**"How do I know what you know? Patent examiners and the generation of patent citations"**

Juan Alcácer
Stern School of Business
New York University
jalcacer@stern.nyu.edu

Michelle Gittelman
Stern School of Business
New York University
mgittelm@stern.nyu.edu

First draft: December, 2003
This draft: May 2004

JEL Classification Numbers:
Keywords: Technology, patents, prior art

We acknowledge the helpful comments of Luis Cabral, Bill Greene, Tom Hemnes, Jim Hirabayshi, Anita McGahan, Kevin Oliver, Gonçalo Pacheco-de-Almeida, Joe Porac, Claire Preisser, Rachelle Sampson, Til Schuermann, Edlyn Simmons, Scott Stern, Don Walters, Bernard Yeung, and Minyuan Zhao as well as seminar participants at University of Michigan, NYU and Organization Science Winter Conference 2004. Thanks also to our research assistants Nisha Bhalla, Jignasa Doshi, Jack Nguyen, Aakash Patel and Neeti Shah. Errors remain our own.

**Abstract**

There is a large and growing body of work that relies on patent data to study patterns of technological evolution, knowledge creation and diffusion, and firm technology strategy. Analysis of prior art – citations to patents by other patents – has been a core methodology in the literature, in which citations are the "paper trails" tracking social, organizational, and geographic pathways of knowledge flows. However, in many instances researchers have been limited in their interpretations of their findings because citations made by patent examiners have not been separated out from citations made by inventors. We leverage a recent (2001) change in the reporting of patent data that indicates whether prior art citations are made by inventors or examiners. Our data consist of citing-cited pairs of patents generated from a large, random sample of patents issued over the period 2001-2003. We find the magnitude of examiner citations to be quite high: for all US patents granted over the period, 40 per cent of all citing-cited patent pairs are generated by examiners; on a per patent basis, examiners imposed 67 per cent of all prior art citations. Moreover, some 40 per cent of patents granted in this period have *all* citations imposed by examiners, and about 70 per cent of patents have at least half or more of their prior art citations introduced by examiners. We hypothesize that inventors are more likely than examiners to cite technologies that are near to the citing patent in space, technology class, organizational and social boundaries, and time. We find this to be the case for geography. However, the magnitude of the difference in geographic citing patterns is so small as to be potentially economically insignificant. Regarding technology and vintage effects, examiners are more likely to proximate citations than inventors, reversing the expected pattern. Overall, our results do not change the presumption that patents trace out knowledge flows: inventors face strong legal pressures to reveal all they know, and citations do contain a signal of knowledge flows. However, our results indicate that examiners are not adding random noise to a core of inventor knowledge but may be amplifying the signal attributed to inventors, raising the possibility of Type 2 error for hypotheses about inventor knowledge flows. We also find differences between inventor- and examiner-citations accruing to highly-cited patents, indicating that these groups select different patents for citation. However, differences between examiner and inventor citation streams attenuate over time, which is suggestive of a learning process between examiners and inventors that has not been previously considered in the literature.

**Introduction**

In their seminal paper on knowledge spillovers, Jaffe, Henderson and Trajtenberg (1993, p. 578) write that "Krugman. . .perceives that [k]nowledge flows. . . are invisible; they leave no paper trail by which they may be measured and tracked…But knowledge flows do sometimes leave a paper trail, in the form of citations to patents." Since that time, analysis of prior art – citations to patents by other patents – has been a core methodology in the technology strategy and economics of innovation literatures. This increasing use of patents can be traced to both a growing interest in knowledge as a driver of organizational performance and economic growth and the lower costs of accessing and analyzing large quantities of patent data. [1]

A risk associated with the rapid growth in the number and scope of patent studies is that the application of patent data to measure economic, organizational, and social phenomena could outpace understanding of how the data are generated and what they actually mean. We know that patents are an appropriate measure of innovative activities in only a few industries (Levin, Klevorick, Nelson, Winter, Gilbert and Griliches 1987). Further issues arise regarding the meaning of patent citations. Patent citations are used to measure knowledge, but whose knowledge? The prevailing assumption is that patents codify the knowledge of individuals and organizations who invented the patented technologies. However, other actors are important in shaping the contents of patents. Patent examiners and attorneys are involved in drafting the contents

[1] Through the US Patent Office and other patent offices, all patents are freely available online, and there are additionally some 900 proprietary and non-proprietary electronic patent databases that provide online patent search tools. In 2001, researchers at the NBER, who pioneered many of the early patent-based studies of knowledge transfer, made their patent database available for free to the public. The data contain not only information about all US patents up to1999, but also patent citations, along with a number of economically interesting variables coded by the NBER team (see Hall, Jaffe and Trajtenberg 2001, for a description of the database).

of patents and generating citation lists, and their influence on the data is likely to be considerable.

