how they differ from 'traditional' patent citations reported on the front page of patent documents.

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 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.

2.1 A legal perspective on in-text patent citations

The justifications 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. Appendix A gives real examples of citations in each of these contexts.

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 reference this in the text of the patent specification.[6] 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 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.[7] For the most part, they are argued for implicitly through Information Disclosure Statements submitted by the applicant throughout the application and patent prosecution processes—these are the citations that appear on the front page of a patent.[8] 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), citations may be added

[6] 37 CFR 1.57

[7] 35 USC 102; 35 USC 103

[8] 37 CFR 1.56

to allay doubts that, for example, a claimed function of the invention is physically possible. The latter is unlikely to be questioned by an examiner (Machin, 1999).

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 patentability requirements laid out in U.S. law, as discussed above. Some of these reasons overlap with those that require applicants to submit Information Disclosure Statements, 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 Information Disclosure Statement 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.[9] 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).[10]

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 (see Appendix A). In either case, the inventor was likely aware of this prior art during the invention process.

Second, those citations that are *not* duplicated on the front page are most likely

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

[10] Manual of Patent Examining Procedure, Section 2120

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 '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.[11] 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 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

[11] 35 USC 112

in the final product and are cited to demonstrate how the invention is used in practice. In-text citations are also more likely to come from inventors themselves, perhaps independently from the motives for citing. For these reasons, the interpretation of a patent accumulating a large number of in-text forward citations is 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 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."*

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 reasons for making 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 know 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 result from 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;

with both data and code publicly available, future research can build on this work to add the context to in-text citations and, ultimately, better understand what a highly-cited patent represents in this setting.

3 Methods

In this section, we describe the data sources and the different steps of the processing pipeline. We want to provide extensive insights into our technical choices in order to stimulate and enable future extensions or improvements.[12]

3.1 Data

The processing pipeline starts with the full-text of 16,781,144 patent documents filed at the U.S. Patent and Trademark Office (USPTO) since 1790.[13] We extracted the full-text data from the IFI CLAIMS dataset, made available by Google Patents as part of its public datasets.[14]

The text we are considering is the specification of the patent. The specification is a written description of the invention and of the manner and process of making and using the invention. It also includes information about related applications and government interest statements (de Rassenfosse et al., 2019a). It does *not* include the patent's claims or the information on the front-page.

The starting point is a long chain of characters without any structure and indication about which characters might refer to a patent citation.

3.2 Extraction task

The first step involves identifying the relevant strings of characters referring to a patent citation in the full text. An early attempt to do so dates back to Galibert et al. (2010), who combined a set of regular expressions to identify the cited patent number itself (e.g., country codes followed by a series of digits) based on the neighbouring

[12] Readers who are not specifically versed into technical considerations can skip this section without much harm to their understanding of the nature of the data.

[13] The first extracted citation is in 1846.

[14] https://console.cloud.google.com/marketplace/partners/patents-public-data

text (e.g. "herein described by"). A similar approach was implemented by Berkes (2018) for U.S. patents published before 1947. Although intuitive, these approaches lead to moderately satisfying results. Galibert et al. (2010) report a precision of 64.4 percent, a recall of 61 percent and a f1-score of 62.9 percent while Berkes (2018) does not report performance metrics. The fundamental reason behind these low scores is that language is highly variational and there are many ways of citing a patent. On this point, Adams (2010) warned the community about the complexity of the extraction task. Using a random sample of USPTO patents, he found an "alarming" (p. 26) degree of variation in the form of in-text patent citations. In this context, any attempt to use a list of predefined rules is likely to have mixed results and, above all, to lack generalisation.

In order to overcome this limitation, NLP researchers have developed statistical models that can learn to find and tag entities, such as cited patents, using a training set of annotated documents, where a researcher has labeled the presence (or not) of the entities of interest. Although an in-depth presentation of the related Named Entity Recognition (NER) literature is out of the scope of this paper, we summarize the general working principles of these models below.[15]

The key is to see a text as two sequences: a sequence of tokens and a corresponding sequence of latent labels (e.g. "PATCIT" for patent citations versus "O" for other). The task is to predict the sequence of labels. The algorithm is trained on an annotated set of documents, that is, a set of documents for which we know both the sequence of tokens and the sequence of labels. The probability of each token to belong to a given label is a recursive function of the token itself and its features (digits, capital letters, etc), the neighbouring tokens (its context) and the *neighbouring labels.* The overall goal of the algorithm is to predict rightly the full sequence of latent labels for a given sequence of tokens. If a token (or a sequence of tokens) is unknown or deviates from the learning examples, the algorithm can still leverage the other attributes to decide which sequence of labels is the most probable for the whole sentence, leading to a considerable generalization improvement.

For example, let us assume that the algorithm has been trained on a corpus of texts

[15] See Li et al. (2020) for a recent survey of this literature.

where citations come in the following form (with *d* denoting any digit): "described by patent *d,ddd,ddd*" and where the corresponding sequence of labels is [O, O, O, PATCIT]. Let us further assume that the algorithm is supplied a new text with a slightly different form of citation such as "described by Pat 9,535,657". Although the algorithm has never seen the token "Pat", it has learnt from the training data that the sequence of token "described by" frequently precedes a PATCIT label by two tokens. Combined with the fact that the token "9,535,657" exhibits the features frequently associated with a PATCIT (digits and commas), then the algorithm is expected to override the absence of the "patent" token and still to predict the right sequence of labels, [O, O, O, PATCIT].

The aforementioned limitations and improvement opportunities have been well identified by the machine learning community in the second half of the 2000s. In particular, Lopez (2010) developed the Grobid library in 2008 (and has been continuously improving it since then) with the goal of overcoming the limitations of a "rule-based" approach using a statistical approach. Grobid has now become an open source project leveraging modern NLP to efficiently structure scientific documents in general, but retains a specific focus on patents. It includes models trained at extracting and structuring bibliographical references (scientific articles, books, proceedings, etc.) and patents from full-text documents. The algorithmic backbone of Grobid is the Conditional Random Fields (CRF) model. This model belongs to the family of sequence labeling models described above and was first introduced in 2001 (Lafferty et al., 2001). The CRF model has been widely used in various fields and applications.[16]

Grobid's patent citations' extraction model was originally trained on 200 annotated full-text patents.[17] This training set included 62 percent of EPO patents, 19 percent of WIPO patents and the remaining 19 percent of USPTO patents. As for the rest of Grobid models, the patent extraction model is a CRF model. The specific features entering the CRF model to support patent citation detection include the relative position of the current token in the document, the matching of a common country code (e.g., US, EP, WO, etc) and the matching of a common kind code (e.g., A1, A2, B1, B2, etc).

[16] See Sutton and McCallum (2006) for a survey.

[17] The training set was enriched since that time and now includes 270 patents, including 51 percent of EPO patents, 33 percent of WIPO patents and the remaining 26 percent of USPTO patents.

The output of the extraction tasks is a set of text spans that were tagged as patent citations (e.g., “United States Patent 9,535,657”). The information extracted at this stage is not structured and, therefore, improper for researchers.

3.3 Parsing task

The next step involves parsing the extracted patent citation strings. We take the raw span of the extracted citation as an input, with the goal of obtaining the following normalized attributes: the country code of the patent authority, the patent number and the type of the patent. This task is challenging due to the many forms in which patent citations occur in the text. Typically, the patent authority can appear as a code or a name (e.g “US Patent 9,535,657” or “United States Patent 9,535,657”) either immediately next to the patent number or relatively far from it (e.g., “US Patent number 9,535,657” or “US Patents 9,911,050, 9,607,328, 9,535,657”).

Lopez (2010) proposes an efficient solution for tackling this task. The fundamental idea is that both the sets of possible inputs and outputs for each patent attribute are finite (e.g., the list of patent organisation names and the list of their codes respectively). In addition, each element of the input vocabulary should be mapped with a unique element of the output vocabulary (e.g. “United States” with “US” or “European Patent Office” with “EP”). In the end, for any given patent attribute, the parsing operation can be thought of as a translation operation between two languages with a finite vocabulary. If this still seems a bit abstract, the reader can simply consider that the aforementioned task consists in regular expression matching followed by string rewriting.[18] This task perfectly fits the usage of Finite State Transducers (FST) which appeared early in the history of automated translation.[19] Importantly, FSTs have been developed with computational efficiency in mind in the early ages of computer science, making them highly efficient in todays’ context.

The output of this task is a well-structured set of attributes describing the cited patent.

[18] Let us assume that we are interested in the organisation attribute and that we have extracted the following span “United States Patent 9,535,657”. This span would trigger a match for “United States” which would then be rewritten as “US”.

[19] See Roche and Schabes (1997) for an in-depth review of Finite State Transducers.

3.4 Consolidation task

The final task consists in matching each extracted patent citation to a unique and consolidated identifier, in order to connect each cited patent document to commonly used patent dataset. For patents, the identifier common to most (if not all) patent datasets is the DOCDB publication number.[20] On this point, note that we depart from Grobid which relies on the European Patent Office (EPO) search API[21] to perform the matching process and uses the EPO document number as its target and consolidation device.

Unfortunately, in a large majority of cases, in-text patent citations do not report the kind code of a patent, or report the original patent number rather than the version used in the DOCDB publication number, making it impossible to assemble the DOCDB publication number using the parsed attributes only. In order to overcome this limitation, we have relied on the Google Patents Linking Application Programming Interface (API).[22] Taking various kinds of inputs, such as the patent office code, the patent number and kind code, the API returns the associated DOCDB publication number. At a high level, the internal mechanism of this service is the following.[23] First, a large number of variations of each publication number was generated. For each variation, the original patent office and DOCDB formatted versions were indexed. Variations include adding and removing 0 padding, two and four digit year dates inside of patent number, Japanese emperor year variants and different combinations of country code, patent number and kind code. Altogether, these variations constitute a large lookup table linking many variations of a publication number to its DOCDB formatted version. Then, at the time of lookup, punctuation is stripped and the country code, number and kind code are searched for before being used to look-up for matches in the large variation table. Note that there are two distinct services, one for applications and one for patents.[24] We decide which one to call based on the status attribute parsed by Grobid which can take four values: “application”,

[20] For the sake of simplicity, we use the "publication number" terminology for both the publication number (for published patents) and the application number (for patent applications).

[21] http://v3.espacenet.com/publicationDetails/biblio

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

[23] We thank Ian Wetherbee from Google Patents for this explanation.

[24] Applications: https://patents.google.com/api/match?appnum
Patents: https://patents.google.com/api/match?pubnum

"provisional", "patent" and "reissued". The first two trigger the application service, while the last two trigger the patent service.

Using the unique publication number returned by the Google Patents Linking API we were able to connect each cited document with richer information from patent datasets generally used by researchers (e.g., PATSTAT, PatentsView, IFI CLAIMS, etc.). We enriched each cited patent with the following attributes: publication date, application identifier, patent publication identifier, INPADOC and DOCDB family identifiers.

3.5 Pipeline

Let us illustrate the process using an example. Consider the following excerpt from the description of US-9606907-B2, which cites two U.S. patents:

> "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."

After the Grobid processing, we know that the patent US-9606907-B2 cites two patents from the U.S. patent office ("US" patent authority code) and that their original numbers are 7,289,386 and 7,532,537. Using the Google Patents Linking API, we find that the two patent citations embedded in the text can be uniquely identified by their publication numbers, namely US-7532537-B2 and US-7289386-B2.

The above pipeline was deployed remotely on a large-size compute engine from Amazon Web Services.[25] In order to increase speed, we used multi-processing, a technique consisting in running multiple processes in parallel at the same time. This technique is especially useful for 'cpu bound' rather than 'io bound' operations, that is when computation is the main limiting factor, not internal communication. Processing documents at an average pace of 400,000 to 500,000 per day, this operation took us approximately one month for a total cost of about 120 USD.[26] Overall, from

[25] We used a t2.xlarge (4 cores and 16Gb of Ram) located in the "USA East Ohio" computing zone.

[26] Note that we simultaneously extracted in-text Non Patent Literature citations (scientific articles, books, proceedings, etc.) and tried to match them with Crossref at runtime. To do so, we used biblio-glutton, a high performance bibliographic reference matching service, and an ElasticSearch index hosted on a separate engine. It appeared that the major processing speed limitation came from ElasticSearch queries. Processing only in-text patent citations would certainly take significantly less time and resources.

the 16,781,144 patent documents that we processed, we were able to extract 63,854,733 in-text patent citations. These citations point to 13,611,323 unique patent documents. We matched 49,409,629 of the extracted in-text cited documents with a publication number.

4 Technical validation

In order to assess the quality of the citation dataset, we undertook a thorough validation exercise of the data and the extraction, parsing and matching tasks. To do so, we relied on Prodigy, a scriptable annotation tool.[27] Lopez (2010) reports performance metrics for all these tasks, however the set of documents we are considering partly differs from the corpus he used. In particular, a significant part of the patents in our corpus is much older than any document considered for Grobid training and evaluation. We also carried out detailed error analyses as a way to support future improvement efforts.

4.1 Data consistency

USPTO patent documents' format and the quality of the scanned document (for older patents) has changed throughout the years. Before 1971, patents were largely unstructured with no clear delimitation between the metadata and the specification text itself (see Figure 1). The modern patent format was introduced in 1971 and progressively replaced the old format before becoming the unique format after 1976. This format is semi-structured and clearly distinguishes between the metadata sections and the specification section *inter alia* (see Figure 2). These specificities of the source data have some notable implications on our output data.

First, the text of patents published in the old format includes the header of the patent. The header summarizes the main attributes of the patent, including its technological classes, title and most importantly its number. In this case, the extraction algorithm is likely to extract a patent citation which does not correspond to the kind of object we are looking for. Fortunately, this specific pitfall is relatively easy to spot

[27] Prodigy (2018-2020) https://prodi.gy/.

as the citation appears very early in the text. Figure 3 reports the distribution of the rank of the first character of the extracted citations before and after 1976. We observe a clear excess mass between 0 and 50 characters before 1976. Building on this observation, we focused on the corpus of patents published before 1971 and randomly drew 50 citations starting before character 50. Confirming our doubts, we found that 88 percent were self references, 8 percent were technological classes and 4 percent were dates. In this context, we chose to flag all citations detected in a patent published before 1976 and starting before character 50 to make it easy to exclude them from analysis.

