Patent Citations and Empirical Analysis[*]

Jeffrey M. Kuhn[†]
Kenan-Flagler Business School
University of North Carolina at Chapel Hill
jeffrey_kuhn@haas.berkeley.edu

Kenneth A. Younge[‡]
College of Management of Technology
École Polytechnique Fédérale de Lausanne
kenneth.younge@epfl.ch

Alan Marco
School of Public Policy
Georgia Tech
_____@gatech.edu

July 2017

Abstract

Many studies of innovation rely on patent citations to measure intellectual lineage and impact. In this paper, we use a vector space model of patent similarity and new data from the USPTO to show that the nature of patent citations has changed dramatically in recent years. Today, far more citations are created per patent, and the mean technological similarity between citing and cited patents has fallen. We link these developments to changes in the data generating process for patent citations and the rise of cross-citing amongst large patent families. We also demonstrate that data limitations have generated overstated results in several important papers using patent citations. Overall, we argue that the use of patent citations needs to be thoroughly re-validated for future empirical work.

[*] We thank Andrew Toole, Ludovic DiBiaggio, Neil Thompson, Noam Yuchtman, Rui de Figueiredo, and Tony Tong for helpful comments, and seminar participants at the Searle Center Conference on Innovation Economics, the United States Patent and Trademark Office, Skema Business School, and the Academy of Management.
[†] The Hoover Institution Working Group on Intellectual Property, Innovation, and Prosperity at Stanford University provided generous financial support during the preparation of this paper.
[‡] The author thanks Google for a generous research grant of computing time on the Google Cloud.

1. Introduction

A substantial amount of research adopts patenting as an empirical measure of innovation. Seminal work by Griliches measured patent counts (Griliches 1981), 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 1990). In this vein, studies have used patent citations to measure private patent value, (Lanjouw and Schankerman 2001, Harhoff, Scherer et al. 2003), firm market value (Hall, Jaffe et al. 2005), cumulative innovation (Caballero and Jaffe 1993, Trajtenberg, Henderson et al. 1997), geographic spillovers (Jaffe, Trajtenberg et al. 1993), technology life-cycles (Mehta, Rysman et al. 2010), social importance (Moser, Ohmstedt et al. 2013), originality (Jung and Lee 2015), and technological impact (Corredoira and Banerjee 2015). Overall, a search for "patent citations" on Google Scholar returns over 19,000 results.[4]

Recent research, however, has begun to call into question a straightforward interpretation of patent citations. Abrams, Akcigit et al. (2013) 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, Lampe 2012). Citations may suffer from significant noise and measurement error (Gambardella, Harhoff et al. 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 2015). 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 et al. 2010).

In this article, we highlight an important change in the data generating process of patent citations in the last five years that dramatically changes the statistical nature of patent citation measures and that impacts the appropriateness of using raw counts for such measures in 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, split by the number of citations made by each patent. In recent

[4] Conducted on January 6, 2017.

years, a large percentage (46.8%) of references derive from a very small percentage (less than 5%) of patents with more than 100 backward citations.[5]

--- Insert Figure 1 about here ---

To examine how the aforementioned changes affect the information content and quality of citations, we compute a vector space model that compares the text of every patent, to every other patent, granted by the USPTO.[6] 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 in recent years. 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). 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 weakening is continuing to accelerate.

--- Insert Figure 2 about here ---

We also employ new data, including both publicly available data and internal USPTO data, to better understand the data generating process for patent citations. We show that publicly available data misattributes the source of a small but significant percentage of citations, identifying the patent examiner (or patent applicant) as submitting the citation when in fact the patent applicant (or patent examiner) actually submitted the citation first.[7] We demonstrate the large and growing importance of patent citations made to pending patent applications, and we show how to update citation data to include these missing references. Finally, we use internal USPTO data on the timing of applicant and examiner disclosures to show that citations are frequently submitted long after a patent application is filed – a fact that challenges the view of those patent citations as an indicator of knowledge inheritance – and we explore the empirical consequences of such delays.

The main contribution of this paper is to provide *prima facie* evidence that the citation generating process is now generating significant measurement error for many academic studies. We demonstrate that changes to the data generating process are leading to widespread violation of several assumptions that are central to the empirical literature on innovation, and that these violations are leading to tenuous and

[5] Conversations with patent examiners suggest that reviewing more than 100 citations is extremely difficult to do, given the time constraints of the patent office. As such, the 46.8% of references that originate from less than 5% of patents, do not represent what the literature typically portrays as a "patent citation."

[6] We show in related work that patent-to-patent similarity can be a powerful predictor of institutional features such as shared classification and patent priority (Younge & Kuhn, 2015). Patent similarity also correlates highly with both common sense and expert comparisons of patent pairs. Patent similarity, however, cannot identify whether any particular citation is meaningful or informative.

[7] Although potentially misleading, this identification is not technically erroneous since the patent reads "cited by examiner".

potentially unsubstantiated conclusions. Moreover, we replicate several recent studies and show how a more detailed analysis of the data can support substantially different conclusions. As such, we argue that future research with patent citation-based measures will require new empirical methods and measures. At the same time, we also predict that the diverse empirical objectives of innovation research will likely preclude a one-size-fits-all correction to patent citations that will work reliably across different research contexts.[8]

To make the case outlined above, the following sections are organized as follows. Section 2 discusses the institutions and incentives that underlie the generation of patent citations at the USPTO. Although the U.S. patent system is not the only driver of patent citations, it clearly is an important factor with a disproportionate effect on the world-wide patenting regime; many studies also rely solely on USPTO data. Section 3 describes several core assumptions underlying the use of patent citations in empirical analysis. Although every scholar cannot be expected to understand all of the inner workings of the U.S. patent examination process, researchers nevertheless do rely on several core assumptions about how that process works. We re-examine the patent examination process for clarity and to identify ways to test the validity of commonly held assumptions. Section 4 introduces the many new sources of data used in our analysis. Researchers can now benefit from many new and updated sources of data, although many studies still rely on the NBER dataset (Hall, Jaffe et al. 2001).[9] Section 5 describes our results, which demonstrate many of the ways in which the assumptions described in Section 3 are systematically violated. Section 6 discusses examples of the empirical consequences of these violations, while Section 7 describes various types of robustness checks that may alleviate such concerns. Section 8 concludes.

2. Institutions and Incentives

Several institutional features in the patenting regime may be leading to a proliferation of low-information citations. (Kuhn 2011). According to the duty of disclosure owed by all patent applicants, an applicant must disclose to the patent office any reference of which the applicant is aware and which the applicant believes to be relevant to the examination of the patent application. Recent changes to the duty of disclosure – coupled with the high cost of failing to comply, wherein a patent can be invalidated – may be leading applicants to "over disclose." The effects of the duty of disclosure can be particularly pronounced in large patent families, where the number of potentially relevant references increases dramatically relative to isolated applications. Faced with uncertainty regarding the duty of disclosure and being sensitive to the costs and risks associated with manually evaluating each potentially relevant reference, applicants with large

[8] We discuss several robustness tests in the Appendix to bolster conclusions derived from patent citations; we also propose a new method based on similarity data for comparing the information quality of different corrections and share the data for the correction through the Patent Research Foundation.

[9] For example, the Office of the Chief Economist at the USPTO provides updated research data at http://PatentsView.org.

patent families appear to be responding by automatically "cross-citing" (i.e., copying) large sets of references across entire families of applications. This section describes the institutions and incentives that give rise to patent citations, including those related to patent examiners (Section 2.A), patent applicants (Section 2.B), and patent families (Section 2.C).

A. Patent Examiners

The primary purpose of patent citations is not to serve as a data source for academics but rather to document and facilitate the examination of patent applications at the USPTO. A patent applicant is entitled to a patent only if the patent application describes and claims a new and non-obvious invention. A patent examiner determines whether the patent application meets these requirements by comparing the application's claims with the prior art. Patent citations constitute the set of prior art references that the patent examiner considers when making this determination. Patent citations thus act as an information source that facilitates the gatekeeping function of patent examiners by helping them to avoid granting non-innovative or overly broad patents.

When a patent examiner determines that the claims of a patent application fail to encompass a novel and non-obvious advance over the identified prior art, the examiner rejects the claims. A patent applicant whose claims are rejected may amend the claims to overcome the rejection, may argue that the rejection is improper, or may abandon the application. In the U.S., no rejection is ever truly final, and most issued patents are rejected at least once before being granted. (Carley, Hegde et al. 2015). A patent examiner is required to support a rejection by identifying the specific portions of cited prior art documents that disclose the features recited in the claims. The applicant and examiner evaluate and debate the significance of these "rejection citations," often resulting in the applicant narrowing the claims to distinguish a new invention from the prior art. (Kuhn and Thompson 2017). About 80% of patent applications ultimately issue as patents (Carley, Hegde et al. 2015), and about 70% of issued patents are narrowed before they are issued. (Kuhn and Thompson 2017).

Patent examination thus takes the form of a negotiation between the applicant and examiner regarding the scope of the claims, with the application being abandoned if the parties do not reach a consensus. Within the negotiation between the applicant and the examiner, the citations selected by a patent examiner to support rejections are the subset of citations that most actively influences the patent's scope. Patent citations therefore serve a public good function in that they provide the raw materials that patent examiners use to identify the rejection references that help to prevent the issuance of unwarranted or overly broad patents. In Section 4, we show that examiners use only about 11.4% of all cited patents and patent applications to support rejections.

Cotropia, Lemley et al. (2013) present empirical evidence that rejection references are typically submitted by patent examiners rather than patent applicants. This argument makes logical sense. The examiner is required to search the prior art for citations to potentially use to support rejections and has incentives to find those rejection references, to document the examiner's search efforts, and to record citations for possible use later in the examination process. In contrast, the applicant is less likely to submit references that invalidate the application's claims. If an applicant knows that a reference is likely to influence the claim scope, then the applicant might reasonably submit different claims to the patent office to start with. If instead the applicant is submitting references to the patent office without carefully reviewing them, then these references are likely to be relatively unrelated to the application and unlikely to lead to rejections. Scholars have argued that in some cases applicants strategically withhold citations likely to be influential. (Lampe 2012).

B. Patent Applicants

The examiner is required to conduct their own search for prior art, but doing so is often difficult and time-consuming. The applicant often already possesses documents that may be relevant to the examination of the patent, but disclosing such information may not appear to be in the applicant's interest. Accordingly, the USPTO imposes on "[e]ach individual associated with the filing and prosecution of a patent application . . . a duty of candor and good faith in dealing with the Office" (37 Code of Federal Regulations 1.56). The duty of candor is not merely a negative obligation to abstain from fraud, but also is an *affirmative* "duty to disclose to the Office all information known to that individual to be material to patentability" (37 CFR 1.56).

The "duty of disclosure" described above provides strong incentives for applicants to submit citations with their patent applications. If an applicant fails to cite a reference that the applicant knew about and believed to be relevant, then future infringers of a patent can allege that the applicant committed inequitable conduct during the patent examination. If a court finds evidence of inequitable conduct then the court may rule the patent altogether unenforceable, even if the patent is otherwise actually valid and being infringed upon. Defendants in infringement cases therefore strongly pursue claims of inequitable conduct during their defense and use a powerful document discovery process to try to unearth evidence of omissions by the applicant, whether those omissions were truly intentional or not. Therefore, if an applicant knows of something – *anything,* really – then it may make more sense to "disclose" it, even if it has little or nothing to do with the patent application at hand. At the same time, applicants pay no penalty or cost whatsoever for citing too many references, even if references are included with little relevance to a particular application.

The duty of disclosure, however, does not by itself explain an increase in citations from applicants, or the increasing concentration of citations by a small portion of patent applications. Kuhn (2011) argues that changes in the case law have amplified the incentives to disclose in several ways. 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. The contours of the duty of disclosure derive from ambiguous case law (not a "bright line" rule), and so applicants face considerable ambiguity regarding how the law will be enforced. 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* when it is forced to go to court to stop infringement (Lemley and Shapiro 2005). Given the long-term trend toward a stronger duty to disclose, and given the 20-year lifetime of a patent over which rights must be protected, and given examples in the media where infringers find a 'smoking gun' in email or other sources during discovery, applicants must anticipate strict enforcement of the duty to disclose in the future.

