RAND Journal of Economics
Vol. 51, No. 1, Spring 2020
pp. 109–132

Patent citations reexamined

Jeffrey Kuhn*

Kenneth Younge**

and

Alan Marco***

Many studies rely on patent citations to measure intellectual heritage and impact. In this article, we show that the nature of patent citations has changed dramatically in recent years. Today, a small minority of patent applications are generating a large majority of patent citations, and the mean technological similarity between citing and cited patents has fallen considerably. We replicate several well-known studies in industrial organization and innovation economics and demonstrate how generalized assumptions about the nature of patent citations have misled the field.

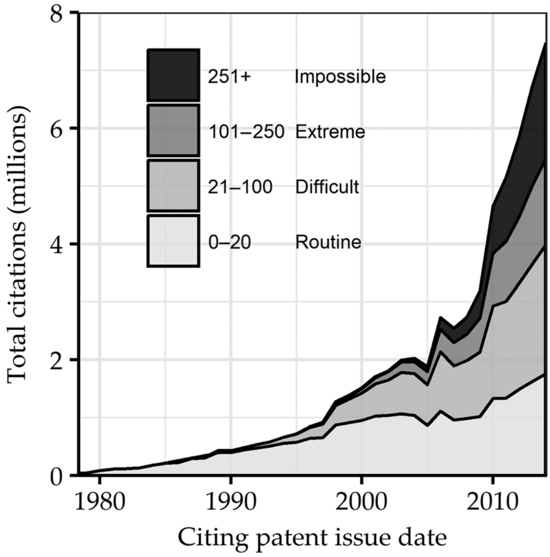
1. Introduction

■ Scholars often use patenting as an empirical measure of innovation. Seminal work by Griliches (1981) measured patent counts, and later work weighted such counts by the number of forward citations received by each patent on the view that more important patents receive more references (Trajtenberg, 1990a). Later studies used patent citations to measure private patent value (Lanjouw and Schankerman, 2001; Harhoff et al., 1999), firm market value (Hall, Jaffe, and Trajtenberg, 2005), cumulative innovation (Caballero and Jaffe, 1993; Trajtenberg, Henderson, and Jaffe, 1997), geographic spillovers (Jaffe, Trajtenberg, and Henderson, 1993), technology life-cycles (Mehta, Rysman, and Simcoe, 2010), social importance (Moser, Ohmstedt, and Rhode, 2017), originality (Jung and Lee, 2016), and technological impact (Corredoira and

* University of North Carolina at Chapel Hill; jeffrey_kuhn@kenan-flagler.unc.edu.
** École Polytechnique Fédérale de Lausanne; kenneth.younge@epfl.ch.
*** Georgia Tech; alan.marco@pubpolicy.gatech.edu.
We thank Ian Cockburn, Rui de Figueiredo, Jeff Furman, Michelle Gittelman, Dietmar Harhoff, Robert Merges, Michael Meurer, Atul Nerkar, Gaetan de Rassenfosse, Katja Seim, Tim Simcoe, Neil Thompson, Tony Tong, Andrew Toole, Brian Wright, Noam Yuchtman, and anonymous reviewers for helpful comments. We also thank seminar participants at the 2016 Searle Center Conference on Innovation Economics, the 2017 Boston University Law Schools Technology Policy and Research Initiative Conference, the 2016 Munich Summer Institute, the United States Patent and Trademark Office (2015), Skema Business School (2015), the 2016 DRUID Conference, and the 2016 Academy of Management Annual Meeting. The authors thank the United States Patent and Trademark Office for providing access to data. The authors also thank Google for a generous research grant of computing time on the Google Cloud.

FIGURE 1

TOTAL CITATIONS MADE BY YEAR, DIVIDED BY NUMBER OF BACKWARD CITATIONS MADE BY CITING PATENT



Banerjee, 2015). Overall, a search for "patent citations" on Google Scholar (conducted January 1, 2018) returns over 21,000 results.

Recent research, however, has begun to call into question a straightforward interpretation of patent citations. Abrams, Akcigit, and Popadak (2018) find an inverted-U relationship between patent citations and market value, suggesting that high citation counts may indicate strategic use of the patent system (instead of impactful innovation). Evidence also has emerged that the search for and disclosure of prior art varies between applicants in systematic ways (e.g., Sampat, 2010). Citations may suffer from significant noise and measurement error (Gambardella, Harhoff, and Verspagen, 2008; Roach and Cohen, 2013), and the comparison of patents between cohorts can be problematic because citation counts have inflated substantially over time (Marco, 2007). Failing to correct for time period, technology, and geographic region can introduce significant bias into an analysis (Lerner and Seru, 2017). Correcting citation counts to return to the original goal of devising a measure of innovation activity that is broadly comparable across contexts, however, is problematic due to endogeneity in the pendency, citation lags, and filing years of a given sample (Mehta, Rysman, and Simcoe, 2010).

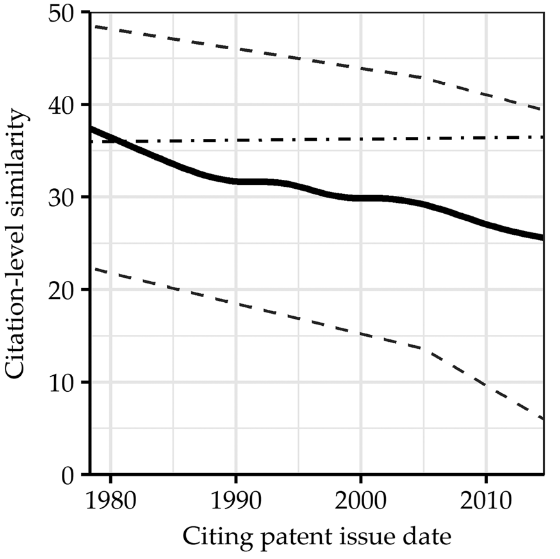
In this article, we highlight an important change in the data generating process of patent citations that changes the statistical nature of patent citation measures and the appropriate use of patent citations in current and future research. Specifically, we observe a dramatic increase in the number of citations generated per year and relate that change to a small proportion of patents flooding the patent office with an overwhelming number of references. Figure 1 shows the number of backward citations over time, segmented by the number of citations made by each patent. In recent years, 46.8% of references derive from less than 5% of patents with more than 100 backward citations.[1]

To examine how the aforementioned changes in citation patterns affect the information content and quality of citation-based measures, we use a vector space model based on Younge and Kuhn (2015) that compares the text of every patent to every other patent granted by the

[1] Conversations with examiners suggest that reviewing more than 100 citations is extremely difficult, given time constraints.

FIGURE 2

MEAN AND INTERQUARTILE RANGE OF SIMILARITY FOR CITATIONS BY CITING YEAR. DOT-DASHED LINE PLOTS OLS ESTIMATE OF SIMILARITY, CONTROLLING FOR COUNT OF BACKWARD CITATIONS



United States Patent and Trademark Office (USPTO). The measure allows us to determine the technological distance between each citing/cited pair of patents based on the technical description of each patent. We find that the information quality of patent citations has changed dramatically. Patents today have a much larger pool of relevant prior art to draw from than patents in the past—and one therefore might expect that the average similarity of citing and cited patents would be *increasing* over time (or at least not decreasing)[2]—but that has not been the case. Figure 2 shows that the mean similarity of patent citations has declined significantly from 1985 to 2014. The trend suggests that the technological relationship reflected by the average patent citation has weakened over time, and that the decline is continuing to accelerate.

We argue that the cause of falling *average* technological similarity is a mixture problem; the cause is not *what* is cited, but rather *how many* citations are submitted by a certain portion of the population. As a preview of our findings, Figure 2 plots an OLS regression line (dot-dashed) regressing textual similarity on citing patent issue year, controlling for the number of backward citations; the precise estimate is reported in Model 2 of Table 2. After including fixed effects for each integer count of backward citations (i.e., a method to estimate the trend *within* that fixed level of generating citations), textual similarity is modestly *increasing* over time. Thus, the number and scope of cited references has expanded dramatically over time to now include citations that are much further afield from the citing patent than was the case in the past.

We also employ new, internal data from the USPTO to better understand the data generating process for patent citations. We show: that publicly available data misattributes the original source (i.e., applicant or examiner) for a small but significant proportion of patent citations; that a large and growing number of patent citations are made to pending patent applications (not issued patents) and therefore are missed by scholars; and that citations are frequently submitted by

[2] One might expect the opposite—that textual similarity will decline over time due to an ever-expanding pool of prior art. Discussions with patent attorneys indicate that patent applicants are only required to cite the prior art that is believed to be material to patentability, and are not required to cite prior art that is “cumulative” or common knowledge in the field.

applicants long after a patent application is filed, potentially challenging the view that patent citations are an indicator of direct knowledge inheritance.

The main contribution of this article is to provide *prima facie* evidence that the use of patent citations is increasingly generating significant measurement error for many academic studies, leading to tenuous and unsubstantiated inferences. Given the diverse empirical objectives of studies in economics that rely on patent citations, we conclude that a "one-size-fits-all" correction for counts of citations is unlikely to work reliably across each and every research context. As such, we argue that future research with patent data will require new measures and methods, and we encourage the development of measures beyond patent citations.

Section 2 describes many core assumptions that justify the use of patent citations in empirical analysis, and Section 3 introduces the data sources. In Section 4, we demonstrate the widespread violation of the core assumptions identified in Section 2. To illustrate the problem for a general audience, in Section 5, we replicate the analysis of foundational studies in industrial organization and show how a more detailed analysis of more comprehensive data supports substantially different conclusions. Section 6 discusses how patent families and cross-citing further complicate citation data and then examines potential corrections, such as collapsing citations by patent family, weighting citations, and/or using new metrics that do not derive from citations. Section 7 concludes.

