Patent Citations: An Examination of the Data Generating Process

Jeffrey M. Kuhn
Haas School of Business
University of California Berkeley
jeffrey_kuhn@haas.berkeley.edu

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

February 1, 2016

Abstract

Existing measures of innovation often rely on patent citations to indicate intellectual lineage and impact. We show that the data generating process for patent citations has changed substantially since citation-based measures were originally validated. Today, far more citations are created per patent, and the mean technological similarity between citing and cited patents has fallen significantly. These changes suggest that the use of patent citations for scholarship needs to be re-validated. We develop a novel vector space model to examine the information content of patent citations, and show that methods for sub-setting and/or weighting informative citations can substantially improve the predictive power of patent citation measures. We make data for a basic correction available for future scholarship through the Patent Research Foundation.

I. Introduction

A substantial amount of research examines patenting as an empirical measure of innovative activity. 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 theory that the impact of a patent increases as subsequent innovation references a patent (Trajtenberg 1990). Studies have used patent citations used to measure geographic information spillovers (Jaffe, Trajtenberg et al. 1992), cumulative innovation (Caballero and Jaffe 1993, Trajtenberg, Henderson et al. 1997), firm market value (Hall, Jaffe et al. 2005), private patent value, (Lanjouw and Schankerman 2001, Harhoff, Scherer et al. 2003), and social importance (Moser, Ohmstedt et al. 2013). Broadly speaking, citations have been used as indicators of technological impact, reliance, diversity, accumulation, distance, quality, value, and much more.

Recent research, however, calls into question a straightforward interpretation of patent citation counts. Abrams, Akcigit et al. (2013) find an inverted-U relationship between patent citations and market value and suggest that high citation counts indicate strategic uses 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 (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 (Marco 2007) because citation counts have inflated substantially over time. Failing to correct for time period, technology, and geographic region can introduce significant bias into an analysis (Lerner and Seru 2014). Correcting citation counts, however, can be challenging 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 5 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 an exponential increase in the number of citations generated per year, and relate this 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 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. Conversations with patent examiners suggest that reviewing more than 100 citations is extremely difficult given the time constraints of the patenting system.

--- Insert Figure 1 about here ---

To understand 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.[1] The measure allows us to determine the technological distance between each citing/cited pair 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. One

[1] 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.

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 ---

Our analysis provides *prima facie* evidence that the citation generating process is now generating significant measurement error for many academic studies, such that future research with patent citation-based measures will require new empirical methods and measures. We discuss potential corrections for the problem, and although developing and validating every potential correction is beyond the scope of this article, we propose one straightforward correction that may suffice in a majority of empirical contexts. Along the way, we also propose a new test—based on similarity data—for comparing the information quality of different groups (or corrections) of patent citations, and share the data for the correction through the Patent Research Foundation.

II. The Duty to Disclose and Patent Families

Several institutional changes in the patenting regime may be leading to a proliferation of low-information citations. First, patent applicants owe a duty of disclosure to the patent office under which an applicant must cite any reference the applicant knows about and believes to be relevant. 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." Second, 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 patent families appear to be responding by automatically "cross-citing" (copying) large sets of references across entire families of applications.

A. Patent Applicants

A patent provides the right to exclude others from making, using, and selling an invention, but only if the patent application describes and claims a new and non-obvious invention. To determine whether an application meets those requirements, an examiner at the patent office compares an application's claims to the prior art. The examiner is required to conduct their own search for prior art, but doing so is often difficult and time-consuming. The applicant, however, often possesses documents discovered during their own prior art search, or documents identified in other patent applications, that may be relevant to the examination of the patent. Since disclosing such information may not seem to be in an applicant's interest, 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 invalidate the patent altogether, 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 and 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 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.

B. Patent Families

Changes in the law regarding the duty of disclosure, however, do not alone 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 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. 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.

Filing families of patents on the same technology is an expensive proposition, and scholars have found that patent family size is evidence of private patent value (Harhoff, Scherer et al. 2003). Because patent families are valuable, applicants have greater incentives to avoid unnecessary risks and accordingly greater incentives to cite references. At the same time, patent families unearth more potential citations since the USPTO performs a new prior art search for every member of the family. The easiest and safest way to ensure that the duty of disclosure is met is to simply copy every citation made in any member of the family into every other member of the family—a strategy referred to by practitioners as "cross-citing."

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.