C. Patent Families

Changes in the law regarding the duty of disclosure, however, do not by themselves explain the observed increase in citations by a small portion of patent applications. Incentives to disclose are sharpest in situations where the patent application is likely to produce a valuable patent and where many potential citations are unearthed. Both are true when an applicant files a group of interrelated patent applications called a 'patent family.' Patent families provide possibly redundant protection for a given technology – if one patent in the family fails to provide adequate protection, other patents in the family may still provide sufficient coverage to block others from using the technology. Patent families also improve claim scope by allowing the patent applicant to obtain similar – but still somewhat different – patent claims based on the same initial specification. A single patent in a complex technology may be "invented around," but inventing around an entire family of patents is much more challenging. Finally, patent families provide flexibility. By effectively splitting an initial application and keeping at least one part of the patent family pending at the patent office, the applicant can retain an option to modify the claims for many years after the initial filing. Or, the applicant may leave only the allowed claims in a parent patent application while preserving the ability to seek an allowance on the rejected claims in a continuation.[10]

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 et al. 2003). Because patent families are valuable, applicants have greater incentives to

[10] We define a patent family to include a group of patent applications linked by priority claims to a common priority document. We define priority claim broadly to include linkages established by provisional priority, foreign priority, divisional, continuation, and continuation-in-part relationships. We note that our definition of "family" is very different from that of many European scholars, who reserve the term for the filing of the same patent application in different countries.

avoid unnecessary risks and accordingly greater incentives to cite references. At the same time, patent families unearth more potential citations since the USPTO performs a new prior art search for every member of the family, even if the child application is assigned to the same examiner (which is typically the case). 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."

The institutional factors above, including the duty to disclose knowledge about prior art, the rise in the strategic importance of patent families, and the cross-citing behavior of patent families, all lead to the conclusion that backward citation counts should be increasing in the last several years, and that the effect is stronger as the size of the patent family increases – a hypothesis that we test in the empirical analysis that follows.

3. Empirical Assumptions in the Literature

Under the conventional narrative in the patent citations literature, inventors standing on the shoulders of giants develop a new invention based in part on past knowledge. Some of this past knowledge is known explicitly to the inventor, while other knowledge exists as background information that implicitly informs the inventor's efforts. Prior to filing the patent application, the applicant may search the prior art, which may bring to light some of this background information. The applicant, who owes to the patent office "a duty to disclose . . . all information known to that individual to be material to patentability,"[11] includes with the patent application filing papers a list of technologically related patents and other documents known to the inventor or produced by the search. The patent examiner supplements these documents with the results of an independent search, and the patent is issued only after a careful comparison of the patent's claims to the prior art. At issuance the patent office tabulates the resulting citations, which are then made available for analysis by scholars.

In this section, we argue that buried in this narrative are a set of implicit assumptions that have important implications for empiricists. It is important to note that scholars tend to make these assumptions from necessity, not by choice, as data constraints tend to limit even the most careful empirical analysis. Further, scholars have long known that citation-based measures are noisy, and make these assumptions for the typical citation in large sample analysis rather than for particular citations.

[11] 37 C.F.R. 1.56.

A. Assumptions about Technological Relatedness

The most fundamental assumption in the citations literature is that the typical citation indicates a degree of technological relatedness between the citing and cited patents. (Jaffe, Trajtenberg et al. 1993). Although we do not dispute this assumption in its most general formulation, scholars frequently employ methodology validated using citations generated by patents issued many years ago to citations generated by patents today. Thus, the literature implicitly assumes that the technological relatedness between citing and cited patents is relatively stable over time. If this implicit assumption does not hold, then the initial research validating citations as a measurement tool would need to be revisited.

> *ASSUMPTION A1: A patent citation indicates technological relatedness between the citing and cited patent in a way that is stable for citations over time.*

A closely related assumption is that for the typical citation, the applicant decides to a cite a patent based on an evaluation of the technological relatedness between that patent and the invention. Although one view of the world might assume a virtuous and conscientious applicant, Lampe (2012) argued that applicants strategically and systematically withhold citations in an attempt to influence the patent examination process. In either case, the literature implicitly assumes that the applicant first compares the invention to a set of documents identified by the inventor or resulting from a prior art search and then decides which documents to cite based on the relevance of each document to the invention's patentability.

> *ASSUMPTION A2: The technological content described in the patent forms the basis for the applicant's decision as to whether to cite the patent.*

B. Assumptions about Citation Provenance

The next assumption relates to a citation's provenance. While the early literature generally assumed that examiners were responsible for patent citations,[12] scholars have more recently focused on separating the contributions of applicants and examiners (Alcacer and Gittelman 2006, Alcacer, Gittelman et al. 2009, Hegde and Sampat 2009). Much of the empirical knowledge spillovers literature rests on the assumption that the prototypical patent citation indicates that the inventors of the citing patent had in mind the invention described in the cited patent when developing the invention described in the citing patent. (Thompson and Fox-Kean 2005). After all, a citation added by a patent examiner years after a patent application is filed is surely a poor indicator of a knowledge flow from the inventors of the cited patent to the inventors of the

[12] For example, Jaffe et al. (1993) state: "It is the patent examiner who determines what citations a patent must include."

citing patent. For this reason, recent scholarship differentiates between examiner-submitted and applicant-submitted citations as a measure of knowledge spillovers (Thompson 2006). Scholars typically rely on an USPTO-specified indicator to identify examiner-submitted citations and have long had little choice but to assume the indicator's accuracy. At the same time, scholars have found that citations identified in the publicly available data as having been submitted by an examiner can be more indicative of impact than applicant-submitted citations. (Alcacer, Gittelman et al. 2009, Hegde and Sampat 2009).

> *ASSUMPTION B1: Citations indicated in publicly available data as examiner-submitted were originally submitted by examiners.*

As discussed above, the literature on knowledge spillovers often uses patent citations to identify a form of intellectual inheritance, wherein the inventors of the citing patent explicitly build on the knowledge underlying the cited patent when developing the invention. Of course, the duty of disclosure is not limited to the inventors, but also extends to other employees of the firm that employs the inventors and to the attorney responsible for the application. Nevertheless, scholars' inability to observe the actor responsible for a particular citation has forced the literature to assume that the typical applicant-submitted citation is indeed identified by the inventor.

> *ASSUMPTION B2: Citations identified as applicant-submitted were identified by inventors.*

C. Assumptions about the Patent Examination Process

The next assumption is definitional in nature. Prior to 2001, a patent citation was a reference made by an issued U.S. patent to another issued U.S. patent. However, a change in the law caused patents filed in 2001 or later to be published by default after 18 months. On publication, an application becomes a new source of prior art for other patent applications to reference, since an applicant or examiner can search for, find, and cite the publication even if the application has not issued as a patent. A patent's claims can change considerably between filing and issuance, but the description of the invention in the final patent is virtually identical to the publication document because patent law precludes the applicant from adding new matter to the application after it is filed.[13] We therefore believe that it is logical to treat citations to published applications that ultimately issue as patents as equivalent to citations to the issued patents themselves. Nevertheless, all citation-based measures (to our knowledge) ignore publications because the USPTO does not update citations made to publications to link to the eventually issued patent, instead leaving citations

[13] 35 U.S.C. 132(a).

made to publication documents in their original format. Thus, the literature implicitly assumes that citations are linkages between issued patents, which ignores the potentially important citations made to co-pending (and therefore contemporaneous) patent applications.

ASSUMPTION C1: A patent citation is a reference from a pending patent application (which later issues as a patent) to an earlier-filed issued patent.

The final assumption relates to the legal impact of a citation within the examination process. As discussed in Section 2, an applicant is entitled to a patent only if their application describes and claims a new and non-obvious invention, and patent examiners are tasked with ensuring that patent applications meet these requirements. Because data constraints have long prevented scholars from peering inside the patent examination process, the literature has been effectively forced to assume that all citations are given equal weight in the patent review process and have an equivalent effect on the issuance and scope of the citing patent.

ASSUMPTION C2: All citations have the same legal impact on the citing patent, controlling for whether a citation was submitted by the applicant or the examiner.

4. Data

The data derive from several sources. For the period from 2005 to 2015, we collected patent citations and patent bibliographic data from bulk data files published by the USPTO and aggregated by Google (2016). For the period from 1976 to 2004, we employed the NBER patent data file (Hall, Jaffe et al. 2001) to identify patent citations and patent bibliographic data. Throughout this time period, we also employed citations data not publicly available derived from internal USPTO citation submissions forms. Patent-to-patent textual similarity data is drawn from related work in which Younge and Kuhn (2015) computed a pairwise similarity measure for all 14 trillion possible pairs of patents issued from 1976 to 2014. We thank Hanley (2015) for firm disambiguation data to link each patent to an assignee identifier.

Table 1 shows variable definitions, with the citation (i.e., the citing-cited pair) as the unit of analysis and correlations between variables presented in Table 2. Our sample of patents includes bibliographic information for 5,027,882, patents issued from January 6, 1976 to March 17, 2015. Our patent citation sample includes 62,104,091 patent citations from January 6, 1976 to December 30, 2014. For much of the regression analysis, we restrict the sample of citations to those made by patents issued 2005 or later in order to benefit from the more expansive bibliographic data available for patents issued during this time period.

There are 35.4 million citations made by patents issued 2005 or later, and 19.2 million citations made by patents issued prior to 2005.

--- Insert Table 1, Table 2, and Figure 3 about here ---

The *Similarity* variable indicates the textual similarity of the citing and cited patent, where a higher value indicates a greater level of similarity of the text of the patent specification between the citing and cited patent. Younge and Kuhn (2015) show that the textual 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. Additional details regarding the construction of the similarity data are provided in the Appendix.

As a very important improvement over how citations have been collected and counted in the past, we include in our sample approximately 7.6 million citations to patent publications, which make up 18% of citations made by patents issued 2005 or later. We identified these citations by first extracting the cited patent publication number from the USPTO bulk data files and then linking each publication number forward to the patent that ultimately issues. We ignore citations to patent publications associated with applications that do not ultimately issue as patents because the objective of almost all analysis to date has been to examine *realized* patent citations, not citations to applications that fail to issue.[14] The variable *Is Publication Citation?* indicates whether the cited patent was identified as a patent publication at the time the citation was made. Figure 3, which plots publication citations as a percentage of all citations over time, highlights the increasing importance of citations to patent publications. Publication citations make up 25% of all citations as of 2015, a proportion growing at a rate of about two percentage points each year.

The variables *Is Examiner Citation (Bulk Data)?*, *Is Examiner Citation (Internal Data)?*, and *Is Duplicate Citation?* provide information about who submitted the citation. Both the patent applicant and the patent examiner can cite documents relevant to the examination of a patent application, and the USPTO began identifying examiner-cited references for patents issued after 2001. *Is Examiner Citation (Bulk Data)?* is a dummy variable that indicates whether the citation is identified as being submitted by the patent examiner in the bulk data, which we collect for citations made by patents issued 2005 or later. Examiner citations (bulk data) make up 27% of citations made by patents issued 2005 or later.

The variable *Is Examiner Citation (Internal Data)?* provides the same information as the variable *Is Examiner Citation (Bulk Data)?,* except that it is constructed directly from internal USPTO citation forms used by patent applicants (USPTO Form 1449) and patent examiners (USPTO Form 892) to record patent citations. Applicants and examiners enter document numbers for cited patents on these forms and add them to the formal record of a patent application in order to enter a citation. The USPTO manually adds the US

[14] An analysis of citations involving patents that do not issue, however, might be an interesting avenue for future research.

patent references and US pre-grant publication references to an electronic data file in order to support examination procedures. Because the data are hand-entered by administrative assistants, errors occur and the data are not considered to be authoritative.[15] Therefore, the information to be contained in the official patent document goes through a review when applications are approved for issuance as a patent. This review includes interaction between USPTO personnel and individuals from the USPTO's contracted publisher. In the end, the publisher generates the data that is made available in the bulk data downloads, and that data then serves as the source for the official patent document. As such, the bulk data on prior art references is created anew by the publisher based on the documents (including the 1449 and 892 forms) in the patent application file, or "file wrapper." We returned to records of the original forms to examine how the process of generating data for public consumption has affected empirical research. The process we used for cleaning and validating the raw USPTO citation form data is described in additional detail the Appendix.

