# Beyond the front page: In-text citations to patents as traces of inventor knowledge

Cyril VERLUISE
Gabriele CRISTELLI
Kyle HIGHAM
Gaétan DE RASSENFOSSE

October 2025

Innovation and Intellectual Property Policy
Working Paper series no. 30

Available at: https://ideas.repec.org/p/iip/wpaper/30.html

EPFL Working Paper Series
STiP lab

# Beyond the front page: In-text citations to patents as traces of inventor knowledge*

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

**Abstract**

This study introduces in-text patent-to-patent citations—references embedded in the body of patent documents—as a novel data source to trace knowledge flows. Unlike front-page citations, which often reflect legal requirements, in-text citations are more likely to originate from inventors and signal meaningful technological linkages. We show that they exhibit stronger geographic and semantic proximity, greater self-referentiality, and closer alignment with inventor knowledge. Though less frequent than front-page citations, they yield robust results in models of knowledge diffusion. We release a validated dataset and reproducible code to support future research. Our findings offer new opportunities for strategy scholars interested in the microfoundations of innovation, the geography of knowledge flows, and the role of inventors in shaping firms' knowledge trajectories.

**Keywords**: citation, patent, knowledge flow, open data, spillover

*We thank the editor and two anonymous reviewers for their constructive feedback. We are also grateful to Philippe Aghion, Antonin Bergeaud, Bronwyn Hall, Adam Jaffe, and Francesco Lissoni for helpful comments, and to Patrice Lopez (WIPO) and Ian Wetherbee (Google) for their technical support. Lucas Violon provided outstanding research assistance. All remaining errors are ours.

†QuantumBlack, Collège de France, and Paris School of Economics

‡London School of Economics and Political Sciences

§Motu Economic and Public Policy Research

¶École polytechnique fédérale de Lausanne, gaetan.derassenfosse@epfl.ch (corresponding author)

# INTRODUCTION

Innovation depends on how effectively firms generate, combine, and deploy knowledge, but the traces that these processes leave behind are faint. Despite its central role, knowledge remains notoriously difficult to observe. In this desert of data, patent data "loom up as a mirage of wonderful plentitude and objectivity" (Griliches, 1990, p. 1661). One prominent use of patent data was proposed by Jaffe et al. (1993), who established patent citations as a proxy for knowledge flows. Since then, a large body of strategic management scholarship has used patent citations to trace how ideas flow across organizations and inventors (*e.g.,* Mowery et al., 1996; Almeida and Kogut, 1999; Rosenkopf and Nerkar, 2001; Singh, 2005; Singh and Marx, 2013).

Although patent citations are often viewed as an appealing proxy for the elusive "paper trail" of knowledge (Krugman, 1991), they suffer from well-documented shortcomings. Chief among them is that inventors exert only limited influence over which references appear on a patent's front page—the traditional source of patent citation data. Front-page citations emerge from a multi-stage and highly mediated process in which the examiner has final say. The USPTO's post-2001 distinction between examiner- and applicant-supplied citations is a welcome advance (Thompson, 2006), yet it only partially resolves the problem. Some examiner citations were initially suggested by the applicant, and not all applicant citations have been provided by inventors, leaving the true origin of the citation ambiguous.

A growing body of evidence confirms that front-page citations to patents only imperfectly capture knowledge flows. Alcácer et al. (2009) and Sampat (2010) show that "applicants" (including inventors) supply only a minority of patent citations and that citation behavior varies widely across firms. Lampe (2012) adds a strategic dimension, documenting deliberate withholding of material prior art—a pattern contested by Kuhn et al. (2023). Fadeev (2024) suggests that front-page patent citations are less about individual inventors' knowledge dissemination and more about the strategic relationships between firms. Adding to the noise, patent attorneys frequently insert citations on behalf of their clients (Jaffe et al., 2000;

Wagner et al., 2014). Moreover, unlike citations in scientific papers, patent references serve a legal purpose rather than a cognitive purpose (Meyer, 2000). Finally, front-page citations are only available for granted patents, leaving out close to half the patent applications that are withdrawn or abandoned (Carley et al., 2015). Together, these factors render front-page patent citations a notoriously noisy proxy for knowledge diffusion (Duguet and MacGarvie, 2005) and, some argue, a metric that may miss such flows altogether (Arora et al., 2018).

The limitations of front-page patent citations invite exploration of alternative indicators of knowledge flows. This paper examines in-text patent-to-patent citations—references that appear within the body of patent documents—as one such alternative. Although datasets containing these "internal" patent citations have been available for some time (Berkes, 2018, pre-1947; Verluise and de Rassenfosse, 2020; Marx and Fuegi, 2022), their distinctiveness and empirical value have not yet been systematically assessed.

Our analysis demonstrates that in-text patent-to-patent citations better reflect inventors' knowledge set than front-page citations. Descriptive evidence shows that in-text citations are (a) more geographically concentrated than front-page references—regardless of whether those references were added by examiners or applicants; (b) textually closer to the patents they cite, indicating tighter thematic proximity; and (c) substantially more self-referential, pointing back to the focal firm's earlier inventions. These patterns suggest that in-text citations are more likely to originate with inventors themselves rather than with patent attorneys, and thus offer a sharper lens on knowledge diffusion. Evidence from a survey of patent attorneys backs this claim: a randomly selected in-text patent-to-patent citation is 19 to 44 percent more likely to originate with the inventor than is a randomly selected front-page "applicant" citation—the USPTO's catch-all label for any reference supplied under the applicant's name, whether by inventors, in-house counsel, or outside attorneys.

In addition to these analyses, we assess the data's practical utility by replicating Balsmeier et al. (2023). This study is one of the most recent and stringent studies of knowledge diffusion, leveraging variation from inventor death. Although in-text patent-to-patent citations are

less numerous than front-page references, their relative scarcity need not be a limitation. The replication reveals that in-text citations yield statistically robust estimates, even when applied in such a restricted empirical setting.

We further systematically validate our in-text patent-to-patent citation dataset. Using an open-source machine-learning parser to locate references within the specification, we measure its precision and recall, document residual error patterns, and release the entire reproducible workflow. Our publicly available code cleans the parser's raw output, disambiguates citation strings, and maps them to standardized patent numbers—giving scholars a transparent, auditable foundation for future research. The dataset covers 49.5 million in-text patent citations matched to close to 8 million unique patent documents (with best coverage starting from 1976).

The rest of the paper is organized as follows. We start by providing background information on in-text patent-to-patent citations, highlighting their main differences with front-page citations. This section explains why inventors are more likely to exert more influence over the selection of in-text citations than over front-page citations. We then provide a statistical analysis of front-page vs. in-text citations. Next, we replicate a recent study using citation data to track knowledge flows. We conclude by offering avenues for future research. The details of the data creation pipeline and validation metrics are provided in the Online Appendix, together with the survey results and additional supporting information.

# THE EPISTEMOLOGY OF IN-TEXT CITATIONS

A U.S. patent document is legally composed of three segments. The front page provides bibliographic data, the abstract, and the official list of cited prior art. It is followed, when required, by drawings that visually support the disclosure. The remainder constitutes the specification, which houses the full narrative of the invention—background, summary, detailed description—and the claims. Because patent offices digitize and release front-page

citations, they have become the default raw material for empirical research. Yet, the specification itself contains additional references embedded in the text, which we refer to as in-text citations. These citations may point to any genre of prior art, notably earlier patents and scientific publications. Our study concentrates on the subset of references to the patent literature, noting that in-text citations to scholarly literature have been examined elsewhere (*e.g.,* Bryan et al., 2020; Marx and Fuegi, 2022).

In-text citations are inserted to satisfy the statutory patentability criteria in U.S. patent law. Applicants use them to (i) differentiate the invention from prior art to argue novelty (35 U.S. Code §102) and non-obviousness (§103), (ii) demonstrate enablement under §112 by guiding readers to supporting technical detail, and (iii) illustrate utility (§101) through concrete applications (see, *e.g.*, Barton, 2003; Feit, 2011). Because these rhetorical functions overlap only partly with the examination-oriented purpose of front-page references, in-text citations plausibly carry information not reflected in front-page citations. Further, we contend that this incremental signal is shaped by the inventor's input during specification drafting, making in-text patent citations a promising indicator of underlying knowledge flows (Bryan et al., 2020).

## A legal perspective on in-text patent citations

The justifications for adding in-text citations listed above map directly onto the statutory requirements that an application must satisfy to be patentable. While novelty and non-obviousness are primarily assessed by the examiner through direct comparison with prior art, enablement and utility are mainly argued by the applicant within the specification. Table 1 illustrates the variety of in-text citations and their (often explicit) legal purposes.

[INSERT TABLE 1 HERE]

**Novelty and non-obviousness.** Applicants disclose the prior art they know by filing one

or more Information-Disclosure Statements (IDSs), and most of these references later appear on the patent's front page.[1] That front-page list, however, does not include all citations in the IDSs, nor is it limited to those in the IDSs. Examiners run their own searches and routinely add references that were never in an IDS, marking them "cited by examiner." Conversely, an applicant-supplied reference may be retained yet re-labelled examiner-cited once the examiner confirms its relevance. The resulting front-page catalogue, therefore, blends multiple sources of prior art and only imperfectly represents the applicant's original knowledge set—even for the so-called "applicant" citations.

Applicants may also advance novelty and non-obviousness within the specification itself by contrasting the invention with specific prior patents. In that case, the same reference typically appears both in the text and on the front page. Note that some patent lawyers simply add any in-text citations onto the IDS just in case any are considered material to patentability. When this occurs, front-page citations are a superset of in-text citations.

**Enablement.** Section 112 is the legal requirement that a patent's specification describe the invention clearly and completely enough for any skilled practitioner in the field to make and use it without undue experimentation. Applicants often streamline that task by incorporating earlier patents by reference, citing them in the specification. If these citations are not relevant to novelty or non-obviousness, applicants are not required to disclose them via an IDS. Therefore, 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).

**Utility.** The invention must be "new and useful" to be patentable. The first part

[1] See 37 CFR §1.56 and 37 CFR §1.97.

of this clause is covered by the novelty and non-obviousness requirements described above. The second, usually referred to as the "utility" requirement, is particularly open to interpretation, but generally requires the patented invention to work (Machin, 1999). 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). While there is no burden on the applicant to prove that the invention works (Cotropia, 2009), applicants sometimes cite prior patents or scientific work to demonstrate that the claimed function is physically achievable. These references, like enablement citations, are more likely to remain confined to the specification.

## In-text patent citations are more likely to originate from inventors

Because in-text citations arise for a broader set of reasons than those that populate the IDS, they capture portions of the inventive knowledge base that never surface on the front page, particularly those items deemed necessary to meet the enablement or usefulness requirement.[2] Although in-text citations fulfill legal objectives that differ from scholarly attribution, inventor influence is nevertheless likely to be stronger here than on the front page, for two reasons.

First, the in-text citations that are duplicated on the front page, as prior art material to patentability, are likely the most relevant pieces of prior art against which the invention needs to be judged as novel and non-obvious. The fact that these citations are also in the patent description would imply that they either fulfilled multiple requirements, or were so technologically close to the citing patent that applicants need to make explicit arguments for novelty in the description with reference to specific items in the prior art. In either case, the inventor was likely aware of this art—or worked closely with counsel to frame the necessary technical distinctions during drafting.

Second, citations that remain only in the text likely serve utility or enablement functions.

[2]See Manual of Patent Examining Procedure, Sections 2164 and 2107.02.

The enablement requirement states that a hypothetical "person skilled in the art" should be able to make and use the invention, and applicants add in-text citations to assist these hypothetical persons. Consequently, this information was almost certainly necessary during the invention process, and the inventors were, therefore, aware of it. It is difficult to imagine attorneys alone supplying such technical scaffolding without substantial input from the inventors themselves.

Both of these arguments imply that inventors are more likely to exert more influence over the selection of in-text citations than over front-page citations. Accordingly, we propose that in-text patent-to-patent citations constitute a promising indicator of knowledge flows.

# SEARCHING FOR TRACES OF INVENTOR'S INPUT

This section compares in-text and front-page patent citations along geographic, semantic, and bibliographic dimensions and summarizes survey evidence on citation provenance. The results presented in this section collectively suggest that in-text citations to patents provide a stronger signal of knowledge flows than front-page citations. We begin with broad descriptive statistics for each citation type.

## Descriptive statistics

Table 2 describes the full sample of 16,781,144 U.S. patents and patent applications in our dataset published between the first patent grant in 1790 and August 2019.[3] Of these, 7,869,894 (about 46.90%) contain at least one patent citation in the body of the specification, whereas 76.79 percent make at least one front-page patent citation.