III. Data

To investigate changes in the pattern of patent citations, we constructed a patent-level sample with four components: 1) bibliographic data on each patent; 2) disambiguation of the patent family (if any); 3) citing/cited pairs of citations; and 4) patent-to-patent textual similarity.

A. Patents

Our sample of patents includes bibliographic information for 5,027,882, patents issued from January 6, 1976 to March 17, 2015. For the period from 1976 to 2004, we employ the NBER

patent data file (Hall, Jaffe et al. 2001). For the period from 2005 to 2015, we collect bibliographic data from extensible markup language (XML) files published by the USPTO and aggregated by Google (2016). These files contain a wealth of information about each issued patent, including patent citations, patent priority claims, and inventor names. We thank Douglas Hanley (2015) for firm disambiguation data to link each patent to an assignee identifier.

B. Patent-to-Patent Similarity

We calculated the textual similarity of each citing/cited dyad in our sample. The patent similarity value is drawn from related work in which we compute and validate a pairwise similarity measure for all 14 trillion possible pairs of patents (Younge and Kuhn 2015). We show in the related work that the computed similarity measure correlates strongly with both expert and lay person evaluations, and that it also predicts such characteristics as shared patent class and patent family. On a scale of 0 to 100, citations in our sample have a mean similarity value of 27.9. Figure 3 shows the distribution of similarity values for different types of citations, as well as the distribution of similarity values for randomly selected pairs of patents. As expected, the similarity of patent citations is much higher than random, but citation similarity can differ on many aspects and Figure 3 does not reveal how different types of citations have changed over time. We examine these issues in our analysis below.

--- Insert Figure 3 about here ---

C. Patent Families

We calculated an identifier for each patent family based on both "explicit" and "implicit" families, where applicable. We define "explicit" families 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 an "explicit patent family" as including all patents linked by any priority relationship.

Members of a patent family, however, do not need be linked explicitly by a priority claim. For example, a patent applicant can file virtually identical patent applications that claim different aspects of the same invention but are not linked by any priority relationship. Such applications are not identified as belonging to the same patent family in commonly used data sets such as PatStat. We therefore identify "implicit" families by using similarity data (described below and validated in Younge and Kuhn 2015) to identify highly similar patents issued to the same firm. In related work, we show that nearly all pairs of patents in the same family exhibit similarity values of greater than 95, while such similarity values are exceptionally rare for non-family members. We therefore define pairs of patents whose similarity exceeds 95 as being in the same implicit family.

--- Insert Table 1 about here ---

D. Patent Citations

Our patent citation sample includes 62,104,091 patent citations from January 6, 1976 to December 30, 2014. For the period from 1976 to 2004, we take citations from the NBER patent data file (Hall, Jaffe et al. 2001). For the period from 2005 to 2014, we extract citations from

extensible markup language (XML) files published by the USPTO and aggregated by Google. 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. As an important improvement over how citations have been collected and counted in the past, we also include approximately 7.6 million citations to patent publications. The citation data includes several citation-level indicator fields. For example, in 2001 the USPTO started to identify whether a citation was provided by the patent examiner or the patent applicant. We collect this information from the raw XML data for patents issued 2005 or later. Examiner-citations make up 24.6% of citations made by patents issued 2005 or later. We also use the unique assignee identifier associated with each patent to construct an indicator for each citation as to whether the citing and cited patent are 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 6.5% of our sample.

--- Insert Table 2 about here ---

IV. Results

As shown in Figure 4, 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 4, 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 4 and 5 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 5 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 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 primarily 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. Moreover, it is troubling that the citations being produced by highly citing patents appear to be particularly uninformative, even for the tiny fraction of the phenomenon that they do represent. Figure 6 shows 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 three or less references is 34, the similarity for patents with hundreds of cites falls to 16.

--- Insert Figure 6 about here ---

To better understand the nature of the dramatic changes identified above, we now analyze several related factors. First, we find that patent applicants – not patent examiners – are the actors adding citations to patents with hundreds of references. Figure 7 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. Second, an increase in volume of patents issued per year is compounding the problem. Figure 8 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 applicant citation behavior mean that citations generated today are far more numerous and far less informative than citations in years past. We suspect that these changes are likely to become even more profound in the coming years.