In particular, the fact that every patent passes across the desk of an examiner, who adds some unknown number of patent citations to it, raises concern that patent citations may not be a good measure of direct knowledge transfer between inventors but could instead reflect the administrative and institutionally mediated process of patent examination. Until recently, examiner citations have not been separated from citations made by inventors.[2] Notwithstanding a few important attempts to analyze examiner citing patterns and understand whether citations capture knowledge spillovers (Jaffe, Trajtenberg and Fogarty 2000; Meyer 2000; Michel and Bettels 2001; Thompson 2003; Breschi and Lissoni 2004; Cockburn, Kortum and Stern 2004), little is known about the magnitude of examiner citations and whether they differ from inventor citations. As a result, researchers have been limited in their interpretations of their findings and have been forced to treat aggregate citations as a "noisy signal" of knowledge flows, without being able to specify much about the actual degree of noise versus signal in the data.

This paper provides a comprehensive analysis of patent citations that accounts for differences between inventors and examiners in generating cited-citing patent pairs. We leverage a recent (2001) change in the reporting of patent data that indicates whether prior art citations are made by inventors or examiners. Our data consist of some 16,000 citing-cited pairs of patents generated from a random sample of 1,500 US

[2] Unless we specify otherwise, we use the term "inventor citations" to mean prior art citations made by patent applicants as opposed to patent examiners. These might be made by individual inventors or by a firm's attorneys or other individuals representing patent owners.

citing patents issued over the period 2001-2003.[3] We supplement our analysis with interviews with patent attorneys and patent professionals. Our empirical strategy is twofold. We first show the magnitude of examiner citations. We find it is quite high. Based on an analysis of citations from *all* patents granted in the US between January 1 2001 and August 31 2003 – 442,839 citing patents and 5.4 million citations – 40 per cent of all citing-cited patent pairs were generated by examiners; on a per patent basis, examiners imposed 67 per cent of all prior art citations. Moreover, some 40 per cent of patents granted in this period have *all* citations imposed by examiners, and about 70 per cent of patents have at least half or more of their prior art citations introduced by examiners.

We then set out to understand whether examiner citations differ statistically from inventor citations in order to answer the following question: are there systematic differences in inventor and examiner citation streams that might bias inferences made from aggregate citation data? We estimate the likelihood of examiner citations along of number of dimensions: self-citation by individual inventors and by firms; geographic proximity; similarity of technological class; type of assignee;[4] and vintage effects. We find that examiners differ from inventors in their propensity to cite along these dimensions. However, the difference between mean values in the two citation streams is in many cases so small as to be economically insignificant. In several cases, the differences are in the opposite the direction one would associate with aggregate citations as measures of inventor knowledge.

---

[3] We do not consider non-patent references in our analysis. All references to prior art and citations are to patents issued in the United States

[4] Patent assignees are organizations listed as patent owners.

A second part of the analysis disaggregates inventor and examiner citations for *forward* citations. Our objective is to learn whether highly-cited patents – which are associated with patents of unusually high economic and technological value – "earn" their status through citation by examiners (an administrative process) or citation by inventors (an evolutionary, or diffusion, process). We find evidence of both, with surprisingly little overlap between them. We see a pattern in which inventor and examiner forward citations streams are initially negatively correlated, but converge over time. The pattern raises the possibility of a channel of learning between examiners and inventors previously unexplored in the literature. We draw conclusions for the use of patent citations to measure knowledge and high-impact technologies.

**What is the practical meaning of prior art?**

Patent data have been used by researchers to measure and track technologies and knowledge; in practice, they are generated by a complex legal and institutional process that is primarily aimed at classifying and proving the patentability of individual claims. We first define the practical meaning of prior art citations, then turn to a discussion of assumptions made in the literature about their meaning as proxies for knowledge flows.