In total, the full sample contains 49,542,360 in-text patent-to-patent citations, roughly

[3]The data are available for both granted patents and patent applications. They can be accessed at https://doi.org/10.5281/zenodo.3710993. Additional information is provided on the project website: https://cverluise.github.io/PatCit/.

one-fifth of the 265,659,106 front-page citations made by the same patents. Although the distribution of the number of in-text citations per patent is highly skewed, the unconditional mean is 2.95 in-text citations per patent, rising to 6.29 among patents that make at least one citation. For completeness, note that the 49.5 million figure relates to in-text citations that we were able to match to a DOCDB publication number. As further explained in the Online Appendix, we identified over 60 million traces of in-text patent citations, and we were able to associate 49.5 million of these with a standardized patent number.

The (unconditional) number of in-text patent-to-patent citations is of the same order of magnitude as that of in-text patent-to-article citations (3.51 in Bryan et al., 2020, Table 1). However, front-page patent-to-patent citations are significantly more numerous than front-page patent-to-article citations (4.60 in Bryan et al., 2020, Table 1 vs. 15.83 in our Table 2).

[INSERT TABLE 2 HERE]

We next examine the overlap between in-text and front-page citations. Figure 1 reports the proportion of citing patents by citation type. Before 1947, front-page citations did not exist; only in-text citations were available. However, fewer than five percent of patents from that era include in-text citations. In more recent decades, the share of patents with at least one in-text citation stabilizes around 60 percent, with most of these patents containing both in-text and front-page citations.

A notable shift occurs after November 2000, when the USPTO began publishing patent applications. This change allows researchers to observe the specification—including in-text citations—even for applications that are ultimately abandoned. As a result, the proportion of patent documents featuring only in-text citations increases during this later period.

[INSERT FIGURE 1 HERE]

Turning now to overlap in terms of citing–cited patent pairs, we assign each pair to one of three mutually exclusive sets: citation pairs found only in the text, only on the front page, or in both locations. We observe 11,799,723 pairs appearing in both places, accounting for just 5.79 percent of all front-page citations but 25.83 percent of all in-text citations. We observe that 33,883,406 in-text citations never appear on the front page, accounting for 14.25 percent of all 237,619,091 patent-to-patent citations recorded (not reported).

## In-text citations are more geographically concentrated than front-page citations

In this section, we compare the geographic properties of in-text and front-page citations. We begin by plotting the geographic distance between all pairs of citing and cited inventors' geocoded addresses using data from de Rassenfosse et al. (2019) for both in-text and front-page citations. We consider citing patents granted between 1980 and 2010 and exclude self-citations at the INPADOC family level.[4] Figure 2 shows the probability distribution functions of the distance, in kilometers, between citing and cited inventor dyads for the entire sample (panel A) and for citation pairs within 200 kilometers (panel B). Both graphs portray in-text citations as slightly more localized than those on the front page. Panel B, in particular, shows a higher share of in-text citations within a 50-kilometer distance.

[INSERT FIGURE 2 HERE]

We further compare in-text citations with two subgroups of front-page citations: those added by the applicant and those added by the examiner. The data for these comparisons are available for patents granted since 2001. Applicant front-page citations have been argued

[4]INPADOC families group together all documents that share at least one priority filing, either directly or indirectly (*e.g.*, via a third document). This "extended" definition contrasts with the narrower DOCDB definition (Martínez, 2011)

to be a less biased proxy of knowledge flows than examiner ones (Jaffe et al., 2000; Alcacer and Gittelman, 2006). As in Marx and Fuegi's (2022) study of patent-to-article citations, we adopt Thompson's (2006) approach, regressing a measure of geographic distance between citing and cited patents on a citation category indicator and citing patent fixed effects.

Table 3 reports our results. The sample includes patents granted between 2001 and 2010, and we continue to exclude self-citations at the INPADOC family level. In line with Thompson (2006), we find that applicant front-page citations are more localized than those added by the examiner. This result holds for both a coarse outcome, the probability of citing a patent originating from the same country (column 1, panel A), and a fine-grained outcome, the logarithmic transformation of the distance between citing and cited patents (column 1, panel B).

Comparing examiner and in-text citations (column 2), we find that in-text citations are also more localized than examiner front-page ones, and to a much greater extent than applicant front-page citations. Column (2) in panel A shows that in-text citations are, on average, 7 percentage points more likely to connect patents from the same country than examiner citations, whereas applicant citations are only about 2 percentage points more likely to do so. Column (2) in panel B indicates that in-text citations are, on average, approximately 67 percent closer than examiner front-page citations, compared to a difference of about 14 percent between applicant and examiner front-page citations.

Column 3 shows that in-text citations are also significantly more localized than applicant front-page citations, although the coefficients are, as expected, slightly smaller than those estimated in comparisons with examiner citations. The economically and statistically significant greater localization of in-text citations persists when we restrict our sample to citation pairs within 200 kilometers (panel C) or focus only on citation pairs within the United States (panel D).[5]

Our results are also robust to the use of restricted samples that exclude citations older

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

than ten years (column 4), patents with more than 100 front-page citations (column 5), and applicant self-citations (column 6).[6] The persistence of in-text citations' greater localization when we exclude applicant self-citations is particularly noteworthy, as it suggests that in-text citations may better capture technological knowledge flows not only within but also between organizations.

Our results consistently indicate that patent-to-patent in-text citations are decisively more localized than front-page ones. Just as the greater localization of applicant front-page citations relative to examiner ones may reflect the presence of fewer references unknown to the inventors, the strong geographic concentration of in-text citations—around the places where inventors work and live, and likely source a large share of their knowledge—suggests that in-text citations are a cleaner proxy for inventors' knowledge than front-page citations.

[INSERT TABLE 3 HERE]

## In-text citations are textually more similar than front-page citations

To further assess the distinct nature of in-text patent-to-patent citations, we examine the textual similarity between citing and cited patents. Prior research has established that text-based similarity is a useful proxy for technological relatedness (*e.g.*, Younge and Kuhn, 2016; Arts et al., 2021). In this context, a higher average similarity for in-text citations compared to front-page ones would suggest that they tend to connect more closely related inventions. Such a result would be consistent with the idea that in-text citations are more grounded in the substantive content of the invention and, therefore, more likely to reflect meaningful knowledge linkages.

We calculate the semantic similarity for a given patent pair as the dot product of Google

[6] The results in columns (4)–(6) indicate a smaller difference between in-text and front-page citations than our baseline estimates, and a larger one between applicant and examiner front-page citations. Nevertheless, in-text citations remain substantially more localized than both groups of front-page citations.

Patents' document embedding vectors, which are made available to researchers through the Google Patents Public Datasets.[7] The embeddings are trained to predict CPC categories from each patent's full text using a WSABIE algorithm (Weston et al., 2010).

Table 2 shows that the median pairwise similarity between patents cited on the front page and the citing patent is 0.71. In contrast, the median similarity between patents cited in the specification and the citing patent is 0.80. Higher values indicate greater similarity, offering *prima facie* evidence that in-text citations connect conceptually closer patents than front-page citations.

Figure 3 plots the pairwise similarity distributions for in-text and front-page citations, alongside two reference distributions. The first reference distribution ("Within art unit") is based on the similarity between randomly chosen pairs of patents examined by the same USPTO art unit.[8] The second reference distribution ("Random") is based on the similarity between in-text cited patents matched to a random citing patent. Specifically, we construct this set by randomly reassigning the cited patent in each in-text citation to another patent from the in-text citation pool, preserving the original set of citing and cited patents but randomly reconfiguring the citation links.

To ensure consistency, we restrict the analysis to citing and cited patents granted by the USPTO between 2000 and 2009 and exclude all within-INPADOC-family citations (N = 325,247). Self-family citations are much more frequent in the patent text than on the front page (as discussed in the next section); removing them improves the comparability of the similarity distributions. Pairs of patents used for the "Within art unit" and "Random" distributions are randomly sampled to match the sample size.

[INSERT FIGURE 3 HERE]

[7] See https://tinyurl.com/googlepatentdata. A variety of similarity measures exist (see Ganguli et al., 2024, for a review); however, for the present work, we required a low-dimensional vector form that could quickly and intuitively estimate the semantic distance between a large set of patents. Google Patents' embeddings are perfect for this purpose.

[8] An art unit is a group of patent examiners organized around specific technology areas. There are more than 600 art units at the USPTO.

We draw two main observations from the graphical comparison. First, the modal peak of the in-text citation distribution is shifted toward higher similarity values compared to the front-page citation distribution. This shift indicates that patents cited in-text are, on average, more similar to the citing patent than those cited on the front page. Such a pattern is consistent with the idea that inventors reference closely related prior art in the specification—often to distinguish their invention from it or explain its relevance, as explained above. This pattern is also consistent with applicants (and attorneys) erring on the side of caution by over-disclosing known prior art on IDSs and may also reflect a deliberate attempt to overwhelm the patent office to hide actual relevant references (Taylor, 2012; Bryan et al., 2020).

Second, the in-text citation distribution displays a heavier left tail, with more citations at lower similarity levels, particularly around the level observed among patents examined by the same art unit. This is expected: in-text citations are not constrained by legal relevance for patentability and may include prior art from a broader, yet still related, technological space. Importantly, Figure 3 suggests that these citations constitute a small minority of in-text citations, at least to the extent that semantic similarity can measure this kind of relationship.

Taken together, these findings reinforce the view that in-text citations are more tightly coupled to the technical content of the citing invention than front-page citations. While not immune to noise, they appear to offer a more selective and content-driven signal of knowledge connection, supporting their potential as a valuable proxy for tracing knowledge flows.

## In-text citations are more self-referential than front-page citations

To further understand the origins and informational content of in-text citations, we examine their degree of self-referentiality. A higher prevalence of self-citations among in-text references, compared to front-page ones, may reinforce the view that they are more tightly linked to the knowledge base actively used by inventors, rather than serving purely legal or

administrative functions.

We consider two forms of self-citations: family-level self-citations and applicant-level self-citations. To assess family-level self-citations, we map each citing and cited patent to its corresponding INPADOC patent family and compute the share of citations in which the cited patent belongs to the same family as the citing one.[9] We find that 10.51 percent of in-text citations are family self-citations, whereas the corresponding value for front-page citations is 1.63 percent (see Table 2). This source of information is not particularly informative—the patent is usually citing an earlier version of itself, typically in the "related applications" section of the description. From a practical standpoint, these citations are easy to identify and omit. Importantly, they represent only a small share of in-text citations overall.

Turning to applicant-level self-citations, we compute the share of citations in which the citing and cited patents share at least one inventor or one assignee. We rely on harmonized names provided in the Google Patents dataset, labeling a citation as "same-applicant" when at least one inventor or assignee name is shared between the citing and cited patents. This approach casts a wide net: inventors may move between firms, and the name harmonization process is more likely to merge distinct entities (due to common names or errors) than to split identical ones. As a result, our estimates likely overstate the true number of applicant-level self-citations. However, this issue affects front-page and in-text citations equally, limiting concerns about its impact on the comparative analysis.

The differences we observe are substantial. Among in-text citations, 17.43 percent share at least one inventor and 22.46 percent share at least one assignee with the citing patent (see Table 2). For front-page citations, the corresponding figures are 5.98 percent and 9.26 percent. These findings highlight the importance of self-reliance in knowledge creation and suggest that in-text citations more prominently capture this phenomenon. More broadly, they provide further evidence that in-text citations more directly reflect the knowledge set of inventors than do front-page citations.

[9]Performing the analysis on the DOCDB family leads to substantially similar conclusions.

Self-citations are a foundational tool for measuring knowledge flows in strategic management research, being commonly used to capture knowledge reuse (*e.g.,* Berry, 2014; Melero et al., 2020) or to operationalize constructs such as generative appropriability (Ahuja et al., 2013; Argyres et al., 2025). Scholars have employed variations of self-citation-based metrics to examine knowledge transfer across a firm's divisions (Miller et al., 2007), knowledge "inheritance" from prior employers (Chatterji, 2009), and the "specificity" of a firm's knowledge base (Wang et al., 2009, 2016), among other applications. Our findings call for greater attention to in-text citations as a lens for analyzing intra-organizational knowledge dynamics.

## Patent attorneys believe that in-text citations are more likely to originate from inventors

To complement our empirical analyses, we conducted a survey of U.S. patent attorneys to gain insight into the likely origin of citations found in patent applications.[10] Attorneys are uniquely positioned to answer this question: they directly observe the references they add and those that the applicant supplied. In the latter case, they may also directly exchange information with inventors, allowing them to observe the entire chain of events.

The survey focused on two types of citations: those listed on the IDS (from which "applicant" citations are drawn) and those embedded in the body of the specification (in-text). Respondents were asked to estimate how often each type of citation is supplied by the applicant (vs. themselves) and, among those supplied by applicants, how frequently the inventor is the source.

The results suggest that in-text citations are significantly more likely than front-page citations to originate with inventors. According to respondents, approximately 29–38 percent of in-text patent citations are inventor-supplied, compared to 20–29 percent for citations listed in the IDS. This implies that a randomly selected in-text citation to a patent is between

[10] We refer readers to Section D of the Online Appendix for a complete account of the survey methodology, sample characteristics, and detailed estimation procedures.