2. Assumptions

■ Strong assumptions about the patent system have helped economists to shed light on what might otherwise be an intractable problem in the presence of severe limitations on data availability. As Krugman points out, "knowledge flows are invisible; they leave no paper trail by which they may be measured and tracked, and there is nothing to prevent the theorist from assuming anything about them that she likes" (Krugman, 1993; also cited in Jaffe, Trajtenberg, and Henderson, 1993). At the risk of overgeneralizing, the dominant view of patenting by economists is that it is a process by which inventors stand on the shoulders of giants to create inventions based on prior knowledge. Sometimes that knowledge is known explicitly to the inventor; other times it is known implicitly, as background information. The inventor reveals knowledge of what came before by disclosing it in a patent application. The patent examiner conducts an independent search to find prior art unknown to the inventor and grants a patent only after evaluating the applications' claims in light of the prior art.

The above narrative offers a way out of a difficult measurement problem by highlighting patent citations as a tool for measuring knowledge flows. Although scholars have long known that measures based on patents and patent citations are noisy in practice, the belief has prevailed that a large-sample analysis can model an unbiased "typical" case such that a valid signal will rise above the noise. However, the prevailing narrative also reveals many implicit yet commonly held assumptions about the data generating process for patent citations. In this section, we parse the narrative above to examine three broad categories of assumptions, which in later sections we demonstrate are systematically violated in ways that have important implications for empiricists.

□ **Assumptions about the duty of disclosure.** The patent applicant owes to the US patent office "a duty to disclose . . . all information known to that individual to be material to patentability" (37 C.F.R. 1.56). The duty is not merely a negative obligation to abstain from fraud, but rather an *affirmative* duty to disclose "all information" known to the individual. If an applicant fails to cite a reference it knew about and believed to be relevant, then a future infringer of the patent can allege that the applicant committed inequitable conduct and a court may rule the patent to be unenforceable, even if the patent is otherwise valid. Because failure to disclose can put the entire patent at risk, including any business strategies and products built around the patent, researchers assume that applicants generally comply with the duty of disclosure, and that they do so in a similar way.

Assumption A1. Applicants comply with the duty of disclosure in a uniform manner.

The duty of disclosure is not limited to inventors—it also applies to other employees of the firm, such as an attorney who is responsible for preparing the patent application. Nevertheless, the inability to determine the actual individual responsible for submitting a particular citation has led scholars to assume that applicant-submitted citations are identified and provided by the inventor. This assumption has been critical for many areas of research—the empirical literature on knowledge spillovers, for example, often uses patent citations to identify intellectual inheritance by inventors.

Assumption A2. Citations marked as applicant-submitted were identified by inventors (not legal or administrative personnel).

□ **Assumptions about the technological relatedness of patent citations.** Another core assumption is that the applicant decides to cite a patent based on an active evaluation of the technological relatedness of two inventions (Jaffe, Trajtenberg, and Henderson, 1993). For example, Lampe (2012) suggests that applicants may try to influence the examination process by withholding citations that they have reviewed and identified as strategically important. Whether one believes that applicants act in good faith or with strategic intent, the general assumption has been that applicants submit citations based on the technological content of each document.

Assumption B1. Applicants actively review and decide which patents to cite based on the technological content of the potentially cited patent.

Related to Assumption B1 is the implicit assumption that the technological relatedness of citations is stable over time. After all, scholars rely on validation work performed using now historical data when drawing conclusions about analysis performed on more modern data.

Assumption B2. A patent citation indicates a level of technological relatedness between the citing and cited patent, and that relatedness is independent of the filing date and issue date of the citing and cited patents.

□ **Assumptions about the patent examination process.** Given the complexity of the US patent system, scholars have had to make simplifying assumptions about the patent examination process in order to gain traction on other questions more fundamental to economics. The early literature generally assumed that patent examiners generated patent citations. More recently, however, scholars have been able to separate citations based on who submitted them—that is, the patent applicant or examiner (Alcacer and Gittelman, 2006). As such, research on knowledge spillovers assumes that inventors hold knowledge about the cited invention, except when the citation is added by the examiner (Thompson, 2006).[3] Scholars also have found that citations submitted by examiners are associated with greater patent impact than those submitted by applicants (Hegde and Sampat, 2009; Alcacer, Gittelman, and Sampat, 2009). To draw these conclusions, scholars have relied on one indicator variable from the USPTO.

Assumption C1. Citations indicated in publicly available data as cited by examiner were originally submitted by examiners.

An often-overlooked aspect of the patenting system is that patent examination and prosecution constitute a back-and-forth process that unfolds over time. Somewhat surprisingly, the

[3] Researchers often conflate references that are "cited by examiners" and references that are "added by examiners." We return to this point in Section 3.

examination process typically does not begin until a year or more after filing, and pendency varies wildly both over time and between examining divisions in the patent office (i.e., "art units"). Patent citations can be added to the record at any time during examination, but scholars have lacked data on when citations are submitted. Scholars therefore have been forced to assume a date for each citation—typically the filing date (or year) for the application.

Assumption C2. Patent citations originate on the filing date of the citing patent.

In summary, we again stress that researchers have relied on these assumptions not by choice, but due to limitations in available data. In the next section, we compile and construct several new sources of data to empirically evaluate the validity of Assumptions A through C.

3. Data

■ The data for our analysis derive from several sources. For the period from 1976 through 2004, we import data from the updated National Bureau of Economic Research (NBER) patent data file (Hall, Jaffe, and Trajtenberg, 2001).[4] For the period from 2005 through 2015, we collected patent citations and patent bibliographic data from the USPTO bulk data repository. Throughout both periods, we also collected new data from internal USPTO citation submissions forms for publicly available applications.[5] Our final sample includes data for 5,027,882, patents issued between January 6, 1976 and March 17, 2015, and 62,104,091 patent citations made by patents issued between January 6, 1976 and December 30, 2014. For most regressions, we restrict the sample to citations made by patents issued 2005 through 2014 in order to exploit the more expansive bibliographic data for patents issued during this time period. Table 1 summarizes definitions, source, and summary statistics for all variables. Given the large number of variables, and the technical importance of each variable for the particular analysis to which it is related, we define and discuss each variable when it appears in the analysis.

As an improvement over how citations have been counted in the past, we include 7.6 million citations to pre-grant publications. A change in the patent law (American Inventors Protection Act, 1999) led to most patent applications being published well before they actually issue, typically 18 months after filing. Consequently, a pre-grant publication can be cited as prior art. Although a patent's claims can change between filing and issuance, the technical description of the invention rarely changes, because patent law precludes the applicant from adding new matter to the application after it is filed (35 U.S.C. 132(a)). All citation-based measures that we know of, however, ignore citations made to patent publications because the USPTO does not update citation records to link references to publications forward to issued patents. The publication document is typically the only version of the application that is cited (not the patent that ultimately issues from the publication); failure to roll the publication citation reference forward to the resulting patent therefore misses these citations in most analyses. As of 2015, publication citations make up 25% of all citations.

4. Results from testing assumptions

■ The objective of this section is to evaluate the degree to which the assumptions set forth above are systematically violated. First, we show that applicant behavior has changed dramatically in recent years, leading to the emergence of extremely high-citing patents. Then, we show that the textual similarity of patent citations has fallen over time, and that the drop has been driven primarily by high-citing patents. Next, we show that the USPTOs designation of examiner citations includes citations that were first submitted by applicants. Further, many applicant (and

[4] As updated by Jim Bessen at sites.google.com/site/patentdataproject/ (data accessed June 1, 2017).

[5] We thank the USPTO for access to this data. These data were collected when one of the coauthors was Chief Economist at the USPTO. The internal data was purged of any nonpublic applications prior to being cleared for use.

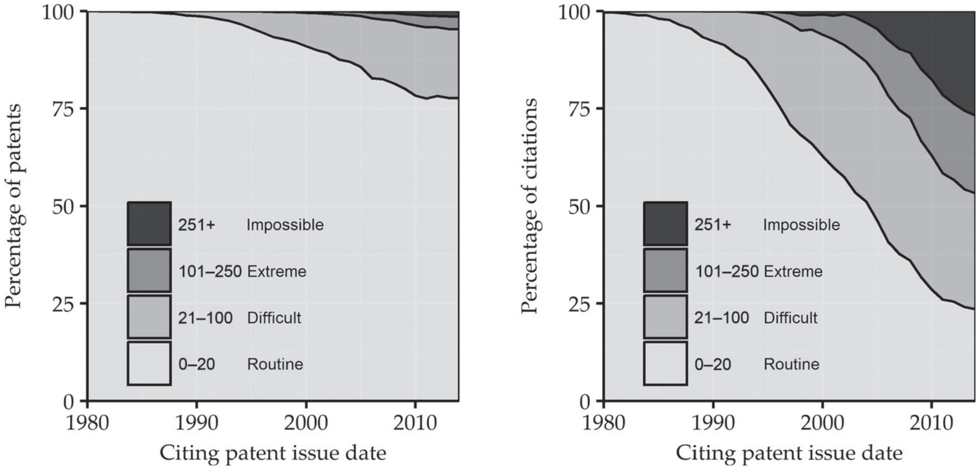
TABLE 1 Summary Statistics

Variable	Definition	Source	Minimum	Maximum	Median	Mean	Standerd Deviation
Similarity	Textual similarity between the citing and cited patent	Younge and Kuhn (2015)	0	100	25.98	26.90	21.17
Is Publication Citation?	Dummy = 1 if the cited patent was originally identified as a patent publication	USPTO Bulk Data System	0	1	0	0.18	0.38
Is Examiner Citation (Bulk Data)?	Dummy = 1 if citation was made by the patent examiner (XML data)	USPTO Bulk Data System	0	1	0	0.27	0.44
Is Examiner Citation (Internal Data)?	Dummy = 1 if citation was made by the patent examiner (Form data)	Internal USPTO data	0	1	0	0.28	0.45
Is Duplicate Citation?	Dummy = 1 if citation was made by both the patent examiner and the applicant	Internal USPTO data	0	1	0	0.05	0.21
Citation Lag	Difference in years between filing year of citing patent and filing year of cited patent	USPTO Bulk Data System	−10	39	9.01	10.44	7.70
Submission Lag	Time in years between application filing and citation submission	Internal USPTO data	0	15	0.98	1.40	1.43
Is Submitted At Application Filing?	Dummy = 1 if citation is submitted within 90 days of application filing	Internal USPTO data	0	1	0	0.28	0.45
Form Order	The index of the form on which the citation was submitted	Internal USPTO data	1	51	1.00	1.57	1.51
Is Self-Citation?	Dummy = 1 if the citing and cited patent were originally assigned to the same firm	Hanley (2015)	0	1	0	0.05	0.23
Backward Citation Count	Total number of citations made by the citing patent	USPTO Bulk Data System	1	6526	54.00	162.03	286.81
Family Size	Total number of patents in the family of the citing patent	USPTO Bulk Data System	1	478	2.00	8.25	27.41

Notes: The unit of the analysis is a citing-cited citation dyad. The sample includes all citations made by patents issued 2005-2014 found in either USPTO bulk data or USPTO internal data.