Second, in the old format, what we now call 'front-page citations' were printed *after* the patent specification, and these are also sometimes mistakenly included in our source data as part of the full-text of the patent. Since all patents have a different number of characters, looking at the distribution of citations by starting character does not make sense. However, we can still look at the relative place of the detected citations. Figure 4 shows their distribution as a function of their relative place (expressed in percentile) in the full text. Comparing the distribution before and after 1976 reveals a sizable excess mass for the pre-1976 distribution in the last four percent of the full-text characters. Additionally, looking at a random sample of 100 citations extracted from patents published before 1976 and occurring in the last four percent of the characters, we find that 99 percent belong to the 'front page' citations section. Hence, for patents published before 1976, we chose to flag all citations detected in the last 4 percent of the full-text and encourage the user to exclude them from their analysis.

Third, during the transition period between the old and modern formats (approximately throughout 1971–1975) there were two patent formats in use, complicating the delineation of the specification text section during this time period. As a result, we observed that 'full-texts' from this time, mistakenly include the front-page of patents that are in the modern format. This can lead to the incidental extraction of 'in-text' citations that are actually front-matter, including front-page citations and references to the patent itself (including priority filings). Unfortunately, there is no straightforward solution to this problem. We encourage data users to systematically ignore patents that are both in text and front page citations during this time span.

All figures reported above and below exclude flagged patent citations as they are most likely not to correspond to real in-text patent citations (unless explicitly specified).

4.2 Extraction task

Lopez (2010) reports high performance metrics for the extraction task. Using cross-validation, a technique consisting in training the model ten times using 80 percent of the sample and testing it on the remaining 20 percent, the author reports the following average performance metrics: 94.66 percent of precision, 96.16 percent of recall and a f1-score of 95.4 percent. As far as we know, these are the best performances reported in the literature to date. Although this motivated our choice to use Grobid, we are fully aware that our dataset partly differs from the Grobid training set and performance could thus be affected.

In order to evaluate the quality of the extraction in our specific case, we randomly sampled 160 U.S. patents and annotated them by hand. As previously discussed, the citation of a patent can come in various ways. For instance, the country of the patent office can be reported as a code preceding the patent number, as a name anywhere in the surrounding of the patent number, etc. In this context, the only stable element of a patent citation is the patent number itself. That is why Grobid returns the first and the last character of the patent number of detected patent citations. Hence, our validation exercise consisted in comparing the spans detected by Grobid as a patent number and the spans labelled by humans as a patent number. Each patent was annotated by a single human annotator using the platform featured by Figure 5a.[28] The body of the text is displayed together with annotations from Grobid predictions and the annotator goes through the text to correct missing and wrong annotations. The tagged spans are saved upon exit.

As depicted by Figure 6, the validation sample and the universe of citing patents display very similar distributions by publication year.

From the 160 random U.S. patents in the validation set, human annotators found

[28] "Human annotators" are not undergraduate RAs but coauthors of this paper.

that 103 (64.4 percent) patents cited at least one patent for a total of 470 in-text patent citations. Table 3 reports the extraction performance that we obtained together with the Galibert et al. (2010) and Galibert et al. (2010) benchmarks. Comparing 'gold' annotations from human annotators with the predictions obtained from Grobid, we find that Grobid exhibits a satisfying 97 percent precision and 82 percent recall (f1-score nearing 90 percent). Importantly, these results largely outperform Galibert et al. (2010), who used regular expressions for the same patent citations extraction. They reported a precision of 64.4 percent and a recall of 61 percent. This result clearly confirms that a statistical approach to in-text citation extraction is much more relevant than a regular expression approach. Interestingly, the performance obtained by Grobid on our extended corpus is very similar to the benchmark reported by Lopez (2010) regarding precision (97.44% vs 97%) but lower in terms of recall (97.74% vs. 82%). This difference means that, applied to our extended corpus, Grobid is as reliable as reported in Lopez (2010) when it has detected a patent citation. However, it misses patent citations more often in our extended corpus due to older forms of citations appearing in early-twentieth century patents.

The error analysis suggests that both false positives and false negatives exhibit patterns that could be specifically addressed by future improvements of the Grobid training set. Table 4 provides examples for each category of errors that we were able to identify. Starting with false negatives, that is patent citations that were not detected by Grobid, we find three categories of context generating this type of errors: 1) the context does not clearly mention "patent" or "application" but rather implicitly suggests a patent citation; 2) the patent is cited in the form "inventor (date) <PATCIT>" and 3) the patent is cited as "Serial Number <PATCIT>". While category 1) could have been expected and would certainly be hard to correct without generating a large number of false positives, categories 2) and 3) might certainly be partly addressed by augmenting the training dataset with older patents that tend to adopt this form of citations more often. Now, looking at false positives, that is text spans that were wrongly identified by Grobid as patent citations, we can find three categories of errors as well: 1) technological classes reported as "dd/ddd", 2) date and 3) docket number. Note that the categories 2 and 3 have only one occurrence each.

4.3 Parsing task

Grobid FST was built manually based on 1,500 patent citation examples. It was then evaluated on 250 references which were unseen before. Lopez (2010) reports a 97.2 percent accuracy for the full parsing task (patent organisation code, number and kind code). Once again, we thought that it was important to confront those results with our specific dataset.

In order to validate the quality of the parsing, we randomly sampled 300 extracted citations with their parsed attributes. As already discussed, the attributes can be relatively far from the patent number that serves as the citation anchor. Hence, it was necessary to provide the human annotators with a contextualized citation. In practice, using the patent number reported by Grobid as an anchor, we extracted a chunk of text containing a window of ten tokens on the right and left of the detected patent. This text and the tagged patent were then displayed to the annotator together with the Grobid parsed attribute as illustrated by Figure 5b. The annotator would then accept or reject the attribute depending on what he actually found in the text. Each example was validated by a single annotator whose decisions were saved upon exit.

Lopez (2010) reports an *overall* 97.2 percent accuracy. Since the attributes can be used independently, we believe that a detailed understanding of the performance and errors for each attribute is also valuable for the community. Hence, we performed three distinct validation exercises, one for each attribute. Our results are summarized in Table 6.

Considering the parsing of the patent organisation, we first checked for sample representativity. Table 5 reports the distribution of the patent organisations in the validation sample. It appears that two-thirds of the citations in the sample were mapped to the U.S. patent office. This result is very much in line with the results that we report on the full dataset (see Section 5). Similarly, the patent organisations in the remaining third of the validation sample are also the most represented organisations at scale, including the Japanese Patent Office, the World Intellectual Property Organisation, the European Patent Office and the German Patent Office *inter alia*. On the 300 examples that we validated, we found only five errors, leading to a 98.3 percent accuracy score. Errors spread over five distinct patent offices and we do not observe any systematic confusion between patent offices, which suggest that errors generate noise rather than a systematic bias.[29]

When it comes to the parsing of the patent number, there is no specific way of checking sample representativeness. Overall, on the 300 examples that we validated, we found thirteen errors, leading to a 95.7 percent accuracy score. Among the errors, we find two recurring cases. First, patent citations in their Paris Cooperation Treaty (PCT) form (e.g., PCT/EP2005/008238) generate patent numbers mixing part of the letters in the prefix and the patent number itself (e.g., PTEP2005008238). Second, as already reported in Lopez (2010), we found that Grobid removes the first letter of the patent number of Japanese applications with date prior to 2000 (e.g., H08-193210 where H stands for the Heisei era that spanned from 1989 to 2019). However, this indication is key to uniquely identify the application. This letter refers to the era and acts as the time marker. Note that this specific issue is partly fixed by the Google Patent matching API as explained in Section 3.

Lastly, we validated the parsing of the so-called kind code, that is the code indicating the specific kind of document the citation refers to (granted patent, application, reissue, design, etc.). Over the 502 random examples, we obtain an accuracy of 97.6 percent. Note, however, that this measure includes a large proportion of null results as the kind code is in fact rarely reported in the text. In order to further characterize the quality of the parsing, we drew a sample of 50 citations where the parsed kind code was not null. We found 7 mistakes, meaning a 'conditional' accuracy of 86 percent. Specifically, we found three groups of parsing errors: errors due to unconventional formatting, OCR issues and Grobid mistakenly interpreting 'Cl' (class abbreviation) for the 'C' kind code. Importantly, every instance in standard form was correctly parsed.

4.4 Matching task

The matching task involves associating the extracted attributes with a unique identifier, which is the DOCDB publication number in our case. In order to validate this

[29] The five offices are: SA (Saudi Arabia), AL (Albania), CH (Switzerland), DE (Germany) and BE (Belgium).

step of the process, we randomly sampled 200 citations from our final dataset and we compared the concatenation of the parsed attributes with the publication number provided by the Google Patent's Linking API. The annotator's task was to answer the following questions: i) if there is a matched publication number, is it the right one? ii) if there is no match, would it be possible to find one for a human reasonably well trained in the task? A single human annotator fulfilled this validation exercise. Based on that, we can assign each annotated example to a standard classification outcome category and derive the associated performance metrics. Table 7 summarizes these categories, their contents and the results from the validation exercise.

On the 200 examples in the validation sample, we find that 147 were matched and 53 remained unmatched. Among the 147 matches, 137 were correct (True positives) and 10 were incorrect (False positives) including six patents that could have been matched and four non-patent items that should not have been matched. Among the 53 unmatched examples, we found that 17 could have been matched (False negatives) while no match could be found for the remaining 36 (True negatives). Overall, we find that the matching procedure achieves a 93.2 percent precision and a 88.96 percent recall, leading to a 91.06 percent f1-score.

Next, we delved into the nature of the errors and non-matches. Tables 8 and 9 respectively detail errors occurring during matching and cases classified as unmatchable by the human annotator. We find that errors arising at this final step of the processing pipeline are partly inherited from upstream steps. Among the ten incorrect matches, half are due to either a parsing error or an extraction error. In the same way, among the thirty-six unmatched citations that were judged unmatchable, 56 percent were directly related to either a parsing error or an extraction error. Another group of errors seems to arise from the specificities of in-text citations and their intrinsic ambiguities. This group includes citations of provisional patent applications (which might well never appear in standard patent datasets) and partial citations that even a human cannot match.[30] This family of errors represent 41 percent of the thirty-six unmatchable detected citations in our validation sample. Eventually, focusing on the

[30] A provisional application is a legal document filed at the patent office that establishes an early filing date, but does not mature into an issued patent unless the applicant files a regular non-provisional patent application within one year.

unmatched citations that a human can match reveals some blind spots of the Linking API. Over the seventeen cases in this category, 52 percent are caused by missing zeros after the country code/year or a Japanese publication number reporting the year after the serial number rather than before it as is usually expected.

While the previous step can characterize the performance of the matching procedure with high precision, due to the small size of the validation sample it cannot uncover rare irregularities that might still be of sizable magnitude at large scale.

Considering the full dataset, Figure 7 show, the yearly number (7a) and share (7b) of citing patents according to the matching status of the extracted in-text citations.[31] Patents with all in-text citations matched to a publication number represent 42.7 percent of the total, whereas those with only some in-text citations matched represent 32.7 percent. Patents with no in-text citations matched account for the remaining 24.6 percent. From 1947 to 1964, patents with all in-text citations matched report an increasing yearly share, from around 40 percent to almost 70 percent. For patents published between 1965 to 1975, the performance of our matching procedure worsens, as the proportion of patents with only some citations matched or no citation matched grows. From 1976 onwards, the share of patents with all citations matched returns to be the largest (around 77 percent in 1976), although it progressively decreases for patents published during the following years in our dataset.

These aggregate figures mask high variation depending on the patent office of the cited patent documents. Table 10 reports the number of extracted in-text citations and the number and relative share of matched citations for the top five patent offices in our dataset. More than half of in-text citations are made to patents filed at the USPTO (about 58% of the total). We are able to match 89 percent of them to their correct publication number. Patents filed at the World Intellectual Patent organisation (WIPO) and the Japan Patent Office (JPO), with respectively around 6.5 millions (10% of the total) and 5.7 millions (9% of the total) citations are the second and third largest groups. We match almost 82 percent of the citations to WIPO patent filings and around 77 percent to JPO patent filings. We obtain a similar match rate (i.e., 73%) for citations to patents filed at the German Patent and Trade Mark Office (DPMA),

[31] We consider only citing patents with at least one extracted in-text citation.

around 1.4 million of extracted citations. We obtain less satisfactory match rates for citations to EPO patent filings. They are 2.2 millions and we match only 51 percent of them.

5 A first look into in-text citation data

Front-page patent citations have been extensively used over the past decades and multiple studies have assessed their validity as indicators and discussed their pitfalls. As far as we can ascertain, we are the first to introduce a consistent and validated dataset of in-text patent citations covering all U.S. patents. The purpose of this section is to provide an overview of the characteristics of in-text citations as compared to 'traditional' front-page citations.

We find that in-text and front-page patent citations are two largely distinct sets. We also find that in-text citations are semantically and technologically more similar to the citing patents than their front-page counterparts. This result suggests that in-text patent citations might be a better proxy for knowledge flows than front-page citations, as argued in Section 2. We report that the forward citations counts obtained from the front page and the in-text citations are only weakly correlated. Additionally, we find that in-text citations are more internationalized and reveal a higher degree of self reliance. Table 11 summarizes the key figures of the section. We use our dataset for in-text citations and the IFI CLAIMS dataset for front-page citations. Unless specified, we consider all U.S. patents published from 1790 to 2018.

5.1 Order of magnitudes

From the 16,781,144 U.S. patents in our dataset, we find that 9,453,181 U.S. patents cite at least one patent in the body of their description, corresponding to 56.3 percent of all U.S. patents. Looking at the same set of patents, we observe that 11,923,551 patents (71.3% of the total) exhibit at least one front-page patent citation. In-text citations exhibit high variability over time. The share of U.S. patents citing at least one patent has increased from less than five percent in the second half of the nineteenth century to 70 percent in the 2010s.