--- Insert Figures 7 and 8 about here ---

Next we turn to patent families. Table 3 reports 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 9 plots these results, showing 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 Tables 3 and 4 and Figure 8 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 citation 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 10 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 11 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 10 and 11 about here ---

In summary, the data generating process for patent citations appears to have changed significantly in the last 10 years. The number of citations generated per year has increased exponentially, the distribution of backward citations among citing patents has become much more skewed, and the average citation similarity has decreased significantly. The evidence therefore suggests that simple counts of patent citations may no longer be adequate for high quality empirical work; instead, researchers may have to sample the data in an informed way. Patent citations can proxy for many phenomena – including firm market value, knowledge flows, technological relatedness, and patent impact – and so it is beyond the scope of this article to propose and validate a correction for each area. Nevertheless, in the next section we propose criteria for researchers to follow when making such corrections, and offer a simple correction that may work well for many purposes.

V. Sample Correction

In this section we propose a simple correction to address the concerns identified earlier. The correction samples citations that are more likely to be informative (in terms of legal impact, cumulative innovation, and knowledge flows), and drops citations that are more likely to be

unrelated and overrepresented. Although any sampling procedure can be a source of bias (including one that simply takes *all* citations), we show that the proposed correction improves the sample on several important factors. In short, we start from a pool of all patent citations – including citations initially made to publications – and then retain citations from the following groups: (1) examiner citations, (2) self citations, and (3) citations made by patents that cite 20 or fewer references.

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 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 12 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-provided citations fell steadily from 2005 to 2014, the similarity of examiner-provided citations has held steady at approximately 30%. We also show in Appendix B that examiner citations are much more likely to lead to rejections, and thus have a more substantial legal impact on the citing patent.

--- Insert Figure 12 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 13 demonstrates that self-citations are on average more similar than citations to others' work 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 13 about here ---

Routine Citations. A large body of work in the literature suggests citations can be a reasonable indicator of the impact of a cited patent on the citing patent. Figure 6, however, shows that the similarity of citations declines dramatically as the number of backward citations increases. We therefore investigated whether a subsample of citations – such as those making a "routine" number of 20 or fewer citations – did *not* exhibit the same decline over time. Figure 14 tests that question by plotting the mean similarity over time for citations made by patents citing fewer than 20 references and fewer than 100 references – as well as the average overall. Given the relative stability in similarity of the first (up to) 20 citations, and evidence that similarity falls as more citations are included, we adopt a simple cut off of 20 backward citations, after which we exclude applicant-submitted citations that were not otherwise self-citations or examiner added citations.

--- Insert Figure 14 about here ---

Results from the correction. As expected, we find that our sample of corrected citations is much more representative of citing patents than uncorrected citations. Figure 15 plots Lorenz curves for citations made in 2014: in the uncorrected sample, 25% of patents contribute 75% of citations; in the corrected sample, 25% of patents contribute 50% of citations. We also find that the corrected sample exhibits a representativeness that is stable over time. Figure 16 plots Gini coefficients for citations made by patents issued in a given year: the Gini coefficient for the uncorrected sample rises steeply between 1995 and 2014; the Gini coefficient for the corrected sample is both lower and nearly constant over time. The improvements in the Lorenz curve (demonstrating less inequality and greater representativeness), in conjunction with the stable Gini coefficients over time, suggest that the correction is much more representative of citing patents than otherwise.

--- Insert Figures 15 and 16 about here ---

We also find that citations in the corrected sample are much more similar than citations in the uncorrected sample, with a roughly constant level of similarity over time. Figure 17 plots the similarity of patents in the corrected ("keep") and uncorrected ("exclude") samples over time, and includes the set of all citations for reference. The similarity of the corrected sample is stable at about 33, in line with pre-1995 levels, while the similarity of the citations excluded from the sample is much lower, dropping to nearly 20.

--- Insert Figure 17 about here ---

As noted earlier, every backward citation becomes a forward citation when tabulated in the forward direction from the perspective of cited patents. To understand how our correction

affects the calculation of forward citations (the measure most often used in the literature), Figure 18 plots the corrected forward citation count on the vertical axis compared to the uncorrected forward citation count on the horizontal axis. As the correction drops cases and therefore always attenuates the count of citations, results are mechanically constrained to the area below the 45-degree diagonal. We observe in Figure 18 considerable dispersion from the correction, where some cases demonstrate little to no correction, while others demonstrate high levels of correction and fall close to zero. Manual inspection suggests that cases going to zero are associated with big families that cross-cited frequently. We also observe that the correction leaves the resulting distribution of forward citations somewhat less skewed.