19 and 44 percent more likely to have been provided by the inventor than a comparable front-page applicant citation.

The survey also asked about scientific references. Respondents believe that inventors supply around 54–57 percent of in-text citations to scientific papers, compared to 50–57 percent for IDS-listed references. We further estimate that a randomly selected in-text citation to a scientific article is between 10 and 17 percent more likely to have been provided by the inventor than a comparable front-page applicant citation. The strength of the "inventor signal" is weaker for in-text articles than for in-text patents relative to front-page references. However, according to survey respondents, in-text and front-page scientific references are more likely to be suggested by the inventor than patent references.

These findings provide additional evidence that in-text patent-to-patent citations better reflect the knowledge set that inventors actively draw upon during the invention process, compared to front-page patent-to-patent citations. While attorneys often add references to comply with patentability standards, citations embedded in the technical narrative are more likely to be anchored in the inventor's own understanding of the relevant prior art.

# DO IN-TEXT CITATIONS PROVIDE A STRONG ENOUGH SIGNAL OF KNOWLEDGE FLOWS?

The previous section established that in-text citations are more likely to originate from inventors than front-page citations. Accordingly, they provide a clearer signal of knowledge flows than the citation data scholars have traditionally relied upon. A potential limitation, however, is their relative sparseness compared to front-page citations, as shown in Table 2. The signal may be cleaner, but is it strong enough to be practically useful? For narrowly focused studies that aim to rigorously assess the use of citations as indicators of knowledge diffusion, the answer might well be no. To explore this possibility, we replicate one of the most recent and methodologically demanding studies in this area: Balsmeier et al. (2023).

## Empirical approach

A large body of research on the localization of knowledge spillovers builds on the matching approach developed by Jaffe et al. (1993). Starting from a given sample of patents, this approach involves matching patents that cite the focal sample with patents that are similar in application date and technology field, and estimating differences in the two citing populations' geographic locations. While very influential, this approach has been subject to a number of criticisms, as extensively discussed by Thompson and Fox-Kean (2005).

Balsmeier et al. (2023) propose an alternative identification strategy based on co-invented patents in which one of the inventors died after the application date but before the patent was granted. They identify 1,621 such patents filed between 1976 and 2005, drawn from a universe of approximately 3 million granted patents over that period. These patents involve 1,621 deceased inventors and 3,870 surviving co-inventors, all located in geographically dispersed areas. Although the resulting sample is highly selective, the setting enables a natural experiment: because the deceased and surviving inventors contributed to the same invention, their relative influence on subsequent citations can be cleanly compared.

To measure local knowledge spillovers, the authors track the geographic distance between the hometowns of inventors on citing patents and those of the cited inventors—both deceased and surviving. They then compare the volume of citations originating from within concentric distance bands around each inventor's hometown. Because the cited invention is held constant across all observations, any systematic reduction in citations near the deceased inventor's location—relative to that of the surviving co-inventors—can be attributed to the loss of that individual's contribution to knowledge diffusion.

## Results

Using the original replication code, we recover 34,586 front-page citations (slightly fewer than the 34,749 reported in the original study), corresponding to 28,398 citing patents and 1,621 cited patents. In contrast, we identify 6,360 in-text citations, representing 5,605 citing

patents and 988 cited patents. Despite the smaller sample, there is no evidence of bias in the timing of citations: the distribution of citation ages is nearly identical across front-page and in-text citations (not reported).

However, the spatial distribution of citations differs substantially. In-text citations are notably more localized. For instance, while the original study reports that 15 percent of front-page citations occur within 10 miles of the cited inventor, our replication using in-text citations yields a corresponding figure of 22 percent. This eight-percentage-point difference persists across distances under 150 miles (not reported). Given our earlier findings on the geographic concentration of in-text citations, this result does not come as a surprise.

Table 4 compares the original results from Balsmeier et al. (2023, Table 2) with our replications using both front-page and in-text citations. Our replication with front-page citations closely matches the original estimates, though not exactly.

[INSERT TABLE 4 HERE]

For the full sample, front-page citations fall by 15–25 percent within a 30-mile radius of the deceased inventor's location, with the effect diminishing to approximately 6.5 percent at 60 miles before losing statistical significance. In contrast, in-text citations fall by 9–15 percent within 30 miles, tapering to about 5.5 percent at 60 miles, where they also lose significance.

In the long-distance sample, in which the deceased inventor lived at least 500 miles from all surviving co-inventors, front-page citations fall by 63–76 percent within 30 miles, and by 39 percent at 150 miles (remaining statistically significant). In-text citations in this group fall by 64–73 percent within 30 miles, dropping to 22 percent at 150 miles, a result that is no longer statistically significant.

Overall, the findings are similar across front-page and in-text citations, with two notable differences. First, the drop in in-text citations within 20 miles in the main sample

is smaller than the corresponding drop in front-page citations. This is consistent with the idea that applicant-submitted front-page citations often include a broader set of useful but non-essential references, which may be more marginal to the inventive contribution. When a co-inventor passes away, the more peripheral front-page citations—which may rely more on informal local pointers—drop more sharply.

Second, in-text citations in the large distance sample exhibit a smaller drop beyond the 100-mile mark compared to front-page citations. The original study finds a significant long-distance effect for front-page citations, which is somewhat surprising. The lack of a comparable effect for in-text citations aligns more closely with prior expectations about the localized nature of knowledge spillovers.

# CONCLUDING REMARKS

This study introduces and analyzes in-text patent-to-patent citations as a promising data source for examining knowledge flows. By distinguishing these citations from traditional front-page references, we document that in-text citations carry a stronger inventor-driven signal of technological influence. This finding suggests that in-text citations can serve as a valuable tool for scholars seeking to understand the microfoundations of innovation (*e.g.*, Grigoriou and Rothaermel, 2014; Felin et al., 2015), intra-organizational knowledge dynamics (*e.g.,* Miller et al., 2007; Wang et al., 2009) and the strategic behavior of firms in the patenting process (*e.g.*, Ceccagnoli, 2009; Somaya, 2012). We highlight five broad avenues for future research.

First, in-text citations are available for abandoned (and published) applications— a feature that front-page citations lack. Accordingly, they open new possibilities for studying, for example, the antecedents of invention novelty, and lack thereof (*e.g.*, Jung and Lee, 2016; Chen et al., 2021), and the decision-making processes behind patent abandonment (Somaya, 2012). Furthermore, these data provide a window into applicant citations before 2000. All

studies from before this time can now be reassessed using this information, particularly if they focus on historical contexts such as the impact of the Bayh-Dole Act (*e.g.,* Sampat, 2006) or the introduction of software patents (*e.g.,* Hall and MacGarvie, 2010).

Second, researchers can leverage the fact that in-text patent-to-patent citations offer a cleaner lens on knowledge flows than traditional front-page references. Because earlier studies relied on noisier data, many promising research efforts may have succumbed to the well-known "file-drawer problem," where null or ambiguous results remain unpublished (Rosenthal, 1979), potentially distorting our understanding of knowledge diffusion. Our findings suggest that some of these shelved analyses—or hypotheses not supported by the data—may yield significant effects if re-estimated using in-text citations. Therefore, we encourage scholars to revisit abandoned paths and capitalize on this sharper signal to strengthen the empirical foundations of innovation research.

Third, while this study has focused on citations as proxies for knowledge flows, patent citations have long been used to measure a broader range of constructs, including most notably patent "importance" (Jaffe and de Rassenfosse, 2017). Definitions of "importance" vary across contexts (capturing economic value, technical merit, or legal robustness), and scholars readily acknowledge the limitations of existing patent metrics. Yet, the absence of viable alternatives has led to an over-reliance on front-page citations. In-text citations could be a helpful complement. A systematic exploration of in-text patent-to-patent citations as indicators of patent impact is a promising line of inquiry, enabling improvements in the precision of long-standing measures such as patent generality, breakthroughs, or patent centrality (Trajtenberg et al., 1997; Ahuja and Morris Lampert, 2001; Gilsing et al., 2008)—or producing entirely new measures.

Fourth, in a related vein, much remains to be understood about the relationship between in-text and front-page citations, especially where they overlap. For example, when an in-text citation also appears on the front page—particularly when added by an examiner—it may signal especially high relevance. Such overlap implies that both the applicant (perhaps

even the inventor) and the examiner deemed the prior art pertinent, whether to satisfy disclosure obligations, frame novelty claims, or evaluate patentability. If these cited patents originate from firms operating in similar technological domains, they may offer a rich signal of competitive positioning or technological rivalry (McGahan and Silverman, 2006; Arts et al., 2025).

Finally, from a data perspective, future work could focus on extracting the context in which in-text citations appear. Not all references serve the same legal or rhetorical purpose, and recent work in bibliometrics has shown the value of analyzing citation contexts in scientific literature (*e.g.*, Jurgens et al., 2018; Nicholson et al., 2021). We see strong potential for in-text data to enrich research on established phenomena where citation context matters. Because in-text citations embed more of the inventor's voice, they offer a window into the cognitive underpinnings of inventive activity, helping advance strategic management's broader ambition to understand the microfoundations of innovation.

# References


Ahuja, G., Lampert, C. M., and Novelli, E. (2013). The second face of appropriability: Generative appropriability and its determinants. *Academy of Management Review*, 38(2):248–269.

Ahuja, G. and Morris Lampert, C. (2001). Entrepreneurship in the large corporation: A longitudinal study of how established firms create breakthrough inventions. *Strategic Management Journal*, 22(6-7):521–543.

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

Alcácer, J., Gittelman, M., and Sampat, B. (2009). Applicant and examiner citations in us patents: An overview and analysis. *Research Policy*, 38(2):415–427.

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

Argyres, N., Rios, L. A., and Silverman, B. S. (2025). On the heels of giants: Internal network structure and the race to build on prior innovation. *Strategic Management Journal*.

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

Arts, S., Cassiman, B., and Hou, J. (2025). Technology differentiation, product market rivalry, and m&a transactions. *Strategic Management Journal*.

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

Balsmeier, B., Fleming, L., and Lück, S. (2023). Isolating personal knowledge spillovers: Coinventor deaths and spatial citation differentials. *American Economic Review: Insights*, 5(1):21–33.

Barton, J. H. (2003). Non-obviousness. *IDEA: The Journal of Law and Technology*, 43:475–508.

Berkes, E. (2018). Comprehensive universe of u.s. patents (CUSP): Data and facts. *Unpublished manuscript, Ohio State University*.

Berry, H. (2014). Global integration and innovation: multicountry knowledge generation within mnc s. *Strategic Management Journal*, 35(6):869–890.

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

Carley, M., Hedge, D., and Marco, A. (2015). What is the probability of receiving a us patent. *Yale Journal of Law & Technology*, 17:203–223.

Ceccagnoli, M. (2009). Appropriability, preemption, and firm performance. *Strategic Management Journal*, 30(1):81–98.

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

Chen, T., Kim, C., and Miceli, K. A. (2021). The emergence of new knowledge: The case of zero-reference patents. *Strategic Entrepreneurship Journal*, 15(1):49–72.

Chien, C. V. and Wu, J. Y. (2018). Decoding patentable subject matter. *Patently-O Patent Law Journal 1*, 1:10–19.

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

61:65–129.

de Rassenfosse, G., Kozak, J., and Seliger, F. (2019). Geocoding of worldwide patent data. *Scientific Data*, 6(1):260.

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

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

Fadeev, E. (2024). Creative construction: Knowledge sharing and cooperation between firms. Technical report, Working paper, Duke University.

Feit, I. N. (2011). Does a utility that is unproved at the time of filing violate sec. 112. *Journal of the Patent & Trademark Office Society*, 93:1–18.

Felin, T., Foss, N. J., and Ployhart, R. E. (2015). The microfoundations movement in strategy and organization theory. *Academy of Management Annals*, 9(1):575–632.

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

Ganguli, I., Lin, J., Meursault, V., and Reynolds, N. F. (2024). Patent text and long-run innovation dynamics: The critical role of model selection. *NBER Working Paper*, (w32934).

Gilsing, V., Nooteboom, B., Vanhaverbeke, W., Duysters, G., and Van Den Oord, A. (2008). Network embeddedness and the exploration of novel technologies: Technological distance, betweenness centrality and density. *Research Policy*, 37(10):1717–1731.

Grigoriou, K. and Rothaermel, F. T. (2014). Structural microfoundations of innovation: The role of relational stars. *Journal of Management*, 40(2):586–615.

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

Hall, B. H. and MacGarvie, M. (2010). The private value of software patents. *Research Policy*, 39(7):994–1009.

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

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

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

Jung, H. J. and Lee, J. 2016). The quest for originality: A new typology of knowledge search and breakthrough inventions. *Academy of Management Journal*, 59(5):1725–1753.

Jurgens, D., Kumar, S., Hoover, R., McFarland, D., and Jurafsky, D. (2018). Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

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

Kuhn, J., Younge, K., and Marco, A. (2023). Strategic citation: A reassessment. *Review of Economics and Statistics*, 105(2):458–466.