When we consider the total number of citations, we find that the number of in-text citations reaches one-third of the front-page citations. We extracted 63,854,733 in-text patent citations while the total number of front-page citations listed by U.S. patents during the same period amounts to 203,557,2015. On average, the body of a patent contains 3.8 patent citations, 6.7 patent citations conditional on citing at least one patent. Once again, there is high variability over time, from less than one in-text patent citation until the early 1960s to more than five since the beginning of the twenty-first century (unconditional on having at least one in-text patent citation).

5.2 Overlap between in-text and front-page patent citations

A natural question is how large is the overlap between in-text and front-page patent citations. To answer it, we list all unique pairs of citing-cited patents, called 'citations' thereafter, for both in-text and front-page citations. Comparing the two lists of citations yields three exclusive and exhaustive sets: citations appearing in the text only, citations observed on the front page only, and citations recorded in both.

We find that citing-cited patent pairs resulting from in-text and front page citations are largely exclusive from one another. Figure 8 depicts the number of patent citations appearing in the text only, on the front page only, and in both. There are 11,868,037 patent citations appearing both in the text and on the front page, which represents only 5.79 percent of all front-page citations and 24.2 percent of all in-text citations.[32] 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. In the end, over the whole period, considering in-text patent citations adds 37,541,592 citations that are not found among front-page citations. Focusing only on front-page patent citations leads to missing 15.34 percent of all patent citations. That is what we call *the missing 15 percent of patent citations*.

A potential explanation of the missing 15 percent of patent citations could be patents cited as a "Translation of Patent ..." or as "Patent Abstracts ...". These references are listed in the front page as part of the non-patent literature (NPL). A legitimate question, therefore, is whether these missing 15 percent are (at least partly)

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

available as front-page NPL. If true, extracting in-text patent citations would not bring more information than parsing 'patent' citations reported in the front page NPL section, a much simpler task. To delve deeper into this question, we trained a text classifier to determine whether a front-page NPL citation contains information on a patent. This classifier achieves a sufficient 78.31 percent precision and 89.04 percent recall on the test set. We then applied it to the universe of front-page NPL citations recorded in the DOCDB database. In the end, we estimate that there are 1,714,260 such 'patent' citations reported in the front page NPL sections of U.S. patents since 1947. Making the bold assumption that all these citations appear in the text as well, these patent citations would reduce the missing patents to 14.63 percent (-0.71 percentage points).

It is now clear that in-text and front-page patent citations exhibit very little overlap. Thus, quantitatively, considering in-text patent citations does bring new information. Next, we try to understand whether and how their qualitative characteristics differ.

5.3 Textual similarity between citing and cited patents

Figure 9 shows distributions of semantic similarity between citing and cited documents for in-text and front-page citations. Semantic similarity is calculated as the dot product of Google Patent's document embedding vectors, which were recently made available to researchers.[33] The embeddings are trained to predict CPC categories from each patent's full-text with a WSABIE algorithm (Weston et al., 2010). Figure 9 also shows two reference semantic similarity distributions. The first one ('Within art unit') is based on the similarity between randomly chosen pairs of patents examined by the same art unit. The second one ('Random') is based on the similarity between cited in-text patents matched to a random citing patent.

To produce Figure 9 we considered only citing and cited patents that were granted by the USPTO in the years 2000–2009 (for ease of interpretation). In Figure 9a we removed 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 sample size. INPADOC

[33] https://cloud.google.com/blog/products/gcp/google-patents-public-datasets-connecting-public-paid-and-private-patent-data. Note that a myriad of similarity measures exist, including Younge and Kuhn (2016); Arts et al. (2018, 2020).

families, also known as 'extended patent families', include all patents that can be linked through their priorities (but not necessarily to a single common priority filing).[34] 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 9b we report the same similarity distributions, excluding citations between patents belonging to the same DOCDB family. These families, also known as 'simple patent families,' consist of sets of patents linked to a common priority filing. They are smaller and more selective than INPADOC families. The in-text citation similarity distribution shown in 9b 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 9a), as they are more comparable to front-page citations.

One can make a number of observations from this graphical comparison of similarities. First, in agreement with our validation measures, there are unlikely to be a large portion of in-text citations that 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 roughly similar. Second, the in-text citation distribution is shifted to slightly higher levels of similarity when compared to the distribution for front-page citations. This shift indicates that patents cited in-text are, on average, more technologically similar to the citing patent than patents cited on the front page. 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 are drawn from a wider (but still related) set of prior art.

This evidence reinforces our view of in-text citations as a promising indicator of

[34] https://www.epo.org/searching-for-patents/helpful-resources/first-time-here/patent-families/inpadoc.html

knowledge flows, potentially less "noisy" than front-page ones (Jaffe et al., 1998; Corsino et al., 2019; Kuhn et al., 2020) and more closely related to the focal inventors' prior knowledge, less affected by the complex patent examination procedure (Choudhury et al., 2020) through examiners' (Alcacer and Gittelman, 2006) or patent attorneys' practices (Jaffe et al., 2000).

5.4 Forward citations

The count of forward citations, that is, the number of times a patent is cited by another patent, has been widely used in various contexts as a way to measure the quality of a patent, but also as an output measure in settings where innovation or knowledge flows are susceptible to be affected by another economic variable (see Jaffe et al., 1993; Almeida, 1996; Kerr, 2008; Agarwal et al., 2009 *inter alia*). Due to the large interest of the community for this forward citations count and its central role in innovation research, it seems natural to use our dataset to compute the forward citations count based on in-text citations rather than the usual front page citations. Further, we compare the forward citations counts obtained using in-text citations and front page citations.

To do so, we consider U.S. patents and their in-text and front-page citations. We slightly depart from raw forward citations counts in two ways. First, in order to make our results immune to potential variations between in-text and front-page citing patterns, we compute the forward citations count at the invention level, as defined by the DOCDB family, rather than at the publication level.[35] Second, we exclude citations from patents belonging to the same extended invention family as defined by the INPADOC family. This is a conservative choice aiming at excluding self-references in a broad sense.

The first observation is that in-text and front page citations are 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. This result further confirms the fact that in-text

[35] The kind of pitfall that we want to avoid is, for example, a higher tendency to cite applications in the text instead of granted patents on the front page and vice versa, including for the exact same invention.

citations bring additional and distinct information from their front page counterparts.

The second observation is certainly the most important and puzzling one: the count of forward citations based on in-text citations is 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 correlation between the two measures is 0.23. Figure 10 shows these two forward citations count for a random sample of 10 percent of the patent families cited both in the text and on the front page. The regression line corresponds to a univariate model where the dependent variable is the front page forward citations count and the in-text count is the independent variable. The associated R-squared is close to zero (0.03), highlighting the poor predictive power of the dependent variable over the independent variable. Roughly speaking, the two forward citations counts are almost orthogonal. The above result is puzzling and raises a host of questions about the use of in-text forward citations count.

The third set of observations relates to the distribution of forward citations counts. Figure 11 compares the empirical probability (panel 11a) and cumulative (panel 11b) 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. 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.

Among our results, the orthogonality puzzle is certainly the most challenging to grasp for the community.

5.5 The internationalization of patent citations

International patent citations, that is, citations to and from patents granted at a foreign patent office, have been used as a way to measure countries' contribution to the creation and diffusion of innovation and ultimately productivity growth. For example, Eaton and Kortum (1996a,b, 1997, 1999) have used international patent citations to infer the direction and magnitude of the international diffusion of technology. Such studies have a profound impact on our representation of who are the main contributors of technological progress and worldwide productivity growth. Here we look at the distribution of U.S. citations by country of patent office obtained using the 'traditional' front-page citations and the newly available in-text citations.

We find that in-text patent citations are almost three times as much internationalized than front-page citations. Figure 12 represents the number of citations from U.S. patents by country of the cited patent for front-page (panel 12a) and in-text (panel 12b) citations. Since 1947, the share of in-text patent citations to non-U.S. patents has reached 28.82[36] percent while it is 11.0 percent for front-page citations. Thus, considering only front-page citations leads to a more U.S.-centric view of knowledge flows. Going further, we find that some countries are dramatically under-represented in front-page citations as compared to in-text citations. For example, the share of Japanese patents in-text citations is almost three times as large as their share in front-page citations. We believe that our representation of the direction and magnitude of international knowledge flows might well improve in light of our newly available data.

5.6 Self-reliance

In the context of the present study, we call 'self-reliance' the citation of one or more patents belonging to the same family or originating by the same patentees as the citing patent itself. There are two main reasons to be interested in the role of self-reliance in patent citations. First, the diffusion of a piece of knowledge is likely to be conveyed primarily by the persons and organisations who created it. Second, one

[36] This result is immune to the exclusion of "self-citations". Excluding within-INPADOC-family citations, citations to non-U.S. patents represent 31.38 percent of all citations.

might be worried that in-text citations are mostly self-reference, that is citations of patents belonging to the same family of invention. Consequently, they would not bring much information compared to already available patent family information.

Starting with same-family citations, we map each citing and cited patent to its patent family and compute the share of citations citing a patent belonging to its own family. We consider both the DOCDB families and the INPADOC families. As previously explained, INPADOC families are more permissive as they include in the same group all the documents sharing directly or indirectly (e.g., via a third document) at least *one* priority. We find that the share of in-text citations belonging to the same DOCDB (INPADOC) family is 6.42 (10.65) percent. This is higher than front-page citations' self-references figures, which are 0.69 (1.63) percent. That being said, even considering the most permissive definition of invention families, 90 percent of in-text citations are not self-references, bringing useful information of patented inventions' knowledge background outside their respective patent family.

Turning to same-patentee citations, 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. We find that 17.43 (22.46) percent of in-text patent citations have at least one inventor (assignee) in common with their citing patent, against 5.98 (9.26) percent for front-page citations, that is almost three (two) times as much. 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.

5.7 Geographic distribution

A large literature has documented how geography restricts knowledge flows' breadth (Jaffe et al., 1993; Audretsch, 1998; Peri, 2005; Belenzon and Schankerman, 2013). Scholars have pointed to labor mobility within regional labor markets (Almeida and Kogut, 1999) and localized co-invention networks (Breschi and Lissoni, 2009) as leading mechanisms of knowledge flows' geographic concentration. Patent (front-page) citations have been a crucial data source for these studies, proxying the elusive "paper trail" of knowledge (Krugman, 1991) connecting patented inventions.

In this section, we compare in-text and front-page citations in the geographic space. Specifically, we take citing-cited inventor dyads in the two citation groups and calculate the distance between the two inventors' geocoded addresses (de Rassenfosse et al., 2019b), comparing their geographic distribution. Despite being a mere descriptive exercise, this analysis can provide useful insights about differences between in-text and front-page citations along the geographic dimension.

Figure 13 shows the probability distribution function and cumulative distribution function of in-text and front-page citing-cited inventor dyads. The x-axis quantifies distance in kilometers. All graphs using all kind of citations portray in-text citations as more localized than front-page ones. Panel 13e in particular, shows a higher share of citations within 25km of distance from the cited inventor's location for in-text citations, relative to front-page ones. We also report the same distributions excluding all self-citations between patents appearing in the same INPADOC family and all self-citations at the assignee-level.[37] While in-text citations seem to still be slightly more localized, the difference with front-page ones is minimal and substantially less sharp than suggested by unconditional figures, mostly the result of a higher share of in-text citations occurring at "zero" distance (see panel 13f). The higher geographic localization of in-text citations portrayed in Figure 13 when considering all citations seems to be explained by a larger occurrence of self-citations for in-text relative to front-page citations.

At the descriptive level, in-text and front-page citations do not display particular differences in terms of their geographic distributions. Nevertheless, we believe that an econometric investigation will be needed to probe this question properly (e.g., following the approach pioneered by Jaffe et al., 1993).

6 Concluding remarks

This paper introduces a novel dataset on patent citations. It provides 63,854,733 million citations identified in the full-text of 16,781,144 million U.S. patent documents

[37] We identify self-citations using the same procedure employed in section 5.6.

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 in various fields of the social sciences, we expect these data to be of considerable interest to the scientific community.

Three main messages are particularly noteworthy. First, we found little overlap between the 'traditional' front-page citations and the novel in-text citations. We estimate that the inclusion of in-text citations adds a net 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 and provided tentative evidence that in-text citations offer a particularly relevant trace of knowledge flow compared to front-page citations. We have also explained why in-text citations represent valuable signals about patent importance. Capturing knowledge flow and measuring patent importance are two of the most popular uses of patent citations and, therefore, we encourage researchers to explore the present data.

Finally, we have relied on best-in-class techniques from NLP and have performed in-depth validation exercises to ensure the quality of the data, achieving highly satisfactory results. We see these results as a proof of the considerable potential offered by the open source community and more particularly applications of modern NLP to information extraction in applied economics and management. In this context, we have made the codebase and the replication material (including code and validation data) natively open source and the data open access.[38] We encourage the community to contribute to the continuous improvement of the dataset. Of particular interest will be the deployment of our pipeline to other jurisdictions.

In conclusion, we hope that the public release of the dataset will enable the community to shed new light on studies exploiting citation data to track knowledge flows and measure patent importance. Furthermore, the data may open new research questions related, e.g., to strategic knowledge disclosure (*à la* Lampe, 2012) or knowledge sourcing (*à la* Wagner et al., 2014). On a more technical level, we see value in lever-

[38] The code is licensed under the MIT license https://opensource.org/licenses/MIT. The data are licensed under the CC-BY-4 license https://creativecommons.org/licenses/by/4.0/legalcode.

aging the context of patent citations to determine citation intent (enablement, usefulness, non obviousness, improvement, etc). Such contextual information could lead to a more accurate usage of patent citation data.

References

Adams, Stephen, "The text, the full text and nothing but the text: Part 1–Standards for creating textual information in patent documents and general search implications," *World Patent Information*, 2010, *32* (1), 22–29.

Agarwal, Rajshree, Martin Ganco, and Rosemarie H Ziedonis, "Reputations for toughness in patent enforcement: Implications for knowledge spillovers via inventor mobility," *Strategic Management Journal*, 2009, *30* (13), 1349–1374.