--- Insert Figure 18 about here ---

Conclusion

In this paper, we have argued that many of the assumptions used by researchers to support empirical measures based on patent citations are no longer as valid as they once were. 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 skewed, the distribution has become even less representational over time, such that now a small fraction of patents contribute the vast majority of all backward patent citations. Straightforward approaches to correcting these problems – such as including time controls in regressions or focusing on examiner citations – are unlikely to convincingly address the problem and risk introducing other, as yet unseen biases (see Appendix C for an in-depth discussion of

corrections, past and present). We proposed and validated a simple approach for correcting many of the empirical problems we identify in this article.

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 particle data impacted 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, 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.

Changes to the patent system – and in how applicants interact with it – have changed considerable over the last decade and have thrown the original assumptions 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 1975 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 science exploration provides for a deeper, more nuanced understanding of empirical research, with tighter connections between the empirical analysis and the theory that it is designed to test.

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). In a similar vein, we provide access to the data for the correction 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.

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.

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). Firm Name Disambiguation Project. University of Pittsburgh.

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.

Jaffe, A. B., M. Trajtenberg and R. Henderson (1992). Geographic localization of knowledge spillovers as evidenced by patent citations, National Bureau of Economic Research.

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.

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 (2014). The use and abuse of patent data, Working Paper.

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.

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

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.

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. M. Kuhn (2015). "Patent-to-Patent Similarity: A Vector Space Model." Available at SSRN.

Table 1. Number and percentage of patents by family size.

Size	Patents	Percent
1	1,162,730	56%
2	491,437	23.7%
3-5	294,834	14.2%
6-10	74,425	3.6%
11-20	30,457	1.5%
21-100	19,108	0.9%
100+	3,250	0.2%

Table 2. Data sources and characteristics for citing/cited pairs of patent citations.

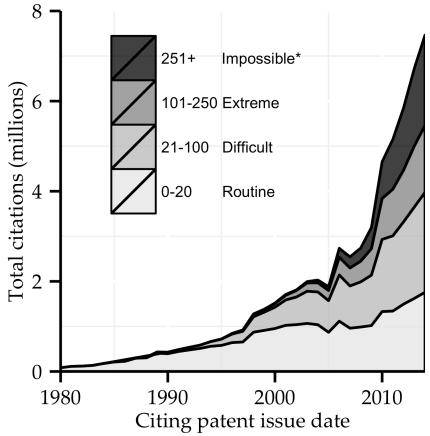
	Source	Observations	Self	Other	Examiner	Applicant	Rejection	Mean.Similarity
1	XML Patent	35,391,516	4.8%	95.2%	24.6%	75.4%	4.1%	25.8%
2	XML Publication	7,555,527	8.2%	91.8%	31.4%	68.6%	-	31.1%
3	NBER Patent	19,157,048	9.2%	90.8%	-	-	-	30.5%
4	Total	62,104,091	6.5%	93.5%	17.8%	51.3%	2.5%	27.9%

Table 3. Least squares regressions on the backward citation count of a focal patent.

	Dependent variable:				
	Backward citation count (logged)				
	OLS				*glm: quasipoisson link = log*
	(1)	(2)	(3)	(4)	(5)
Family size (logged)	0.591***	0.614***	0.628***	0.656***	0.790***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
Patent year		0.100***		0.101***	0.151***
		(0.0005)		(0.0005)	(0.001)
Family size			−0.005***	−0.005***	−0.007***
			(0.0001)	(0.0001)	(0.0001)
Constant	−0.393	−0.565	−0.397	−0.572	−0.172
	(0.830)	(0.821)	(0.829)	(0.820)	(6.108)
Observations	2,076,241	2,076,241	2,076,241	2,076,241	2,076,241
R^2	0.189	0.206	0.190	0.208	
Adjusted R^2	0.189	0.206	0.190	0.208	

Note: *p<0.1; **p<0.05; ***p<0.01
USPTO Technology Center and application year fixed effects

Figure 1. Count of backward citations, separated by total citations made by citing patent (in millions).



*Given patent examiner time constraints.

Figure 2. Mean similarity of patent citations by citing patent issue date (inner-quartiles as dashed lines).