We include indicator variables for both *Is Examiner Citation (Bulk Data)?* and *Is Examiner Citation (Internal Data)?* to differentiate how artifacts of the data generating process affect data used by researchers for empirical analysis. For example, the bulk data is in some cases misleading or inaccurate, attributing a citation to the examiner (applicant) when it was actually submitted by the applicant (examiner). In addition, another type of attribution problem arises when both the patent applicant and the patent examiner submit the same reference. We refer to such situations as duplicate citations (*Is Duplicate Citation?*) and attribute the citation to the *Is Examiner Citation (Internal Data)?* variable based on whichever party submitted it first.

Citation Lag, *Submission Lag*, and *Filing Citation* provide information about citation timing. *Citation Lag* indicates the difference in years between the filing year of the citing patent and the filing year of the cited patent. We employ filing years rather than filing dates for calculating the citation lag because the data does not provide filing date for cited patents issued prior to 2005. *Submission Lag* indicates the time in years between the filing date of the citing 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 Index* indicates the ordering of the USPTO citation form, with the first form submitted for a patent having an index of 1, the second having an index of 2, and so on. Applicant and examiner forms are numbered independently.

Is Self Citation? indicates whether the citing and cited patents were assigned to the same firm at the time they are issued. We refer to citations where the citing and cited share an assignee identifier as a "self citation" and all others as "other citations." Self citations make up 7% of our sample. *Is 102 Rejection Citation?* and *Is 103 Rejection Citation?* indicate whether the cited patent was used by the patent examiner to support a rejection of the claims of the citing patent. Identifying citations used in rejections by the

[15] For instance, there may be simple typographical errors, US references may be confused with foreign references, and entire pages or forms may be omitted.

USPTO is very difficult because they are not specifically identified as such in any publicly available tabulated data set.[16] Accordingly, we follow Cotropia, Lemley et al. (2013) and analyse the raw text of communications (known as "office actions") sent from USPTO patent examiners to applicants to identify rejection citations. To obtain a much more comprehensive sample than Cotropia et al., we examine office actions by the USPTO between 2005 and 2008, inclusive (the limited range of data available for the analysis). We use optical character recognition (OCR) to convert more than 50 million pages of documents from images to text, and then construct a sample of office actions for more than 1 million U.S. patents where the patent examiner explains their rejection in the text. We then used natural language processing techniques and regular expressions to identify patent numbers used to support rejections. We also identify for each rejection whether it is based on an alleged violation of U.S. patent law based on the novelty requirement (35 U.S.C. 102) or non-obviousness requirement (35 U.S.C. 103). For the resulting sample of patents, rejection citations make up 7.3% of all citations.

Backward Citation Count and *Family Size* provide important contextual information about each citation. *Backward Citation Count* indicates the total number of citations made by the citing patent. The distribution of backward citation counts is highly skewed, with a median count of 29 backward citations and a mean count of 121 backward citations. *Family Size* indicates the number of patents in the family of the citing patent. We determined a family identifier for each patent family by analysing patent priority claims. A patent priority claim is a legal mechanism by which a patent applicant can establish a priority date for a patent that pre-dates the filing date by limiting the claims of the later-filed patent to subject matter described in the earlier-patent. By establishing an earlier priority date, the applicant can remove from consideration prior art that was published after the priority date, but before the filing date. Patent law provides for several types of priority relationships, but for simplicity we define a patent family as including all patents linked by any domestic priority relationship. Figure 4 shows a density plot of family size on the patent level. Most patents have very small families, but the right tail of the distribution is long, with about 1.1% of patents being in families of more than 21 patents.

--- Insert Figure 4 about here ---

5. Results

This section analyzes the data described in Section 4 and reports how the assumptions identified in Section 3 are systematically violated. Section 5.A shows that citation similarity has fallen dramatically over

[16] The European Patent Office identifies "X" and "Y" citations as citations that may "block" claims when designating prior art Czarnitzki, D., K. Hussinger and B. Leten (2011). "The market value of blocking patent citations." ZEW-Centre for European Economic Research Discussion Paper(11-021), Von Graevenitz, G., S. Wagner and D. Harhoff (2011). "How to measure patent thickets—A novel approach." Economics Letters **111**(1): 6-9.. However, a European "blocking" citation does not necessarily indicate a rejection, and the concept of an "X" or "Y" blocking citation does not exist in the U.S. patent system.

time and that this phenomenon is largely driven by high-citing patents. Section 5.B argues that applicant cross-citing – the copying of citations between related patents owned by the same firm – explains much of the fall in technological relatedness between citation over time. Section 5.C presents evidence that the attribution of citations to applicants and examiners in publicly available citation data is incorrect in a small but important subset of citations. Section 5.C also shows that many citations are submitted long after the application is filed, calling into question the assumption that an applicant citation represents knowledge the inventor possessed at the time the invention was made. Finally, Section 5.D examines the citation similarity of different types of citations.

A. Declining Citation Similarity

As shown in Figure 5, most patents include 20 or fewer citations – a level we label as "routine." The proportion of routine patents has fallen over time, but more than 75% of patents still have 20 or fewer citations as of 2014. As the number of citations rises from 20 to 100, discussions with patent attorneys and examiners suggest that patent applications become much more "difficult" to review. Although there are cases with 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 5, 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, since 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.

--- Insert Figures 5 and 6 about here ---

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 6 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.

The citations generated by highly citing patents appear to be particularly uninformative. Figure 7 plots the mean and inner quartile similarity values of patent citations by the number of backward citations made

by the citing patent. While 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.

--- Insert Figure 7 and Table 3 about here ---

The low similarity of citations made by high-citing patents coupled with the small but increasing frequency of such patents seems to be driving the overall decline in citation similarity. Table 3 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 .282 points per year ($p<0.001$), a result consistent with Figure 2. Model 2 adds fixed effects for the number of backward citations made by the citing patent. With these fixed effects included, citation similarity is nearly constant over time, increasing slightly at a rate of .014 points per year. ($p<0.001$). Simply controlling for the log of the backward citation count produces a similar result, as shown in Model 3. Figure 8 presents these 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 3.

--- Insert Figure 8 and Figure 9 about here ---

We conclude from this analysis that the data generating process for patent citations has changed over time in a way that systematically and substantially violates Assumption A1, 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 large numbers of relatively unrelated references and not by a secular trend in the textual similarity of patents. The implication of the skewed mixture problem described above is 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 our data suggest that the data generating process began to change most significantly after 2005, it is important to note that the measurement effects are not themselves isolated to years after 2005. Because backward citations are "caught" as forward citations by patents back in time, even measures for patents back in 1995 or 2000 are now being affected by the new citation patterns. An increase in volume of patents issued per year is compounding the problem. Figure 9 shows that the number of patents issued per year increased slowly from 75,000 in 1976 to about 150,000 in 2009, and then doubled to 300,000 only 5 years later. Together, the rise in patenting volume and the change in citation similarity mean that citations generated today are far more numerous and far less informative than citations in years past. We suspect that these trends are likely to become even more profound in the coming years.

B. Applicant Cross-Citing

The previous section presented evidence regarding how the data generating process for patent citations has changed, but it did not address the reasons for those changes. Here, we argue that many of the low similarity citations generated by highly citing patents appear to be mechanically selected, rather than individually evaluated based on technological relevance. That is, citations frequently seem to be copied from one patent application to another (or "cross-cited"), often in groups of hundreds or more, in an effort to meet legal compliance obligations. To better understand the nature of these dramatic changes, we analyze several related factors in this section. First, we find that patent applicants – not patent examiners – are the actors adding citations to patents with hundreds of references. Figure 10 shows the percentage of citations added by patent examiners as the number of patent citations increases: patent examiners contribute the *majority* of citations in patents that cite only a handful of references, but the percentage of citations added by examiners trails off quickly as the total number of references increases.

--- Insert Figure 10 about here ---

Next we turn to patent families. Table 4 reports ordinary least squares regression results predicting the number of backward citations made by a focal patent, including fixed effects for USPTO technology center and patent filing year. The first four models are estimated by ordinary least squares, and for robustness the fifth model is estimated by quasi-maximum likelihood estimation (Wooldridge 2014). Both the explanatory and dependent variables are logged, so estimated coefficients are log-log elasticities. We find strong empirical support for the argument that larger patent families generate more backward citations per patent, than otherwise. Family size (logged) is a large, consistent, and highly significant predictor of backward citation counts (logged) in all models; backward citation counts have also increased over time, even when controlling for family size. Figure 11 shows these results graphically, plotting the mean backward citation count against family size. Isolated patents and patents in very small families cite fewer than 20 references on average, while patents in families of 20 or more members cite over 100 references on average.

--- Insert Table 4 and Figure 11 about here ---

We also hypothesized that families copy the same citations across the entire family. To examine the extent of "cross-citing," we computed a citation-level variable to indicate whether the cited patent in a given citing-cited pair also appeared as the cited patent for other family members. For patents issued 2005 or later, 48.1% of citations are duplicated across family members. Figure 12 plots the probability that a citation made by a focal patent also appears as a citation made by an explicit family member of the focal patent against the number of patents in the family. This likelihood is mechanically zero when the family includes only the focal patent but reaches 75% when the family size reaches four patents. Moreover, any citation made by a patent in a large patent family is likely to be made many times across the family. Figure 13 plots the mean

number of times that a given citation occurs across the family of the citing focal patent against the number of patents in the family. Each citation made by a patent in a family of 16 members appears approximately 10 times.

--- Insert Figures 12 and 13 about here ---

Filing large families of related applications and taking extra precaution to cite every potentially relevant reference would seem to require significant effort and expense for patent applicants. Figure 14 plots the likelihood that the patent owner pays the first and second maintenance fee payments (due 3.5 and 7.5 years after issuance) for a patent as a function of the number of backward citations. Indeed, as backward citation count increases, the likelihood that the patent's maintenance fees are paid also increases.

--- Insert Figure 14 about here ---

Together these results suggest that, in contrast to Assumption A2, the applicant's decision as to whether to cite a patent is increasingly made not based on an individual evaluation of the technological content described in the patent but rather on mechanical reasons such as the patent having been cited in a related application. Although patents cited in applications related to the focal patent are of course likely to be at least somewhat similar to the focal patent, this mechanical cross-citing seems to be driving the substantial decline in mean citation similarity over time. Further, the widespread phenomenon of applicant cross-citing also suggests that changes to the patent citation data generating process have led to a systematic violation of Assumption B2, which states that citations identified as applicant-submitted were identified by inventors.

In conclusion, while it is reasonable to believe that inventors may review and submit a handful of citations to the patent office, it is unlikely that inventors are individually evaluating and selecting hundreds of citations for submission, especially citations that are copied across large families of patents. Consequently, we conclude that the cross-citing explosion in the number of patent citations is largely disconnected from the realities of what inventors actually do, which is likely to have tremendous consequences for empiricists who employ citations to measure innovation-related phenomenon. Figure 1 suggests that in recent years ***the majority*** of citations are likely to be the result of cross-citing, rather than careful review by someone directly involved in the innovation process for the focal patent.

C. Citation Source and Timing

Table **5** presents summary statistics comparing citation attribution in the USPTO bulk data underlying most modern citation data sets with the raw citation submission form data, revealing surprising discrepancies. Of 7.6 million citations identified in the USPTO bulk data as being submitted by the patent

examiner, about 506 thousand (or around 6.6%) were actually submitted first by the patent applicant and only later by the patent examiner.

--- Insert Table 6 about here ---

Citation similarity varies not only with the number of backward citations submitted, but also with citation timing. Table 6 includes estimates of six models that regress various timing-related control variables on citation similarity, with Models 1, 2 and 3 estimated on a subsample restricted to applicant citations and Models 4, 5 and 6 estimated on a subsample restricted to examiner citations. Model 1 shows that applicant citations submitted within the first 90 days of filing the application are about 2.3 percentage points more similar (p<0.001) than other citations. Each successive year between application filing and citation submission is associated with a .487 (p<0.001) reduction in citation similarity. Figure 15 plots this result graphically for all citations (i.e., both examiner and applicant). The magnitudes of these effects are reduced when controlling for the log of backward citation count in Model 2, but the effects remain statistically significant (p<0.001). Model 3 shows that each successive year between the issue date of the citing patent and the issue date of the cited patent is associated with a decrease in similarity of about .29 percentage points (p<0.001). At the same time, an applicant citation that is later resubmitted by the patent examiner is about 8.6 percentage points more similar to the citing patent than other applicant citations (p<0.001). The estimates for examiner citations in Models 4-6 are similar to those for applicant citations, although the coefficient for backward citation count is closer to zero because applicants contribute the bulk of citations in high-citing patents.