Akcigit, Ufuk, John Grigsby, and Tom Nicholas, "Immigration and the rise of american ingenuity," *American Economic Review*, 2017, *107* (5), 327–31.

Albert, Michael B, Daniel Avery, Francis Narin, and Paul McAllister, "Direct validation of citation counts as indicators of industrially important patents," *Research policy*, 1991, *20* (3), 251–259.

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

Almeida, Paul, "Knowledge sourcing by foreign multinationals: Patent citation analysis in the US semiconductor industry," *Strategic management journal*, 1996, *17* (S2), 155–165.

_ **and Bruce Kogut**, "Localization of knowledge and the mobility of engineers in regional networks," *Management Science*, 1999, *45* (7), 905–917.

Andrews, Michael, "Historical patent data. A practitioner's guide," *Available at SSRN: https://ssrn.com/abstract=3415318*, 2020.

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

_ **, Jianan Hou, and Juan Carlos Gomez**, "Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures," *Research Policy*, 2020, p. 104144.

Audretsch, Bruce, "Agglomeration and the location of innovative activity," *Oxford Review of Economic Policy*, 1998, *14* (2), 18–29.

Belenzon, Sharon and Mark Schankerman, "Spreading the word: Geography, policy, and knowledge spillovers," *Review of Economics and Statistics*, 2013, *95* (3), 884–903.

Benson, Christopher L and Christopher L Magee, "Quantitative determination of technological improvement from patent data," *PloS one*, 2015, *10* (4), e0121635.

Berkes, Enrico, "Comprehensive universe of US patents (CUSP): data and facts," *Unpublished, Ohio State University*, 2018.

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

Bryan, Kevin A, Yasin Ozcan, and Bhaven Sampat, "In-text patent citations: A user's guide," *Research Policy*, 2020, *49* (4), 103946.

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

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

Chien, Colleen V and Jiun Ying Wu, "Decoding Patentable Subject Matter," *Patently-O Patent Law Journal 1, Santa Clara University Legal Studies Research Paper*, 2018, *1.*

Choudhury, Prithwiraj, Evan Starr, and Rajshree Agarwal, "Machine learning and human capital complementarities: Experimental evidence on bias mitigation," *Strategic Management Journal*, 2020, *41* (8), 1381–1411.

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

Cotropia, Christopher A, "The folly of early filing in patent law," *Hastings Law Journal*, 2009, *61*, 65–129.

de Rassenfosse, Gaétan, Adam Jaffe, and Emilio Raiteri, "The procurement of innovation by the US government," *PloS one*, 2019, *14* (8), e0218927.

_ , Jan Kozak, and Florian Seliger, "Geocoding of worldwide patent data," *Scientific data*, 2019, *6* (1), 1–15.

Eaton, Jonathan and Samuel Kortum, "Measuring technology diffusion and the international sources of growth," *Eastern Economic Journal*, 1996, *22* (4), 401–410.

_ and _ , "Trade in ideas Patenting and productivity in the OECD," *Journal of International Economics*, 1996, *40* (3-4), 251–278.

_ **and** _ , "Engines of growth: Domestic and foreign sources of innovation," *Japan and the World Economy*, 1997, *9* (2), 235–259.

_ **and** _ , "International technology diffusion: Theory and measurement," *International Economic Review*, 1999, *40* (3), 537–570.

Freilich, Janet, "Prophetic Patents," *UC Davis Law Review*, 2019, *53*, 663–731.

Galibert, Olivier, Sophie Rosset, Xavier Tannier, and Fanny Grandry, "Hybrid Citation Extraction from Patents.," in "Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010" 2010, pp. 17–23.

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

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

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

_ , _ , and Rebecca Henderson, "Geographic localization of knowledge spillovers as evidenced by patent citations," *the Quarterly journal of Economics*, 1993, *108* (3), 577–598.

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

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

Kerr, William R, "Ethnic scientific communities and international technology diffusion," *The Review of Economics and Statistics*, 2008, *90* (3), 518–537.

Krugman, Paul R, *Geography and trade*, MIT press, 1991.

Kuhn, Jeffrey, Kenneth Younge, and Alan Marco, "Patent citations reexamined," *The RAND Journal of Economics*, 2020, *51* (1), 109–132.

Lafferty, John, Andrew McCallum, and Fernando CN Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

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

Lanjouw, Jean O and Mark Schankerman, "Patent quality and research productivity: Measuring innovation with multiple indicators," *The Economic Journal*, 2004, *114* (495), 441–465.

Li, Jing, Aixin Sun, Jianglei Han, and Chenliang Li, "A survey on deep learning for named entity recognition," *IEEE Transactions on Knowledge and Data Engineering*, 2020, pp. 1–1.

Lopez, Patrice, "Automatic extraction and resolution of bibliographical references in patent documents," in "Information Retrieval Facility Conference" Springer 2010, pp. 120–135.

Machin, Nathan, "Prospective Utility: A New Interpretation of the Utility Requirement of Section 101 of the Patent Act," *California Law Review*, 1999, *87*, 421–456.

Marx, Matt and Aaron Fuegi, "Reliance on Science by Inventors: Hybrid Extraction of In-text Patent-to-Article Citations," *National Bureau of Economic Research, Working Paper 27987*, 2020.

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

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

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

Porter, Alan L, Jan Youtie, Philip Shapira, and David J Schoeneck, "Refining search terms for nanotechnology," *Journal of nanoparticle research*, 2008, *10* (5), 715–728.

Righi, Cesare and Timothy Simcoe, "Patent examiner specialization," *Research Policy*, 2019, *48* (1), 137–148.

Roche, Emmanuel and Yves Schabes, *Finite-state language processing*, MIT press, 1997.

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

Sutton, Charles and Andrew McCallum, "An introduction to conditional random fields for relational learning," *Introduction to statistical relational learning*, 2006, *2*, 93–128.

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

Verluise, Cyril and Gaétan de Rassenfosse, "PatCit: A Comprehensive Dataset of Patent Citations (Version 0.15) [Data set]," *Zenodo. http://doi.org/10.5281/zenodo.3710994*, 2020.

Wagner, Stefan, Karin Hoisl, and Grid Thoma, "Overcoming localization of knowledge?the role of professional service firms," *Strategic Management Journal*, 2014, *35* (11), 1671–1688.

Weston, Jason, Samy Bengio, and Nicolas Usunier, "Wsabie: Scaling up to large vocabulary image annotation," in "Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Three" 2010, pp. 2764–2770.

Younge, Kenneth A and Jeffrey M Kuhn, "Patent-to-patent similarity: a vector space model," *Available at SSRN: https://ssrn.com/abstract=2709238*, 2016.

Tables

Table 1: Composition of the dataset

Kind code	**Kind of document**		**Number**	**Share**
	Pre 2001	*Post 2001*		
A	Patent	Patent application	11,909,035	0.71
B	Reexamination certificate	Patent	4,188,597	0.25
S	-	Design patent	613,050	0.04
P	Plant patent	Plant patent & Plant patent application	34,852	2.00E-3
E	-	Reissued patent	32,226	2.00E-3
H	-	Statutory invention registration (SIR)	2,255	1.00E-4
I	-	-	1,129	6.00E-5

Table 2: Composition of the dataset: focus on patents and applications

Kind code	**Kind of document**		**Number**	**Share**
	Pre 2001	*Post 2001*		
A	Patent	-	6,145,197	0.37
A1	-	Patent application publication	5,753,613	0.34
A2	-	Patent application publication (republication)	1,742	1.00E-4
A9	-	Patent application publication (corrected publication)	8,483	5.00E-4
B1	-	Patent (no pre-grant publication)	776,074	0.04
B2	-	Patent	3,412,523	0.2

Notes: Share of full dataset.

Table 3: In-text patent citations extraction performance

	Number of patents in the test set	Avg number of patent tags per patent	Precision	Recall	F1-score
Galibert et al. (2010)	760	12.75	64.4%	61.0%	62.6%
Lopez (2010)	20	9.96	97.44%	97.74%	97.68%
Verluise et al (2020)	160	2.93	97%	82%	89.2%

Table 4: In-text patent citations extraction error analysis

Error type	Category	Example
False negative	1	"introduced into a mold (as in Example 1 of 2,154,639) wherein it is polymerized to form a A"
	2	"Aug. 20, 1935 2,255,030 Tholstrup Sept. 2, 1941 2,394,733 Wittenrnyer Feb. 12, 1946 2,433,349 Drewell Dec. 30"
	3	"Filed May 25, 1973, Ser. No. 364,196 Int. Cl. Blk 1/00, 3/06; C01b"
False positive	1	"US. Cl ..29/492, 29/497, 29/498, 29/502, 29/589, 29/628 [51] lnt.Cl."
	2	"Aug. 12, 1941. ALKAN&#39; emoumnmrc COMPASS I iled July 15, 1936 3"
	3	"No. 09/808,790, (Attorney Docket No. 20468-000110), previously incorporated herein by reference. FIG"

Notes: The underlined span of text triggered the error. In the false negative case, it was not detected by Grobid as a patent citation while it should have been the case. In the false positive case, it was detected by Grobid as a patent citation while it is not.

Table 5: Distribution of U.S. patent citations by patent office

Patent office	**Number of occurrences in validation sample**	**Share in validation sample**	**Share in universe of U.S. patents**
US	203	0.67	0.61
JP	52	0.17	0.09
WO	18	0.06	0.10
DE	9	0.03	0.02
EP	5	0.02	0.03
KR	4	0.01	7.00E-3
FR	4	0.01	6.00E-3
BE	2	7.00E-03	3.00E-3
SA	1	3.00E-03	3.00E-3
CH	1	3.00E-03	3.00E-3
AL	1	3.00E-03	0.02

Table 6: In-text patent citations parsing accuracy

	Number of examples in the test set	**Organisation name**	**Original number**	**Kind code**	**All**
Lopez (2010	250	-	-	-	97.2%
Verluise et al (2020)	300	98.4%	95.7%	97.6%	-

Notes: Lopez (2010) does not distinguish between the accuracy on the three attributes and reports the overall accuracy of the Finite State Transducers to translate the natural language citation into a fully structured citation represented by its three attributes.

Table 7: In-text patent citations matching performance

	True		**False**	
	Content	*Number*	*Content*	*Number*
Positive	A publication number was correctly matched	137	A publication number was incorrectly matched	10
Negative	No matched publication number and no match found by the annotator	36	No matched publication number but a match was found by the annotator	17

Table 8: In-text patent citations matching error analysis

Error type	**Category**	**Sub-category**	**Example**	**Number of occurrences**
False match	Incorrect patent	Badly formatted pre-2000 Japanese patent	JP5064281 instead of JPS5064281	5
		Incorrect extraction of pre-1970 U.S. patent due to bad OCR	CA-8465T-T (from 2,936,846 5/60 Tyler et al, in reference list)	1
	Non patent	Garbled table	-	2
		Technology class	US-32537 extracted from "... U.S. Cl. 325/392, 325/37..."	1
		Date	US-312012 extracted from "...filed Aug. 31, 2012, . . . "	1
False no-match	Formatting	Missing leading zeros after country code or date	EP592106 instead of EP0592106	6
		Year reported after instead of before patent number	JP3518222000 instead of JP2000351822	3
		Incorrect extraction of country code	SU-14553625 extracted from "U.S. Utility application Ser. No. 14/553,625"	1
	Wrong service call	-	-	7

Notes: Error analysis based on 200 random examples.

Table 9: Extracted citations judged unmatchable by the annotator

Category	**Example**	**Number of occurrences**
Garbled tables	AL-1226-C extracted from "...AL C 257 75.108 67.122 6.016 1..."	11
Provisional patent applications	US-60723639 extracted from "U.S. provisional application Ser. No. 60/723,639"; provisional patent applications are not public information	8
Incorrect and ambiguous number formats	EP-87309853 extracted from "European patent specification No 87309853.7" (non-standard format of a non-searchable application number)	4
Incorrect parsed attributes	WO-PTS0767103 instead of WO-PTUS07067103	5
Non searchable	DE-19654649 (not indexed by Google Patents)	3
Non patents (technological class, dates, etc)	US-32128 extracted from "... U.S. Cl. 322/79, 310/68 D, 321/28, ..."	10

Notes: The Number of occurrences includes both matched and unmatched examples.

Table 10: Number and share of citations matched by patent organisation (selected)

Patent organisation	**Total number of citations**	**Share of citations matched**
USPTO	37,072,526	89.14
WIPO	6,453,099	81.89
JPO	5,659,300	77.22
EPO	2,228,096	51.27
DPMA	1,371,114	73.46

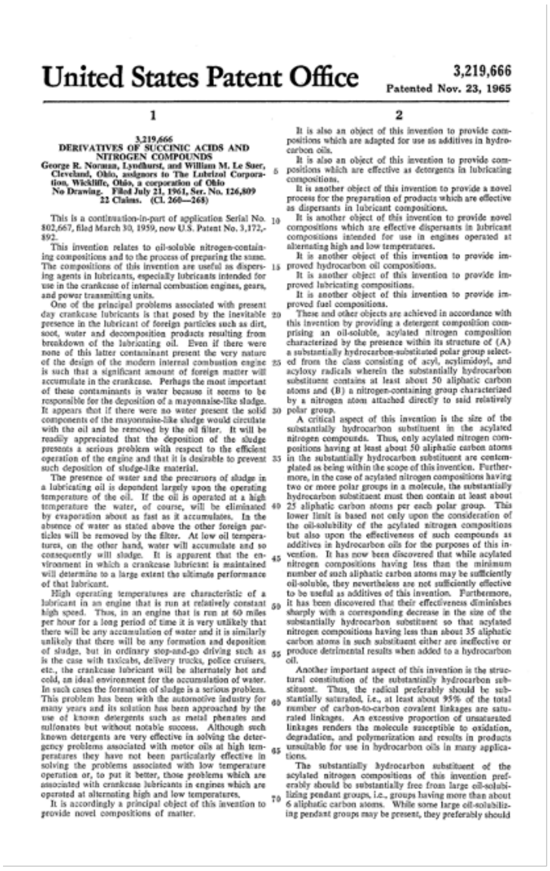
Table 11: In-text and front page citations at-a-glance

	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.