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

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

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

Marx, M. and Fuegi, A. (2022). Reliance on science by inventors: Hybrid extraction of in-text patent-to-article citations. *Journal of Economics & Management Strategy*, 31(2):369–392.

McGahan, A. M. and Silverman, B. S. (2006). Profiting from technological innovation by others: The effect of competitor patenting on firm value. *Research Policy*, 35(8):1222–1242.

Melero, E., Palomeras, N., and Wehrheim, D. (2020). The effect of patent protection on inventor mobility. *Management Science*, 66(12):5485–5504.

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

Miller, D. J., Fern, M. J., and Cardinal, L. B. (2007). The use of knowledge for technological innovation within diversified firms. *Academy of Management Journal*, 50(2):307–325.

Mowery, D. C., Oxley, J. E., and Silverman, B. S. (1996). Strategic alliances and interfirm knowledge transfer. *Strategic Management Journal*, 17(S2):77–91.

Nicholson, J. M., Mordaunt, M., Lopez, P., Uppala, A., Rosati, D., Rodrigues, N. P., Grabitz, P., and Rife, S. C. (2021). scite: A smart citation index that displays the context of citations and classifies their intent using deep learning. *Quantitative Science Studies*, 2(3):882–898.

Rosenkopf, L. and Nerkar, A. (2001). Beyond local search: boundary-spanning, exploration, and impact in the optical disk industry. *Strategic Management Journal*, 22(4):287–306.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3):638.

Sampat, B. N. (2006). Patenting and us academic research in the 20th century: The world before and after bayh-dole. *Research Policy*, 35(6):772–789.

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

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

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

Somaya, D. (2012). Patent strategy and management: An integrative review and research agenda. *Journal of Management*, 38(4):1084–1114.

Taylor, R. B. (2012). Burying. *Michigan Telecommunications & Technology Law Review*, 19:99–130.

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

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

Trajtenberg, M., Henderson, R., and Jaffe, A. (1997). University versus corporate patents: A window on the basicness of invention. *Economics of Innovation and New Technology*, 5(1):19–50.

Verluise, C. and de Rassenfosse, G. (2020). Patcit: A comprehensive dataset of patent

citations (version 0.15) [data set]. *Zenodo*. *http://doi.org/10.5281/zenodo.3710994*.

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

Wang, H., He, J., and Mahoney, J. T. (2009). Firm-specific knowledge resources and competitive advantage: the roles of economic-and relationship-based employee governance mechanisms. *Strategic Management Journal*, 30(12):1265–1285.

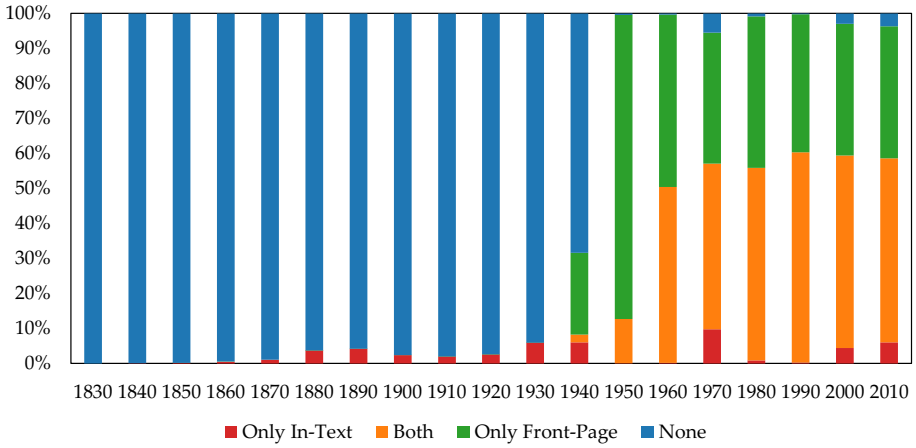
Wang, H., Zhao, S., and He, J. (2016). Increase in takeover protection and firm knowledge accumulation strategy. *Strategic Management Journal*, 37(12):2393–2412.

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

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

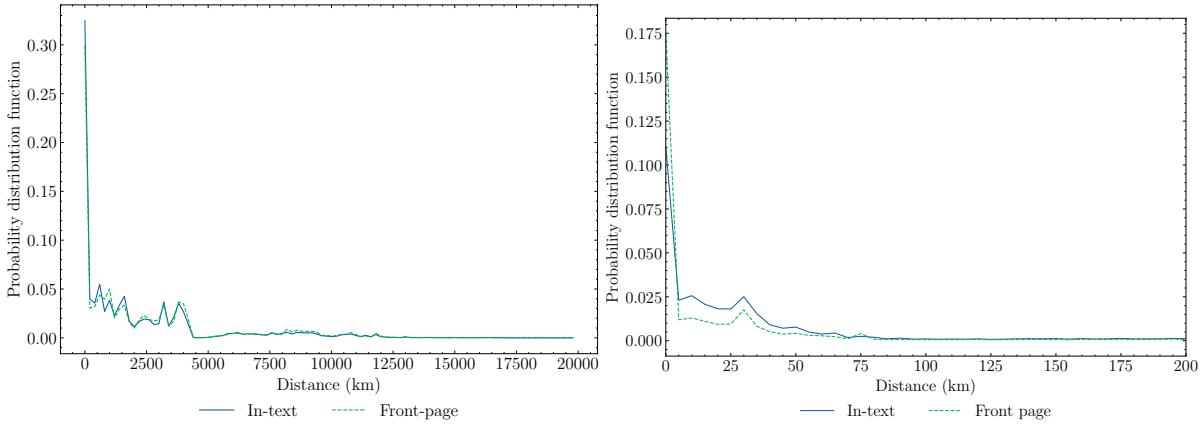
# Figures

Figure 1: Citation types by decade




**Notes:** Proportion of patents by type of citation available.

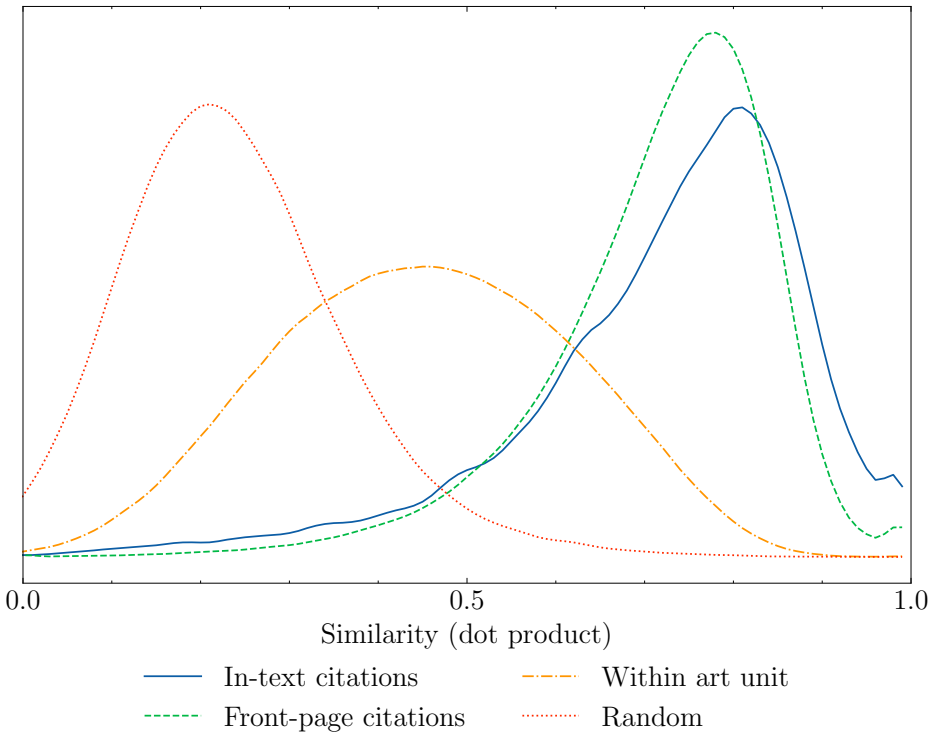
Figure 2: Citations distribution across cited inventors' location




(a) All (b) Distance $\leq$ 200km

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

Figure 3: Citing-cited patent pairwise similarity distribution




**Notes:** Within-INPADOC-family citations omitted.

# Tables

Table 1: Typology of in-text citations

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

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

| | **Front page** | **In-text** |
|---|---|---|
| Number of patents | 16,781,144 | 16,781,144 |
| Number of patents with at least one citation | 12,887,079 | 7,869,894 |
| Share of patents with at least one citation | 76.79% | 46.90% |
| Number of citations (DOCDB) | 265,659,106 | 49,542,360 |
| Average number of citations per patent | 15.83 | 2.95 |
| Average number of citations per patent - conditional on citing at least one patent | 20.61 | 6.29 |
| Median pairwise similarity (dot product) between citing and cited patent [lower quartile, upper quartile][a] | 0.71 [0.62, 0.78] | 0.80 [0.68, 0.88] |
| Share of cited patents in the same INPADOC family | 1.63% | 10.51% |
| Share of cited patents with at least one common inventor | 5.98% | 17.43% |
| Share of cited patents with at least one common assignee | 9.26% | 22.46% |

**Notes**: [a]: After removing within-DOCDB family citations (and after 1947 only).

Table 3: Citations' geographic localization: In-text vs. applicant- & examiner-front-page

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

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

Table 4: Localization of knowledge flows

| | Cites from within X miles | | | | | | |
|---|---|---|---|---|---|---|---|
| | 10 | 20 | 30 | 40 | 50 | 100 | 150 |
| *Main sample* | | | | | | | |
| Original values | −0.246 | −0.299 | −0.190 | −0.101 | −0.072 | −0.016 | −0.031 |
| — *front-page* | (0.080) | (0.065) | (0.045) | (0.031) | (0.030) | (0.028) | (0.025) |
| Replicated values | −0.238 | −0.292 | −0.185 | −0.097 | −0.070 | −0.013 | −0.027 |
| — *front-page* | (0.079) | (0.064) | (0.046) | (0.031) | (0.029) | (0.028) | (0.025) |
| Replicated values | −0.098 | −0.139 | −0.160 | −0.074 | −0.063 | −0.007 | −0.021 |
| — *in-text* | (0.105) | (0.074) | (0.063) | (0.032) | (0.026) | (0.030) | (0.027) |
| *Large distance sample* | | | | | | | |
| Original values | −1.391 | −1.225 | −0.997 | −0.954 | −0.804 | −0.604 | −0.512 |
| — *front-page* | (0.287) | (0.257) | (0.234) | (0.234) | (0.218) | (0.210) | (0.208) |
| Replicated values | −1.402 | −1.231 | −1.011 | −0.969 | −0.815 | −0.607 | −0.492 |
| — *front-page* | (0.290) | (0.261) | (0.237) | (0.236) | (0.220) | (0.211) | (0.215) |
| Replicate values | −1.070 | −1.293 | −1.041 | −1.042 | −0.971 | −0.410 | −0.246 |
| — *in-text* | (0.466) | (0.468) | (0.439) | (0.421) | (0.393) | (0.643) | (0.672) |

**Notes**: Estimates are based on a Poisson regression model. Standard errors in parentheses. The "large distance" sample refers to cases where all co-inventors are located at least 500 miles apart. We report Panels A and C from the original study. Consistent with the original findings, we find no significant effects of premature deaths when using in-text citation data (Panel B in the original study) and omit these estimates for readability.

*Online Appendix for:*

# Beyond the front page: In-text citations to patents as traces of inventor knowledge

October 13, 2025

This Online Appendix provides supporting information for its companion paper. It provides a detailed overview of the data and describes the survey of patent attorneys. Section A explains the data processing pipeline, from data acquisition to matching patent numbers to the standardized DOCDB format. Section B reports the results of detailed validation tasks. Section C will be of interest to most readers, as it explains the data structure. Finally, section D explains the survey of patent attorneys, whose main results are discussed in the paper.

## Contents

# A Citation extraction and cleaning

To construct the dataset we present in this work, it was necessary to accurately identify, extract, and validate citations to other (potentially foreign) patent documents from the full text of over 16 million USPTO patent applications and grants. This task was not trivial and, as such, this section details each stage of this undertaking.

## A.1 Data

The processing pipeline starts with the full text of 16,781,144 patents and patent applications available in the Google Patents public data. The USPTO was created in 1790, but the Google Patents database contains relatively complete data from 1836 (and the first extracted citation is in 1846).[1] However, reliance on optical character recognition for early patents means that extracted citation numbers only reach sufficiently high reliability thresholds, for most use-cases, from about 1976.

The data hosted by Google Patents as part of its public datasets come from the full-text data of IFI CLAIMS Patent Services.[2] The text we consider is the specification section of the patent, *not* including the patent's claims. We do not process the information on the front page.