FIGURE 3

AREA PLOT OF CITATIONS BY BACKWARD CITATION COUNT OF CITING PATENT OVER TIME



(a) Percentage of patents **(b)** Percentage of citations

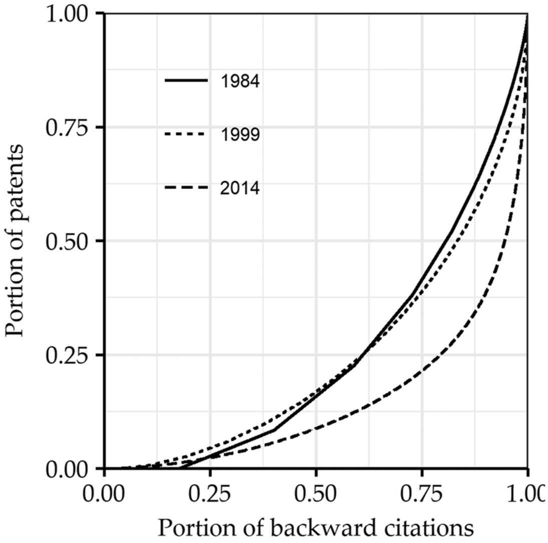
examiner) citations are submitted long after the application is filed, and citations exhibit substantial heterogeneity in technological relatedness across a number of dimensions.

□ **Duty of disclosure.** Assumptions A1 and A2 suggest that applicants comply with the duty of disclosure in a uniform manner and that citations submitted by applicants were in fact identified by *inventors* (as opposed to an attorney or other actor). We now assess the extent to which that is the case. As shown in Figure 3a, most patents include 20 or fewer citations—a level we label as "routine." The proportion of routine patents has fallen over time, but as of 2014, more than 75% of patents still have 20 or fewer citations. As the number of citations rises from 20 to 100, discussions with patent attorneys and examiners suggest that reviewing each citation becomes much more "difficult" given time constraints. Although there are cases of particularly complex technologies where documentation for 50 or even 100 citations may be required, reviewing every citation in such a list imposes a substantial burden on the patent examiner. As seen in Figure 3a, the portion of patents citing a "difficult" number of references has grown steadily over time, from 0 in 1980 to approximately 20% in 2014. Next, we label as "extreme" patents with backward citation counts of 101 to 250 citations, because it is difficult to imagine anyone (at either the applicant's office or the patent office) reviewing that many documents in detail. Finally, we label backward citation counts of 251 or more as simply "impossible" given the time constraints of examiners. We also note that "extreme" and "impossible" levels of citations strongly suggest that those citations were not identified by inventors, for doing so would leave little (to no) time for actual invention.

Although patents that cite "difficult" or "extreme" numbers of references form a relatively small percentage of all patents, they contribute disproportionately to the number of backward citations generated in a particular year, and that influence is growing quickly. Figure 3b shows the percentage of patent citations made in a given year attributable to patents in each category (routine, difficult, extreme, impossible). By 2014, patents that cite an "extreme" or "impossible" number of citations are responsible for more than 46% of all patent citations, even though they comprise less than 5% of all patents issued in that year. In contrast, the 75% of patents that make a "routine" number of 20 or fewer citations, are responsible for less than 24% of the total number of citations. Figure 4 illustrates this change starkly by plotting Lorenz curves for backward citation counts across patents issued in 1984, 1999, and 2014.

FIGURE 4

LORENZ CURVES OF CUMULATIVE PROPORTION OF PATENTS BY BACKWARD CITATION COUNT, BY YEAR



We conclude that the data generating process for patent citations has changed over time in a way that now systematically and substantially violates Assumptions A1 and A2. The empirical evidence is clear that there is great heterogeneity in the way that citations are generated now, and that a pooled analysis of all citations will overrepresent a relatively small portion of the population of citing patents. In Section 6 and in an online web Appendix, we argue that several institutional features in the patenting regime lead some applicants to overcite the prior art (Kuhn, 2010), which in turn leads to a proliferation of low-information citations. Specifically, the duty to disclose prior art, in conjunction with the rise in the strategic importance of patent families, has led certain applicants to mechanically "cross-cite" anything and everything they have ever come across.

□ **Technological relatedness of patent citations.** To better understand how the technological relatedness of citations has changed over time, we leverage the work of Younge and Kuhn (2015), who compute and validate a textual similarity measure for all 14 trillion possible pairwise combinations of USPTO patents. The *Similarity* variable measures the textual similarity between the technical specifications of two patents, where a higher value indicates a greater level of similarity between the citing and cited patent. Although textual similarity is not a perfect representation of technological relatedness—for technologies can be related in implicit ways that are not represented in the text—we show in related work that the computed similarity measure correlates strongly with both expert and lay person evaluations, and that it also predicts such characteristics as shared patent class and patent family (Younge and Kuhn, 2015).

To compute *Similarity*, each patent is given a vector that positions the patent in a vector space with more than 700,000 dimensions, one for each substantive word that appears in the text of an issued patent. Each vector is determined via a term-frequency inverse-document-frequency (TFIDF) approach that up-weights rare words and down-weights common words. Although the TFIDF approach is relatively simple, a benchmarking study on patent data shows that TFIDF performs well in situations of long, extended, and highly granular text (Shahmirzadi, Lugowski, and Younge, 2018). The similarity measure is then computed for each pair of patents

FIGURE 5

MEAN AND INTERQUARTILE RANGE OF SIMILARITY FOR CITATIONS BY COUNT OF BACKWARD CITATIONS BY CITING PATENT

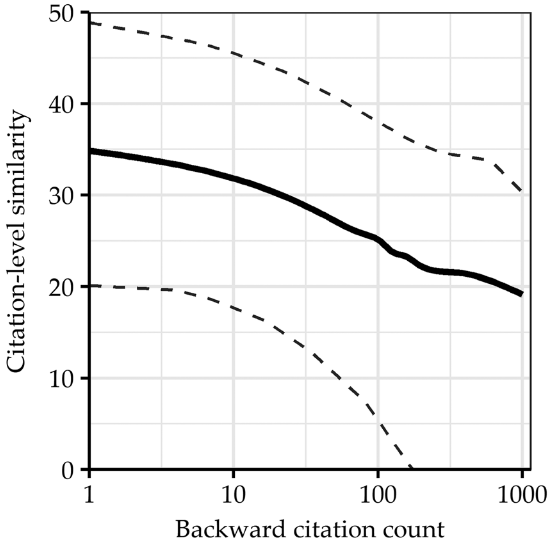


TABLE 2 OLS Model of Citation Similarity Over Time

	(1)	(2)	(3)
Citing Patent Issue Year	−0.282***	0.014***	0.028***
	(0.003)	(0.003)	(0.003)
Backward Citation Count (logged)			−2.887***
			(0.015)
Constant	36.512***		37.404***
	(0.090)		(0.089)
Backward citation count fixed effects	*No*	*Yes*	*No*
Observations	1,000,000	1,000,000	1,000,000
R^2	0.010	0.051	0.044

Notes: Dependent variable is citation similarity. Sample includes one million randomly selected patent citations made by patents issued 1976–2014. Model 2 includes fixed effects for the count (not categories) of backward citations. In unreported results, the coefficients are similar to Model 2 and Model 3 when running a subsample analysis limited to each of the four categories of patents identified in Figure 1 (i.e., having 0–20, 21–100, 101–250, 250+ citations). Standard errors in parentheses. Two-tailed tests: ***p < 0.001, **p < 0.01, *p < 0.05.

by determining the angular distance between them using a cosine measure between the patents' two vectors. On a scale of 0 to 100, citations in our sample have a mean similarity value of 27.9.

Based on textual *Similarity*, citations generated by highly citing patents appear to be particularly uninformative. Figure 5 plots the mean and interquartile range of *Similarity* for patent citations by the number of backward citations made by the citing patent. Whereas the citation similarity for patents citing one, two, or three references is 34, the mean citation similarity for patents citing hundreds of references falls to 20. The low similarity of citations made by high-citing patents, coupled with the small but increasing frequency of such patents, appears to be driving the overall decline in citation similarity. Table 2 presents estimates of ordinary least squares regressions of citing patent issue year on citation similarity. As shown in Model 1, citation similarity has declined over time at a rate of about 0.282 points per year ($p < 0.001$), a result consistent with Figure 2. Model 2 adds level-by-level fixed effects for the exact number of backward citations made by the citing patent. With fixed effects included, citation similarity is nearly constant over time, increasing slightly at a rate of 0.014 points per year. ($p < 0.001$). That is, there is little difference in citation similarity between two patents issued at different points in time that have the same number of backward citations.[6] Controlling for the log of the backward citation count produces a similar result, as shown in Model 3. Figure 2 presents the results graphically by plotting the mean similarity of patent citations over time along with a dotted line plotting the estimate of Model 2 from Table 2.[7]

We conclude from the above analysis that the data generating process for patent citations has changed over time in a way that systematically and substantially violates Assumption B2, which assumes a stable level of technological relatedness between citing and cited patents. Further, the results suggest that the decline in citation similarity is driven by the small but growing portion of patents that cite very large or extreme numbers of relatively unrelated references, and not by a secular trend in the textual similarity of all citations. As such, we also conclude that for many patent citations, the evidence suggests a violation of Assumption B1, which assumes that applicants actively decide which patents to cite based on a review of the technological content of each patent.