--- Insert Table 7 and Figure 15 about here ---

Despite applicant and examiner citations exhibiting similar trends in terms of the relationship between citation similarity and citation timing, examiner citations are much more similar overall to the citing patent. Table 7 presents estimates of five models that compare examiner and applicant citations directly. Each citation is either first submitted by the examiner or the applicant as indicated by the dummy variable *Is Examiner Cite (Internal Data)?*. Each citation is classified as a duplicate if it is eventually submitted by both the applicant and the examiner. The constant term thus 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 percent (p<0.001), and examiner citations are more similar than applicant citations by between 2.2 and 6.2 percentage points (p<0.001). Duplicate citations are the most similar category, at between 11.9 and 12.5 percentage points (p<0.001) more similar than non-duplicate citations.

As in Table 6, citation submission delays in Table 7 are associated with substantially decreased similarity. Each additional year between patent filing and citation submission is associated with a reduction in citation similarity of between .74 and 1.26 percentage points (p<0.001), and each successive citation

submission form is associated with a reduction in citation similarity of between .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. Figures 16 and 17 present several of these results graphically. Importantly, the higher similarity associated with examiner citations seems to be driven entirely by the fact that citation similarity declines with the number of citations submitted and that applicants submit so many more citations than examiners. 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.

--- Insert Table 6, Figure 16, and Figure 17 about here ---

Together these results suggest a systematic violation of Assumption B1 on citation provenance. A small but significant percentage of citations identified in the bulk data as being examiner-submitted were in fact first submitted by applicants. Moreover, these dual citations seem particularly relevant, exhibiting substantially higher technological similarity on average than citations submitted by examiners or applicants alone. Further, many applicant citations seem to be submitted long after the application is filed, in successive rounds of citation submissions associated with lower similarity citations. Such a citation pattern again seems indicative of mechanical cross-citing rather than individual citations being selected by inventors on the basis of technological relevance, which reinforces the evidence in Section 5.B of widespread violation of Assumption B2.

D. Citation Type and Impact

Publication Citations. In Section 3 we argued that the definition of patent citations should be expanded to include citations to published patent applications that ultimately issue as patents. As Figure 4 demonstrates, publication citations constitute a large and growing portion of all patent citations – nearly 25% of citations made by patents issued in 2014. Moreover, as shown in Figure 18, the mean similarity between the citing and cited reference is much higher for citations made to publications than to issued patents.[17] The higher mean similarity is likely due in large part to citation lag. As shown in Section 5.C, citations made to more recent patents tend to be more similar than citations made to older patents. Of course, publication citations still exhibit declining similarity over time for the reasons described in Sections 5.A and 5.B. However, together Figure 4 and Figure 18 illustrate a widespread violation of Assumption C1, which states that citations are made to issued patents. Instead, nearly 25% of patent citations are made to published

[17] We calculate the similarity between the citing patent and the patent that ultimately issues from the cited publication document. However, the similarity calculation includes only the description of the invention, and not the claims. Since the description of the invention is almost identical between filing and issuance, the similarity value is almost certainly nearly identical as well.

applications that ultimately issue as patents, and these citations exhibit substantially higher technological similarity on average than citations made to issued patents.

--- Insert Figure 18 about here ---

Examiner Citations. Earlier we establish that applicants, not examiners, are responsible for generating the flood of citations we observe in recent years. Patent examiners conduct a search for prior art with a particular goal in mind – finding references that suggest the examiner may need to issue a rejection (after which a negotiation between examiner and applicant ensures, and claims are narrowed or the rejection stands). Moreover, patent examiners are evaluated and promoted based in part on their examination quality, which depends in part on search quality. Because patent examiners are time-constrained and operate under a carefully constructed quality control system, they face incentives to find relevant references but disincentives to spend time listing irrelevant references. We believe that these incentives are likely to cause references identified by patent examiners to be more relevant on average than those identified by applicants. Figure 19 supports the argument that examiner citations are on average more similar than applicant citations and do not suffer from the same time trend effects. While the similarity of applicant-submitted citations fell steadily from 2005 to 2014, the similarity of examiner-submitted citations has held steady at approximately 32%. Citations submitted by both the applicant and the examiner (as indicated by USPTO Forms 1449 and 892), therefore, are particularly similar (similarity score of approx. 38%) and also maintain the greatest stability over time.

--- Insert Figure 19 about here ---

Self-Citations. Self-citations are those made by a patent applicant to prior patents by the same applicant. Although self-citations may also be cross-cited, just like other citations, such multiplicity is not necessarily spurious. Instead, frequent self-citations may simply indicate instances in which the applicant builds upon prior innovation. Figure 20 demonstrates that self-citations are on average more similar than citations to others' work (approx. 45% for self-citations, relative to approx. 25% for all other citations, in 2014) and that they are not subject to the same decline in similarity over time. Although self-citation similarity declined somewhat prior to 1995, similarity held relatively constant after that date.

--- Insert Figure 20 about here ---

Rejection Citations. Rejection citations are those cited by a patent examiner to support the examiner's rejection of the application's claims as either not novel or obvious. Rejection citations should be much more similar on average than non-rejecting citations. After all, patent law generally requires that the examiner show that each feature recited in a claim be taught in a rejection reference in order for a rejection (based on either novelty or non-obviousness grounds) to be properly supported. Figure 21 shows that this is indeed the case.

--- Insert Figure 21 about here ---

Altogether, the results on citation type and impact suggest ways in which Assumption C2 is now systematically violated. Different citations exert different types of impacts on patent applications in ways that are observable in the data. At one extreme, an individual citation made by a patent that cites hundreds or thousands of other references seem unlikely to have been closely evaluated by anyone involved in the development of the invention or the preparation and examination of the patent application. At the other extreme, citations selected by patent examiners to support novelty or non-obviousness rejections tend to directly influence the scope of the patent's claims and thus the legal right to exclude afforded by the patent. (Thompson and Kuhn 2016, Kuhn and Thompson 2017). Between these two extremes lie a range of citations, including a substantial portion currently ignored in most empirical analysis (i.e., citations to published patent applications), that vary in characteristics such as technological similarity, source (i.e., applicant vs. examiner), and context (e.g., the number of other citations submitted).

In summary, looking across all of our results, it appears that the data generating process for patent citations has changed significantly in the last 10 years, and it is no longer appropriate to (explicitly or implicitly) make empirical conclusions based on the assumptions outlined in Section 2. The number of citations generated per year has increased dramatically, the distribution of backward citations among citing patents has become much more skewed, and the average citation similarity has decreased significantly. Further, publicly available USPTO bulk data is misleading researchers about the true source of citations (i.e., applicant vs. examiner) for a subset of cases that were likely to be the most influential to the examination process. In Section 6 below, we illustrate some of the specific consequences of these findings. In general, however, the evidence suggests that simple counts of patent citations can no longer be relied on for high quality empirical work, at least without extensive robustness and sensitivity checks tailored to each specific research question. Because patent citations can proxy for many phenomena – including private market value, knowledge flows, technological relatedness, and impact – it is beyond the scope of this article to propose and validate a general-purpose correction for every research area using patent citations. This paper has attempted to make the case that high-quality empirical work cannot continue down the track that it is on, using simple patent citation counts, but clearly more research is required.

6. Consequences

Secular trends in the way patent citations are generated, and problems with publicly available data, are of little interest to scholars if such issues do not have a meaningful impact on the conclusions we draw from empirical research. In this section, we show how several recent and influential articles would have arrived at very different conclusions if they were run on a large sample of data spanning the changes in patent citation

patterns covered in this paper, and/or employing more detailed data that has not been publicly available. We stress that we are not arguing that these articles were sloppy or wrong at the time; our argument is only that recent changes in both patent citations and the quality, quantity, and interpretation of data, have significant consequences for the conclusions presented in previous empirical work and now widely referenced by scholars.

A. Applicant Citations Do Matter

There has been some discussion in the literature as to whether patent citations submitted by applicants matter (Cotropia, Lemley et al. 2013, Frakes and Wasserman 2016). Clearly, whether applicant-submitted patent citations influence the decision to issue a patent, and the scope of the patents that do issue, is an important policy issue in patent law. While most countries do not impose any requirement to disclose prior art, patent applicants in the U.S. incur considerable risk and expense to submit citations. Determining whether such efforts are actually beneficial to the patenting system, therefore, also is a matter of practical importance to many applicants. To this end, Cotropia, Lemley et al. (2013) investigated the extent to which patent examiner prefer their own search results (i.e., examiner citations) over applicant-submitted citations to support rejections. They selected a sample of 1,000 patents and manually identified which citations were used to support rejections, and they found that rejection citations are overwhelmingly examiner-cited rather than applicant-cited. Many observers, therefore, have concluded that applicant-submitted citations simply do not matter.

To re-examine this question, we perform a similar analysis with a much larger sample and more detailed data. Table 8 presents results for a logistic regression, reporting odds-ratios, of whether a citation is selected by the patent examiner to support a rejection of the claims, based on the source of that citation (i.e., provided by examiner or applicant). An examiner-submitted citation is more likely to be selected to reject the patent's claims across all models. As shown in Model 1, a novelty rejection citation (i.e., a 102 rejection) is 4 times as likely to be examiner-submitted than applicant-submitted in a patent with 20 or fewer total backward citations ($p<0.001$). This result makes sense given that the patent examiner is tasked with conducting a *targeted* search of the prior art with the purpose of determining whether to reject the claims; in contrast, the applicant is required to cite any information already known to the applicant that may be relevant (i.e., a much bigger net). Moreover, given that the applicant may have already focused the claims to avoid the cited prior art, many of the citations the applicants provide may not be associated with rejections precisely because the claims already reflect the prior art. One in five of citations submitted by applicants, however, are used in novelty rejections. Applicant citations, therefore, do seem to play a significant role in patent examination in Model 1. Model 2, however, performs the same analysis using the USPTO bulk data, in line

with the analysis of Cotropia, Lemley et al. (2013). In Model 2, the role of applicants appears to be much smaller, a decrease that is therefore due to the fact that in the USPTO bulk data credits key citations initially submitted by applicants to examiners, and duplicate citations are particularly likely to be rejection citations.

--- Insert Table 8 about here ---

Models 3 and 4 perform a similar analysis on an unrestricted sample (i.e., on that includes all citations, regardless of the number of backward citations from the source). Although relaxing this restriction more than triples the sample size of *citations*, the actual increase in the number of *patents* in the sample is much less substantial due to the inclusion of high-citing patents. In the unrestricted sample using updated data, a novelty rejection citation is 8.7 times more likely to be examiner-submitted than applicant submitted ($p<0.001$). This result makes sense because an examiner is unlikely to rely on any particular applicant citation as a rejection citation when the applicant submits hundreds of citations for a single patent. But using the USPTO bulk data to attribute citation gives an even more dramatic increase in the magnitude of the coefficient, with a novelty rejection citation being 15.2 times as likely to be examiner-cited than applicant-cited ($p<0.001$). The results for non-obviousness (103) rejection citations in Models 5-8 are similar to those for novelty (102) rejection citations in Models 1 to 4.

Together these results show that confounding factors, such as misleading data and an increasingly skewed distribution of backward citation counts, make patent citations submitted by applicants appear to be much less influential than they actually are. When accounting for applicant cross-citing (by restricting the sample to citations coming from patents with 20 or fewer total backward citations) and when using more accurate internal USPTO data, applicants contribute about 1 in 5 novelty rejection citations; that ratio falls to just 1 in 15 if these issues are ignored, greatly overstating the irrelevance of applicant citations.

B. Applicant Citations Predict Patent Value

Empirical innovation scholars have long used forward patent citations as a proxy for a patent's value (Trajtenberg 1990, Trajtenberg 1990, Hall, Jaffe et al. 2005), and once the patent office began to indicate the source of a citation (Alcacer and Gittelman 2006, Alcacer, Gittelman et al. 2009), scholars employed the data to refine citation-based analysis. Hegde and Sampat (2009), for example, studied the relationship between forward patent citation counts and an important indicator of patent value: whether the patent applicant pays the three renewal fees imposed by the patent office 4, 8, and 12 years after a patent is issued. They find that examiner-submitted citations are much more predictive of these renewal fee payments than applicant-submitted citations. We perform a similar analysis in Table 9. Like Hegde and Sampat (2009), we select three cohorts of patents for analysis; however, the passage of time allows us to select more recent years with more complete data (i.e., patents issued in 2000, 2004, and 2008) than the cohorts they used (i.e.,

patents issued in 1992, 1996, and 2000). Further, unlike Hegde and Sampat (2009), we count applicant and examiner citations using more accurate internal USPTO citation form data not publicly available.