## A.2 Extraction task

Our starting point is effectively a long chain of characters without any obvious structure, nor any indication about which characters might refer to a patent citation. As such, the first step involves identifying the strings of characters that refer 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) in combination with neighboring text cues (*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 only 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 the lack of formatting requirements for in-text citations, which in turn leads to a large number of both citation cues and patent number formatting, which are difficult to capture with regular-expression-based extraction techniques. On this

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

[2] https://console.cloud.google.com/marketplace/product/google_patents_public_datasets/google-patents-public-data.

point, Adams (2010) warned the community about the challenge of the extraction task. Using a random sample of USPTO patents, he found an "alarming" (p. 26) range of in-text patent citation formats. This problem severely limits citation extraction exercises that are dependent on lists of predefined rules, generally leading to mixed results and, above all, a lack of generalizability.

In order to overcome this limitation, natural-language processing (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, wherein 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 outside the scope of this Online Appendix, we summarize the general working principles of these models below and direct interested readers to the excellent survey conducted by Li et al. (2020) for further reading.

The key to this approach is to view a text as two sequences: a sequence of tokens and a corresponding sequence of latent labels (for example, in this supplement, we label patent citations "PATCIT" and label everything else "O" ). The task is, therefore, to predict the sequence of labels with the sequence of tokens as the input. The algorithm is trained on an annotated set of documents: a set of documents for which we know both the sequence of tokens and the sequence of labels. The probability that each token belongs to a given label is a recursive function of the token itself and its features (digits, capital letters, etc.), as well as the neighboring tokens (its context) and the *neighbouring labels*. The overall goal of the algorithm is to predict, correctly, the whole sequence of latent labels for a given sequence of tokens. If a token (or a sequence of tokens) is unknown or otherwise deviates from the examples in the annotated set, 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 improvement in generalizability relative to rule-based approaches.

For example, let us assume that the algorithm has been trained on a corpus of texts 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), the algorithm is expected to override the absence of the "patent" token and still to predict the correct sequence of labels, [O, O, O, PATCIT].

The aforementioned limitations and improvement opportunities are well-known to the

machine-learning community. In particular, Lopez (2010) developed the Grobid library in 2008 with the goal of overcoming the limitations of rule-based approaches by 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.[3] 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, first introduced in 2001 (Lafferty et al., 2001) and belonging to the family of sequence-labeling models described above. The CRF model has been widely used in various fields and applications.[4]

We rely entirely on Grobid's patent citations' extraction model, which was initially trained on 200 annotated full-text patents.[5] The specific features entering the CRF model that is implemented by Grobid to support patent citation detection include the relative position of the current token in the document, the matching of a common country code indicating the issuing office (*e.g.*, US, EP, WO, etc.) and the matching of a common kind code indicating the document type (*e.g.*, A1, A2, B1, B2, etc.).

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, requires significant post-processing before it is usable.

## A.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 attributes: the country code of the patent authority, the patent document number, and the type of the patent document. This task is challenging due to the many formats 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 the sets of possible outputs for each patent attribute are finite (*e.g.*, the list of patent organization names and the list of their codes, respectively). In addition, each element of the input vocabulary should be mapped to a unique element of

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

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

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

the output vocabulary (*e.g.*, "United States" with "US" or "European Patent Office" with "EP"). In short, for any given patent attribute, the parsing operation can be thought of as a translation operation between two languages with a finite vocabulary. In practice, Grobid implements this task with a Finite State Transducer (FST), a process that appeared early in the history of automated translation.[6]

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

## A.4 Matching task

The final task matches each extracted patent citation to a unique and consolidated identifier, in order to connect each cited patent document to commonly-used patent datasets. For patents, an identifier common to many patent datasets is the DOCDB publication number.[7] At this point, we depart from Grobid, which relies on the European Patent Office (EPO) search API to perform the matching process and uses the EPO document number as its target and consolidation device.[8]

Unfortunately, in the vast 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 API.[9] Taking various 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.[10] First, a large number of variations of each publication number are generated. For each variation, the original patent office and DOCDB formatted versions are indexed. Variations include adding and removing padding zeroes, two and four-digit year dates inside the patent number, Japanese era 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 checked for matches in the large lookup table. Note that there are two distinct services, one for applications and one for patents.[11] We decide which one to call based on the status

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

[7] For simplicity, we use the term "publication number" for both the publication number (for published patents) and the application number (for patent applications).

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

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

[10] We thank Ian Wetherbee from Google Patents for his kind explanation.

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

attribute parsed by Grobid which can take four values: "application", "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, and INPADOC and DOCDB family identifiers.

## A.5 Pipeline Summary

To illustrate the above process in its entirity, 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.

Overall, from the 16,781,144 patent documents that we processed, we were able to extract 64,185,636 in-text patent citations, of which 49,542,360 were matched with a publication number covering 7,869,894 unique patent documents.

# B Validation of extracted citations

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.[12] Although Lopez (2010) reports performance metrics for all these tasks, the set of documents that we are considering differs somewhat from the corpus in that work. In particular, a significant number of patents in our corpus are much older than any document considered for Grobid training and evaluation. Lastly, we also carried out detailed error analyses to support future improvement efforts.

## B.1 Data consistency

USPTO patent documents' scan quality (especially for older patents) and format have changed throughout the years. Before 1971, patents were largely unstructured with no clear delimitation between the metadata and the specification text that is of interest to us (see Figure B1). The modern patent format was introduced in 1971 and progressively replaced the old format before becoming the only format published from 1976. This new format is semi-structured and clearly distinguishes between the metadata sections and the specification section, *inter alia* (see Figure B2). These peculiarities of the source data have some notable implications for our output data.

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

Figure B1: 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 that 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.

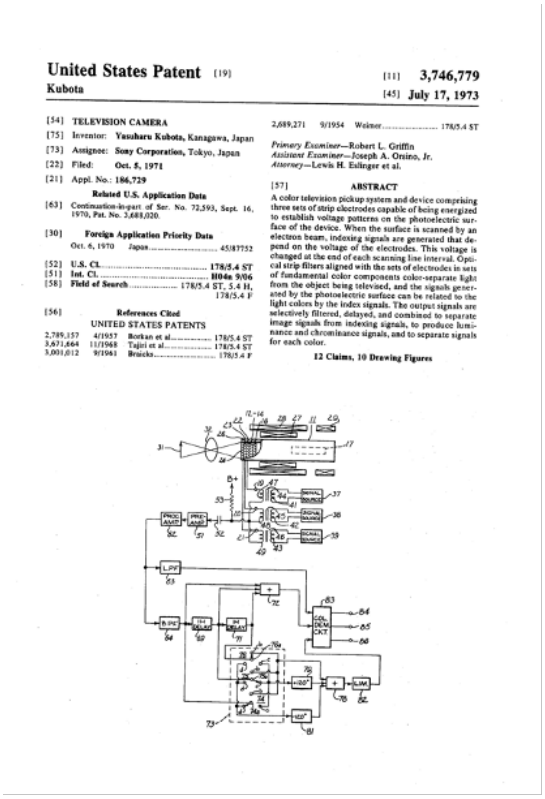
These and other objects are achieved in accordance with this invention by providing a detergent composition comprising an oil-soluble, acylated nitrogen composition characterized by the presence within its structure of (A) a substantially hydrocarbon-substituted polar group selected from the class consisting of acyl, acylimidoyl, and acyloxy radicals wherein the substantially hydrocarbon substituents 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 B2: 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 | Bortan et al. | 178/5.4 ST |
| 3,671,664 | 11/1968 | Tajiri et al. | 178/5.4 ST |
| 3,001,012 | 9/1961 | Bruicks | 178/5.4 F |
| 2,689,271 | 9/1954 | Weimer | 178/5.4 ST |



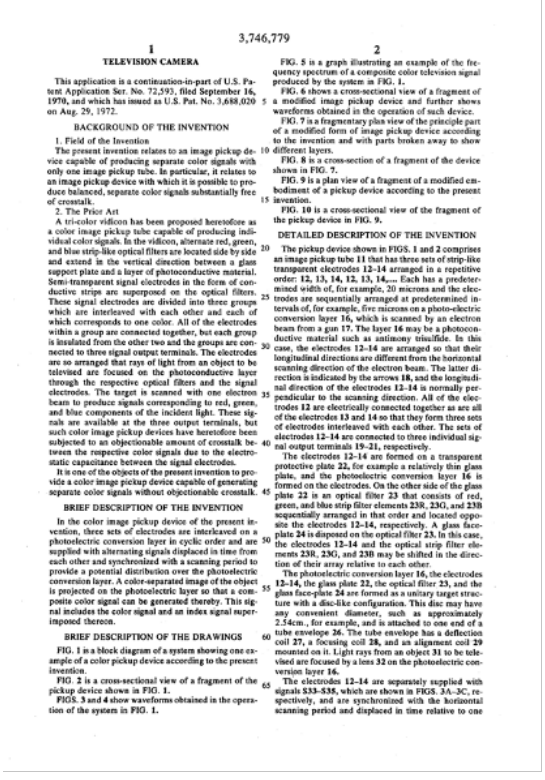
*Primary Examiner*—Robert L. Griffin
*Assistant Examiner*—Joseph A. Orsino, Jr.
*Attorney*—Lewis H. Eslinger et al.


[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 so 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





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 faceplate **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.

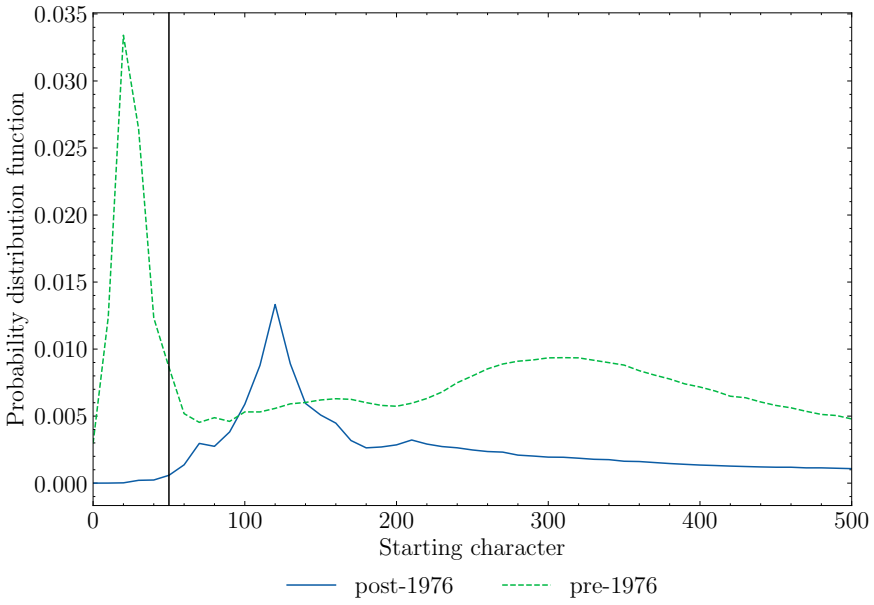
The photoelectric conversion layer **16**, the electrodes **12–14**, the glass plate **22**, the optical filter **23**, and the glass face-plate **24** are formed as a unitary target structure with a disc-like configuration. This disc may have any convenient diameter, such as approximately 2.54cm., for 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

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 that does not correspond to the kind of object we are looking for. Fortunately, this specific pitfall is relatively easy to spot as the citation appears very early in the text. Figure B3 reports the distribution of the rank of the first character of the extracted citations before and after 1976.[13] We observe a clear excess mass between 0 and 50 characters before 1976. To understand this pattern, we randomly drew 50 citations from pre-1971 patents that started before character 50. We found that 88 percent were patent-self references, 8 percent were technological classes, and 4 percent were dates. To address this issue, we chose to flag all citations detected in a patent published before 1976 and starting before character 50 to facilitate their exclusion from analysis.