The implications of the highly skewed mixture problem are stark: measures that count the number of forward patent citations from current years are capturing the contribution of a very small fraction of all patents. Even though the data generating process began to change most significantly after 2005, it is important to note that the observed effects are not themselves isolated to years after 2005. Because backward citations are "caught" as forward citations by patents earlier in time, even measures for patents in 1995 or 2000 are now being affected by the new citation patterns. Moreover, an increase in volume of patents issued per year is compounding the problem. Whereas the number of patents issued per year increased slowly from 75,000 in 1976 to about 150,000 in 2009, it doubled to 300,000 only five years later and continues to grow. Together, the rise in patent volume and the change in citation similarity suggest that citations generated today are increasingly far more numerous and far less informative than in years past.[8]

Importantly, the reduction in similarity over time is driven by applicant-submitted, not examiner-submitted, citations. As shown in Figure 6, the percentage of backward citations submitted by examiners quickly approaches zero as the number of backward citations increases. The precise mechanisms behind this trend are difficult to pin down with certainty. However, Section 6 provides an overview of the institutions and incentives that give rise to citations, whereas the online web Appendix provides a more in-depth discussion supported by evidence suggesting that the increase in quantity and reduction in similarity of applicant citations is driven by cross-citing.

□ **Patent examination process.** Above, we show that applicants are responsible for generating the flood of citations in recent years, not examiners. That is likely due to the fact that examiners search for prior art with a particular goal in mind—finding references to support a rejection. Patent examiners are evaluated and promoted primarily on their throughput (Marco, Sarnoff, and deGrazia, 2019). However, they are evaluated in part based on the quality of their examination, and therefore the quality of their search. Because examiners are time-constrained and operate under a quality control system, they are incentivized to find relevant references, and disincentivized from spending time listing irrelevant ones. In contrast, applicants face strong incentives to disclose any information known to them and essentially no disincentive to cite irrelevant references. We believe such incentives explain why patent examiners submit citations that are, on average, more technologically related than those submitted by applicants (Figure 4).

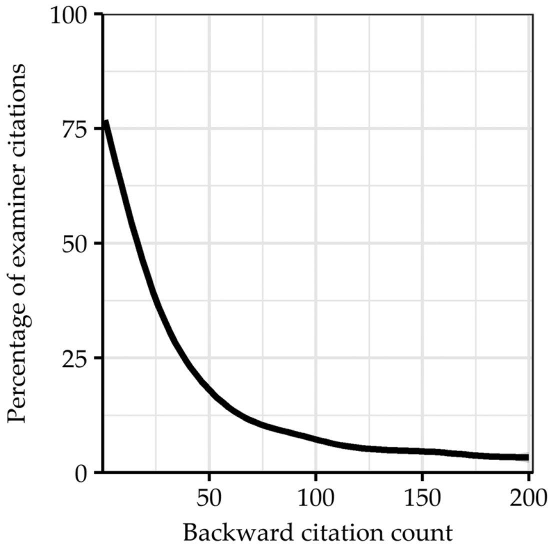
[6] One might be concerned that the effects in Column 2 are only on average within a level of backward citations. Nevertheless, the effect of *Citing Patent Issue Year* is positive within the highest categories of Figures 3A and 3B.

[7] Running the analysis at the cited patent level rather than the citation level produces results consistent with Table 2.

[8] Although scholars often include year fixed effects in an attempt to control for such changes, the dyadic link between two patents at different points in time undermines such a strategy. We discuss including year fixed effects in the third subsection of Section 6.

FIGURE 6

PERCENTAGE OF EXAMINER CITATIONS BY PATENT-LEVEL BACKWARD CITATION COUNT



Although the similarity of applicant-submitted citations fell steadily from 2005 to 2014, the similarity of examiner-submitted citations has remained more stable.

A reduction in the technological relatedness in applicant citations over time could suggest that applicant citations are no longer informative—a conclusion bolstered by publicly available USPTO data. Internal USPTO data, however, reveals a somewhat different picture. The USPTO provides information to the public as to whether an applicant or an examiner submits a citation reference[9]; the indicator, however, is often misunderstood by researchers. In particular, the USPTO identifies a citation as being submitted by the examiner *irrespective of whether the same citation was previously submitted by the applicant*. That is, if the patent applicant first submits a citation and then later the examiner resubmits the same citation (perhaps as a result of an independent search), the bulk data will identify the citation as having been submitted by the examiner (not the applicant).[10] Table 3 presents summary statistics comparing citation attribution in the USPTO bulk data versus internal data. Of 7.6 million citations identified in the USPTO bulk data as being submitted by the patent examiner, about 506,000 (or around 6.6%) were actually submitted first by the patent applicant and only later by the patent examiner.

To evaluate how confusion or misattribution of citations between examiners and applicants might affect empirical analysis, we constructed two indicators for whether a citation is examiner-submitted: *Is Examiner Citation (Bulk Data)?* is an indicator as to whether the citation was flagged as examiner-submitted in the public bulk data file; *Is Examiner Citation (Internal Data)?* is an indicator from internal USPTO records that is coded 1 if the citation was first submitted by the examiner, and otherwise 0. When both the patent applicant and the patent examiner submit the same reference (regardless of who submitted it first), we refer to that situation as a "duplicate citation," as indicated by the variable *Is Duplicate Citation?* For control variables, we calculated *Citation Lag* as the difference in years between the filing year of the citing patent and the filing year of the cited patent[11]; *Submission Lag* as the time in years between the filing date of the citing

[9] See www.patentsview.org/downloads, tables usapplicationcitation and uspatentcitation.

[10] The Manual of Patent Examination and Procedure (MPEP) states that examiners should "not list references which were previously cited by the applicant," but there is no mechanism to correct for duplicate submissions. An indication that a reference was "cited by examiner" is therefore technically, correct even if it was in fact first cited by the applicant.

[11] We use filing year rather than date because the data do not include filing date for patents issued before 2005.

TABLE 3 The Attribution of Patent Citations—Public Data vs. Internal Data

	As Measured By XML Data		
As Measured By Internal USPTO Data	Applicant	Examiner	Total
Applicant	19,218,466	28,021	19,246,487
Examiner	8,133	6,891,903	6,900,036
Both (applicant first)	41,091	506,419	547,510
Both (examiner first)	10,836	122,456	133,292
Total	19,289,260	7,670,385	26,959,645

Notes: This table compares the attribution of patent citations between USPTO public XML bulk data and USPTO internal data. The sample includes each citation made by a patent issued 2005-2014. The Applicant and Examiner columns indicate the attribution of the citation in the USPTO XML files. The Source rows indicate how the citation appears in the raw USPTO citation submission forms. For example, a citation listed as "Applicant first" was initially submitted by an applicant but was later submitted by the patent examiner, according to the raw USPTO citation submission forms. Each citation-source dyad is represented once in the table.

TABLE 4 OLS Estimates of Citation Similarity by Citation Characteristic

	(1)	(2)	(3)	(4)	(5)	(6)
Is Examiner Citation?	4.590***	6.207***	4.292***	5.809***	6.020***	−2.227***
(Internal Data)?	(0.049)	(0.053)	(0.049)	(0.052)	(0.057)	(0.057)
Is Duplicate Citation?	11.933***	12.502***	11.631***	12.402***	12.363***	
	(0.098)	(0.098)	(0.098)	(0.098)	(0.099)	
Submission Lag (Years)		−1.268***			−0.742***	
		(0.016)			(0.021)	
Form Order			−0.872***		−0.417***	
			(0.014)		(0.016)	
Is Submitted At				3.355***	1.723***	
Application Filing?				(0.050)	(0.059)	
Constant	26.062***	27.403***	27.519***	24.788***	26.888***	
	(0.025)	(0.030)	(0.035)	(0.032)	(0.048)	
Citation count FE	*No*	*No*	*No*	*No*	*No*	*Yes*
Observations	1,000,000	1,000,000	1,000,000	1,000,000	1,000,000	1,000,000
R^2	0.020	0.026	0.024	0.025	0.028	0.054

Notes: Dependent variable is the similarity of citing-cited pairs of citations. Sample includes one million randomly selected citations made by a patent issued between 2005 and 2014. Model 6 includes fixed effects for the number of citations submitted for the citing patent by the citing party (i.e., applicant or examiner), controlling for the number of citations submitted, applicants submit more similar citations. Standard errors in parentheses. Two-tailed tests: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

patent and the date on which the citation was submitted as indicated on the USPTO citation form; *Filing Citation* is a dummy variable indicating whether the citation was submitted within the first 90 days of the filing of the citing patent; *Form Order* indicates the sequential order of the USPTO citation form containing the citation (the first form submitted is number 1, the second form is number 2, etc.)—applicant and examiner forms are numbered independently.

Table 4 presents results from five models that investigate the impact of discrepancies in the measurement of applicant-submitted versus examiner-submitted citations. Each citation (i.e., observation in the analysis) must be submitted first by either the examiner or the applicant, as indicated by the dummy variable: *Is Examiner Cite (Internal Data)?* Each citation is classified as a "duplicate citation" if it is eventually submitted by both the applicant and the examiner, as indicated by *Is Duplicate Citation?* The constant term therefore corresponds to the effect where citations are submitted only by the applicant. Across all models, applicant-only citations exhibit similarity values of between 26.1% and 27.5% ($p < 0.001$); examiner citations are more similar than applicant citations by between 2.2 and 6.2 percentage points ($p < 0.001$); and duplicate citations are the most similar category, at between 11.9 and 12.5 percentage points ($p < 0.001$)

more similar than nonduplicate applicant-submitted citations. These results show that a small but significant percentage of citations identified in the bulk data as examiner-submitted (500,000 out of 7.6 million) were in fact first submitted by applicants—a systematic violation of Assumption C1. Moreover, these duplicate citations are particularly relevant, with substantially higher technological similarity on average than citations submitted by either examiners or applicants alone.

Table 4 also reveals new information with respect to the timing with which citations are submitted. Many applicant citations are submitted long after the application is filed, in successive rounds of submitting forms to the USPTO, where citations become less and less similar. Each additional year between patent filing and citation submission is associated with a reduction in *Similarity* of between 0.74 and 1.26 percentage points ($p < 0.001$), and each successive form (i.e., batch of citations received into the system at the USPTO) is associated with a reduction in citation *Similarity* of between 0.74 and 1.23 percentage points ($p < 0.001$). Citations submitted at filing are between 1.72 and 3.36 percentage points ($p < 0.001$) more similar than later-submitted citations. The prevalence of late citations violates Assumption C2, that patent citations originate on the filing date of the citing patent. Late citation does not necessarily imply a lack of knowledge inheritance,[12] but late submission does suggest that knowledge inheritance did not occur through the patent—which casts some doubt on the use of patent citations as a "paper trail" for knowledge flows.