--- Insert Table 9 about here ---

Like Hegde and Sampat (2009), we find that on the margin, an examiner citation is a better indication of renewal fee payment than an examiner citation. In Model 1, for example, each examiner citation is associated with a 0.2 percentage point increase in the likelihood of paying the first renewal fee (p<0.001), while each applicant citation is associated with only a 0.1 percentage point increase in the likelihood of paying the first renewal fee (p<0.001). Unlike Hegde and Sampat (2009), however, applicant citations are highly significant predictors of renewal fee payments across all models, albeit with a magnitude that is lower than it is for examiner citations. That we find a significant relationship between renewal fee payments and forward applicant citations is not entirely surprising given the larger sample size and more complete data afforded by the later cohorts. However, the more accurate attribution allowed by the internal USPTO data may also play a role since under the publicly available data examiners frequently and receive credit for highly similar and influential citations that were in reality first submitted by applicants.

The changes to applicant citation patterns discussed in Section 5 highlight a potential problem in comparing applicant-submitted and examiner-submitted citations. Because the number of applicant citations submitted has increased dramatically in recent years, comparing the marginal effect of a single forward citation submitted by an applicant to the marginal effect of a single forward citation submitted by an examiner is increasingly an "apples to oranges" comparison. In Models 4, 5 and 6, we log the forward citation counts to account for this trend. When considered on a percentage basis, applicant citations are substantially more predictive of renewal fees than examiner citations. For example, in Model 6, a 100% increase in forward examiner citation count is associated with a 1.5 percentage point increase in the likelihood that the third renewal fee is paid (p<0.001), while a similar increase in forward applicant citations is associated with a 4.0 percentage point increase in the likelihood of fee payment. We also note that the R^2 values increase when logging the forward citation counts, suggesting that these models better fit the data.

Together these results show that employing more complete citations data and taking into account changing trends in citation patterns can lead to precisely the opposite conclusion than the literature might suggest. Specifically, applicant citations, and not examiner citations, provide the better proxy for private patent value when considered on a percentage basis.

C. Strategic Citation

Given that applicant patent citations are both influential in the patent examination process and strong predictors of private patent value, it is worth considering how patent applicants decide which citations to

submit to the patent office. Lampe (2012) finds that patent applicants strategically withhold between 21 and 32 percent of relevant citations. He defines a citation as being potentially strategically withheld if it meets two criteria: first, it must be cited in a focal patent by the patent examiner; second, it must also be cited in a different patent owned by the same firm and issued in a year prior to the focal patent.

Table 10 revisits Lampe's analysis. We select from internal USPTO data all citations from 2001 to 2014 that meet Lampe's definition of strategic withholding. We find that 18.1% of citations in the sample, although identified in the USPTO bulk data as being examiner-submitted, were in fact first submitted by the patent applicant when considering internal USPTO data. Further, we find that 14.1% of citations in the sample were actually first submitted in the focal patent rather than a different patent. Because applicants can submit citations long after a patent application is filed, the fact that Patent A issued prior to Patent B does not necessarily imply that a citation submitted in Patent A was actually submitted to the patent office prior to a citation submitted in Patent B. Together, these results show that 30.0% of the citations that meet Lampe's definition of possibly withheld were objectively *not* withheld.

--- Insert Table 10 about here ---

Applicants also may need time to review citations that are unearthed in one application and provide them to the patent office in another application. Allowing applicants at least a year to perform this administrative task seems reasonable given the slow pace of patent examination (typically several years) as well as the burdens of reviewing many thousands of citations made across even a modestly sized patent portfolio. We find that another 14.8% of citations in the sample were likely *not* withheld when one allows for this administrative delay. Finally, applicants sometimes submit no citations at all (possibly indicating a lack of diligence) and other times submit more than 100 citations (likely too many to individually review with great care). Either situation may be problematic, but neither suggests that an applicant strategically chose to withhold particular citations. Together these conditions total 26.3% of citations in the sample. In sum, 65.3% of citations in the sample that meet Lampe's criteria for being possibly withheld were in fact probably not withheld when considering the totality of the circumstances as shown in the data.

Of the 34.7% of citations in the sample for which we can identify no specific problems, many of them could easily have been not disclosed for some other legitimate reason such as lack of inventor involvement in the patent application process, lack of communication between an attorney and an inventor, or "information overload" from having too many references to review and coordinate across a portfolio. We conclude that, in contrast to the findings of Lampe (2012), the data do not support a finding of widespread strategic withholding of citations by patent applicants. As with Sections 6.B and 6.C, these results show that employing more complete citations data and considering the practicalities and patterns at play in citation submission can lead to a substantially different interpretation.

7. Discussion and Conclusion

We have argued that many of the assumptions used by researchers to support empirical measures based on patent citations are no longer valid. Whereas a given citation made by a patent in 1995 was likely to be highly informative, a citation made by a patent in 2015 is likely to be mostly noise. Moreover, although backward patent citation counts have always been somewhat skewed, the distribution has become even less representational over time, and now a small fraction of patents contributes the vast majority of all backward patent citations. Straightforward approaches to correcting the problems – such as including time controls in regressions or focusing on examiner citations – are unlikely to address the problem and may risk introducing other, as yet unseen biases. At an even deeper level, the publicly available citation data itself is misleading regarding the source of a small but particularly important subset of citations. We showed that these issues have significant consequences for empirical scholars using patent citations to study issues of policy, economics, and firm strategy.

To conclude, we make a general point about the nature of data in the social sciences. The foundational research on patent citations was originally developed by Hall, Jaffe, and Trajtenberg in the late 1990s and early 2000s. They showed that patent citations, properly used, can measure such disparate topics as knowledge flows, patent impact, and firm market value. To enable other researchers to use the same data and tools, they released the NBER Patent Citation Data File (Hall, Jaffe et al. 2001), which enabled a rapid expansion of research in the area, and which helped the field to make great advances. It would be no exaggeration to claim that the NBER dataset impacted the field of innovation in comparable ways to how the CERN release of particle data affected physics. The impact and use of such data is instructive in terms of how it reveals both epistemological similarities and differences between the physical and social sciences. In both disciplines, datasets have had a great impact because they have freed researchers from the arduous and frequently redundant task of data collection.

Nevertheless, economics is not physics, and the data used by economists highlights key differences between the physical and social sciences. Most fundamentally, the phenomena studied by the physical sciences are static over time, while phenomena in the social sciences continuously evolve. We expect a particle interaction measured in the year 2000 to behave in the same manner in 2010. In contrast, the pace of innovation and the characteristics of innovative companies have changed dramatically over the same period. Given social evolution, economists cannot assume that empirical conclusions drawn from data collected even 10-20 years ago will *necessarily* hold today.

Measurement in the physical sciences is usually a direct affair, while economists typically have to rely on proxies to measure their constructs of interest. Properties such as temperature, pressure, and mass are both directly relevant to the physical sciences and measurable with a high degree of precision. In contrast,

knowledge flows and innovation impact are nebulous concepts that can be measured only indirectly and, even then, frequently with the aid of many assumptions. For instance, the foundational research in patent citations relies on the assumption that a citation indicates a substantive connection between the intellectual contribution of the citing and the cited patent. At the time the NBER dataset was constructed, systematically providing direct evidence of a substantive connection was probably infeasible, but those familiar with the workings of the patent system did not really doubt it. Further, if the assumption had not been true, then presumably the foundational conclusions would not have held up.

The patent system itself – as well as how applicants interact with it – has changed considerably over the last decade and has thrown many fundamental assumptions made in the empirical innovation literature into doubt, raising questions about whether the meaning of a citation today is the same as it was when the use of patent citations as a measure in innovation scholarship was initially validated. Whereas the data generating processes in the physical science are isolated and controlled, analogous processes in the social sciences are frequently aggregated and uncontrolled. A firm that files patent applications and submits citations does so not under carefully controlled conditions conducive to causal inference, but rather while operating under numerous, changing, and conflicting incentives. Thus, data generating processes in the social sciences cannot be taken for granted and must be constantly evaluated and re-examined.

Data science offers a path forward. Modern computing techniques allow us to analyze the entire population of data, not just a sample. For instance, in this article we analyze all 60 million citations generated by patents between 1980 and 2014, drawing from related work that analyzes 14.2 trillion patent-to-patent similarity values. Data science also allows us to dig below the data, exploring how and why that data is generated, and ideally unearth what that data is likely to mean. Data access is crucial to supporting such efforts. Patent data has helped the field to make great advancements, in large part due to efforts by others to make their data open and easy to use (Hall, Jaffe et al. 2001). The USPTO has recently launched a range of data accessibility and transparency initiatives to provide more information to scholars and applicants (e.g., www.PatentsView.org). In a similar vein, we also provide access to data via the Patent Research Foundation (www.patrf.org) and encourage scholars to use it in their own research.

References

Abrams, D. S., U. Akcigit and J. Popadak (2013). Patent Value and Citations: Creative Destruction or Strategic Disruption?, National Bureau of Economic Research.

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

Alcacer, J., M. Gittelman and B. Sampat (2009). "Applicant and examiner citations in US patents: An overview and analysis." Research Policy **38**(2): 415-427.

Caballero, R. J. and A. B. Jaffe (1993). How high are the giants' shoulders: An empirical assessment of knowledge spillovers and creative destruction in a model of economic growth. NBER Macroeconomics Annual 1993, Volume 8, MIT press**:** 15-86.

Carley, M., D. Hegde and A. Marco (2015). "What is the probability of receiving a us patent." Yale JL & Tech. **17**: 203.

Corredoira, R. A. and P. M. Banerjee (2015). "Measuring patent's influence on technological evolution: A study of knowledge spanning and subsequent inventive activity." Research Policy **44**(2): 508-521.

Cotropia, C. A., M. A. Lemley and B. Sampat (2013). "Do applicant patent citations matter?" Research Policy **42**(4): 844-854.

Czarnitzki, D., K. Hussinger and B. Leten (2011). "The market value of blocking patent citations." ZEW-Centre for European Economic Research Discussion Paper(11-021).

Frakes, M. D. and M. F. Wasserman (2016). "Is the time allocated to review patent applications inducing examiners to grant invalid patents?: Evidence from micro-level application data." Review of Economics and Statistics.

Gambardella, A., D. Harhoff and B. Verspagen (2008). "The value of European patents." European Management Review **5**(2): 69-84.

Google. (2016). "USPTO Bulk Downloads: PAIR Data." Retrieved January 21, 2016, from https://www.google.com/googlebooks/uspto-patents-pair.html.

Griliches, Z. (1981). "Market value, R&D, and Patents." Economics Letters **7**(2): 183-187.

Hall, B. H., A. Jaffe and M. Trajtenberg (2005). "Market value and patent citations." RAND Journal of economics: 16-38.

Hall, B. H., A. B. Jaffe and M. Trajtenberg (2001). The NBER patent Citations Data File: Lessons Insights and Methodological Tools, NBER.

Hanley, D. (2015). Innovation, Technological Interdependence, and Economic Growth. 2015 Meeting Papers, Society for Economic Dynamics.

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

Hegde, D. and B. Sampat (2009). "Examiner citations, applicant citations, and the private value of patents." Economics Letters **105**(3): 287-289.

Jaffe, A. B. and M. Trajtenberg (1996). "Flows of knowledge from universities and federal laboratories: Modeling the flow of patent citations over time and across institutional and geographic boundaries." Proceedings of the National Academy of Sciences **93**(23): 12671-12677.

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

Jung, H. J. and J. Lee (2015). "The quest for originality: a new typology of knowledge search and breakthrough inventions." Academy of Management Journal.

Kuhn, J. M. (2011). "Information Overload at the US Patent and Trademark Office: Reframing the Duty of Disclosure in Patent Law as a Search and Filter Problem." Yale Journal of Law and Technology **13**(1): 3.

Kuhn, J. M. and N. Thompson (2017). "The Ways We've Been Measuring Patent Scope are Wrong: How to Measure and Draw Causal Inferences with Patent Scope."

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

Lanjouw, J. O. and M. Schankerman (2001). "Characteristics of patent litigation: a window on competition." RAND journal of economics: 129-151.