Figure B3: Empirical probability distribution function of citation detection as a function of the starting character

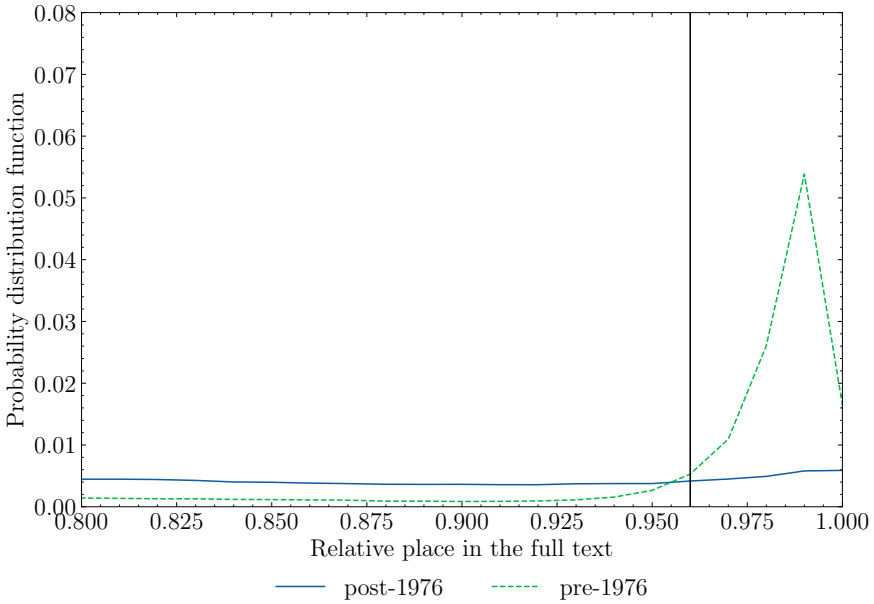



Second, in the old format, what we now call "front-page citations" were printed *after* the patent specification, and these are also sometimes mistakenly included in our source data as part of the full-text of the patent. Since all patents have a different number of characters, we consider the relative location of these citations. Figure B4 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 in the pre-1976 distribution

[13] The spike observed in the post-1976 graph is due to the 'Related Application' section of the patent, which can be addressed with patent family information.

in the last four percent of the full-text characters. To check the nature of these citations, we examine a random sample of 100 citations extracted from patents published before 1976 and occurring in the last four percent of the characters, finding that 99 percent belong to what we would now call the "front-page" citations section. Hence, for patents published before 1976, we flag all citations detected in the last 4 percent of the full-text for easy exclusion.

Figure B4: Empirical probability distribution function of citation detection as a function of the relative place of the starting character




Third, during the transition period between the old and new formats (approximately throughout 1971–1975), two patent formats were being published, 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, with the exception of Figures B3 and B4, exclude flagged patent citations as they are unlikely to correspond to real in-text patent citations (unless explicitly specified).

## B.2 Extraction task

Lopez (2010) reports performance metrics for the extraction task. Using cross-validation,

a technique consisting of 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, therefore, performance could be affected.

To evaluate the quality of our citation extraction, we randomly sampled 160 U.S. patents and manually annotated them. As previously discussed, a patent citation can be presented 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 vicinity 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 B5a.[14] 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 B6, the validation sample and the universe of citing patents display similar distributions by publication year.

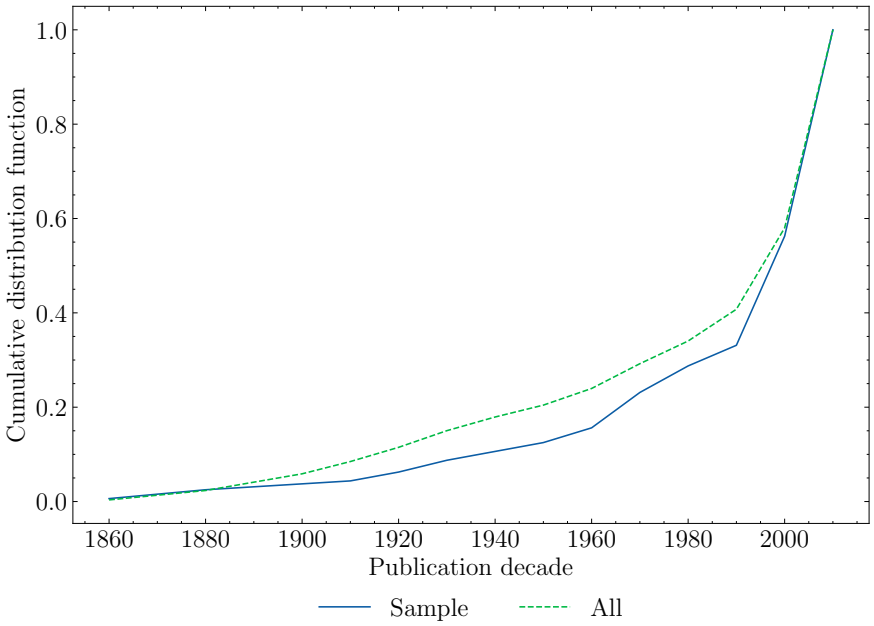
Figure B5: Preview of the annotation platform


(a) Patent extraction validation task


(b) Patent parsing validation task (organisation name)

[14] "Human annotators" are the coauthors of this paper.

Figure B6: Empirical cumulative distribution function of patents in the validation sample and in the universe of U.S. patents (by decade)




From the 160 random U.S. patents in the validation set, we found that 103 (64.4 percent) patents cited at least one patent for a total of 470 in-text patent citations. Table B1 reports the extraction performance that we obtained, together with the Galibert et al. (2010) and Lopez (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). These results significantly outperform Galibert et al. (2010), which used regular expressions to extract citations, reporting a precision of 64.4 percent and a recall of 61 percent. This result clearly confirms that a statistical approach is significantly more effective than a regular expression approach in the context of extracting in-text patent citations. 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, when applied to our extended corpus, Grobid is as reliable as reported in Lopez (2010) when it detects a patent citation. However, it misses patent citations more often in our extended corpus; this is due to older forms of citations appearing in early-twentieth-century patents.

Table B1: 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% |

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 B2 provides examples for each category of errors that we were able to identify. Starting with false negatives (real 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. Looking at false positives (text spans that were wrongly identified by Grobid as citations), we find three categories of errors as well: (1) technological classes reported as "dd/ddd"; (2) date; and (3) docket number. Note that categories (2) and (3) have only one occurrence each.

Table B2: 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' 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.

## B.3 Parsing task

Grobid's FST implementation 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).

In order to validate the quality of the parsing task that we conducted, we randomly sampled 300 extracted citations alongside their parsed attributes. Within the text, 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; 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 B5b. The annotator would then accept or reject the attribute depending on what they actually found in the text. Each example was validated by a single annotator whose decisions were saved upon exit.

Since the attributes can be used independently, a detailed understanding of the performance and errors for each attribute may be valuable for the community. Hence, we performed

three distinct validation exercises, one for each attribute. Table B3 summarizes our results.

Table B3: 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.

We first checked for sample representativeness with respect to the parsing of the patent organisation. Table B4 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 and Trademark Office. This result is consistent with the results that we report for the full dataset in the main body of this article. Similarly, the patent organisations in the remaining third of the validation sample are also the most represented organisations at scale, including the Japan Patent Office, the World Intellectual Property Organisation, the European Patent Office and the German Patent Office. Of the 300 examples that we validated, we found only five errors, leading to a 98.3 percent accuracy score. Errors were 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.[15]

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

Table B4: 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 |

When it comes to the parsing of the patent number, there is no specific way to check sample representativeness. Overall, of the 300 examples that we validated, we found 13 errors, resulting in 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 that mix part of the letters in the prefix with 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 a 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 identifying the application. This letter serves as a time marker for the era. Note that this specific issue is partially fixed by the Google Patent matching API, as explained further below.

Lastly, we validated the parsing of the kind code, which indicates the specific type of document the citation refers to (granted patent, application, reissue, design, etc.). For 502 random samples, we achieve an accuracy of 97.6 percent. Note, however, that this measure includes a large proportion of null results as the kind code is 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 seven mistakes, giving a "conditional" accuracy of 86 percent. Specifically, we identified three groups of parsing errors: errors resulting from unconventional formatting, issues with optical character recognition on scanned documents, and Grobid incorrectly interpreting "Cl" (abbreviation for "class") as the "C" kind code.

Importantly, every instance in standard form was correctly parsed.

## B.4 Matching task

The matching task involves associating the extracted attributes with a unique identifier, that is, the DOCDB publication number. In order to validate this step of the process, we randomly sampled 200 citations from our final dataset and 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 B5 summarizes these categories, their contents, and the results from the validation exercise.

Table B5: 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 |

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 F1 score of 91.06 percent.

Next, we explored the nature of the errors and non-matches. Tables B6 and B7 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 often 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 partial citations that even a human cannot match and citations of provisional patent applications, which may never appear in public patent datasets. (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.) This family of errors represents 41 percent of the thirty-six unmatchable detected citations in our validation sample. Finally, focusing on the unmatched citations that a human can match reveals some blind spots of the Linking API. Among the 17 cases in this category, 53 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.

Table B6: 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 B7: 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 |

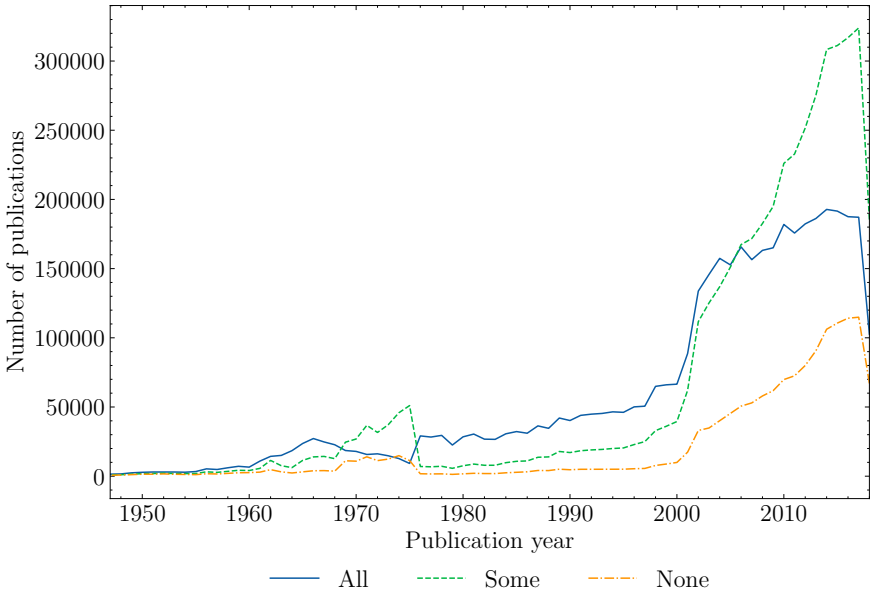
**Notes**: The number of occurrences includes both matched and unmatched examples.

While the previous step can characterize the performance of the matching procedure generally, the small size of the validation sample means that we are unlikely to uncover rare irregularities that might nonetheless be numerous at a large scale.

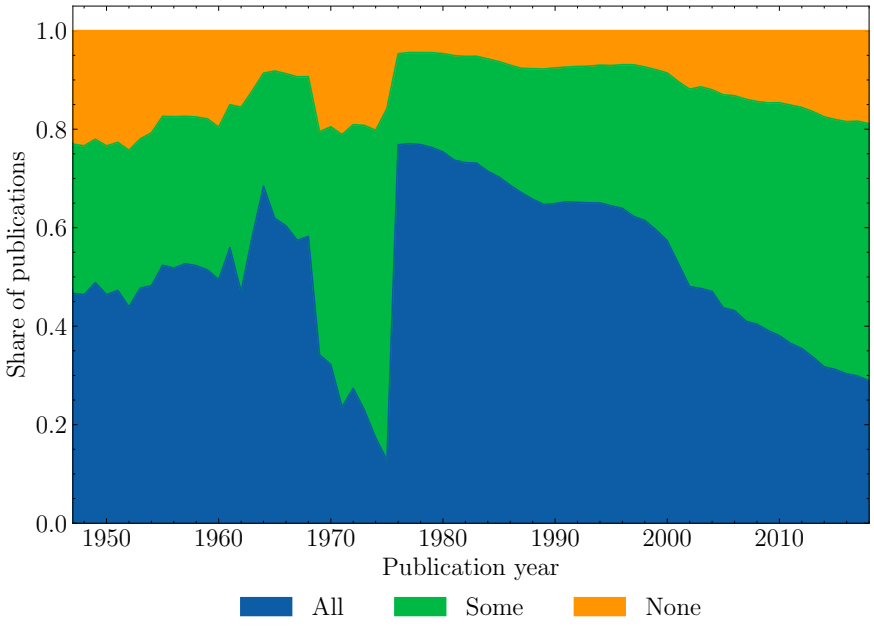
Considering the whole dataset, Figure B7 shows the yearly number (B7a) and share (B7b) of citing patents according to the matching status of the extracted in-text citations.[16] Patents with all in-text citations matched to a publication number represent 42.7 percent of the total, whereas those with 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 and 1975, the performance of our matching procedure worsens, as the proportion of patents with some or no citations matched grows markedly. From 1976 onwards, the share of patents with all citations matched jumps to 77 percent in 1976 before slowly decreasing since that time until the current day.

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

Figure B7: 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.

These aggregate figures mask high variation between the patent offices associated with the

cited patent documents. Table B8 reports the number of extracted in-text citations alongside the share of those that were matched, for the top five patent offices in our dataset. More than half of the extracted 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 Property Organisation (WIPO) and the Japan Patent Office (JPO) with, respectively, around 6.5 million (10% of the total) and 5.7 million (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 (73%) for the 1.4 million extracted citations to patents filed at the German Patent and Trade Mark Office (DPMA). Lastly, we obtain less satisfactory match rates for citations to EPO patent filings; of 2.2 million citations, we match only 51 percent.