Finally, we observe that the higher similarity associated with examiner citations seems to be driven entirely by two facts: first, that citation similarity declines with the number of citations submitted; and second, that applicants submit many more citations than examiners. Model 6 includes fixed effects for the number of citations submitted and an indicator for the submitting party (i.e., the applicant or the examiner). Controlling for the number of citations submitted, applicants submit *more similar* citations than examiners. These results suggest that researchers should be very careful in making inferences about the quantitative impact of examiner and applicant citations; the inferences are likely to be very sensitive to the econometric specification and the quality of the data.

Across all results, the evidence suggests that the data generating process for patent citations changed significantly over the last 10 years. As such, we argue that it is no longer appropriate to rely on the assumptions in Section 2 for high-quality empirical work. The number of citations generated per year has increased dramatically, the distribution of backward citations has become much more skewed, the average citing-cited similarity has decreased significantly, and publicly available data is misleading as to the origination of citations (i.e., applicant vs. examiner) for the cases that are mostly likely to be influential to the examination process. In the next section, we illustrate some of the implications of our findings.

5. Biased results from citations-based measures

■ In this section, we replicate the estimation strategy of several foundational articles and show how a similar analysis on newer and more comprehensive data would arrive at very different conclusions today. The replicated analyses use patent citation data to measure the geographic colocation of knowledge spillovers, the private value of patents, and the market value of firms. The goal is not to replicate the earlier study in exactitude, to test whether the study was executed correctly, or to argue that the studies were incorrect at the time. Instead, our purpose is to show that changes in the quality, quantity, and interpretation of citations data affect the conclusions one would arrive at today, and to show that using such methods today may no longer be valid.

□ **Geographic localization of knowledge spillovers.** Jaffe, Trajtenberg, and Henderson (1993, henceforth, "JTH") used patent citations to demonstrate that knowledge spillovers are geographically localized to a greater extent than one would expect from the distribution of

[12] For example, the applicant may have known of a technology but have been unaware of its description in a patent.

inventive activity. Their identification strategy matched a sample of patent citations (i.e., citations going to an "originating" set of patents), to a counterfactual sample of similar noncitations (i.e., a selected set of patents that did *not* cite the "originating" patents), and then compared the difference between the localization of citations and the localization of matched noncitations.[13] The econometric design of JTH already has been challenged on the basis of their selection of control patents (Thompson and Fox-Kean, 2005). Setting aside questions about the quality of the control, however, it is important to stress that the validity of the JTH method also depends on the core assumption that patent citations are an unbiased indicator of knowledge spillovers. If citations do not indicate knowledge spillovers in the first place, then no degree of improving the matching procedure can unbias the result. The great majority of empirical research in the tradition of JTH continues to assume that patent citations represent actual knowledge spillovers, implicitly relying on a combination of Assumptions A2, B1, and C1.[14]

In this section, we replicate the JTH methodology on newer data and test the assumption that patent citations are an unbiased indicator of knowledge spillovers. We do this by using newer data and comparing the geographic localization rates of a JTH analysis, to an analysis of citations that are more likely to actually be spillovers.[15] Our replication proceeds as follows. First, we set up two analyses of patent citations made to originating patents every five years from 1975 to 2010: (i) a "conventional analysis" that selects all citations to originating patents (plus a standard JTH-matched counterfactual set of noncitations); and (ii) an "improved analysis" that selects only those citations that are more likely to represent a true knowledge flow (plus a standard JTH-matched counterfactual set of noncitations). For the conventional analysis, we selected all citations by US firms, excluding self-citations. For the improved analysis, we selected all citations by US firms, but then excluded several groups of citations, including: (i) self-citations, (ii) citations identified in the USPTO *internal* data as being submitted by an examiner, (iii) citations made by patents citing more than 20 references, (iv) citations submitted more than one year after the citing patent was filed, (v) citations submitted after the initial submission of prior art, and (vi) extremely high *Similarity* citations above a value of 97.5.[16] For both analyses, we limit forward citations to a five-year window,[17] stop the analysis in 2010, and include citations to pending patent applications.

To better understand how well the selection procedures above identify knowledge spillovers, Figure 7 plots the distribution of textual *Similarity* between citing and cited patents for the conventional analysis, and for the improved analysis. The difference between the two distributions is stark: at the low end of the distribution, the conventional approach selects many more citations of very low *Similarity* than the improved approach[18], at the high end of the distribution, the conventional approach selects many more extremely high *Similarity* citations than the improved approach.[19] These differences support the argument that the "improved approach" adjusts the sample in the right direction, making it more likely that the "improved approach" captures more true knowledge spillovers, and less noise.

[13] The measurement of "colocation" requires a geographic boundary. JTH estimated colocation rates at the national, state, and consolidated-Metropolitan Statistical Area (MSA) levels. We estimated colocation at the MSA level.

[14] It should be noted that Thompson (2006) differentiates between citations, assuming that *applicant*-submitted citations can proxy for spillovers whereas *examiner*-submitted citations can proxy for control.

[15] Thompson and Fox-Kean (2005) challenged the validity of the control for JTH, while holding treatment fixed; our replication challenges the validity of treatment, while holding control fixed. As such, we examine the effect of changing patent citation, but we do not attempt to estimate a ***true*** effect for knowledge spillovers.

[16] Manual inspection indicates that citations with a *Similarity* score over 97.5 either share an inventor (and thus are not "spillovers") or are self-citations that were missed due to ambiguities in patent assignment and firm disambiguation.

[17] We select a five-year forward window and stop the analysis in 2010 in order to limit the effect of rightward truncation bias. If all citations are used, then the distribution of citation time-lags would trend downward by cohort, as cohorts advance from 1976 to the present, due to the increasing rightward truncation of patent citations.

[18] As discussed in the first subsection of Section 4 and the online web Appendix, very low *Similarity*; citations are disproportionately generated by applicants that overdisclose and submit citations that the firm comes across tangentially.

[19] Extremely high *Similarity* citations typically share an inventor or are from the same firm, and thus are not spillovers.

FIGURE 7

SIMILARITY DENSITY FOR A CONVENTIONAL AND IMPROVED SELECTION OF CITATIONS TO INDICATE SPILLOVERS

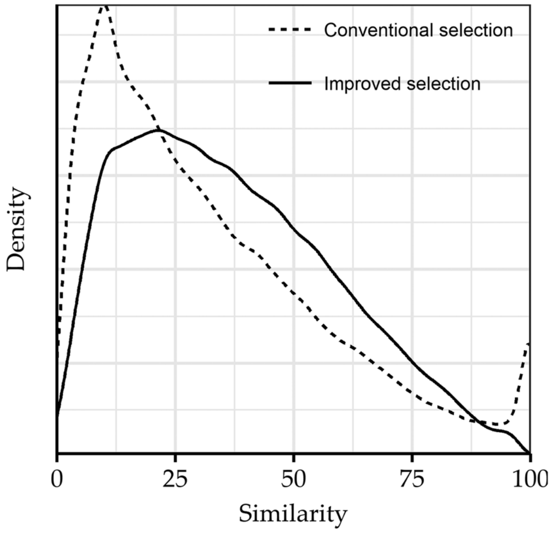
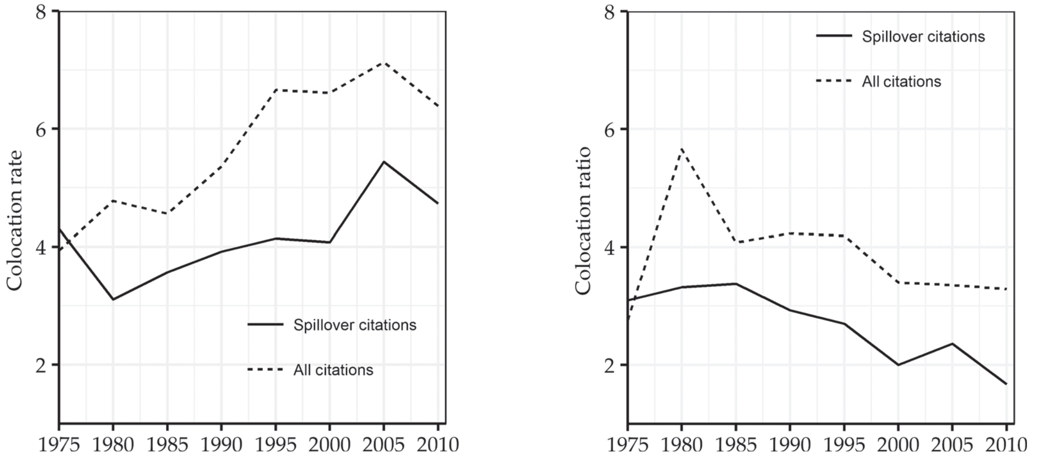


FIGURE 8

GEOGRAPHIC COLOCATION AND KNOWLEDGE SPILLOVERS



(a) Rate of geographic colocation of citations, for all citations versus likely spillover citations

(b) Average Treatment Effect (ATE Ratio), for all citations versus likely spillover citations

To better understand why the estimation of spillovers from patent citations may be biased, we compare results from the conventional analysis to results from the improved analysis. If the conventional analysis is not biased, then removing noise should not affect the mean results, and any attenuation of the effect toward zero would challenge the view that the conventional analysis captures a true spillover effect. Figure 8 plots the results of this comparative analysis. In Figure 8a, we observe that the rate of geographic colocation for patent citations in the conventional analysis trends upward over time, going from a colocation rate at the MSA level of 3.9% in 1976, to 6.4% at the end of the analysis, for a proportional increase of 62%; the colocation rate for citations in the improved analysis, however, remained more consistent, going from 4.3% in 1976, to 4.7% at the end of the analysis, for a proportional increase of 10%. The divergence in Figure 8a suggests that bias in the estimation of colocation rates may have increased over time.