Lemley, M. A. and C. Shapiro (2005). "Probabilistic patents." Journal of Economic Perspectives: 75-98.

Lerner, J. and A. Seru (2015). The use and misuse of patent data: Issues for corporate finance and beyond. Working Paper. Harvard University.

Lerner, J., M. Sorensen and P. Strömberg (2011). "Private equity and long‐run investment: The case of innovation." The Journal of Finance **66**(2): 445-477.

Marco, A. C. (2007). "The dynamics of patent citations." Economics Letters **94**(2): 290-296.

Mehta, A., M. Rysman and T. Simcoe (2010). "Identifying the age profile of patent citations: New estimates of knowledge diffusion." Journal of Applied Econometrics **25**(7): 1179-1204.

Moser, P., J. Ohmstedt and P. W. Rhode (2013). "Patent Citations and the Size of Patented Inventions: Evidence from Hybrid Corn." Available at SSRN 1888191.

Roach, M. and W. M. Cohen (2013). "Lens or prism? Patent citations as a measure of knowledge flows from public research." Management Science **59**(2): 504-525.

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

Thompson, N. and J. Kuhn (2016). Does Winning a Patent Race Lead to More Follow-on Innovation?

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 M. Fox-Kean (2005). "Patent Citations and the Geography of Knowledge Spillovers: A Reassessment." The American Economic Review **95**(1): 450-460.

Trajtenberg, M. (1990). Economic analysis of product innovation: the case of CT scanners, Harvard University Press.

Trajtenberg, M. (1990). "A penny for your quotes: patent citations and the value of innovations." The Rand Journal of Economics: 172-187.

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

Von Graevenitz, G., S. Wagner and D. Harhoff (2011). "How to measure patent thickets—A novel approach." Economics Letters **111**(1): 6-9.

Wooldridge, J. M. (2014). "Quasi-maximum likelihood estimation and testing for nonlinear models with endogenous explanatory variables." Journal of Econometrics **182**(1): 226-234.

Younge, K. A. and J. Kuhn (2015). Patent-to-Patent Similarity: A Vector Space Model. Working Paper: École Polytechnique Fédérale de Lausanne. Available at SSRN: http://ssrn.com/abstract=2709238.

Tables and Figures

Table 1: Variable definitions and summary statistics.

Variable	Definition	Source	Min	Max	Median	Mean	SD
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 Index	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
Is 102 Rejection Citation?	Dummy = 1 if citation is a rejecting citation (novelty) in an Office Action	USPTO PAIR bulk data	0	1	0	0.03	0.17
Is 103 Rejection Citation?	Dummy = 1 if citation is a rejecting citation (non-obviousness) in an Office Action	USPTO PAIR bulk data	0	1	0	0.06	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 the citation, and the sample includes all citations made by patents issued 2005-2014 and found in either USPTO bulk data or USPTO internal data.

Table 2: Variable correlations.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Similarity 1	1.00													
Is Publication Citation? 2	0.09	1.00												
Is Examiner Citation (Bulk Data)? 3	0.13	0.06	1.00											
Is Examiner Citation (Internal Data)? 4	0.10	0.08	0.94	1.00										
Is Duplicate Citation? 5	0.11	-0.01	0.31	0.11	1.00									
Citation Lag 6	-0.19	-0.46	-0.14	-0.13	-0.04	1.00								
Submission Lag 7	-0.03	0.11	0.36	0.39	0.01	-0.10	1.00							
Is Submitted At Application Filing? 8	0.03	-0.08	-0.33	-0.36	-0.02	0.08	-0.59	1.00						
Form Index 9	-0.07	0.06	-0.10	-0.10	-0.04	-0.07	0.41	-0.22	1.00					
Is Self Citation? 10	0.21	0.06	-0.00	-0.01	0.02	-0.15	-0.04	0.03	-0.01	1.00				
Is 102 Rejection Citation? 11	0.05	-0.05	0.20	0.16	0.10	-0.01	0.08	-0.07	-0.02	-0.02	1.00			
Is 103 Rejection Citation? 12	0.03	-0.07	0.30	0.22	0.13	-0.02	0.12	-0.10	-0.02	-0.03	0.26	1.00		
Backward Citation Count 13	-0.16	-0.03	-0.28	-0.26	-0.07	0.09	-0.02	-0.03	0.15	-0.05	-0.08	-0.11	1.00	
Family Size 14	-0.04	-0.03	-0.12	-0.12	-0.03	0.03	-0.05	0.02	0.04	0.02	-0.03	-0.04	0.25	1.00

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

Table 3: OLS estimates 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: The dependent variable is citation similarity. The sample includes 1 million patent citations made by patents issued between 1976 and 2014, with citations randomly selected from USPTO bulk data. Uncontrolled citation similarity declined significantly during the time period (Model 1), but increased slightly when including fixed effects for backward citation count (Model 2) and when controlling for the log of backward citation count (Model 3). Standard errors in parentheses. Two-tailed tests: ***p<0.001, **p<0.01, *p<0.05.

Table 4: OLS estimates of backward citation counts.

	OLS				*glm: quasipoisson link = log*
	(1)	(2)	(3)	(4)	(5)
Family Size (logged)	0.591***	0.628***	0.592***	0.629***	0.775***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
Family Size		−0.005***		−0.005***	−0.007***
		(0.0001)		(0.0001)	(0.0001)
Patent Year			0.026***	0.026***	0.070***
			(0.0002)	(0.0002)	(0.0005)
Dependent variable logged	*Yes*	*Yes*	*Yes*	*Yes*	*No*
Issue year fixed effects	*Yes*	*Yes*	*No*	*No*	*No*
Observations	2,075,006	2,075,006	2,075,006	2,075,006	2,075,006
R^2	0.193	0.195	0.192	0.194	

Notes: The dependent variable is the logged count of backward citations. Coefficients for family size are log-log elasticities given that family size is also logged. The sample includes all patents issued between 2005 and 2014 and found in the USPTO bulk data. All models include fixed effects for the USPTO Technology Center of the focal patent. Across all models, backward citation count (logged) increases substantially (p<0.001) with family size (logged), even when including fixed effects (Models 1 and 2) or controlling (Models 3, 4, and 5) for the issue year of the focal patent. Standard errors in parentheses. Two-tailed tests: ***p<0.001, **p<0.01, *p<0.05.

Table 5: 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 downloads and USPTO internal data taken directly from application forms. The sample includes each citation (1) made by a patent issued between 2005 and 2014 and (2) that appears in both the USPTO bulk data files and the internal USPTO citation submission data. 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 6: OLS estimates of citation similarity, Applicant vs. Examiner.

	Applicant Citations			Examiner Citations		
	(1)	(2)	(3)	(4)	(5)	(6)
Submission Lag (Years)	−0.487***	−0.395***	−0.703***	−1.167***	−1.177***	−1.336***
	(0.004)	(0.004)	(0.005)	(0.007)	(0.007)	(0.007)
Is Submitted At Application Filing?	2.329***	0.766***	0.865***			
	(0.012)	(0.012)	(0.013)			
Backward Citation Count (logged)		−3.222***	−2.860***		−0.602***	−0.881***
		(0.004)	(0.004)		(0.008)	(0.008)
Citation Lag (Years)			−0.294***			−0.375***
			(0.001)			(0.001)
Is Duplicate Citation?			8.622***			7.724***
			(0.027)			(0.028)
Observations	19,926,317	19,926,317	18,069,080	7,579,866	7,579,866	7,019,598
R^2	0.023	0.061	0.083	0.023	0.023	0.052

Notes: The dependent variable is the similarity of citing-cited pairs of patent citations, as measured by the Vector Space Model. The sample includes every citation that (1) appears in both USPTO bulk data and internal USPTO citation submission forms and (2) was made by a patent issued between 2005 and 2014. All models include application year by technology center fixed effects. Models 1-3 restrict the sample to applicant-submitted citations, while models 4-6 restrict the sample to examiner-submitted citations. Across all models, an increase in the lag between application file and citation issuance is associated with decreased citation similarity, particularly for patent examiners. Citations submitted by both the applicant and the examiner (i.e., duplicate citations) have substantially higher citation similarity, while an increase in backward citation count and an increase in the time between the issuance of the citing and cited patent are both associated with decreased citation similarity. Standard errors in parentheses. Two-tailed tests: ***p<0.001, **p<0.01, *p<0.05.

Table 7: OLS estimates of citation similarity by citation characteristic.

	(1)	(2)	(3)	(4)	(5)	(6)
Is Examiner Citation (Internal Data)?	4.590***	6.207***	4.292***	5.809***	6.020***	−2.227***
	(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 Sequence Number			−0.872***		−0.417***	
			(0.014)		(0.016)	
Is Submitted At Application Filing?				3.355***	1.723***	
				(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 fixed effects	*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: The dependent variable is the similarity of citing-cited pairs of patent citations, as measured by the Vector Space Model. The sample includes 1 million randomly selected citations that (1) appear in both USPTO bulk data and internal USPTO citation submission forms and (2) were made by a patent issued between 2005 and 2014. Across all models, an increase in the lag between application file and citation issuance is associated with decreased citation similarity. Citations submitted by both the applicant and the examiner (i.e., duplicate citations) have substantially higher citation similarity, while an increase in backward citation count and an increase in the time between the issuance of the citing and cited patent are both associated with decreased citation similarity. 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.

Table 8: Logit estimates of examiner selecting citation to reject claims, by citation source.

	Odds ratio of novelty (102) rejection				Odds ratio of non-obviousness (103) rejection			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Is Examiner Citation (Internal Data)?	3.995***		8.726***		4.541***		9.807***	
	(0.053)		(0.094)		(0.048)		(0.083)	
Is Examiner Citation (Bulk Data)?		6.549***		15.242***		7.761***		17.994***
		(0.108)		(0.204)		(0.104)		(0.193)
Citation count	*Any*	*Any*	*≤ 20*	*≤ 20*	*Any*	*Any*	*≤ 20*	*≤ 20*
Observations	944,338	944,338	2,196,305	2,196,305	944,338	944,338	2,196,305	2,196,305

Notes: The dependent variable is whether a patent citation is selected by the examiner to support a rejection of the claims (1 = selected for rejection). Coefficients are reported as odds ratios. The unrestricted sample includes all citations that (1) appear in both USPTO bulk data and internal USPTO citation submission forms and (2) were made by a patent issued between 2005 and 2008. Models 1-4 predict the odds ratio of a novelty (i.e., 102) rejection, while models 5-8 predict the odds ratio of a non-obviousness (i.e., 103) rejection. Models 1, 2, 5, and 6 are estimated on an unrestricted sample, while models 3, 4, 7, and 8 are estimated on a sample restricted to patents that cite 20 or fewer patents in total. An examiner-submitted citation is many times more likely to be selected to reject the patent's claims across all models. However, the magnitude of the effect is much smaller when attributing citations based on the internal USPTO citation submission form data (Models 1, 3, 5, and 7) than attributing citations based on the bulk data. (Models 2, 4, 6, and 8). Standard errors in parentheses. Two-tailed tests: ***p<0.001, **p<0.01, *p<0.05.

Table 9: OLS estimates of paying the renewal fee at 4, 8, and 12 years.

	Dependent variable:					
	Renewed at 4?	Renewed at 8?	Renewed at 12?	Renewed at 4?	Renewed at 8?	Renewed at 12?
	(1)	(2)	(3)	(4)	(5)	(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: The 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 that (1) appear in internal USPTO citation submission forms and (2) were 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.

Table 10: A re-evaluation of the evidence for "strategically withheld" citations.

Problem	Count	Percentage	Description
Inaccurate citation source	65,659	18.05%	The citation source assumption does not hold (i.e. the reference was initially submitted by the applicant, not the examiner).
Citation dates out of order	51,175	14.07%	The citation ordering assumption does not hold (i.e. the reference was not submitted earlier in a patent by the same firm).
Subtotal: Definitely not withheld	**109,054**	**29.98%**	**Inaccurate data.**
The citation date assumption was too strict	53,644	14.75%	The reference was first cited in a patent by the focal firm less than one year prior.
No citations submitted	81,966	22.54%	The applicant submitted no citations.
Too many citations for complete review	13,723	3.77%	The applicant submitted 100 or more citations.
Subtotal: Probably not withheld	**128,344**	**35.29%**	**Incomplete data.**
Total	**237,398**	**65.27%**	**Inaccurate or overstated.**

Notes: The sample includes all citations that (1) appear in both USPTO bulk data and internal USPTO citation submission forms, (2) that were made by a patent issued between 2005 and 2014, (3) that were listed as examiner-submitted in the USPTO bulk data, and (4) that were also made in a different patent (a) that issued in a year prior to the focal patent and (b) owned by the same firm as the focal patent.