Figures

Figure 1: Example of the USPTO “old” patent format (US-3219666-A)



United States Patent Office

3,219,666

Patented Nov. 23, 1965

1

3,219,666
DERIVATIVES OF SUCCINIC ACIDS AND NITROGEN COMPOUNDS
George R. Norman, Lyndhurst, and William M. Le Suer, Cleveland, Ohio, assignors to The Lubrizol Corporation, Wickliffe, Ohio, a corporation of Ohio
No Drawing. Filed July 21, 1961, Ser. No. 126,809
22 Claims. (Cl. 260—268)

This is a continuation-in-part of application Serial No. 802,667, filed March 30, 1959, now U.S. Patent No. 3,172,892.

This invention relates to oil-soluble nitrogen-containing compositions and to the process of preparing the same. The compositions of this invention are useful as dispersing agents in lubricants, especially lubricants intended for use in the crankcase of internal combustion engines, gears, and power transmitting units.

One of the principal problems associated with present day crankcase lubricants is that posed by the inevitable presence in the lubricant of foreign particles such as dirt, soot, water and decomposition products resulting from breakdown of the lubricating oil. Even if there were none of this latter contaminant present the very nature of the design of the modern internal combustion engine is such that a significant amount of foreign matter will accumulate in the crankcase. Perhaps the most important of these contaminants is water because it seems to be responsible for the deposition of a mayonnaise-like sludge. It appears thot if there were no water present the solid components of the mayonnaise-like sludge would circulate with the oil and be removed by the oil filter. It will be readily appreciated that the deposition of the sludge presents a serious problem with respect to the efficient operation of the engine and that it is desirable to prevent such deposition of sludge-like material.

The presence of water and the precursors of sludge in a lubricating oil is dependent largely upon the operating temperature of the oil. If the oil is operated at a high temperature the water, of course, will be eliminated by evaporation about as fast as it accumulates. In the absence of water as stated above the other foreign particles will be removed by the filter. At low oil temperatures, on the other hand, water will accumulate and so consequently will sludge. It is apparent that the environment in which a crankcase lubricant is maintained will determine to a large extent the ultimate performance of that lubricant.

High operating temperatures are characteristic of a lubricant in an engine that is run at relatively constant high speed. Thus, in an engine that is run at 60 miles per hour for a long period of time it is very unlikely that there will be any accumulation of water and it is similarly unlikely that there will be any formation and deposition of sludge, but in ordinary stop-and-go driving such as is the case with taxicabs, delivery trucks, police cruisers, etc., the crankcase lubricant will be alternately hot and cold, an ideal environment for the accumulation of water. In such cases the formation of sludge is a serious problem. This problem has been with the automotive industry for many years and its solution has been approached by the use of known detergents such as metal phenates and sulfonates but without notable success. Although such known detergents are very effective in solving the detergency problems associated with motor oils at high temperatures they have not been particularly effective in solving the problems associated with low temperature operation or, to put it better, those problems which are associated with crankcase lubricants in engines which are operated at alternating high and low temperatures.

It is accordingly a principal object of this invention to provide novel compositions of matter.

2

It is also an object of this invention to provide compositions which are adapted for use as additives in hydrocarbon oils.

It is also an object of this invention to provide compositions which are effective as detergents in lubricating compositions.

It is another object of this invention to provide a novel process for the preparation of products which are effective as dispersants in lubricant compositions.

It is another object of this invention to provide novel compositions which are effective dispersants in lubricant compositions intended for use in engines operated at alternating high and low temperatures.

It is another object of this invention to provide improved hydrocarbon oil compositions.

It is another object of this invention to provide improved lubricating compositions.

It is another object of this invention to provide improved fuel compositions.

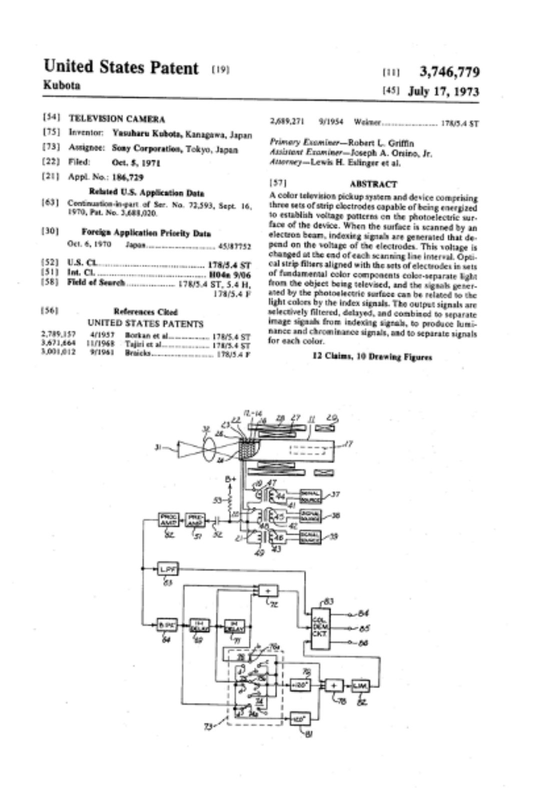
These and other objects are achieved in accordance with this invention by providing a detergent composition comprising an oil-soluble, acylated nitrogen composition characterized by the presence within its structure of (A) a substantially hydrocarbon-substituted polar group selected from the class consisting of acyl, acylimidoyl, and acyloxy radicals wherein the substantially hydrocarbon substituent contains at least about 50 aliphatic carbon atoms and (B) a nitrogen-containing group characterized by a nitrogen atom attached directly to said relatively polar group.

A critical aspect of this invention is the size of the substantially hydrocarbon substituent in the acylated nitrogen compounds. Thus, only acylated nitrogen compositions having at least about 50 aliphatic carbon atoms in the substantially hydrocarbon substituent are contemplated as being within the scope of this invention. Furthermore, in the case of acylated nitrogen compositions having two or more polar groups in a molecule, the substantially hydrocarbon substituent must then contain at least about 25 aliphatic carbon atoms per each polar group. This lower limit is based not only upon the consideration of the oil-solubility of the acylated nitrogen compositions but also upon the effectiveness of such compounds as additives in hydrocarbon oils for the purposes of this invention. It has now been discovered that while acylated nitrogen compositions having less than the minimum number of such aliphatic carbon atoms may be sufficiently oil-soluble, they nevertheless are not sufficiently effective to be useful as additives of this invention. Furthermore, it has been discovered that their effectiveness diminishes sharply with a corresponding decrease in the size of the substantially hydrocarbon substituent so that acylated nitrogen compositions having less than about 35 aliphatic carbon atoms in such substituent either are ineffective or produce detrimental results when added to a hydrocarbon oil.

Another important aspect of this invention is the structural constitution of the substantially hydrocarbon substituent. Thus, the radical preferably should be substantially saturated, i.e., at least about 95% of the total number of carbon-to-carbon covalent linkages are saturated linkages. An excessive proportion of unsaturated linkages renders the molecule susceptible to oxidation, degradation, and polymerization and results in products unsuitable for use in hydrocarbon oils in many applications.

The substantially hydrocarbon substituent of the acylated nitrogen compositions of this invention preferably should be substantially free from large oil-solubilizing pendant groups, i.e., groups having more than about 6 aliphatic carbon atoms. While some large oil-solubilizing pendant groups may be present, they preferably should

Figure 2: Example of the USPTO “new” patent format (US-3746779-A)



United States Patent [19]
Kubota

[11] **3,746,779**
[45] **July 17, 1973**

[54] TELEVISION CAMERA
[75] Inventor: **Yasuharu Kubota,** Kanagawa, Japan
[73] Assignee: **Sony Corporation,** Tokyo, Japan
[22] Filed: **Oct. 5, 1971**
[21] Appl. No.: **186,729**

Related U.S. Application Data
[63] Continuation-in-part of Ser. No. 72,593, Sept. 16, 1970, Pat. No. 3,688,020.

[30] **Foreign Application Priority Data**
Oct. 6, 1970 Japan 45/87752

[52] **U.S. Cl.** .... **178/5.4 ST**
[51] **Int. Cl.** **H04n 9/06**
[58] **Field of Search** .... 178/5.4 ST, 5.4 H, 178/5.4 F

[56] **References Cited**
UNITED STATES PATENTS

2,789,157	4/1957	Borkan et al.	178/5.4 ST
3,671,664	11/1968	Tajiri et al.	178/5.4 ST
3,001,012	9/1961	Braicks	178/5.4 F
2,689,271	9/1954	Weimer	178/5.4 ST

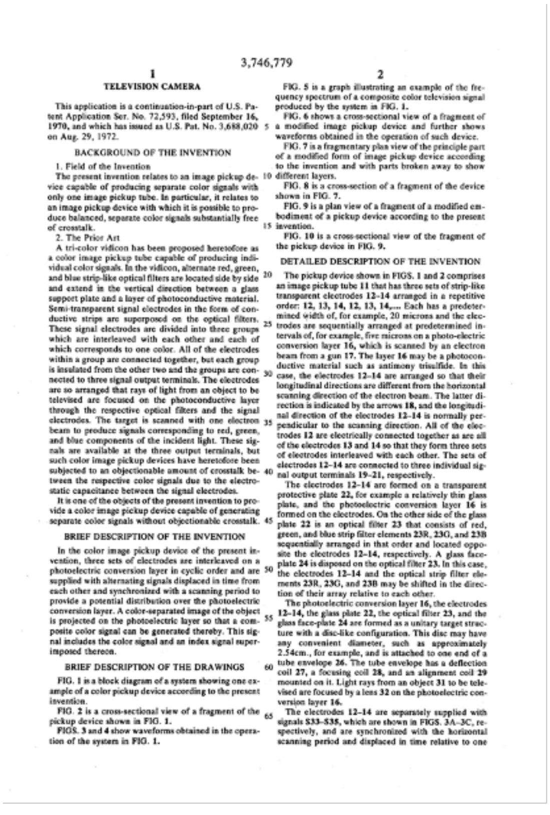
Primary Examiner—Robert L. Griffin
Assistant Examiner—Joseph A. Orsino, Jr.
Attorney—Lewis H. Eslinger et al.

[57] **ABSTRACT**

A color television pickup system and device comprising three sets of strip electrodes capable of being energized to establish voltage patterns on the photoelectric surface of the device. When the surface is scanned by an electron beam, indexing signals are generated that depend on the voltage of the electrodes. This voltage is changed at the end of each scanning line interval. Optical strip filters aligned with the sets of electrodes in sets of fundamental color components color-separate light from the object being televised, and the signals generated by the photoelectric surface can be related to the light colors by the index signals. The output signals are selectively filtered, delayed, and combined to separate image signals from indexing signals, to produce luminance and chrominance signals, and to separate signals for each color.

12 Claims, 10 Drawing Figures

(a) Front page



3,746,779

1

TELEVISION CAMERA

This application is a continuation-in-part of U.S. Patent Application Ser. No. 72,593, filed September 16, 1970, and which has issued as U.S. Pat. No. 3,688,020 on Aug. 29, 1972.

BACKGROUND OF THE INVENTION

1. Field of the Invention

The present invention relates to an image pickup device capable of producing separate color signals with only one image pickup tube. In particular, it relates to an image pickup device with which it is possible to produce balanced, separate color signals substantially free of crosstalk.

2. The Prior Art

A tri-color vidicon has been proposed heretofore as a color image pickup tube capable of producing individual color signals. In the vidicon, alternate red, green, and blue strip-like optical filters are located side by side and extend in the vertical direction between a glass support plate and a layer of photoconductive material. Semi-transparent signal electrodes in the form of conductive strips are superposed on the optical filters. These signal electrodes are divided into three groups which are interleaved with each other and each of which corresponds to one color. All of the electrodes within a group are connected together, but each group is insulated from the other two and the groups are connected to three signal output terminals. The electrodes are so arranged that rays of light from an object to be televised are focused on the photoconductive layer through the respective optical filters and the signal electrodes. The target is scanned with one electron beam to produce signals corresponding to red, green, and blue components of the incident light. These signals are available at the three output terminals, but such color image pickup devices have heretofore been subjected to an objectionable amount of crosstalk between the respective color signals due to the electrostatic capacitance between the signal electrodes.

It is one of the objects of the present invention to provide a color image pickup device capable of generating separate color signals without objectionable crosstalk.

BRIEF DESCRIPTION OF THE INVENTION

In the color image pickup device of the present invention, three sets of electrodes are interleaved on a photoelectric conversion layer in cyclic order and are supplied with alternating signals displaced in time from each other and synchronized with a scanning period to provide a potential distribution over the photoelectric conversion layer. A color-separated image of the object is projected on the photoelectric layer so that a composite color signal can be generated thereby. This signal includes the color signal and an index signal superimposed thereon.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a block diagram of a system showing one example of a color pickup device according to the present invention.

FIG. 2 is a cross-sectional view of a fragment of the pickup device shown in FIG. 1.

FIGS. 3 and 4 show waveforms obtained in the operation of the system in FIG. 1.

2

FIG. 5 is a graph illustrating an example of the frequency spectrum of a composite color television signal produced by the system in FIG. 1.

FIG. 6 shows a cross-sectional view of a fragment of a modified image pickup device and further shows waveforms obtained in the operation of such device.

FIG. 7 is a fragmentary plan view of the principle part of a modified form of image pickup device according to the invention and with parts broken away to show different layers.

FIG. 8 is a cross-section of a fragment of the device shown in FIG. 7.

FIG. 9 is a plan view of a fragment of a modified embodiment of a pickup device according to the present invention.

FIG. 10 is a cross-sectional view of the fragment of the pickup device in FIG. 9.