A granted patent is a novel, non-obvious and manmade invention: an addition to the world's stock of technological knowledge and a stepping-stone for future inventions (the latter a primary intention of the patent system). A patent consists of several components that define the invention, assign rights to individuals and organizations, and delineate the scope of those rights. The description discloses the invention so that it can be understood by others "skilled in the art". The core value of a patent is expressed in its claims, which detail aspects of the invention over which inventors and assignees may exercise ownership rights. Claims cover intellectual property that is not already foreshadowed by existing patents or public knowledge. To make the case that claims are valid (new and non-obvious), patents contain prior art. Prior art may consist of patented and nonpatented information; in fields where patenting is relatively new, much of the prior art may be in the public domain, e.g., published in trade journals.

Prior art citations serve a number of heterogeneous functions: by anticipating the claimed invention, they may be used to limit or reject an individual claim or an entire patent; prior art may strengthen claims, by establishing that earlier versions of the inventions were different from or inferior to the current invention; or they may be "boilerplate" that establish facts described in the patent. Patents that have an unusually high number of prior art citations are likely to contain valuable claims, since the claims were approved *despite* the existence of a large body of prior art. Allison et al (2003) show that valuable patents, measured as litigated patents, cite more prior art than non-litigated patents; Gittelman and Kogut (2003) found that patents that cite a large body of non-patented prior are more highly cited, a measure of patent value. Claims and prior art thus operate together to establish validity and novelty over existing knowledge, and delineate the scope and strength of the intellectual property covered by the claims.

The role of prior art in proving (or disproving) the validity and novelty of claims underscores that the contents of a patent are not just codified knowledge but legal tools that embody and reflect the strategies of a variety of actors: individuals, firms, competitors, and the patent office. While some portion of the prior art contained in patents traces out knowledge flows, citations also reflect the heterogeneous objectives and interests of these different actors.

**What is the meaning of patent citations for studies of technology and knowledge transfer?**

A central premise of the technology strategy and economics of innovation literatures that use patent citation data is that citations reveal evolutionary pathways across innovations, organizations, and time. We may identify two broad streams in the literature. The first seeks to understand patterns of knowledge diffusion, and uses patent citations to trace out knowledge flows across and within organizations, geographic space, and populations of inventors. The second stream, which we group into a category we call productivity studies, is broadly concerned with determinants of technological performance. Here, citation data are used both as a measure of the impact of patents (as captured by forward citation counts) as well as to develop variables that measure different structural characteristics of inventions: breadth, originality, vintage, complexity, and fragmentation of the knowledge base. The two groups of studies make different assumptions about the meaning of patent citation data. We consider each in turn.

***Patent citations as a measure of knowledge flow***

The core assumption in the diffusion literature using patent citations is that "patent citations allow us to observe the patterns and end points of the knowledge transfer process" (Song, Almeida and Wu 2003). This is a particularly strong assumption. We may be confident that a citation from patent A to patent B indicates a technological relationship between them, which may be quite strong (patent A would not be possible without patent B) or relatively trivial (patent A belongs to a class of patents of which B is representative). However, we can be less certain that inventors of patent A actually knew about patent B prior to their own invention.

The legal rules work in favor of the interpretation of citations as revealing knowledge flows from cited to citing patents. Inventors are required to submit patent applications that contain all published art they are aware of (patented and in the public domain) that is relevant to the claims they are presenting. Ultimately, if a patent is approved that does not list relevant prior art (whether by the examiner or inventor), it might be vulnerable to infringement litigation and be found invalid. If, during litigation, it can be shown that inventors *knew* of prior art and failed to disclose it, they are subject to even greater penalties.[5] The following quote, by a patent attorney and former inventor, is illustrative of pressures on inventors and their attorneys for full disclosure:

> The first [time], as an inventor, I was introduced to prior art as a engineer at IBM. There, we were told to disclose and discuss all pertinent publications before they were filed. And failure, we were told by the attorneys, was punished by fraud, imprisonment, and would result in the disbarment of the attorney that was representing us. Basically, the attorneys said that we would have the time in jail to basically explain to them why they could no longer practice law, and so forth, if we didn't give them the right references. Maybe this was unique to IBM, but it's something that I've carried throughout my career in talking with inventors, and so forth, as far as how important I think the duty of disclosure is.[6]