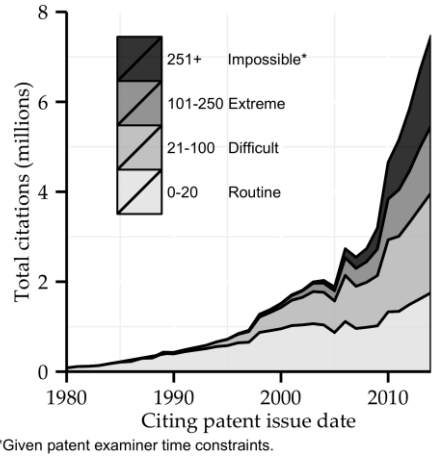

Figure 1. Line plot of total citations made by all patents over time (in millions), divided into areas by the backward citation count of the citing patent.

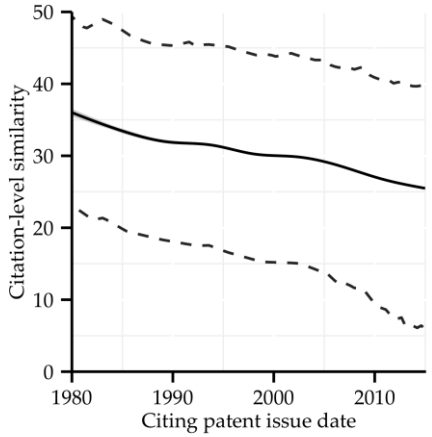

Figure 2. Mean similarity of patent citations by citing patent issue dates, with inner-quartiles shown as dashed lines.

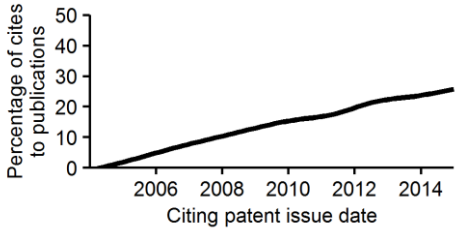

Figure 3. Publication citations as a percentage of all citations over time.

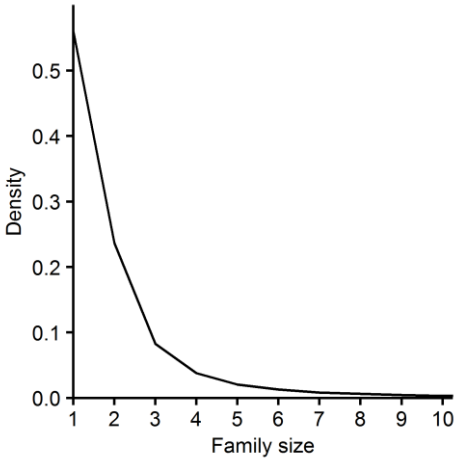

Figure 4. Density plot of patents by family size.

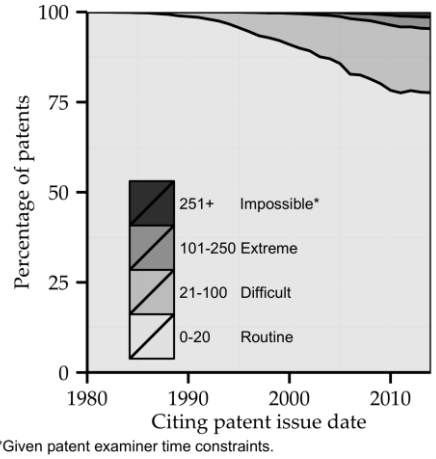

Figure 5. Area plot of the percentage of patents by backward patent citation count over time.

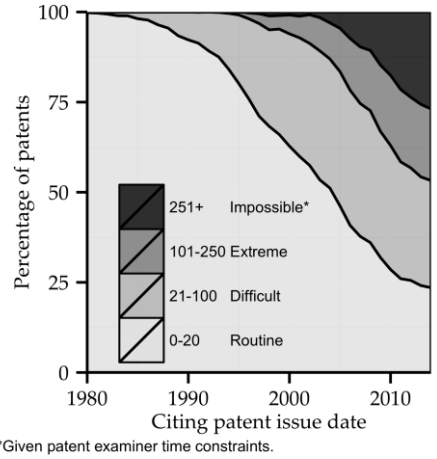

Figure 6. Area plot of the percentage of citations by backward patent citation count of the citing patent over time.

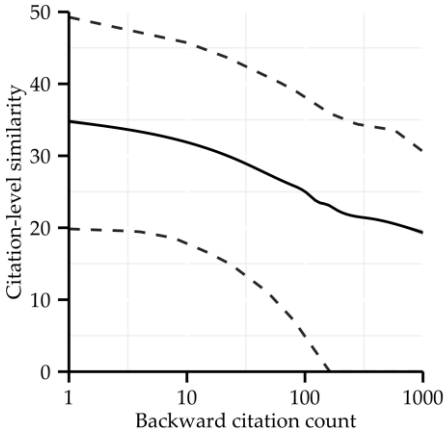

Figure 7. Mean citation similarity by the number of backward citations made by citing patent, with inner-quartiles represented by dashed lines.

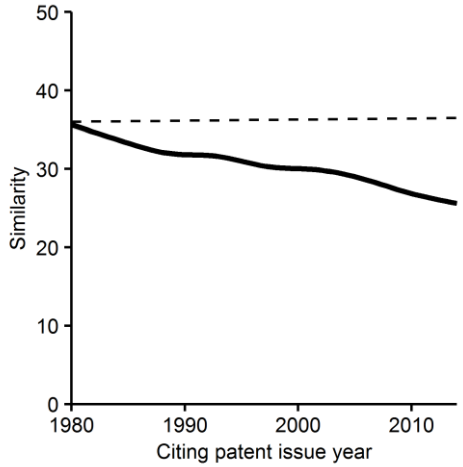

Figure 8. Mean similarity of patent citations by citing patent issue date. The dashed line plots the estimate of an ordinary least squares regression that includes fixed effects for the backward citation count of the citing patent.

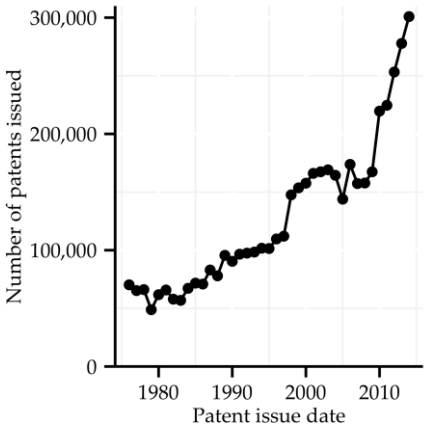

Figure 9. Line plot of the number of patents by issued each year.

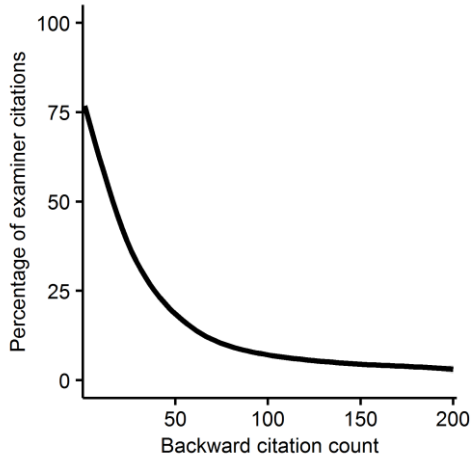

Figure 10. Patent-level plot of the percentage of examiner citations by the backward citation count.

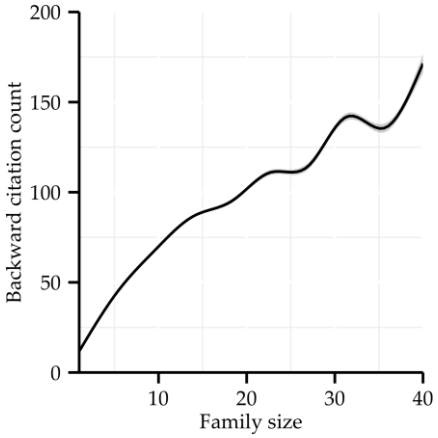

Figure 11. Patent-level plot of the mean backward citation count by family size.

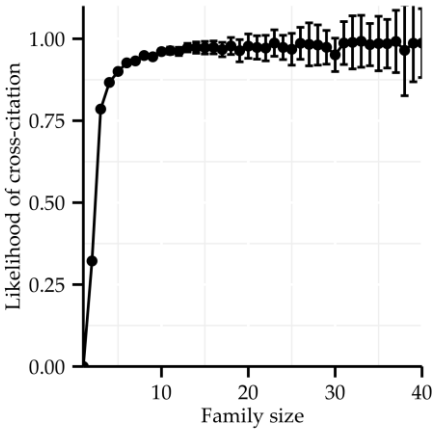

Figure 12. Citation-level plot of the mean likelihood that the cited reference is also cited by a family member of the citing patent, by the family size of the citing patent.

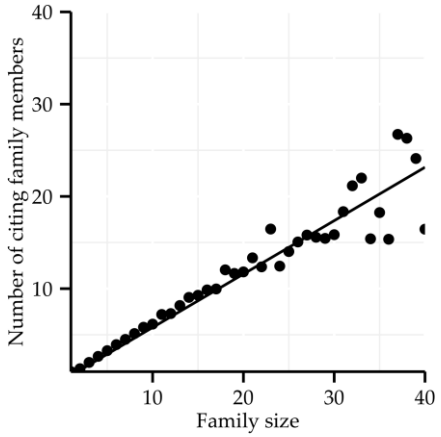

Figure 13. Citation-level plot of the mean number of times the cited patent appears as a citation across a patent family, by the family size of the citing patent.

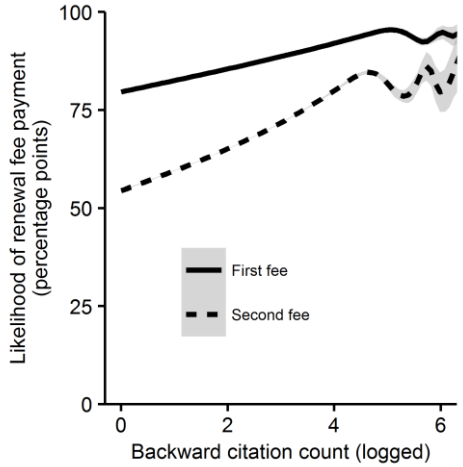

Figure 14. Patent-level plot of maintenance fee payments as a function of backward citation count.

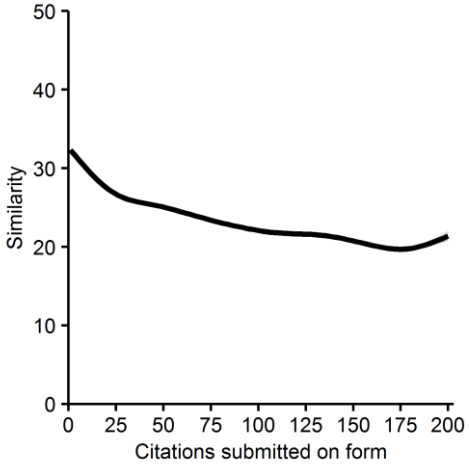

Figure 15. Citation-level plot of mean citation similarity as a function of the number of citations submitted on the same citation submission form.

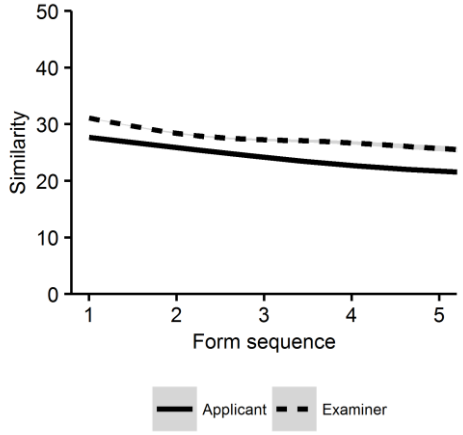

Figure 16. Line plot of similarity by the index of the citation submission form. For instance, applicant form 2 is the second citation submission form submitted by the applicant.