DETAILED DESCRIPTION OF THE INVENTION

The pickup device shown in FIGS. 1 and 2 comprises an image pickup tube **11** that has three sets of strip-like transparent electrodes **12–14** arranged in a repetitive order: **12, 13, 14, 12, 13, 14,...** Each has a predetermined width of, for example, 20 microns and the electrodes are sequentially arranged at predetermined intervals of, for example, five microns on a photo-electric conversion layer **16**, which is scanned by an electron beam from a gun **17**. The layer **16** may be a photoconductive material such as antimony trisulfide. In this case, the electrodes **12–14** are arranged so that their longitudinal directions are different from the horizontal scanning direction of the electron beam. The latter direction is indicated by the arrows **18**, and the longitudinal direction of the electrodes **12–14** is normally perpendicular to the scanning direction. All of the electrodes **12** are electrically connected together as are all of the electrodes **13** and **14** so that they form three sets of electrodes interleaved with each other. The sets of electrodes **12–14** are connected to three individual signal output terminals **19–21**, respectively.

The electrodes **12–14** are formed on a transparent protective plate **22**, for example a relatively thin glass plate, and the photoelectric conversion layer **16** is formed on the electrodes. On the other side of the glass plate **22** is an optical filter **23** that consists of red, green, and blue strip filter elements **23R**, **23G**, and **23B** sequentially arranged in that order and located opposite the electrodes **12–14**, respectively. A glass face-plate **24** is disposed on the optical filter **23**. In this case, the electrodes **12–14** and the optical strip filter elements **23R**, **23G**, and **23B** may be shifted in the direction of their array relative to each other.

The photoelectric conversion layer **16**, the electrodes **12–14**, the glass plate **22**, the optical filter **23**, and the glass face-plate **24** are formed as a unitary target structure with a disc-like configuration. This disc may have any convenient diameter, such as approximately 2.54cm., for example, and is attached to one end of a tube envelope **26**. The tube envelope has a deflection coil **27**, a focusing coil **28**, and an alignment coil **29** mounted on it. Light rays from an object **31** to be televised are focused by a lens **32** on the photoelectric conversion layer **16**.

The electrodes **12–14** are separately supplied with signals **S33–S35**, which are shown in FIGS. 3A–3C, respectively, and are synchronized with the horizontal scanning period and displaced in time relative to one

(b) Specification

Figure 3: Empirical probability distribution function of citation detection as a function of the starting character

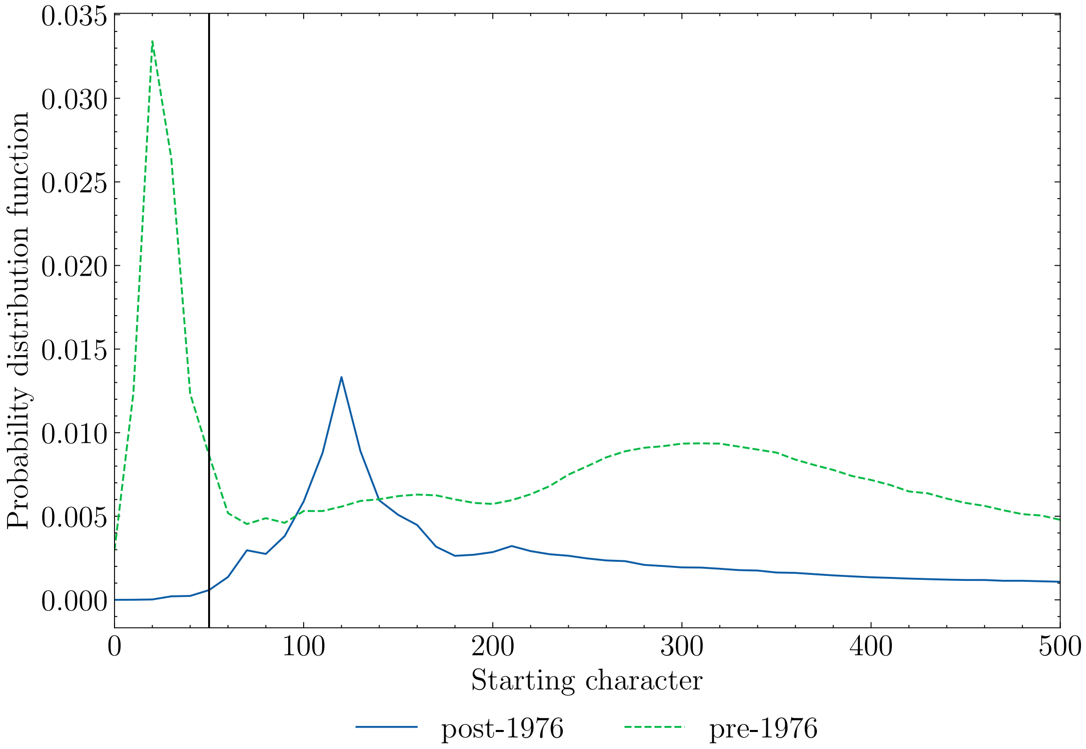


Figure 4: Empirical probability distribution function of citation detection as a function of the relative place of the starting character

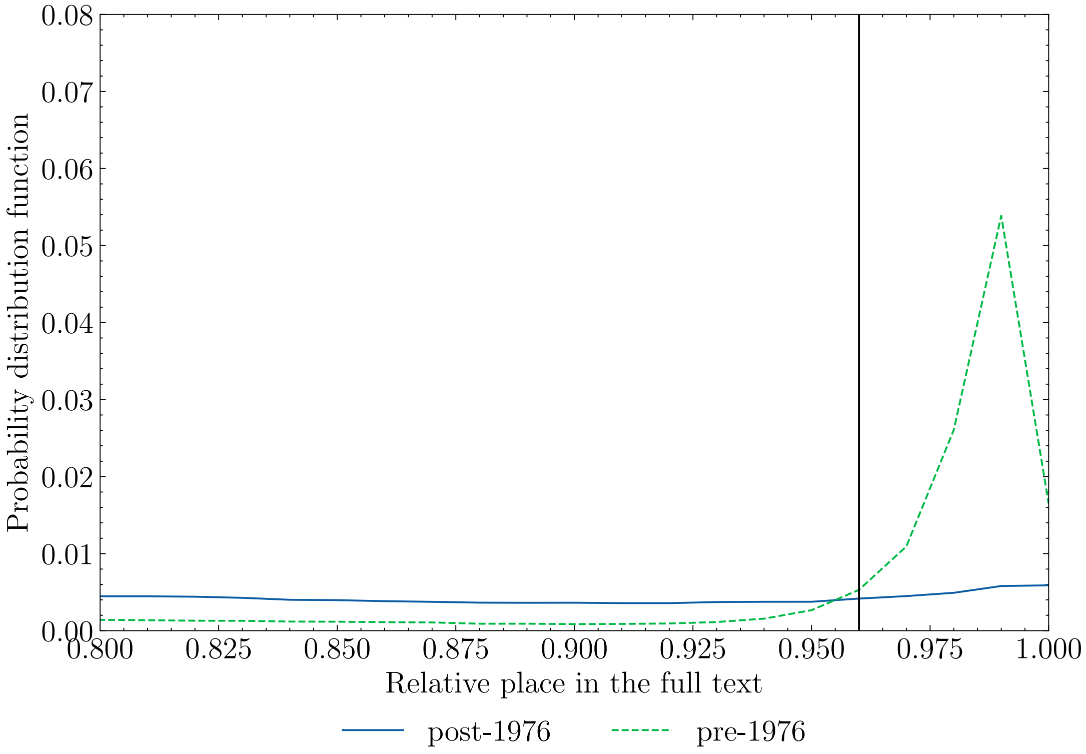
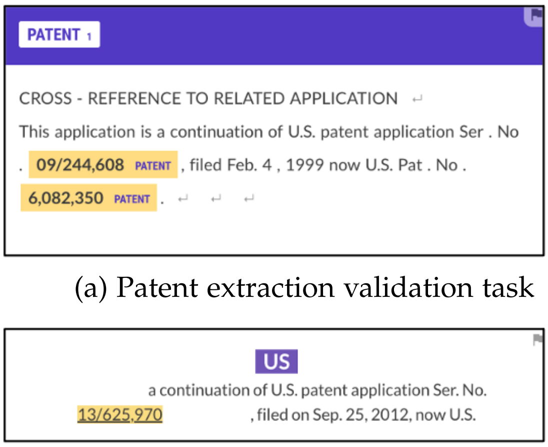


Figure 5: Preview of the annotation platform



CROSS - REFERENCE TO RELATED APPLICATION ↵

This application is a continuation of U.S. patent application Ser . No . 09/244,608 PATENT , filed Feb. 4 , 1999 now U.S. Pat . No . 6,082,350 PATENT . ↵ ↵ ↵

(a) Patent extraction validation task

a continuation of U.S. patent application Ser. No. 13/625,970 , filed on Sep. 25, 2012, now U.S.

(b) Patent parsing validation task (organisation name)

Figure 6: Empirical cumulative distribution function of patents in the validation sample and in the universe of U.S. patents (by decade)

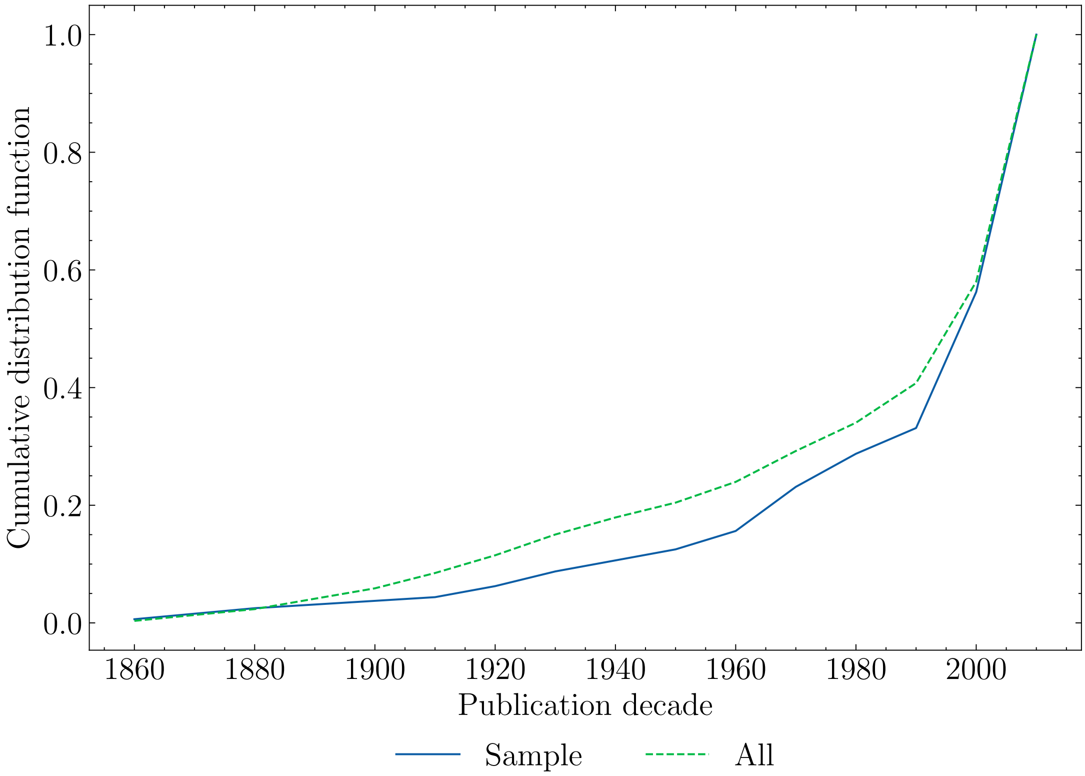
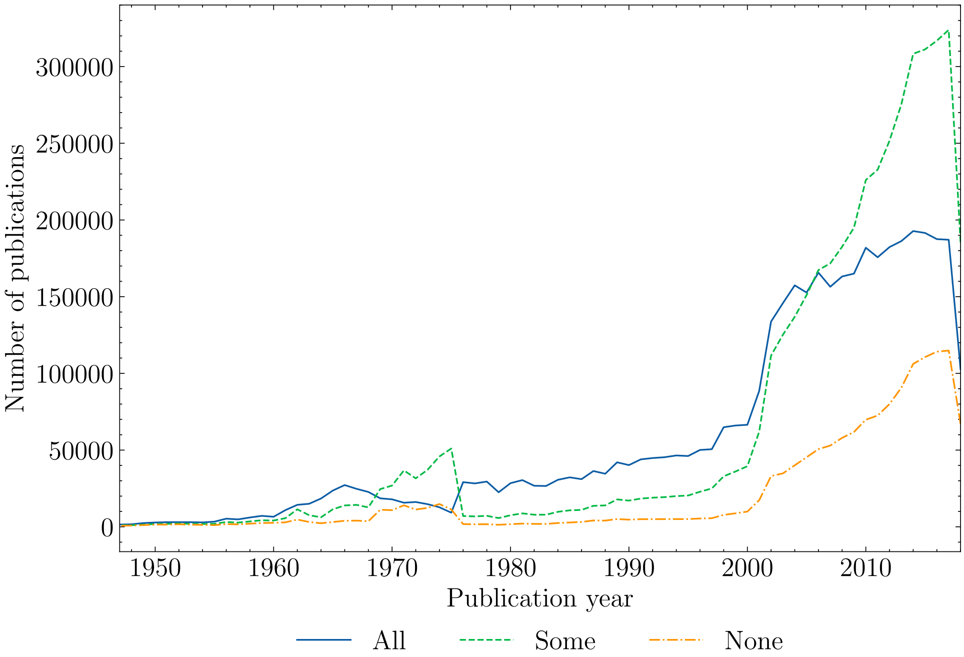
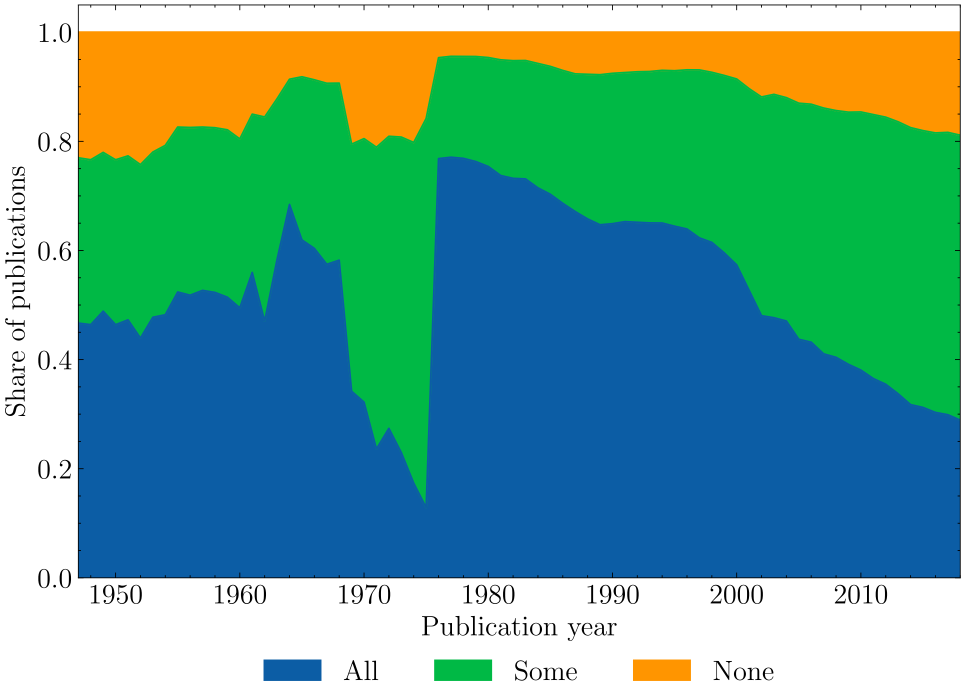


Figure 7: Citing patents over time by in-text citation match status



(a) Number



(b) Share

Notes: "All" (blue solid line) refers to patent publications for which it was possible to match all extracted in-text citations. "Some" (orange dashed line) refers to patent publications for which it was possible to match only some extracted in-text citations. "None" (green dash-dot line) depicts patent publications for which we could not match any extracted in-text citation.