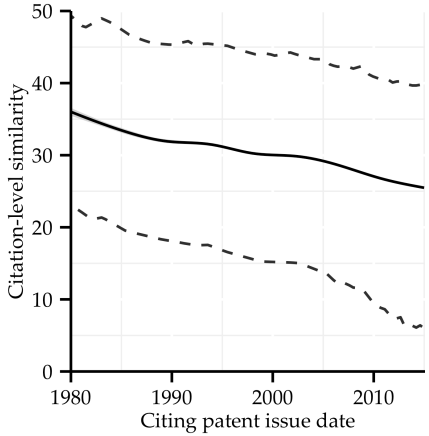


Figure 3. Cumulative similarity separated by type of patent citation and level of similarity.

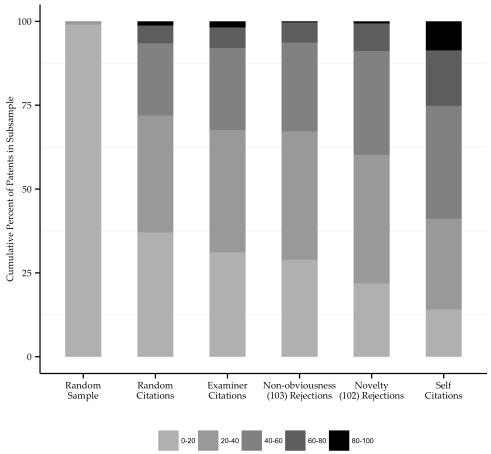
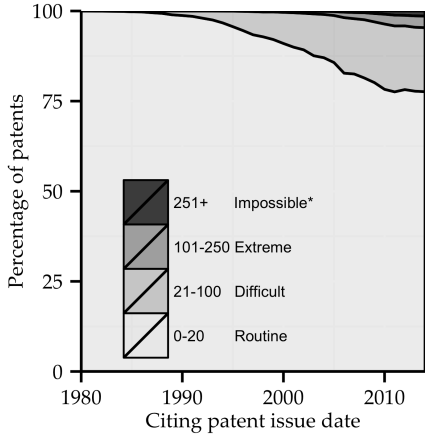
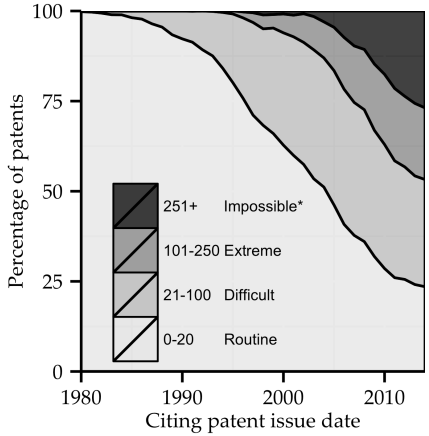


Figure 4. Percentage of patents by backward patent citation count over time.



*Given patent examiner time constraints.

Figure 5. Percentage of citations by backward patent citation count over time.



*Given patent examiner time constraints.

Figure 6. Patent similarity by number of backward citations made by citing patent.

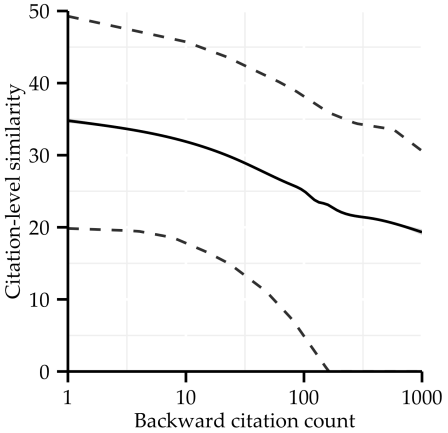


Figure 7. Percentage of examiner citations by backward citation count.

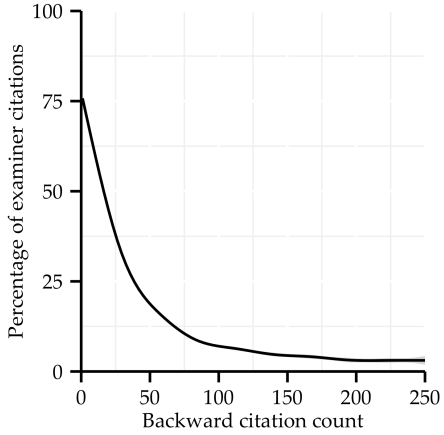


Figure 8. Count of patents by issued year.