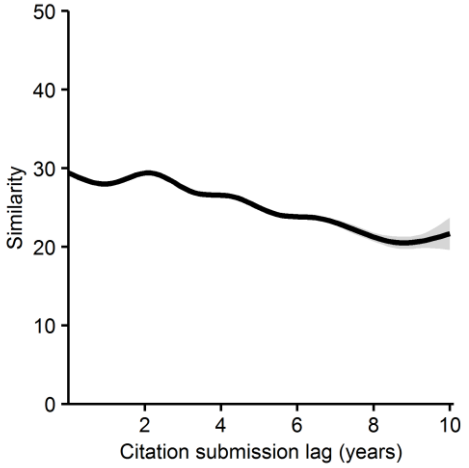

Figure 17. Citation-level plot of mean citation similarity as a function of the time between application filing and citation submission.

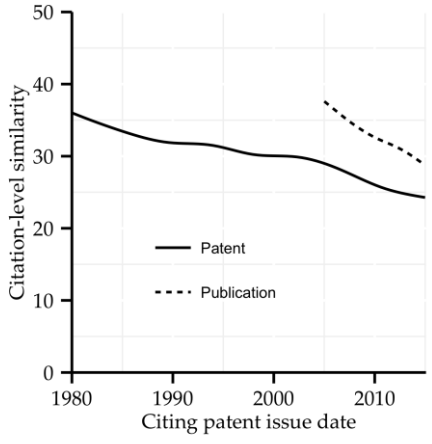

Figure 18. Mean citation similarity over time for citations to published patent applications.

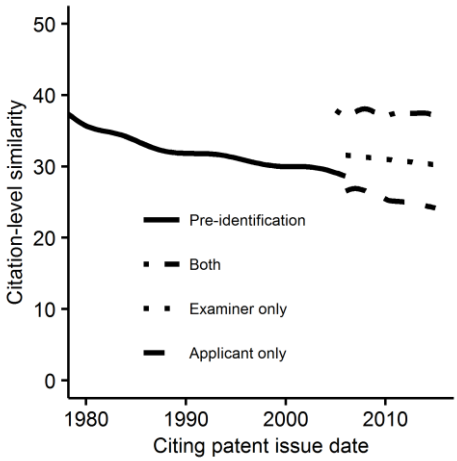

Figure 19. Mean citation similarity by citation source over time.

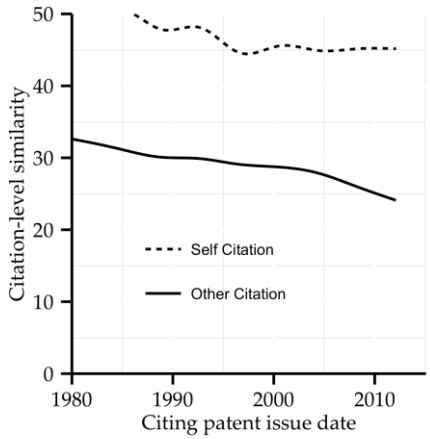

Figure 20. Mean citation similarity by the identity of the firm that originally owned the cited patent.

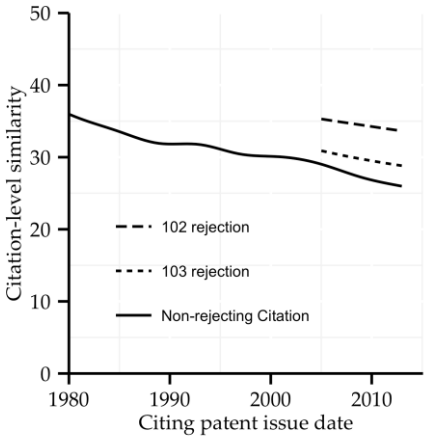

Figure 21. Mean citation similarity over time for rejecting and non-rejecting citations.

Appendix A: Data Construction

A1. Additional details regarding construction of patent-to-patent textual similarity data.

(See Younge and Kuhn, working paper (SSRN)).

A2. Additional details regarding USPTO raw citation form data

Administrative support specialists at the USPTO are tasked with recording all US references (patent grants and pre-grant publications) listed on forms 1449 and 892, along with identification information for the form itself including the application number, the type (1449 or 892), and the date the form was added to the file wrapper. Errors could be made in any field, but they are more likely to occur in the identification numbers for patents and pre-grant publications. This is because applicants (and to a lesser extent examiners) record patent numbers in different ways. Some use a leading "US," and some include the document kind code,[18] for example A1 for a pre-grant publication, B1 for a patent, or C1 for a reexamination certificate. In the cases where foreign patent or application numbers have a similar format to that of US patents (i.e., 7-digit numbers), they may be confused for US patent numbers if the leading jurisdiction code is omitted by the applicant or unnoticed by the data entry clerk.

After removing all punctuation and superfluous marks, we take 7-digit numbers beginning with "US" to be patent numbers. If there is no country identifier, we take them to be US patent numbers as long as they have a publication date earlier than the date of the form. We match application numbers to the resulting patent number for those applications that are granted. Finally, we include citing-cited patent pairs in our sample if the resulting pair matches with a citing-cited patent pair from the USPTO bulk data. We separately identify the form type in order to identify the source of the reference (applicant or examiner). Further, for each cited reference we keep the earliest form date for each form type; this has the effect of dropping duplicates from later forms. Applicants follow different protocols for disclosures. Some repeat all previously cited references when providing new references; some only include the new references. We remove duplicates in order to identify the new information in any disclosure. Note that for a given reference, we track the earliest applicant reference and the earliest examiner reference separately. Thus, we can identify four types of references: those that are first cited by the examiner, first cited by the applicant, only cited by the examiner, or only cited by the applicant.

[18] See https://www.uspto.gov/learning-and-resources/support-centers/electronic-business-center/kind-codes-included-uspto-patent.

Appendix B: Proposed Robustness Checks for Other Empirical Studies

Scholars have long used patent citations in empirical analysis, but the changes to the citation data generating process outlined in this article suggest that scholars should take care to ensure that citations accurately measure the phenomenon of interest in a particular empirical context. However, the broad range of phenomenon patent citations are used to measure coupled with the complexity of citations as an empirical tool precludes a simple, one-size-fits-all correction to account for changes over time. Instead, we suggest that citation-based empirical work be conducted with an understanding of how citations are actually generated, and that substantive conclusions drawn from citation-based empirical analysis be supported by a variety of robustness checks that may depend on the empirical application. In so doing, we echo the recommendations of Lerner and Seru (2015), who emphasize the importance of ensuring that citation-based results are not driven by selection effects, are consistent across the sample, and are robust to alternate empirical approaches.

We propose three criteria on which to evaluate robustness checks when conducting empirical analysis using patent citations, and we evaluate possible robustness checks according to these criteria. First, citation-based measures should, as nearly as possible, represent the full range of patents. Currently, a small proportion of citing patents are responsible for the large majority of citations. Corrected citation measures employed as robustness checks should seek to reduce or eliminate this imbalance. Second, a citation-based measure should be evaluated (and potentially selected) based on the textual similarity between the citing and cited patents. Any particular citation may be important and meaningful without exhibiting a high similarity to the citing patent. However, Younge and Kuhn (2015) show that textual similarity computed using a VSM model is a strong and significant measure of technological relatedness. Thus, patent similarity on average correlates with more informative citations. Third, a robust citation-based measure should reflect the institutional characteristics that define the data generating process. For example, rejection references represent the subset of citations that actively influence the scope of the claims via the process of patent examination, and a citation-based measure that excludes or down-weights them risks discarding what is potentially the most informative set of citations when considering phenomena such as patent impact.

Truncation bias. The most frequently used sample correction to date relates to rightward truncation bias. Lerner and Seru (2015) lay out three such approaches: 1) estimating the future distribution of citations from an earlier, un-truncated distribution (Jaffe and Trajtenberg 1996, Hall, Jaffe et al. 2001, Hall, Jaffe et al. 2005); 2) estimating truncation bias via technology class and year (Hall, Jaffe et al. 2001), and 3) limiting the analysis to citations in a fixed window, such as three years after grant (Lerner, Sorensen et al. 2011). We do not directly examine truncation bias. Nevertheless, the changes to the data generating process that we described in this article would support the third approach as being the most robust. An earlier un-truncated

distribution of forward citations may be a poor predictor of the future distribution when the pattern of backward citations changes dramatically over time; using patent class to construct a counterfactual also risks introducing unanticipated biases (Thompson and Fox-Kean 2005, Younge and Kuhn 2015). Another approach, proposed by Marco (2007), compares the residual of the observed hazard rate of citation with the predicted rate to produce a multiplier. Regardless of the approach used, correcting for rightward truncation remains a first order concern when conducting empirical analysis using patent citations.

Year fixed effects. A common approach for controlling for changes in the patenting system over time is to include year fixed effects in regression models. Patent citations, however, do not easily fit into a year fixed effects control structure. A patent citation is a link between a later-issued citing patent and an earlier-issued cited patent. Measures based on patent citations often use the *cited* patent as the focal point, counting the number of forward citations that the focal patent has received. However, the data is generated by the *citing* patent, which is issued later than the cited patent, frequently by many years. Accordingly, patent citations issued by patents in 2014 effectively "reach back" to the past, affecting the forward citation counts of earlier patents. The fact that patent citations are generated in the backward direction but typically aggregated in the forward direction means that the effect of time controls on regression models can be opaque. Different firms, technologies, or industries can exhibit very different degrees of citation time lags. Thus, depending on the context, including different types and combinations of fixed effects may be a necessary, though not sufficient, condition for demonstrating robustness.

Examiner vs. applicant citations. Another approach to analyzing patent citations is to analyze only those citations generated by the examiner, or only those citations generated by the applicant. The main problem with this approach is that identifying whether an applicant or examiner generated a particular citation is not straightforward. The bulk data provided by the USPTO is in some instances inaccurate and in other instances misleading regarding a citation's source. Further, these inaccuracies are particularly likely for the influential citations that are of most interest to scholars. Also, in unreported results we find that over 10% of citations generated by patents issued 2005 or later are listed as being examiner-cited in one member of a patent family and applicant-cited in another member of a patent family, further muddying the waters regarding the provenance of a citation. Nevertheless, separately analyzing examiner-submitted and applicant-submitted citations may help to indicate the robustness of a particular empirical approach.

Collapsing to family. Another approach might be to collapse patent citations to the patent family, with family defined under either the European or U.S. conception. Recognizing the significance of patent families is important – treating as distinct every citation made to the same patent by different members of the same family seems unwise in light of the cross-citing phenomenon. Simply collapsing to family, however, is unlikely to alone correct for the empirical problems we have identified, and may introduce new

bias into the analysis. Our results indicate that it is not just the *quantity* of patent citations being distorted by patent families; but that the *quality* of citations also change when a citing patent includes an abnormal number of references. Collapsing citations to the patent family level ignores these differences in quality but nevertheless may be an important robustness check.

Citation weighting. Another way to address changes to the patent system over time might be to weight citations when using them in empirical analysis. Hall, Jaffe et al. (2005) showed that citation-weighted patent counts predict firm market value much better than unweighted patent counts because some patents are more important and impactful than others. In a similar fashion, carefully selected citations made by patents that cite only a handful of other references might be more informative than citations made by patents that cite hundreds or even thousands of references. However, the specific weighting scheme to be used is likely to depend on the specific empirical context. For instance, when using citations to measure knowledge flows between inventors, citations made by patents that cite many hundreds of references may need to be weighted close to zero. However, such citations may prove to be more meaningful when using citations to measure generalized technological impact.

Citation similarity. Citation similarity may serve as a criterion with which to evaluate potential robustness checks, but similarity may also be used as an independent robustness check. For instance, citations may be divided into groups based on patent grant date, firm, or some other criteria. If the citations in different groups have substantially different citing-cited similarity, then the overall pool of citations may be subject to selection effects that preclude certain types of inference.

Alternate counting techniques. A final approach to citation robustness checks is to exclude some citations completely when calculate forward citations. For example, in unreported results, we find that citation similarity remains nearly constant over time when excluding citations made by patents that cite more than 20 references. More sophisticated techniques may also be employed, such as counting all examiner-submitted citations as well as all applicant-submitted citations below a designated threshold. Again, although a single, one-size-fits-all selection approach cannot possibly apply to the myriad applications of patent citations, using a reasonable alternate counting technique as a robustness check may do much to bolster conclusions drawn from citation-based empirical analysis.