Patent citations should thus reveal a core of inventor knowledge. But they also likely contain references that are unrelated to inventor knowledge. It is common practice for inventors to hire attorneys and professional patent searchers to research prior art and draft strong claims. The following quote, by the same attorney, illustrates

---

[5] This risk raises the interesting possibility that inventors have incentives *not to know* about related inventions. Where patents are at risk, communication with competitors (for instance, attending conferences or social gatherings) may represent a negative externality for firms not only because engineers could *reveal* too much information to competitors, but because engineers may *learn* too much information from competitors!

[6] Testimony included in "Public Hearing on Issues Related to the Identification of Prior Art During the Examination of Patent Application", June 28, 1999, before the United States Patent and Trademark Office.

how a comprehensive search beyond inventor knowledge is often needed to write strong patents that will stand up to post-issue challenges:

> One of the worst feelings that I've even seen at a licensing table is when you're sitting there trying to license a patent and someone passes across the table a 102(b) reference [a prior art reference] that is completely is out of left field, you've never seen before, that says this patent is invalid and indefensible. It's something that no one, as a practitioner, wants to face, and would rather face, have that reference come up, early on in the prosecution procedure, and be able to be discussed with the examiners, who really know what they're talking about.[7]

The prevalence of lawyers and professional searchers in writing patents is likely to dilute the signal of direct flows contained in inventor citations. These professionals (attorneys and searchers), many of whom were formerly patent examiners, draft claims and search patent databases – a complex, idiosyncratic process – to uncover all potentially relevant prior art; these may include, but not be limited to, works that the inventors were actually aware of or used in their own invention. Indeed, patent professionals have incentives to search as widely as possible beyond the knowledge of inventors to maximize their own value added and the chances that the patent will be approved with strong claims. The addition of prior art by attorneys and professional searchers would presumably add a layer of citations that do not correspond to inventors knowledge.

Finally, all US patents must submit to an examination process by patent examiners, civil servants working at the US Patent and Trademark Office in Washington DC. The role of the examiner is to certify the validity of a patent's claims to novelty and non-obviousness: examiners check the lists of prior art submitted by

---

[7] Ibid.

inventors and attorneys and make changes based on their own search of prior art and reading of the claims. Examiners are often treated in the literature as objective, independent arbiters of prior art who follow uniform procedures, but in fact they communicate with inventors during the examination process, and there is a great deal of heterogeneity in the practices of individual examiners (Cockburn, Kortum and Stern 2004). Furthermore, examiners are subject to administrative pressures that limit the degree to which they search comprehensively. Examiners are expected to search all possible art, including patents (over 5 million in the US alone), non-patented literature such as books and journals, the Internet, and even emails, to understand the state of current art. However, the USPTO imposes production goals on examiners that limit the time for search and examination; it is estimated that examiners can reasonably allot less than eight hours to examine an individual patent.[8]

In summary, aggregate citations observed by researchers do not only capture those inserted by inventors but also citations that lawyers, patent searchers and examiners judged ought to be included. Faced with this problem, researchers have acknowledged that while patent examiners add some citations, there is no reason to expect that they would systematically bias the data. In a study to investigate this type of measurement error, Jaffe, Trajtenberg and Fogarty (2000) surveyed inventors regarding their prior knowledge of cited works contained in their patents. They found that less than one third of inventors had a high degree of familiarity with works cited in their own patents, one third of inventors learned of the cited work *after* completing their own

[8] "Public Hearing On Issues Related To The Identification Of Prior Art During The Examination of a Patent Application", USPTO, July 14, 1999.

invention, e.g., during the process of drafting the patent, and one-third had no prior or post-hoc knowledge of the cited work(s) at all, which the authors attribute to examiner citations. Their survey findings, while supportive of the idea that aggregate citations are a “noisy signal” of knowledge flows, indicated that the ”noise” created by examiner citations may in fact be of greater magnitude than the “signal” of inventor knowledge contained in aggregate citation streams.