To fully understand the substantive effect of these results, one must consider the relative rates in the context of the matched research design. One way to visualize the average treatment effect (ATE) is to plot the colocation rate of citations relative to the colocation rate of noncitation controls (i.e., the "*ATE Ratio*"). An *ATE Ratio* of one (i.e., 1 : 1) indicates no effect—the colocation rate of treatment citations is the same as the colocation rate of control noncitations. Figure 8b plots the *ATE* Ratio for cohorts over time, starting at a level of 1 (the level of no effect). The *ATE Ratio* for the conventional analysis (again at MSA level) varies over time, but averages around 3.9.[20] For the improved analysis, however, we find that the *ATE Ratio* starts at 3.1, but then falls to a low of 1.7. Overall, we conclude that the JTH approach likely overstates the true magnitude of the geographic localization of knowledge spillovers in today's institutional environment.

□ **Private value of patents.** Scholars have long used forward patent citations as a proxy for patent value (Trajtenberg, 1990b). For example, Hegde and Sampat (2009) recently studied the relationship between forward patent citation counts and whether the applicant pays the three renewal fees imposed by the patent office at 4 years, 8 years, and 12 years after a patent is issued (an important indicator of patent value). By differentiating between applicant-submitted and examiner-submitted citations (Alcacer and Gittelman, 2006; Alcacer, Gittelman, and Sampat, 2009) they find that examiner-submitted citations are more predictive of renewal fee payments than applicant-submitted citations.

In Table 5, we perform a similar analysis. Like Hegde and Sampat (2009), we select three cohorts of patents for analysis. The passage of time, however, allows us to select more recent years and analyze more complete data (i.e., patents issued in 2000, 2004, and 2008, instead of patents issued in 1992, 1996, and 2000). Unlike Hegde and Sampat (2009), we count applicant and examiner citations using the more accurate, internal data from the USPTO that is not publicly available. Like Hegde and Sampat (2009), we find that on the margin, an examiner citation is a stronger indication of renewal fee payment than an applicant citation, for the linear specifications (Models 1–3). In Model 1, each examiner citation is associated with a 0.2 percentage point increase in the likelihood of paying the first renewal fee ($p < 0.001$), whereas each applicant citation is associated with only a 0.1 percentage point increase ($p < 0.001$). Unlike Hegde and Sampat (2009), applicant citations in our replication are a highly significant predictor of renewal fee payments across all models, albeit with a lower magnitude than examiner citations for the linear specifications. The statistical significance of this finding may not be surprising, given our larger sample size and more complete data. The use of internal USPTO data, however, may also play a role. Earlier, we demonstrated that examiners frequently receive credit (in public data) for highly similar and influential citations that were first submitted by applicants.

The change in the pattern of patent citations over time, however, highlights an important problem in comparing the magnitude of the effect of applicant-submitted and examiner-submitted citations: because the number of applicant-submitted citations has increased dramatically in recent years, comparing the marginal effect of a *single* forward citation is increasingly an "apples to oranges" comparison. In Models 4, 5, and 6, we log the forward citation counts to better account for such differences and find that applicant citations are substantially *more* predictive (on a percentage basis) of paying renewal fees than examiner citations. For example, in Model 6, a 100% increase in forward examiner citations is associated with a 1.5 percentage point increase in the probability of paying the third renewal fee ($p < 0.001$), whereas a 100% increase in forward

[20] JTH use an amalgam of region definitions. Because the rate of colocation increases as boundaries increase (until all observations are colocated), results from JTH at the consolidated Standard Metropolitan Statistical Area (SMSA) level are at a larger scale with a higher expected rate of colocation than this replication. Nevertheless, JTH report results with a ratio of colocation rates of around 7:1.

TABLE 5 OLS Estimates of Paying the Renewal Fee at 4, 8, and 12 Years

	Renewed at:			Renewed at:		
	4 years? (1)	8 years? (2)	12 years? (3)	4 years? (4)	8 years? (5)	12 years? (6)
2004	0.022***	−0.024***		0.029***	−0.031***	
	(0.001)	(0.002)		(0.002)	(0.003)	
2008	0.007***			0.016***		
	(0.001)			(0.002)		
2000 X Examiner Citations	0.002***	0.002***	0.003***	0.011***	0.010***	0.016***
	(0.0002)	(0.0003)	(0.0003)	(0.001)	(0.002)	(0.002)
2000 X Applicant Citations	0.001***	0.001***	0.001***	0.027***	0.035***	0.040***
	(0.00004)	(0.0001)	(0.0001)	(0.001)	(0.001)	(0.001)
2004 X Examiner Citations	0.002***	0.004***		0.013***	0.026***	
	(0.0002)	(0.0002)		(0.001)	(0.001)	
2004 X Applicant Citations	0.001***	0.001***		0.023***	0.036***	
	(0.0001)	(0.0001)		(0.001)	(0.001)	
2008 X Examiner Citations	0.004***			0.021***		
	(0.0002)			(0.001)		
2008 X Applicant Citations	0.001***			0.024***		
	(0.0001)			(0.001)		
Citation counts logged?	*No*	*No*	*No*	*Yes*	*Yes*	*Yes*
Observations	479,249	280,613	108,105	479,249	280,613	108,105
R^2	0.034	0.040	0.047	0.040	0.048	0.055

Notes: Dependent variable is whether the renewal fee is paid at 4, 8, and 12 years after a patent is issued. All models are OLS linear probability models with (corrected) main class fixed effects. The sample includes all citations made by a patent issued between 2005 and 2014. The forward citation counts are unlogged in columns 1–3 and logged in columns 4–6. The data include citations from USPTO bulk data and USPTO internal data. Some citations appear in one of these data sets but not in the other. In unreported results, we reran the analysis while limiting the sample to include only citations made by patents for which every backward citation is included in both data sets. For each model, the sign and significance of each coefficient is the same, and the magnitudes of most coefficients are quite close. Standard errors in parentheses. Two-tailed tests: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

applicant citations is associated with a 4.0 percentage point increase. R^2 also increases when logging forward citation counts, suggesting a better fit of the model to the data.

From this replication, we conclude that the literature has been too quick to dismiss the importance of applicant-submitted citations (Cotropia, Lemley, and Sampat, 2013). Instead, after accounting for the true origin of a citation, applicant-first-submitted citations are an important source of information, both for the examination process at the USPTO and for research. Researchers should also take care to distinguish between different types of applicant citations, where possible, rather than assuming the same average or marginal effects hold across a sample.

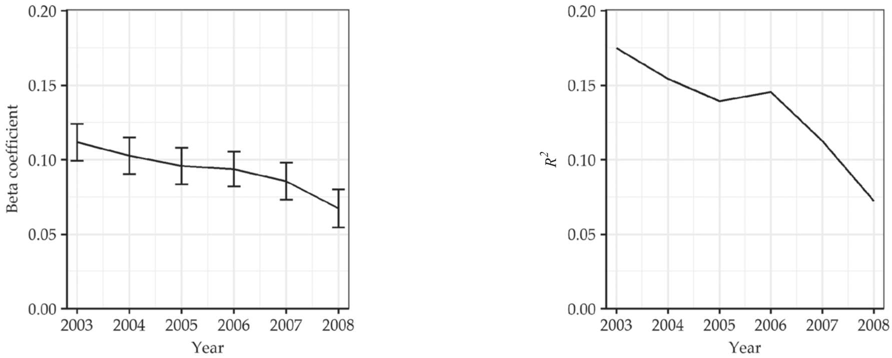
□ **Market value of firms.** Hall, Jaffe, and Trajtenberg (2005) demonstrated that a count of patents, weighted by a count of forward citations, can be a strong predictor of firm market value. In this section, we replicate their estimation strategy on more recent data and evaluate whether their findings continue to hold. We selected patents from US firms that applied for patents between 2003 and 2007 and then linked those firms to financial data from Compustat.[21] We identified all issued patents filed by those firms in those years, as well as all citations made to those patents within a five-year window following the filing date of each patent application.[22] Following Hall, Jaffe, and Trajtenberg (2005), we then calculated a depreciated count of forward citations for each firm-year, by discounting citations received in years after filing at a yearly depreciation rate

[21] We thank Hanley (2015) for data on firm disambiguation to link patents to firms to financial data from Compustat.

[22] We also tested six-year and seven-year windows and obtained similar results. A window of less than 5 is increasingly left-censored due to pendency between filing and grant; more than seven is increasingly right-censored due to availability of data.

FIGURE 9

FORWARD PATENT CITATIONS AND FIRM MARKET VALUE



(a) Magnitude of association between five-year patent forward citations and *Tobin's Q*

(b) Proportion of variance explained in *Tobin's Q* by five-year patent forward citations

of 0.15.[23] We calculated the ratio of market value to asset value (*Tobin's Q*) as our dependent variable (DaDalt, Donaldson, and Garner, 2003).[24] Finally, we estimated equation (1) for each firm-year[25]:

$$ln\big(Tobin's\ Q_{it}\big) = \alpha + \beta\ ln\left(\frac{CiteCount_{it}}{AssetVal_{it}}\right) + \epsilon_{it}. \tag{1}$$

In the early range of our analysis, we find that beta coefficients (Figure 9a) for the strength of the relationship between forward citation count and firm market value were at or above 0.10 values, similar to results in Hall, Jaffe, and Trajtenberg (2005). The strength of the relationship, however, declines considerably over time, from a coefficient of 0.11 in 2003 to a coefficient of 0.06 in 2008. The explanatory power of the model also falls (Figure 9b), from an R^2 of 0.18 in 2003 (again, similar to results reported by Hall, Jaffe, and Trajtenberg, 2005) to a low of 0.07 in 2008.

From this replication, we conclude that the relationship between forward citation counts and firm market value appears to have attenuated considerably over time. Although it may be too soon to say exactly *why* the relationship between forward citation counts and firm market value has lost such predictive power over time,[26] the changes that we document certainly are not helping. In theory, the relationship between forward citation counts and firm market value could be restored at least in part by selecting or weighting citations. However, in unreported results, we were unable to identify such a correction. Moreover, such an approach risks introducing new biases, because which citations to select or how to weight them is not readily apparent, in contrast to the principled selection criteria in the previous two replications.