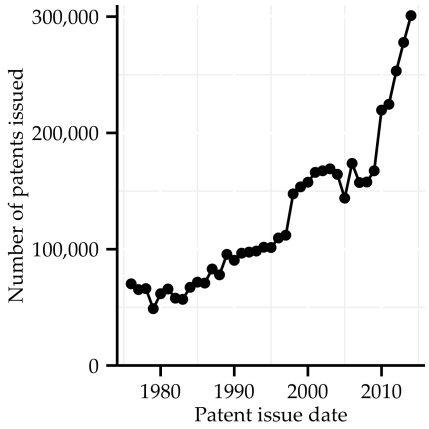


Figure 9. Mean backward citation count by extended family size.

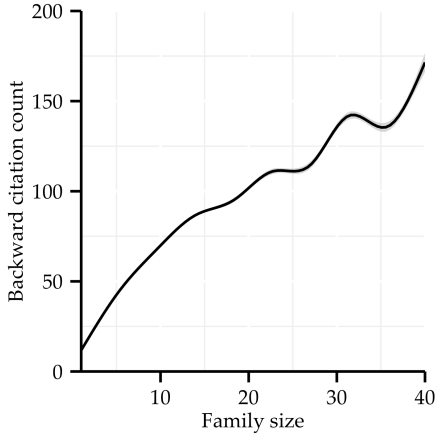


Figure 10. Mean likelihood a cited reference is also cited by a family member, by explicit family size.

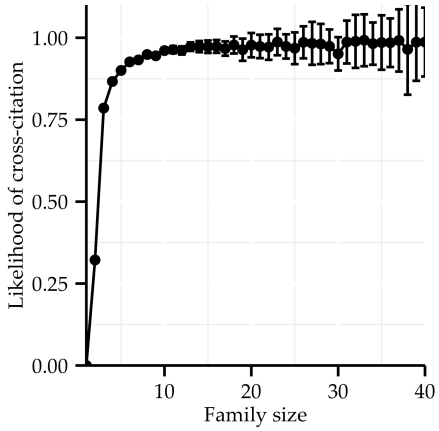


Figure 11. Mean times a cited patent appears as a citation across a patent family by explicit family size.

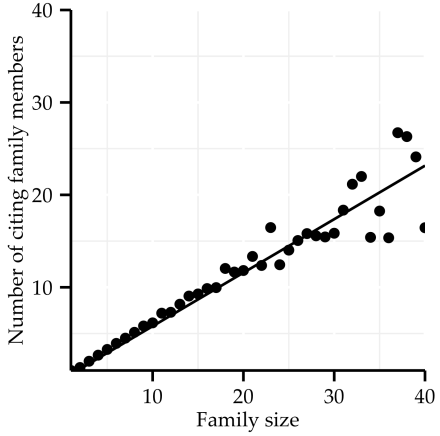


Figure 12. Average similarity by citation source.

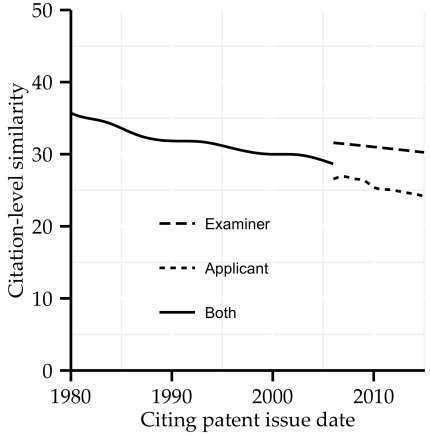


Figure 13. Average similarity by citation recipient.

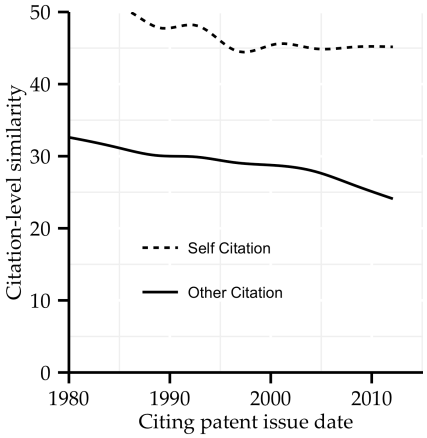


Figure 14. Average similarity of citation pairs by patents making 20 or fewer total citations.