Figure 8: Patent citations by origin

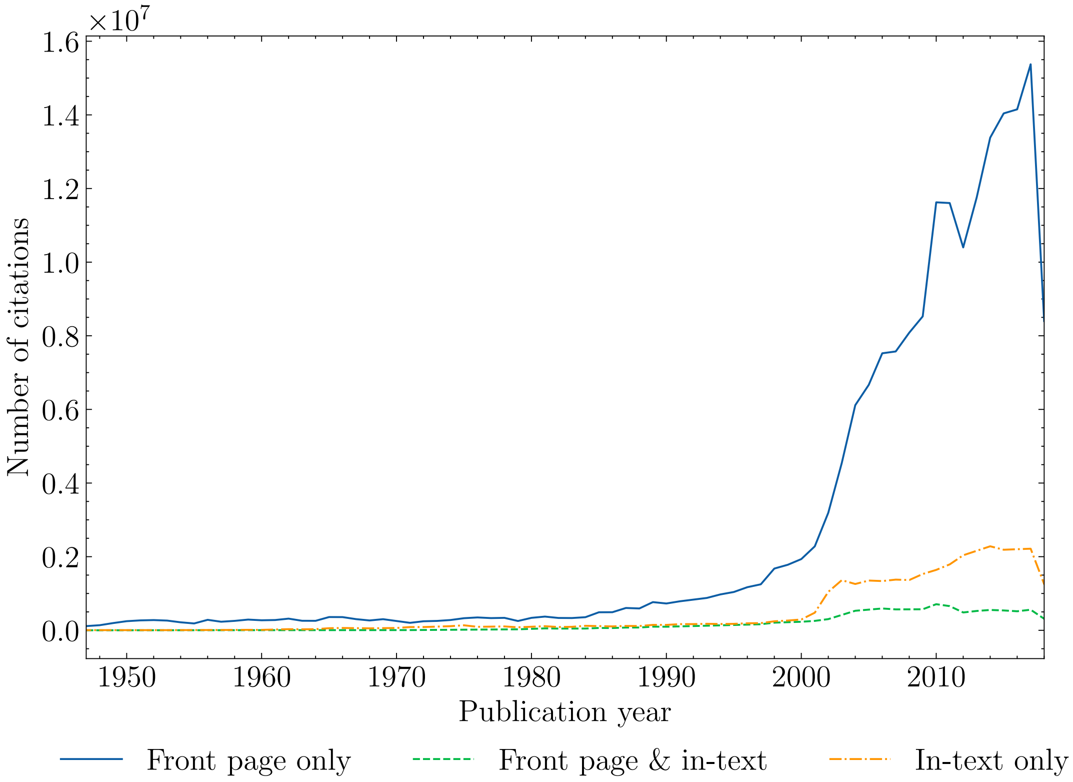
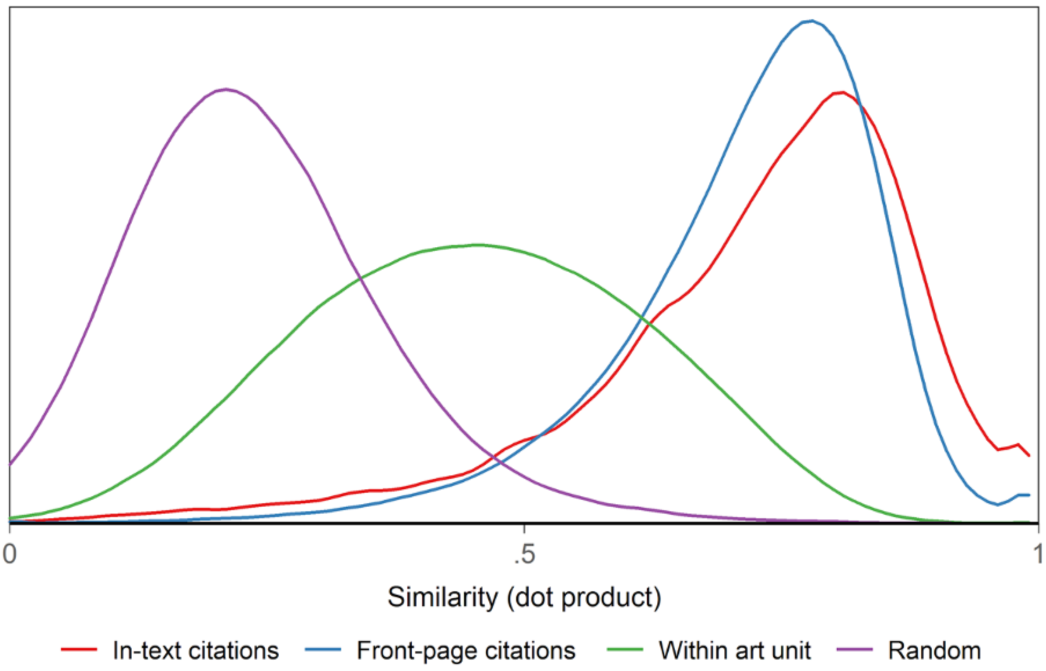
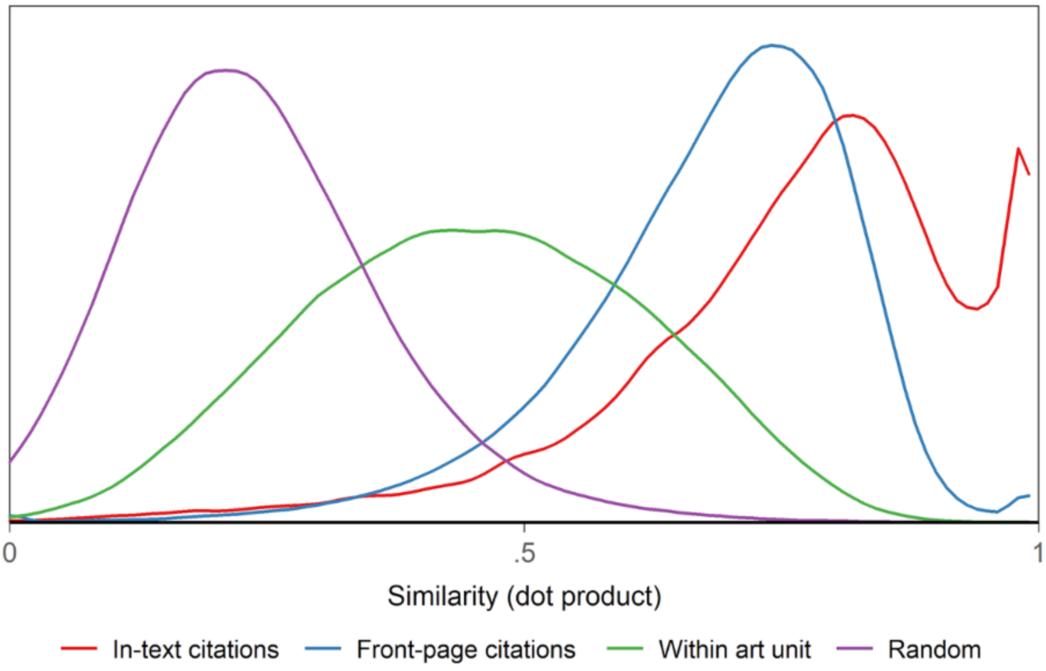


Figure 9: Citing-cited patent pair-wise similarity distribution

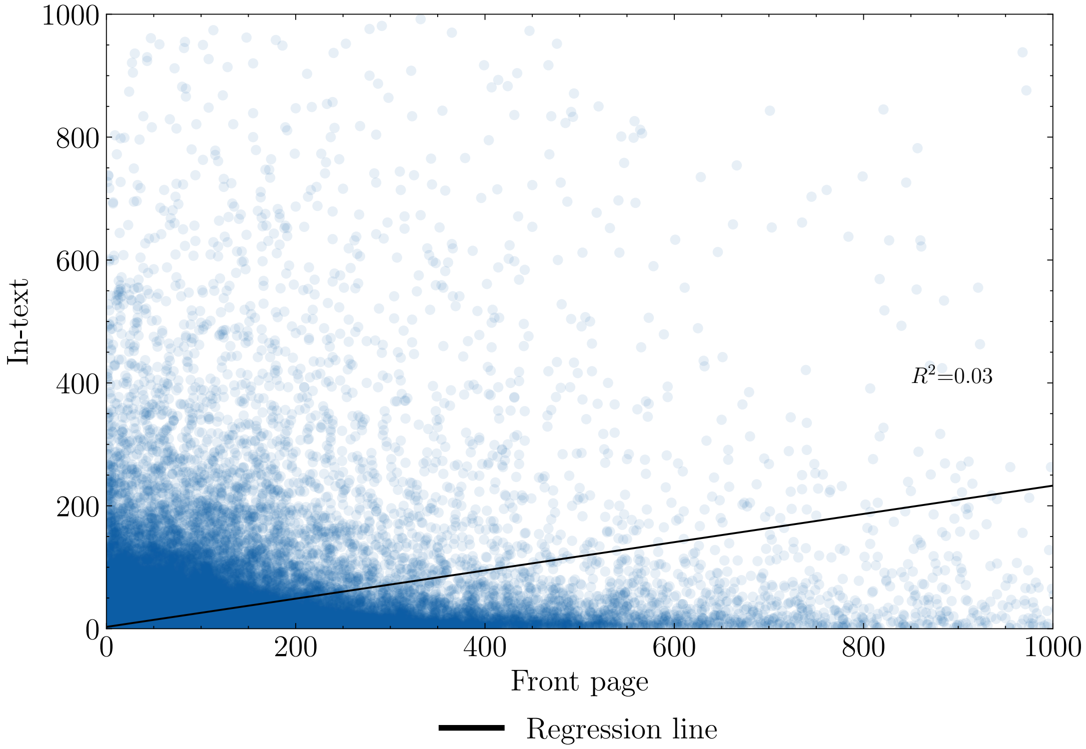


(a) Within-INPADOC-family citations omitted



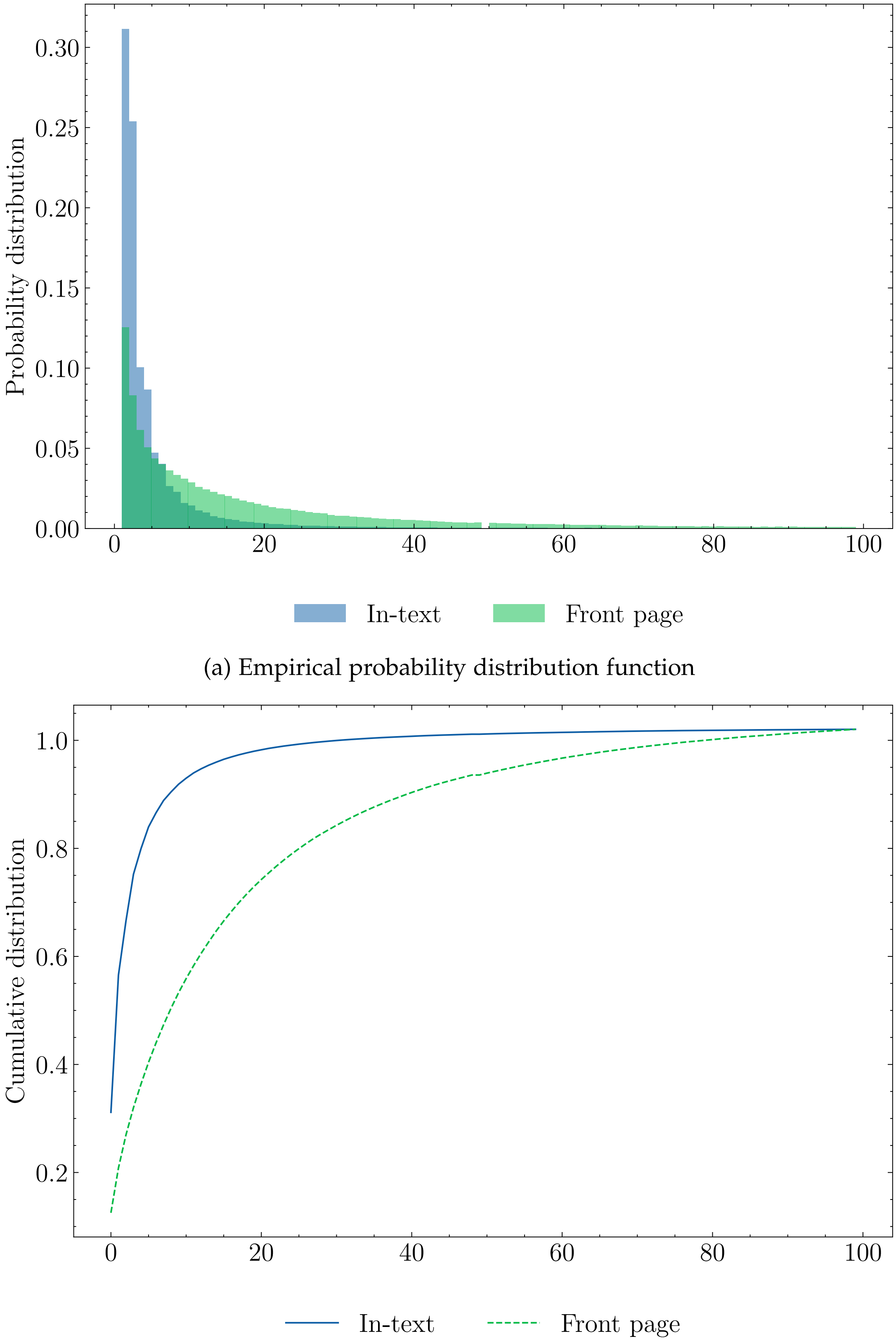
(b) Within-DOCDB-family citations omitted