Table B8: 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 |

# C Data description

Data generation and validation reproducibility are 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 subject to future improvements, exact replication of the data and results detailed above requires the user to choose the tag '0.3.1' of the code.[17]

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 the 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. We also report a flag indicating whether 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 |
| **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 |

[17] [Weblink redacted to preserve anonymity.]

| Name | Description | Type | Nb non null |
|---|---|---|---|
| **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 |
| **___.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 |

| Name | Description | Type | Nb non null |
|---|---|---|---|
| **___.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 |
| **___.flag** | Flag detected citations 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.

Tables C2 and C3 provide an overview of the composition of the dataset.

Table C2: 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 C3: 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 |

# D Attorney survey on citation origin

We surveyed U.S. patent attorneys to gauge who typically supplies the references (applicants, inventors, or attorneys) that appear in the Information Disclosure Statement (IDS) and in the patent specification.

The present Appendix documents every step that underpins the survey evidence presented in the main paper. Section D.1 details the survey method—how we constructed the original attorney sampling frame, collected 1,028 email addresses, and designed the questionnaire. Section D.2 explains the estimation approach and distils the headline patterns that compare applicant- and inventor-supplied citations, before presenting the full set of estimates in Tables D2–D15. Readers who only need the take-away numbers can focus on Table D1 for inventor shares across citation types and Tables D14–D15 for relative likelihood ratios.

## D.1 Method

### D.1.1 Sample frame and participation

The USPTO provides information on attorneys handling patent applications, including first name, last names, and organization. The data are available in PatentsView, table 'g_attorney_disambiguated.' We started by building a list of about 10,000 attorneys who:

- Are associated with at least five patents that include in-text citations
- Are associated with at least one patent from 2015 and before
- Are associated with at least one patent from 2016 and after (*i.e.*, probably still active)
- Remained after the stratified random balancing of patents to include similar numbers of patents in the sample from each technology area

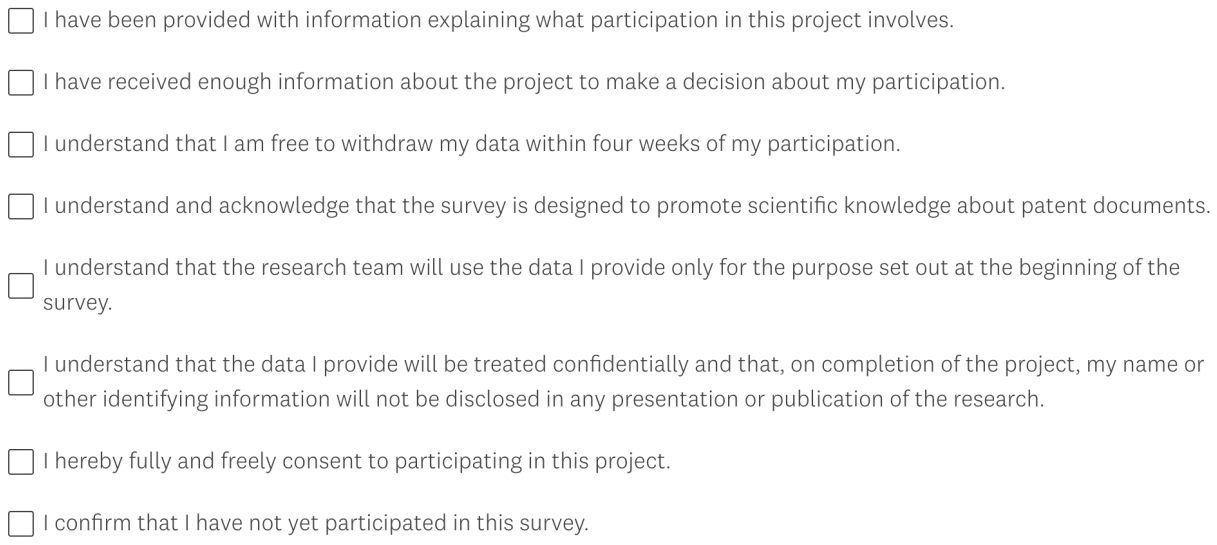
Next, we ordered this list randomly and manually went through individual attorneys, searching for their email addresses. We stopped the process once we reached approximately 1000 addresses (obtaining a total of 1028 contact details).

We programmed the survey in SurveyMonkey and e-mailed it in three waves from August 29, 2024, to September 25, 2024. We explained that the survey was part of an academic research project and offered no monetary incentive. We sent one reminder email a few weeks after the first email.

A total of 133 email addresses bounced or did not reach the relevant person (12.94%). Among the remaining 895 'active' recipients, 115 (12.85%) started the survey.

The survey opens with consent questions, available in Figure D1, during which five respondents left.[18] Out of the 110 respondents who continued with the survey, 92 completed it, leading to a completion rate of 83.65%.

Figure D1: Informed-consent screen shown to all participants




**Notes:** All subjects gave informed consent for inclusion before participating in the study.

### D.1.2 Questionnaire design

We focus on citations appearing in IDS (some of which will ultimately make it to the front page) and in-text citations. The survey considers four groups of citations: scientific references listed in IDS, patent references listed in IDS, scientific references listed in the patent text, and patent references listed in the patent text.

For each group, we have asked respondents to estimate the likelihood that a randomly selected reference was supplied by the applicant (including inventors, in-house counsel, etc.) or by the attorney or a member of their team.[19] Attorneys are presumably knowledgeable about this question since they directly observe whether they added the reference or whether the applicant supplied it.

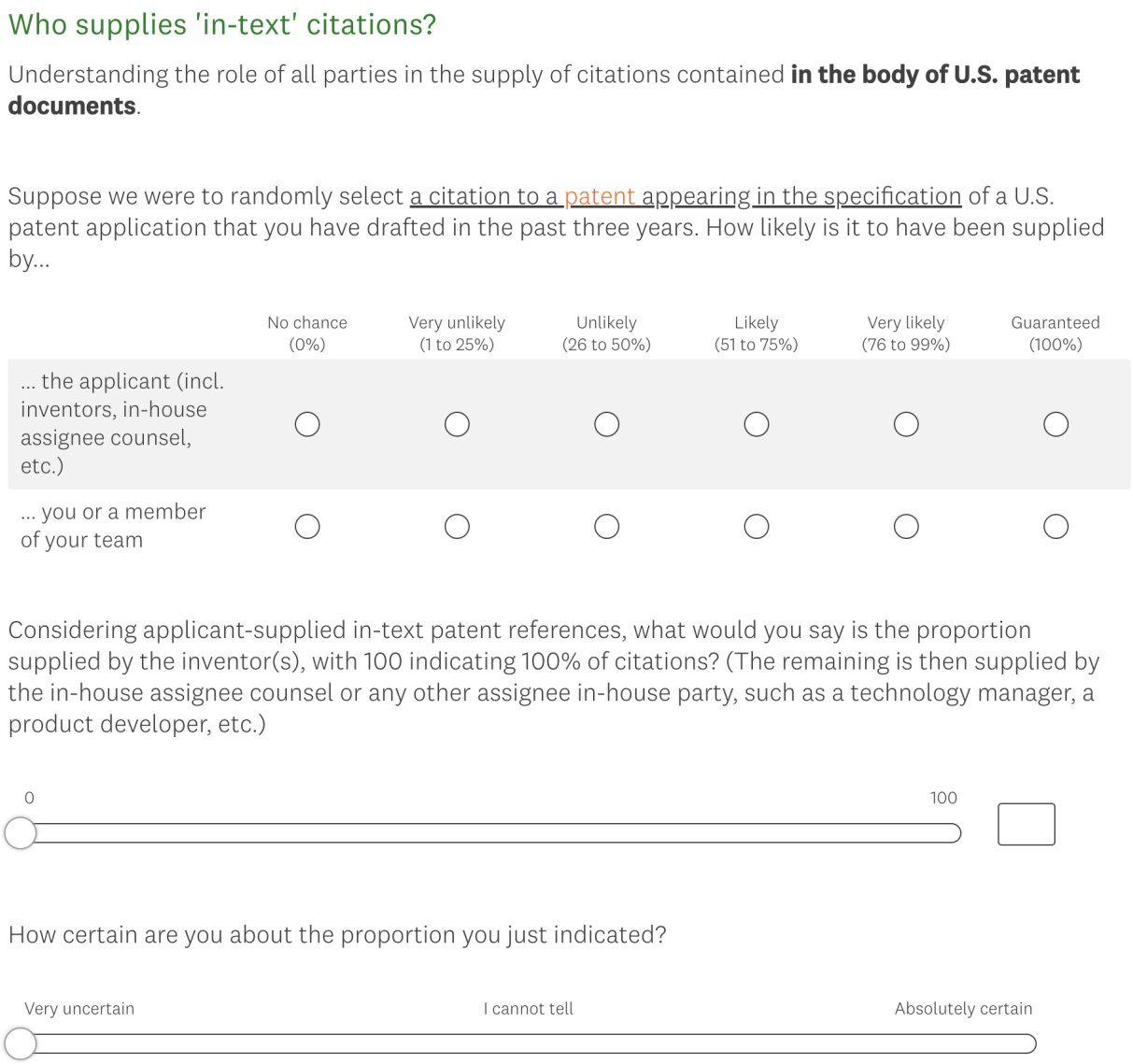
In a follow-up question, we ask the proportion of applicant-supplied references that originate from inventors (as opposed to any other source, such as the in-house counsel). Not

[18] Following the Code of Federal Regulations on the Protection of Human Subjects, this study does not require ethical approval, as it only involves survey procedures, and disclosure of subjects' responses outside research would not reasonably place the subjects at risk (45 CFR §46.104 (d)(2)(ii)).

[19] In the case of IDS, we also ask about a third origin, namely from a foreign attorney as part of a foreign search report from the same patent family.

all patent attorneys will be able to answer that question confidently because they may not observe the origin of applicant-supplied references. For that reason, we have asked them to also qualify the uncertainty of their response with a slider scale ranging from 'Very uncertain' (score of 0) to 'Absolutely certain' (score of 100). Figure D2 provides a screenshot of one view of the survey for illustration.

Figure D2: Screenshot of questionnaire for in-text patent citations




Note that it may seem counter-productive to survey attorneys if our focus is on the role of inventors in supplying references. However, because every reference passes through the attorneys, they observe both their own additions and those from applicants, giving them a 360-degree view that individual inventors rarely have. Furthermore, attorneys typically

track who supplied each reference because they must preserve privilege boundaries, maintain disclosure logs, and be able to document compliance with Rule 56 if questioned later. Inventors, by contrast, seldom know whether additional references were later added by counsel or by co-inventors and cannot see the final IDS once the application is filed. Finally, a single attorney works on dozens of applications each year and often for multiple clients, offering a much broader perspective than obtained if surveying inventors.

To mitigate potential attrition bias, we randomized whether the IDS-citation question block or the in-text citation question block appeared first. That is, half of the respondents will start with the IDS questions, and the other half will start with the in-text questions.

The survey was pre-tested with three patent attorneys. The pre-tests led us to clarify the language of some questions and to add another source of citation origin for IDS (namely, international search reports).

Survey Monkey estimated that the survey would take about 8 minutes to complete. Our respondents took an average of 7:03 minutes, indicating that they replied diligently.[20]

## D.2 Statistical analysis

### D.2.1 Overview of results

**In-text citations to patents**

Considering the origin of citations in the patent text, applicants supply about 42–51% of citations to patents, depending on model specifications (Table D2).

We have considered five specifications to produce this estimate range. *Considering all respondents*, the first specification, reports the mean of answers across all respondents. Next, *Considering respondents with internally consistent answers* only includes responses from respondents for which the sum of applicant and attorney origins is close to 100% (see Figure D2). The third specification, *Using the complement of the attorney answer*, assumes that the proportion of citations originating from applicants is 100 minus the proportion of citations originating from attorneys. The fourth specification, *Using the complement of the attorney answer and internally consistent answers* combines the second and third approaches. Finally, *Normalizing to have a sum of 100* normalizes the proportion of citations originating from applicants by the sum of the proportions of citations originating from applicants and attorneys.

Next, Table D3 reports estimates of the proportion of applicant citations that are supplied by inventors (as opposed to the in-house counsel or any other party). Attorneys

[20] This figure excludes one outlier who took several hours to complete the survey, presumably leaving their browser open.

report that roughly half (50–56%) originate with the inventors themselves, depending on model specifications. There are four specifications. The first two are similar to the first two of Table D2. The third specification, *Considering respondents with high self-assessed certainty*, exploits respondents' self-reported certainty estimate for the answer (see Figure D2). It only considers responses from respondents with a value above 75 (with the slider ranging from 0 to 100). Finally, *Giving more weight to respondents with higher self-assessed certainty* weights the response by the uncertainty score.

**In-text citations to scientific articles**

The reliance on inventors is particularly strong considering scientific papers listed in the patent specifications, with two-thirds of these supplied by applicants (64–70%, Table D5) and, among these, more than two-thirds provided by inventors (70–73%, Table D6).