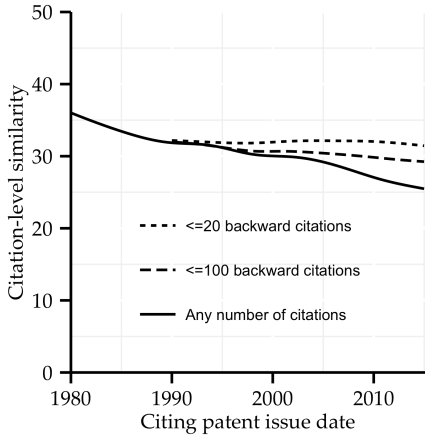


Figure 15. Lorenz curves for corrected and uncorrected citations pools in 2014.

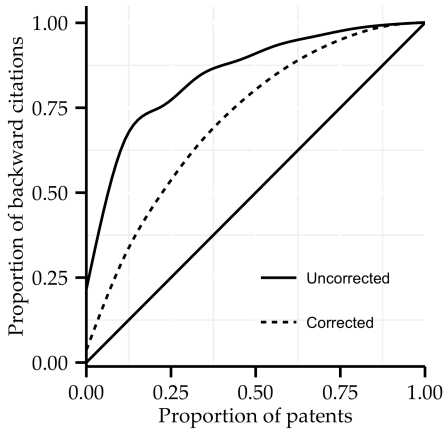


Figure 16. Gini Coefficients by year.

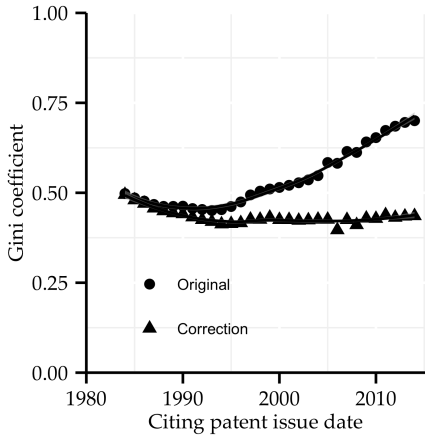


Figure 17. Average similarity of citations in the selected pool.

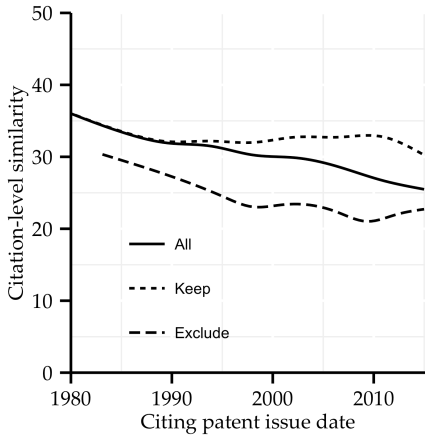
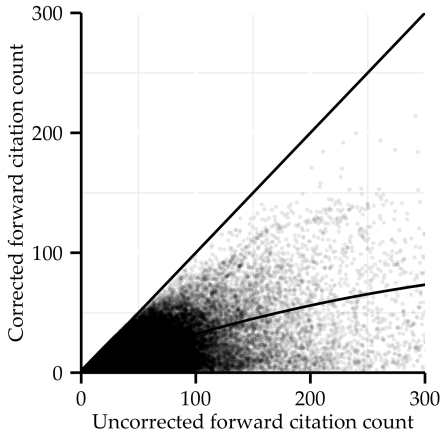


Figure 18. Comparison of uncorrected and corrected forward citation counts.



Appendix B: Patent Rejections

As background for the article, we identified "rejection citations" as a particularly informative type of patent citation. An applicant is entitled to a patent only if their application describes and claims a new and non-obvious invention. When a patent examiner determines that the claims of a patent application fail to make a novel and non-obvious advance over the prior art and they reject the application, the examiner must cite specific portions of prior art disclosing the features recited in the claims. If the applicant has not already cited that prior art, then the patent examiner adds a citation of their own (Alcacer and Gittelman 2006, Alcacer, Gittelman et al. 2009). Patent citations therefore provide the context that patent examiners use to identify rejection references.