---

[23] Forward citation counts suffer from rightward truncation. Whereas Hall, Jaffe, and Trajtenberg (2005) used prior citations to predict censored citations, changing citation patterns suggest such an approach is no longer valid. Instead, we restrict forward citation counts to a five-year window and end our analysis in 2008 due to truncation.

[24] DaDalt, Donaldson, and Garner (2003) demonstrate that calculation of *Tobin's Q* can induce sample-selection bias due to data unavailability and row-wise deletion from the sample. We therefore calculated a simplified approximation of Q, calculated from Compustat variables as $Q = ((PRCC_F * CSHO) + AT - CEQ)/AT$.

[25] Subscripts for firm i, year t. Our estimation accords with the bottom half of column three, Table 2, in Hall, Jaffe, and Trajtenberg (2005).

[26] For example, the attenuation coincides generally with the Great Recession, which is consistent with technological competition losing importance relative to other factors, and therefore citations having less of an effect on firm market value than they did in the past. We thank an anonymous referee for this observation.

6. Discussion

■ Although there is clear evidence that the data generating process for patent citations has changed over time, and that such changes now threaten widely used empirical methods in economics, the precise reasons for the change in patent citations are less clear. In the following section, we discuss some of the factors that may be driving our results, and discuss recommendations for how researchers can handle the problem.

□ **Effect of cross-citing.** A patent examiner is required to conduct their own search for prior art, but doing so is often difficult and time-consuming. The applicant owes no such duty. Although the applicant often possesses documents that may be relevant to the examination of the patent, disclosing such information may not appear to be in their interests. Accordingly, in the United States, applicants are required to cite documents known to be material to patentability. If an applicant fails to comply, then future infringers of a patent can allege that the applicant committed inequitable conduct during patent examination. Further, patent agents and attorneys face disciplinary action—up to and including disbarment—if they knowingly withhold relevant art. If a court agrees, it may rule the patent unenforceable, even if the patent is otherwise valid and infringed. Infringers therefore use a powerful document discovery process to try to unearth evidence of omissions by the applicant, whether or not those omissions were truly intentional. At the same time, applicants pay no penalty whatsoever for submitting excessive or irrelevant references.[27] Thus, if a patent applicant knows of something—*anything*, really—then it may make more sense to "disclose" it, even if its relevance is marginal.[28] Further, by overdisclosing, applicants may gain the benefit of hiding some relevant prior art "in plain sight."

Kuhn (2010) argues that changes in the law have amplified incentives to disclose. The contours of the duty of disclosure derive from ambiguous case law, not a "bright line" rule, so applicants face considerable uncertainty regarding how the law will be enforced. Federal courts have steadily expanded the scope of the duty of disclosure over the last several decades to encompass an ever-larger set of references. Moreover, an applicant must decide whether to cite a reference based not only on what the law is at the time of examination, but also what the applicant believes the law will be at some point *in the future*—that is, the point in time when the applicant will be forced to go to court to stop infringement (Lemley and Shapiro, 2005). Given the 20-year lifetime of a patent, the conservative strategy is to anticipate strict enforcement of the duty to disclose in the future.

The compliance strategy known as "cross-citing" arises at the intersection of the duty of disclosure and a group of interrelated patent applications called a patent family. Legally speaking, in the United States, a patent family is a group of patents linked by priority claims, which are mechanisms that allow the patent applicant flexibility as to when and how to pursue claims of various scope. The motivations for filing patent families (as opposed to individual patent applications) are complex, but in general patent families can provide redundant protection, broader claim scope, streamlined patent examination, administrative simplicity, and compliance with USPTO examination rules. US patents also may belong to an international patent family when applicants seek patent protection in multiple jurisdictions. Similar to cross-citing within US patent families, the generation of international search reports also can propagate citations across all family members.[29]

Filing families of patents on the same technology is an expensive proposition, and scholars have found that the number of countries in which a patent application is filed is evidence of private patent value (Harhoff, Scherer, and Vopel, 2003). Because patent families are valuable,

[27] Although search technology has improved considerably, so too has the number of patents available to be searched. Accordingly, whether applicant search results have increased or decreased in quality over time seems unclear.

[28] Administrative efficiencies also weigh in favor of an aggressive citation strategy. The marginal cost of submitting a citation is close to zero, whereas deciding not to submit a citation requires evaluating the document's relevance.

[29] The USPTO encourages applicants to include references identified in international filings (MPEP 707.05).

applicants have greater incentives to avoid unnecessary risks and accordingly, greater incentives to cite references. At the same time, patent families unearth more potential citations because the USPTO performs a new prior art search for every member of the family. The easiest and safest way to ensure that the duty of disclosure is met is to simply copy every citation made in any member of the family into every other member of the family—a strategy referred to by practitioners as "cross-citing." Cross-citing can then create a network effect, wherein even an irrelevant patent can become highly cited if swept up into an applicant's overzealous disclosure strategy. The online web Appendix provides a more detailed discussion of patent families and presents empirical evidence that both backward citation counts and the duplication of citations across patents increase with patent family size.

□ **Correction for patent families.** Cross-citing is driven, in part, by the growing prevalence of large US and international patent families (Harhoff, Scherer, and Vopel, 2003). Accordingly, collapsing citation counts to the family level might help to correct for the inflationary effects of cross-citing. We examine that proposition by rerunning many of the empirical analyses presented earlier in this article at the family level.[30] We assigned a family identifier to each patent and then count a citation only one time between a citing patent family and a cited patent family. This approach removes every repeated citation, regardless of whether the citation originated through cross-citing or as an evaluation of prior art. Collapsing in this way is an aggressive approach to purging the sample of potentially redundant citations.

To examine the impact of patent families, we replicate at the family level many of the analyses from the main text, including Figures 1–5. We found no evidence that collapsing to patent family corrects for the underlying distributional problem of an extreme number of citations in patent applications in recent years. For example, the online web Appendix Figures 9a and 9b show that the decline in explanatory power of five-year patent forward citations for predicting *Tobin's Q* remains essentially the same when citations are collapsed to the family level.

Figure 6 in the online web Appendix demonstrates that the collapsed sample reduces the misrepresentation of high-citing patents to some degree (i.e., 5% of patents granted in 2012 cite more than 100 references, whereas only 3% of families of patents first granted in 2012 cite more than 100 references), but the collapsed sample does not change the fundamental, underlying problem. Just as 5% of patents generate more than 45% of citations, 5% of families still generate more than 42% of citations, even after collapsing to the family level. The mixture problem persists, because even as the impact of individual high-citing patents is reduced by collapsing them into a high-citing family, the same operation can collapse a collection of individually low-citing patents into a single high-citing family, shifting rather than eliminating the misrepresentation.

We therefore conclude that collapsing to patent family is not an effective correction for the problems identified in the article.[31] At the same time, collapsing citation counts in this way might only generate new biases in the counts of citations based on filing strategies by firms, and/or how inventions are divided into different patent applications or patent families.

□ **Correction strategies and future research.** A general correction for the changing nature of patent citations would be of great value to the field. Unfortunately, we cannot propose a simple and general correction at this time for all of the problems that we identify in this article. In fact, we are skeptical as to whether an all-encompassing correction exists. A concern for any correction is that the problem and solution are moving targets in time, and it is difficult to know whether a correction today (even an imperfect one) would remain valid tomorrow. Moreover, many seemingly sensible corrections may in fact fail to address all of the problems that we

[30] *Collapsing* citations to the patent family (i.e., ignoring "repeated" citations within families) does not make logical sense in every context. Of the assumptions in Section 2, collapsing to patent family might correct for violations of Assumptions B1 and B2, but would not address violations of Assumptions A1, A2, C1, or C2.

[31] It is possible that collapsing to patent families might be more effective if one had better data, but we find no evidence at this time that collapsing the data in this way corrects for any of the problems identified in the article.

identify—or they may just introduce a different bias, as we discuss with family-level citations above. The corrections that have traditionally been employed in the literature generally fall into one of three categories: selection strategies, weighting strategies, and fixed effects strategies.

Selection strategies are common in patent analysis. For example, most studies that use patent citations to measure value exclude self-citations from the analysis. Similarly, the separation of applicant versus examiner citations explicitly acknowledges the potential difference in explanatory power between the particular subsets of citations. In Section 5, we selected citations that were more likely to represent true knowledge flows, but such a strategy is not a general-purpose solution, because the citations that represent knowledge flows are not necessarily the same as those that predict firm market value or private patent value.[32] Selecting citations based on the phenomenon of interest provides a promising avenue for future work, but a selection-based correction imposes a second selection on top of the original selection that is inherent in the very nature of patent citations.

Weighting schemes are commonly used in patent analysis to improve on simple counts. For example, Hall, Jaffe, and Trajtenberg (2005) use citation-weighted patent counts to predict the market value of firms. One can also weight backward citations by the inverse of the number of backward citations in the cited or citing patent to control for the inflation of citations over time or among technologies (e.g., Marco, 2007). Annual deflators can attempt to account for the amount of citations generated by "extreme" or "impossible" backward citing patents, but such attempts are unlikely to account for other violations of the assumptions in Section 2. Similarly, weighting patent citations by textual similarity risks undercounting important cases where inventors have learned from distant technologies or combined facets of unrelated technologies.

Including fixed effects is a prevalent strategy to control for sectoral differences or temporal trends, but simple fixed effects cannot control for within-year heterogeneity or different citing-to-cited time spans. The fact that patent citations are generated in the backward direction, but aggregated in the forward direction, means that the effect of time controls on regression models is not clear. Although including fixed effects can address changes to mean forward citation counts (e.g., by time or by sector), they fail to account for the continuing decline in citation similarity and the increasing misrepresentation of a small proportion of patents in the generation of citations.

Our results highlight a fundamental challenge inherent in the use of patent citations in economic analysis: producing unbiased estimates of economic phenomena requires the researcher to control for multiple, unrelated changes in the patent citation process. Just as the estimation of an economic parameter requires a robust identification strategy that varies by context, the measurement of a proxy for an economic indicator also requires solutions that vary by research question. Thus, we suspect that a one-size-fits-all correction does not exist. Instead, as demonstrated in Section 5, we believe that different phenomena require different empirical techniques, and that good solutions are likely to be tailored to a particular research question and grounded in an understanding of institutional features, including the data generating process.