**IDS citations to patents**

Turning now to the origin of references in IDS, the likelihood that the applicant supplies a patent reference is at least five percentage points lower than citations appearing in the text of the patent (compare the first two rows of Table D2 with the first two rows of Table D8), and possibly much lower (consider the last row of Table D8). Note that the table presents only three specifications. The reason is that there are three possible sources of origin for IDS citations, namely, applicants, attorneys, and international search reports. Accordingly, we cannot define the applicant score as the complement of the attorney score.

Regarding the proportion of applicant-supplied IDS citations, the likelihood that they originate from inventors is also lower than for in-text citations (40–50% in Table D9 to be compared with 50–56% in Table D3).

**IDS citations to scientific articles**

Concerning references to papers listed in IDS, the likelihood that applicants have supplied them is roughly similar to the likelihood obtained for in-text references (compare the first two rows of Table D11 with the first two rows of Table D5). Likewise, roughly two-thirds of them are supplied by inventors (Table D12).

**The prevalence of inventor origin by citation type**

Table D1 reports the probability that a citation is supplied by the inventors for every citation type. To obtain these estimates, we have multiplied the proportion of applicant-supplied references by the proportion of applicant-supplied references believed to originate from inventors. We have implemented several specifications, as defined above.

Table D1: Proportion of citations supplied by inventors

| Citation type | Range across specifications | Source |
|---|---|---|
| Patent citation in-text | [29–38] | Table D4 |
| Paper citation in-text | [54–57] | Table D7 |
| Patent citation IDS | [20–29] | Table D10 |
| Paper citation IDS | [50–57] | Table D13 |

**Notes**: The range reports the lowest and highest point estimates obtained in the associated source.

In a nutshell, comparing in-text vs. IDS citations to patents, our results suggest that in-text citations are more likely to have been supplied by inventors than IDS citations. Furthermore, comparing citations to patents vs. papers, we find that citations to papers are more likely to have been supplied by the inventor than citations to patents.

We can use the data reported in Table D1 to compute the relative likelihood that a citation to a patent is provided by the inventor if it comes from the patent text vs. the IDS form. The result, presented in Table D14, suggests values ranging between 1.19 and 1.44. That is, a randomly chosen in-text patent citation is 19–44% more likely to originate from the inventor than a randomly chosen front-page 'applicant' citation.

Considering references to scientific papers, Table D1 suggests that these are usually supplied by the inventors, in proportions that do not differ significantly between in-text and IDS. If anything, in-text citations to papers are 10 to 17% more likely to be supplied by inventors than a randomly chosen front-page 'applicant' citation, though the point estimate is not statistically significantly different from zero (see Table D15).

### D.2.2 Detailed estimates

#### In-text citations

Table D2: [in-text:patent:applicant] Probability that a citation to a patent appearing in the specification of a U.S. patent application is supplied by the applicant

| Estimate | Method |
|---|---|
| $50.82 \pm 7.80$ (N = 94) | Considering all respondents |
| $43.23 \pm 9.67$ (N = 44) | Considering respondents with internally consistent answers |
| $41.54 \pm 6.62$ (N = 93) | Using the complement of the attorney answer |
| $42.61 \pm 9.65$ (N = 44) | Using the complement of the attorney answer and internally consistent answers |
| $46.25 \pm 5.84$ (N = 90) | Normalizing to have a sum of 100 |

**Notes**: Point estimates are reported with half-widths of 95% confidence intervals in parentheses. See main text for methodological details.

Table D3: [in-text:patent:inventor_share] Among these, the proportion of patent references that are submitted by the inventor(s)

| **Estimate** | **Method** |
|---|---|
| $51.33 \pm 6.91$ (N = 90) | Considering all respondents |
| $49.77 \pm 10.95$ (N = 43) | Considering respondents with internally consistent answer in Table D2 |
| $56.44 \pm 10.85$ (N = 45) | Considering respondents with high self-assessed certainty |
| $53.99 \pm 7.24$ (N = 87) | Giving more weight to respondents with higher self-assessed certainty |

**Notes**: Point estimates are reported with half-widths of 95% confidence intervals in parentheses. See main text for methodological details.

Table D4: [in-text:patent:inventor] Probability that a citation to a patent appearing in the specification of a U.S. patent application is supplied by the inventor (*i.e.*, combining Table D2 and Table D3)

| **Estimate** | **Method** |
|---|---|
| $32.71 \pm 6.39$ (N = 90) | Considering all respondents |
| $28.28 \pm 9.19$ (N = 43) | Considering respondents with internally consistent answer in Table D2 |
| $37.79 \pm 10.08$ (N = 45) | Considering respondents with high self-assessed certainty |
| $34.64 \pm 6.76$ (N = 87) | Giving more weight to respondents with higher self-assessed certainty |

**Notes**: Point estimates are reported with half-widths of 95% confidence intervals in parentheses. See main text for methodological details.

Table D5: [in-text:paper:applicant] Probability that a citation to a paper appearing in the specification of a U.S. patent application is supplied by the applicant

| **Estimate** | **Method** |
|---|---|
| $67.83 \pm 6.50$ (N = 92) | Considering all respondents |
| $70.16 \pm 8.68$ (N = 45) | Considering respondents with internally consistent answers |
| $64.06 \pm 6.02$ (N = 90) | Using the complement of the attorney answer |
| $69.61 \pm 8.69$ (N = 45) | Using the complement of the attorney answer and internally consistent answers |
| $65.57 \pm 5.47$ (N = 87) | Normalizing to have a sum of 100 |

**Notes**: Point estimates are reported with half-widths of 95% confidence intervals in parentheses. See main text for methodological details.

Table D6: [in-text:paper:inventor_share] Among these, the proportion of paper references that are submitted by the inventor(s)

| Estimate | Method |
|---|---|
| $69.46 \pm 6.28$ (N = 92) | Considering all respondents |
| $72.89 \pm 8.37$ (N = 45) | Considering respondents with internally consistent answer in Table D5 |
| $70.51 \pm 8.32$ (N = 59) | Considering respondents with high self-assessed certainty |
| $72.01 \pm 6.28$ (N = 88) | Giving more weight to respondents with higher self-assessed certainty |

**Notes**: Point estimates are reported with half-widths of 95% confidence intervals in parentheses. See main text for methodological details.

Table D7: [in-text:paper:inventor] Probability that a citation to a paper appearing in the specification of a U.S. patent application is supplied by the inventor (*i.e.*, combining Table D5 and Table D6)

| Estimate | Method |
|---|---|
| $53.84 \pm 6.72$ (N = 91) | Considering all respondents |
| $57.25 \pm 9.23$ (N = 45) | Considering respondents with internally consistent answer in Table D5 |
| $56.83 \pm 8.94$ (N = 59) | Considering respondents with high self-assessed certainty |
| $56.60 \pm 6.87$ (N = 87) | Giving more weight to respondents with higher self-assessed certainty |

**Notes**: Point estimates are reported with half-widths of 95% confidence intervals in parentheses. See main text for methodological details.

### IDS citations

Table D8: [ids:patent:applicant] Probability that a citation to a patent appearing in the IDS of a U.S. patent application is supplied by the applicant

| Estimate | Method |
|---|---|
| $45.33 \pm 6.17$ (N = 92) | Considering all respondents |
| $38.17 \pm 8.81$ (N = 41) | Considering respondents with internally consistent answers in Table D2 |
| $28.43 \pm 3.53$ (N = 92) | Normalizing to have a sum of 100 |

**Notes**: Point estimates are reported with half-widths of 95% confidence intervals in parentheses. See main text for methodological details.

Table D9: [ids:patent:inventor_share] Among these, the proportion of patent references that are submitted by the inventor(s)

| Estimate | Method |
|---|---|
| $47.78 \pm 6.74$ (N = 90) | Considering all respondents |
| $40.24 \pm 9.52$ (N = 41) | Considering respondents with internally consistent answer in Table D2 |
| $49.78 \pm 11.07$ (N = 46) | Considering respondents with high self-assessed certainty |
| $49.35 \pm 7.36$ (N = 86) | Giving more weight to respondents with higher self-assessed certainty |

**Notes**: Point estimates are reported with half-widths of 95% confidence intervals in parentheses. See main text for methodological details.

Table D10: [ids:patent:inventor] Probability that a citation to a patent appearing in the IDS of a U.S. patent application is supplied by the inventor (*i.e.*, combining Table D8 and Table D9)

| Estimate | Method |
|---|---|
| $27.34 \pm 5.50$ (N = 90) | Considering all respondents |
| $20.22 \pm 6.56$ (N = 41) | Considering respondents with internally consistent answer in Table D2 |
| $28.74 \pm 8.86$ (N = 46) | Considering respondents with high self-assessed certainty |
| $28.27 \pm 5.97$ (N = 86) | Giving more weight to respondents with higher self-assessed certainty |

**Notes**: Point estimates are reported with half-widths of 95% confidence intervals in parentheses. See main text for methodological details.

Table D11: [ids:paper:applicant] Probability that a citation to a paper appearing in the IDS of a U.S. patent application is supplied by the applicant

| Estimate | Method |
|---|---|
| $65.59 \pm 5.65$ ($N = 92$) | Considering all respondents |
| $64.50 \pm 8.93$ ($N = 42$) | Considering respondents with internally consistent answers in Table D5 |
| $47.21 \pm 4.56$ ($N = 90$) | Normalizing to have a sum of 100 |

**Notes**: Point estimates are reported with half-widths of 95% confidence intervals in parentheses. See main text for methodological details.

Table D12: [ids:paper:inventor_share] Among these, the proportion of paper references that are submitted by the inventor(s)

| Estimate | Method |
|---|---|
| $67.22 \pm 6.08$ (N = 90) | Considering all respondents |
| $66.90 \pm 9.36$ (N = 42) | Considering respondents with internally consistent answer in Table D5 |
| $73.62 \pm 9.20$ (N = 47) | Considering respondents with high self-assessed certainty |
| $68.81 \pm 6.46$ (N = 87) | Giving more weight to respondents with higher self-assessed certainty |

**Notes**: Point estimates are reported with half-widths of 95% confidence intervals in parentheses. See main text for methodological details.

Table D13: [ids:paper:inventor] Probability that a citation to a paper appearing in the IDS of a U.S. patent application is supplied by the inventor (*i.e.*, combining Table D11 and Table D12)

| Estimate | Method |
|---|---|
| $49.78 \pm 6.11$ (N = 90) | Considering all respondents |
| $50.08 \pm 9.68$ (N = 42) | Considering respondents with internally consistent answer in Table D5 |
| $57.12 \pm 9.27$ (N = 47) | Considering respondents with high self-assessed certainty |
| $51.39 \pm 6.48$ (N = 87) | Giving more weight to respondents with higher self-assessed certainty |

**Notes**: Point estimates are reported with half-widths of 95% confidence intervals in parentheses. See main text for methodological details.

### Relative probabilities

Table D14: Relative probability (between in-text and IDS) that a citation to a patent is supplied by the inventor

| Estimate | Method |
|---|---|
| $1.19 \pm 0.24$ (N = 86) | Considering all respondents |
| $1.44 \pm 0.48$ (N = 40) | Considering respondents with internally consistent answer in Table D2 |
| $1.38 \pm 0.44$ (N = 38) | Considering respondents with high self-assessed certainty |

**Notes**: Results obtained by dividing point estimates in Table D4 by point estimates in Table D10. Point estimates with 95% confidence interval half-widths reported. $> 1$ means that inventor origin is more likely for in-text citations than IDS citations. See main text for methodological details.

Table D15: Relative probability (between in-text and IDS) that a citation to a paper is supplied by the inventor

| **Estimate** | **Method** |
|---|---|
| $1.10 \pm 0.12$ (N = 87) | Considering all respondents |
| $1.17 \pm 0.16$ (N = 42) | Considering respondents with internally consistent answer in Table D5 |
| $1.11 \pm 0.35$ (N = 41) | Considering respondents with high self-assessed certainty |

**Notes**: Results obtained by dividing point estimates in Table D7 by point estimates in Table D13. Point estimates of the ratio with 95% confidence interval half-widths reported. $> 1$ means that inventor origin is more likely for in-text citations than IDS citations. See main text for methodological details.

# References


Adams, S. (2010). 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*, 32(1):22–29.

Berkes, E. (2018). Comprehensive universe of US patents (CUSP): Data and facts. *Unpublished manuscript, Ohio State University*.

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

Galibert, O., Rosset, S., Tannier, X., and Grandry, F. (2010). Hybrid citation extraction from patents. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010*, pages 17–23.

Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Li, J., Sun, A., Han, J., and Li, C. (2020). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1.

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

Roche, E. and Schabes, Y. (1997). *Finite-state language processing*. MIT press.

Sutton, C. and McCallum, A. (2006). An introduction to conditional random fields for relational learning. *Introduction to statistical relational learning*, 2:93–128.

# Tables