Identifying citations used in rejections by the USPTO is very difficult. Citations added by an examiner do not necessarily imply a rejection, for they may be added by the examiner simply to document the examiner's search efforts and/or to save citations that can be used later in the examination process. The European Patent Office identifies "X" and "Y" citations as citations that may "block" claims when designating prior art (Czarnitzki, Hussinger et al. 2011, Von Graevenitz, Wagner et al. 2011), but a European "blocking" citation does not necessarily indicate a rejection. Additionally, the concept of an "X" or "Y" blocking citation does not exist in the U.S. patent system.

We followed 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 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 the novelty (35 U.S.C. 102) or non-obviousness (35 U.S.C. 103) requirement of U.S. patent law.

For the resulting sample of patents issued 2005 or later, rejection citations make up just 4% of all citations. We also find that examiners select 11.4% of citations to actually make a rejection. We find strong support for the idea that examiners are primarily responsible for locating rejection references. Model 1 of Table B1 shows that each examiner citation has a 6.5% chance of being used to support a novelty rejection, while an applicant reference has only a 0.3% chance of being used to support a novelty rejection. Model 2 shows that each examiner citation has a 15.0% chance of being used to support a non-obviousness rejection.

Table B1. Linear probability models of a citation being used to support a rejection.

	Dependent variable:	
	Likelihood of being a rejection reference	
	Novelty (102)	Non-obviousness (103)
	(1)	(2)
Is examiner citation	0.062***	0.144***
	(0.0001)	(0.0001)
Constant	0.003***	0.006***
	(0.00004)	(0.0001)
Observations	22,905,341	22,905,341
R^2	0.038	0.094
Adjusted R^2	0.038	0.094
F Statistic (df = 1; 22905339)	903,187.500***	2,376,169.000***
Note:		*p<0.1; **p<0.05; ***p<0.01

REFERENCES FOR APPENDIX B

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.

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).

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

Appendix C: Patent Citation Corrections

A. Previous Corrections

Truncation bias. The most frequently used sample correction to date relates to rightward truncation bias. Lerner and Seru (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).

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 mean 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.

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. For instance, a citation that is independently submitted by both the applicant and the examiner is credited to the applicant. 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.

Collapsing to family. Another approach might be to collapse patent citations to the patent family. Recognizing the significance of patent families is important – simply collapsing to family, however, is unlikely to alone correct for the empirical problems we have identified, and may just 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 explicit patent family ignores differences in quality, and perhaps more importantly, collapsing does not handle *implicit* patent families linked by shared subject matter but not by priority claims. Collapsing citations by explicit, but not implicit, patent families may introduce selection effects wherein firms that rely on implicit family-like relationships, instead of explicit priority claims, are become overrepresented in the data.

B. Criteria to evaluate corrections

We propose three criteria on which to evaluate corrections to patent citation data. 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 should seek to reduce or eliminate this imbalance. Second, a citation-based measure should select on (or up-weight) groups of citations that are more similar to the citing patent. Any particular citation may be important and meaningful without exhibiting a high similarity to the citing patent. However, we show in related work 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 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.

C. Future corrections

Although constructing and validating every possible correction to citation-based measures is beyond the scope of this article, we think it important to discuss some of the techniques that might be used to construct such corrections. One approach to improving citation-based measures is to coalesce citations into patent families. If a single patent is cited in one patent and then copied across an entire family and is, furthermore, only one of hundreds or even thousands of similar references cited by the family of patents in this way, separately counting each citation by a family member seems to be according far too much weight to what is essentially a single citation. However, the definition of a patent family is not clear. Constructing patent families solely from patent priority claims may be insufficient, and a more accurate identification of patent family members may require finding the implicit linkages between highly similar patents that are issued to a single firm.

Another approach to improving citation-based measures is to weight citations by a measure of their information content. For example, we show that a small portion of all patents contributed the large majority of citations, but that the citations made by these patents are unlikely to carry much information content. One possible correction would be to weight each citation by the inverse of the total number of backward citations made by the citing patent. Such an approach would effectively accord each patent an equal "voice" in providing backward citations for analysis. More research is required to explore such options.

REFERENCES FOR APPENDIX C

Hall, B. H., A. Jaffe and M. Trajtenberg (2005). "Market Value and Patent Citations." The RAND Journal of Economics 36(1): 16-38.

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

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.

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.

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.

Younge, K. A. and J. Kuhn (2015). Patent-to-Patent Similarity: A Vector Space Model. É. P. F. d. Lausanne. Available at SSRN: http://ssrn.com/abstract=2709238.