Figure 10: Forward citations count of invention families: front page citations *versus* in-text citations



Notes: We use a 10 percent random sample of all DOCDB patent families with a positive front page and in-text forward citations count. Each data-point represents a DOCDB patent family. The regressions line corresponds to the following model: $in - text\ forward\ citations\ count = a(front\ page\ forward\ citations\ count) + b$

Figure 11: Empirical distribution of forward citations count

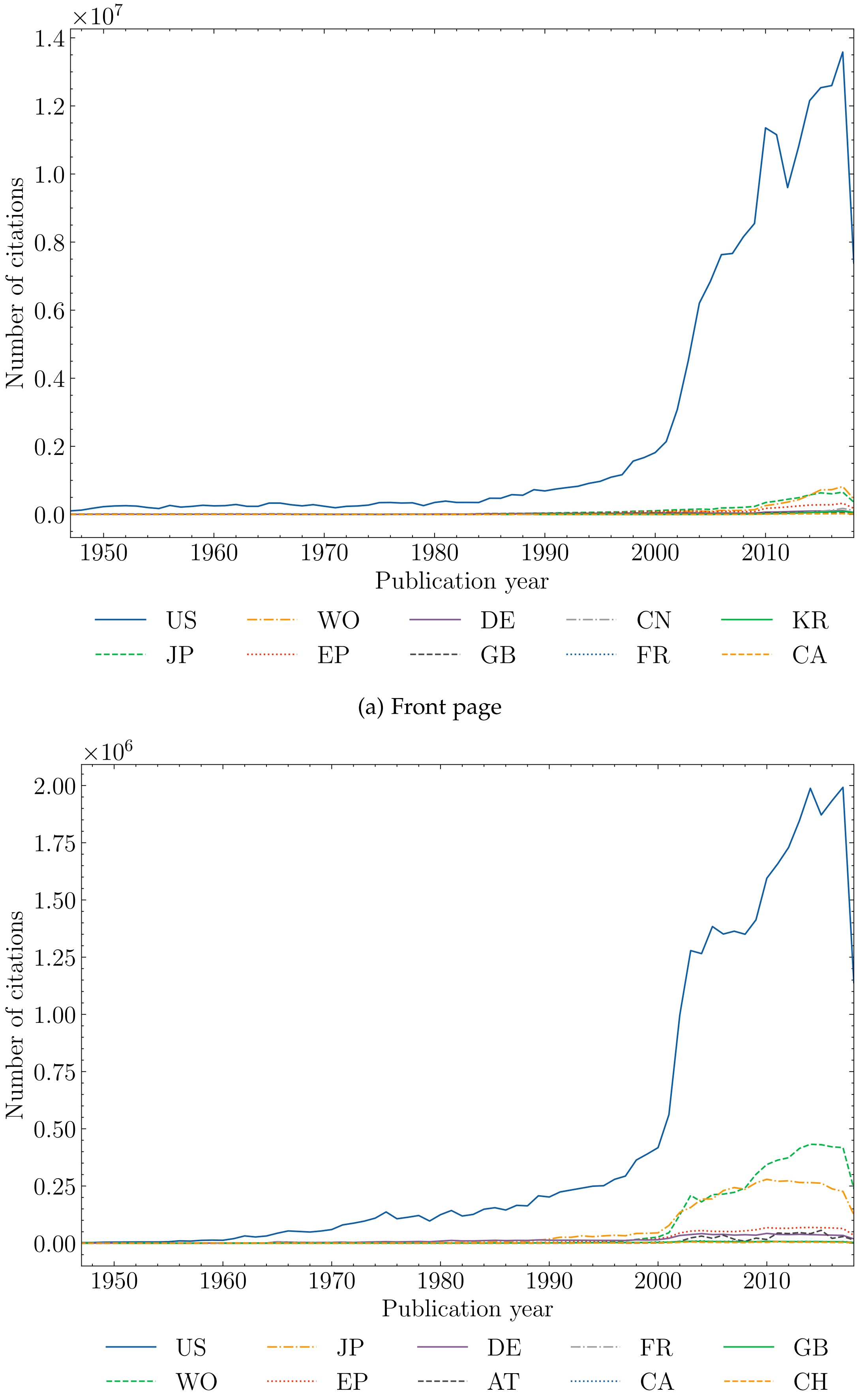


(a) Empirical probability distribution function

(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.

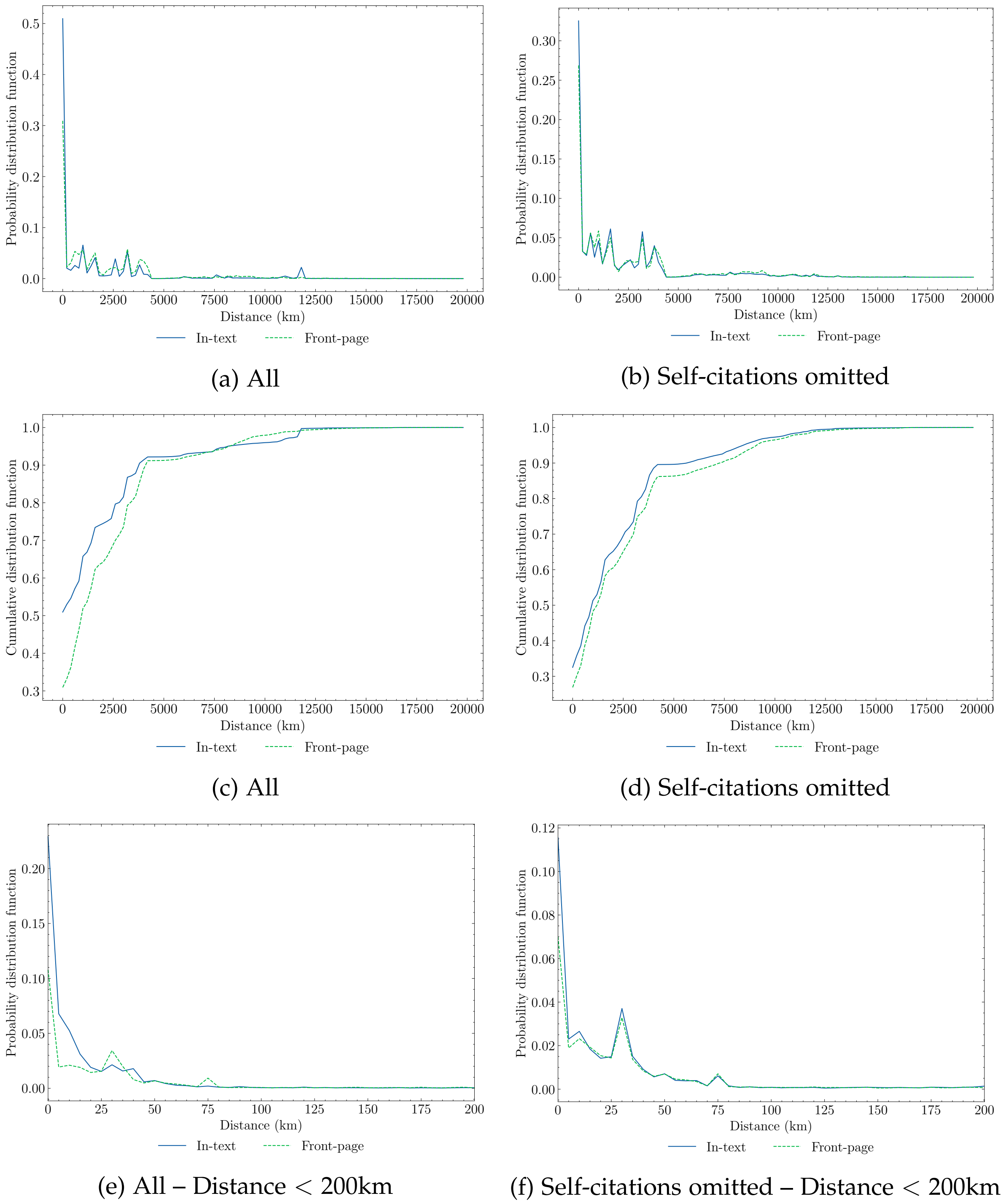
Figure 12: Patent citations by “receiving” country



(a) Front page

(b) In-text

Figure 13: Distribution of citing-cited inventors distance



(a) All

(b) Self-citations omitted

(c) All

(d) Self-citations omitted

(e) All – Distance $<$ 200km

(f) Self-citations omitted – 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. Self citations include within-INPADOC-family citations and same assignee citations. In panel 13a, 13b, 13c and 13d we group observations by 200km bins. In panel 13e and 13f we use 5km bins.

Appendix

A In-text patent citations reasons and examples

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."

B Data record and reproducibility

Data generation and validation reproducibility is guaranteed by the codebase hosted on the project repository. Validation data are supported by Data Version Control (DVC). Since the project is open-source and continuously improving, exact replication of the data and results detailed above requires the user to choose the tag '0.3.1' of the code.[39]

The data are reported as a nested table that is structured as follows:

- Each entry corresponds to the patent document from which we extracted patent citations. Each such patent is identified by a publication number (primary key). In addition to the publication number, we also report its publication date, application identifier, and patent publication identifier. We also include DOCDB and INPADOC family codes, which identify a constellation of inter-related patents that protect the same invention across jurisdictions.

- Each entry has a citation variable in which cited patents are listed and their attributes are nested. Any detected patent is represented by the two attributes parsed by Grobid, the code of its patent office and its original number. When these two attributes can be matched with a publication number, we also report the publication date, application identifier, patent publication identifier and the DOCDB and INPADOC family identifiers. Eventually, we report a flag indicating that the extracted citation is likely to belong to the front matter or the header.

The schema of the table is detailed below.

Name	**Description**	**Type**	**Nb non null**
publication_number	Publication number.	STR	16781144
publication_date	Publication date (yyyymmdd).	INT	15862299

[39] github.com/cverluise/PatCit/tree/0.3.1

Name	**Description**	**Type**	**Nb non null**
appln_id	PATSTAT application identification. Surrogate key: Technical unique identifier without any business meaning	INT	15862299
pat_publn_id	PATSTAT Patent publication identification. Surrogate key for patent publications.	INT	15862299
docdb_family_id	Identifier of a DOCDB simple family. Means that most probably the applications share exactly the same priorities (Paris Convention or technical relation or others).	INT	15862299
inpadoc_family_id	Identifier of an INPADOC extended priority family. Means that the applications share a priority directly or indirectly via a third application.	INT	15862299
citation		REC	16781144
__.country_code	Country code of the cited patent. Parsed by Grobid.	STR	64185636
__.original_number	Original number of the cited patent. Parsed by Grobid.	STR	64185636
__.kind_code	Kind code of the cited patent. Parsed by Grobid.	STR	6096368
__.status	The status of the cited patent. Parsed by Grobid.	STR	64185636
__.pubnum	Concatenation of country code, original number and kind code of the cited patent. Based on attributes parsed attributes.	STR	64185636

Name	Description	Type	Nb non null
__.publication_number	Publication number of the cited patent. Obtained from the google patent linking API.	STR	49542360
__.publication_date	Publication date (yyyymmdd) of the cited patent based on the matched publication_number.	INT	49231609
__.appln_id	PATSTAT application identification of the cited patent. Based on the matched publication_number. Surrogate key: Technical unique identifier without any business meaning.	INT	49231609
__.pat_publn_id	PATSTAT Patent publication identification of the cited patent. Based on the matched publication_number. Surrogate key for patent publications.	INT	49231609
__.docdb_family_id	Identifier of a DOCDB simple family of the cited patent. Based on the matched publication_number. Means that most probably the applications share exactly the same priorities (Paris Convention or technical relation or others).	INT	49231609
__.inpadoc_family_id	Identifier of an INPADOC extended priority family of the cited patent. Based on the matched publication_number. Means that the applications share a priority directly or indirectly via a third application.	STR	49231609

Name	**Description**	**Type**	**Nb non null**
__.flag	Flag detected citations which are likely to be in the header rather than in the specification itself. Flag is True for citations extracted from patents published in the pre-1976 format and with all occurrences detected before character 50 or in the last 4 percent of the text. It is recommended to exclude those citations from most analyses.	BOOL	71407446
__.char_start	First character of the detected cited patent. Refers to description_localized.text in patents-public-data.patents.publications.	INT	71407446
__.char_end	Last character of the detected cited patent. Refers to description_localized.text in patents-public-data.patents.publications.	INT	71407446

Notes: Nested variables are denoted by a dot. For instance, __.country_code is the country code of a cited patent nested in the citation variable.