New methods from machine learning may offer a path forward for constructing general-purpose measures. Such new measures could serve the same purposes as patent citations (Trajtenberg, 1990b; Harhoff et al., 1999; Hall, Jaffe, and Trajtenberg, 2005) without requiring the researcher to control for the many types of temporal trends, institutional features, and other selection effects that now undermine the use of patent citations. Early work in this direction has already begun (Nanda, Younge, and Fleming, 2015; Younge and Kuhn, 2015), and future research in this vein may identify text-based or network-based measures of innovation impact and novelty that correct for limitations in citation measures.

[32] For example, we did not find the selection strategy for assessing spillovers to be an effective correction for replicating original results for Hegde and Sampat (2009) and Hall, Jaffe, and Trajtenberg (2005).

7. Conclusion

■ Scholars have long used patent citations for empirical analysis in economics. In this article, we have reexamined many of the implicit, but commonly held, assumptions that underlie the use of patent citations to proxy for a plethora of phenomena of interest. We have argued that due to systemic changes in the data generation process, many of these assumptions are no longer valid. Although backward patent citation counts have always been skewed, the distribution has become even less representative over time. Moreover, whereas an average citation from a patent issued in 1995 was likely to be highly informative, an average citation from a patent issued in 2015 appears to be much less so. The changes in the data generation process for patent citations present significant problems for applied economists, and the use of long-accepted methods of analysis now pose risks of generating biased or, worse, invalid results. Although we have shown that all of the above issues have important consequences for empirical research, we believe that a singular correction is unlikely to work reliably across every empirical context. Instead, scholars must determine the appropriate criteria for making a correction based on the phenomenon of interest. To support these efforts, we make much of our data and programming code available on our websites for others to use, and call on researchers to continue work in the area.

References

ABRAMS, D., AKCIGIT, U., AND POPADAK, J.A. "Patent Value and Citations: Creative Destruction or Strategic Disruption?" National Bureau of Economic Research Working Paper No. 19647, 2018. https://www.nber.org/papers/w19647

ALCACER, J. AND GITTELMAN, M. "Patent Citations as a Measure of Knowledge Flows: The Influence of Examiner Citations." *The Review of Economics and Statistics*, Vol. 88(4) (2006), pp. 774–779.

ALCACER, J., GITTELMAN, M., AND SAMPAT, B. "Applicant and Examiner Citations in US Patents: An Overview and Analysis." *Research Policy*, Vol. 38(2) (2009), pp. 415–427.

CABALLERO, R.J. AND JAFFE, A.B. "How High Are the Giants' Shoulders: An Empirical Assessment of Knowledge Spillovers and Creative Destruction in a Model of Economic Growth." *NBER Macroeconomics Annual*, Vol. 8 (1993), pp. 15–74.

CORREDOIRA, R.A. AND BANERJEE, P.M. "Measuring Patent's Influence on Technological Evolution: A Study of Knowledge Spanning and Subsequent Inventive Activity." *Research Policy*, Vol. 44(2) (2015), pp. 508–521.

COTROPIA, C.A., LEMLEY, M.A., AND SAMPAT, B. "Do Applicant Patent Citations Matter?" *Research Policy*, Vol. 42(4) (2013), pp. 844–854.

DADALT, P.J., DONALDSON, J.R., AND GARNER, J.L. "Will Any *q* Do?" *Journal of Financial Research*, Vol. 26(4) (2003), pp. 535–551.

GAMBARDELLA, A., HARHOFF, D., AND VERSPAGEN, B. "The Value of European Patents." *European Management Review*, Vol. 5(2) (2008), pp. 69–84.

GRILICHES, Z. "Market Value, R&D, and Patents." *Economics letters*, Vol. 7(2) (1981), pp. 183–187.

HALL, B.H., JAFFE, A.B., AND TRAJTENBERG, M. "The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools." Discussion Paper (2001).

———. "Market Value and Patent Citations." *RAND Journal of Economics*, Vol. 36(1) (2005), pp. 16–38.

HANLEY, D. "Innovation, Technological Interdependence, and Economic Growth." Discussion Paper, 2015. Citeseer.

HARHOFF, D., NARIN, F., SCHERER, F.M., AND VOPEL, K. "Citation Frequency and the Value of Patented Inventions." *Review of Economics and statistics*, Vol. 81(3) (1999), pp. 511–515.

HARHOFF, D., SCHERER, F.M., AND VOPEL, K. "Citations, Family Size, Opposition and the Value of Patent Rights." *Research Policy*, Vol. 32(8) (2003), pp. 1343–1363.

HEGDE, D. AND SAMPAT, B. "Examiner Citations, Applicant Citations, and the Private Value of Patents." *Economics Letters*, Vol. 105(3) (2009), pp. 287–289.

JAFFE, A., TRAJTENBERG, M., AND HENDERSON, R. "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations." *Quarterly Journal of Economics*, Vol. 108(3) (1993), pp. 577–598.

JUNG, H.J. AND LEE, J.J. "The Quest for Originality: A New Typology of Knowledge Search and Breakthrough Inventions." *Academy of Management Journal*, Vol. 59(5) (2016), pp. 1725–1753.

KRUGMAN, P.R. *Geography and Trade*. Cambridge, MA: MIT press 1993.

KUHN, J.M. "Information overload at the US Patent and Trademark Office: Re- Framing the Duty of Disclosure in Patent Law as a Search and Filter Problem." *Yale JL & Tech.*, Vol. 13 (2010), pp. 89–140.

LAMPE, R. "Strategic Citation." *Review of Economics and Statistics*, Vol. 94(1) (2012), pp. 320–333.

LANJOUW, J.O. AND SCHANKERMAN, M. "Characteristics of Patent Litigation: A Window on Competition." *RAND Journal of Economics*, Vol. 32(1) (2001), pp. 129–151.

LEMLEY, M.A. AND SHAPIRO, C. "Probabilistic Patents." *Journal of Economic Perspectives*, Vol. 19(2) (2005), pp. 75–98.

LERNER, J. AND SERU, A. "The Use and Misuse of Patent Data: Issues for Corporate Finance and Beyond." National Bureau of Economic Research Working Paper No. 24053, 2017. https://www.nber.org/papers/w24053

MARCO, A.C. "The Dynamics of Patent Citations." *Economics Letters*, Vol. 94(2) (2007), pp. 290–296.

MARCO, A.C., SARNOFF, J.D., AND DEGRAZIA, C. "Patent Claims and Patent Scope." *Research Policy*, Vol. 48(9) (2019).

MEHTA, A., RYSMAN, M., AND SIMCOE, T. "Identifying the Age Profile of Patent Citations: New Estimates of Knowledge Diffusion." *Journal of Applied Econometrics*, Vol. 25(7) (2010), pp. 1179–1204.

MOSER, P., OHMSTEDT, J., AND RHODE, P.W. "Patent Citations—An Analysis of Quality Differences and Citing Practices in Hybrid Corn." *Management Science*, Vol. 64(4) (2017), pp. 1926–1940.

NANDA, R., YOUNGE, K.A., AND FLEMING, L. "Innovation and Entrepreneurship in Renewable Energy." In A. Jaffe and B. Jones, eds., *The Changing Frontier: Rethinking Science and Innovation Policy*. Cambridge, MA: NBER/University of Chicago Press (2015).

ROACH, M. AND COHEN, W.M. "Lens or Prism? Patent Citations as a Measure of Knowledge Flows from Public Research." *Management Science*, Vol. 59(2) (2013), pp. 504–525.

SAMPAT, B.N. "When Do Applicants Search for Prior Art?" *The Journal of Law and Economics*, Vol. 53(2) (2010), pp. 399–416.

SHAHMIRZADI, O., LUGOWSKI, A., AND YOUNGE, K.A. "Text Similarity in Vector Space Models: A Comparative Study." IEEE - 18th International Conference on Machine Learning and Applications, 2018. https://arxiv.org/abs/1810.00664

THOMPSON, P. "Patent Citations and the Geography of Knowledge Spillovers: Evidence from Inventor- and Examiner-Added Citations." *The Review of Economics and Statistics*, Vol. 88(2) (2006), pp. 383–388.

THOMPSON, P. AND FOX-KEAN, M. "Patent Citations and the Geography of Knowledge Spillovers: A Reassessment." *American Economic Review*, Vol. 95(1) (2005), pp. 450–460.

TRAJTENBERG, M. *Economic Analysis of Product Innovation: The Case of CT Scanners*, Vol. 160. Cambridge, MA: Harvard University Press (1990a).

———. "A Penny for Your Quotes: Patent Citations and the Value of Innovations." *RAND Journal of Economics*, Vol. 21(1) (1990b), pp. 172–187.

TRAJTENBERG, M., HENDERSON, R., AND JAFFE, A. "University versus Corporate Patents: A Window on the Basicness of Invention." *Economics of Innovation and New Technology*, Vol. 5(1) (1997), pp. 19–50.

YOUNGE, K.A. AND KUHN, J.M. "Patent-to-Patent Similarity: A Vector Space Model." Discussion Paper (2015). Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2709238

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Data S1: Online Appendix
Table 1: OLS estimates of backward citation counts.
Table 2: OLS estimates of citation similarity over time.
Figure 1. Mean citation similarity over time for citations from applications, examiners, both.
Figure 2. Percentage of examiner citations by patent-level backward citation count.
Figure 3. Citation-level plot of the mean number of times the cited patent appears as a citation across a patent family, by the family size of the citing patent.
Figure 4. Patent-level plot of maintenance fee payments as a function of backward citation count.
Figure 5. Total citations made by year, divided by number of backward citations made by citing family.
Figure 6. Mean and interquartile range of citation *Similarity* by citing year, collapsed to family. Dot-dash is OLS estimate controlling for count of backward citations.
Figure 7. Area plot of citations by backward citation count of citing patent family over time.
Figure 8. Lorenz curves of cumulative proportion of families by backward citation count, by year.
Figure 9. Mean and interquartile range of *Similarity* for citations by count of backward citations by citing family.
Figure 10. Results from collapsing patent citation data to the family level.