*Examiner citations and inferences about citations as knowledge flows*

We have discussed two reasons why aggregate citations may not approximate inventor knowledge: the participation of attorneys and patent professionals in drafting inventor citations, and the addition of citations by examiners. Nearly all patents in our dataset list lawyers on their front page, attesting to their importance in the process of generating prior art. Insofar as lawyers are likely to be cognitively and behaviorally much closer to examiners than to inventors we expect that, ceteris paribus, inventor citations would approximate examiner citations when lawyers are involved and that the magnitude of the difference between inventor and examiner citations would be greater for patents where lawyers were not involved.

To what extent could examiner citations “contaminate” inferences about knowledge flows made from the aggregate data? In order to answer this, we need to know both the magnitude of examiner citations as well as their distribution with respect to inventor citations. We start with some simple assumptions about the characteristics of prior art we expect inventors “ought to” be adding to their own patents. We expect that inventors’ knowledge of technological antecedents is cognitively bounded, specifically that inventors’ knowledge of prior art is likely to be strongest for technologies that are closest to them along a number of dimensions: social, organizational, technological, geographic, and chronological. We operationalize these dimensions by measuring inventor and examiner citations for the following elements: self-citation at the level of individuals and firms; common technological classes; common geographic locations; and similar time periods. Our a priori expectation is that increasing distance along each of these dimensions will decrease inventor awareness, thereby *increasing* the chances that examiners will add citations. While cognitive awareness of technologies is clearly more complex than these simple assumptions, they allow us to test the degree to which the data deviate from the most basic assumptions about inventor knowledge. If we find that they are rejected, we might question the appropriateness of making more complex assumptions about inventor knowledge from the aggregate data.

Figure 1A shows graphically our distributional priors for inventor citations. The *x*-axis is a given dimension, for example geographic distance, and the *y*-axis represents the expected frequency of citations. Under the hypothesis of knowledge localization, inventor citations would be concentrated towards the origin of the *x* axis (citations to local patents). This is the "true" distribution of inventor knowledge, labeled *INV* in the graph, that researchers seek to measure. However, researchers only estimate knowledge flows with aggregate citations, and cannot tell which citations are added by inventors or examiners. Inferences about patterns of interest, e.g., localization or vintage effects, are made from the aggregate data. How might these inferences change under different distributional assumptions about examiner citations?

We describe two possible scenarios with different assumptions of examiners' behavior. In figure 1B examiners are agents with extensive knowledge in the field whose main purpose is to "fill the gaps" existing in the prior art list submitted by inventors. For example, examiners would add citations to other firms, to patents created in more distant places, patents in other technologies and older patents. Consequently, examiners' distribution of citations, labeled *EXA* in graph 1B, would be skewed toward more distant values on the *x*-axis. The distribution observed by researchers, labeled *AGG* in graph 1B (for aggregate), is the aggregation of inventor and examiner citations. Under moderate to low levels of examiner citations, *AGG* would be flatter than *INV*, making it harder to find any statistically significant relationship between the independent variable in *x* and estimated knowledge flows. This distribution of examiner citations increases the probability of making a Type II error, lowering estimated coefficients, and increasing the chance of accepting the null hypothesis. For example, if *x* represents physical distance, examiner citations work against a finding of localization in the aggregate data. Thus, if significant effects are nonetheless found for the variable of interest, such a pattern would increase confidence in the inference of localization, as the "true" rate is higher than estimated from the aggregate data. Thus, if significant effects are nonetheless found, such a pattern would increase confidence in the inference of localization.

Figure 1C shows the case in which the distributions of inventor citations and examiner citations track each other closely. Behaviorally, such a pattern could emerge if inventors search widely (with the assistance of lawyers and professional searchers) so that their citations anticipate, with some error, what examiners would add. Such a pattern could also emerge if examiners' search is guided by inventors' list of prior art, such that examiners search locally with respect to inventors own prior art searches.

These behaviors are at odds with a "random noise" pattern, insofar as the generation of the two citation streams is highly correlated. Statistically, such a pattern would raise the probability of Type I error, by amplifying the signal of localization in the aggregate data to a greater extent than in the inventor-citation stream. With identical distributions, estimated coefficients would not be biased but significance levels would be inflated without correction for examiner-added citations.

Our empirical approach to explore which of these effects hold is threefold. First, we estimate the magnitude of examiner citations to understand what characteristics of patents result in a high number of examiner-added citations. We then perform univariate difference in means tests for each dimension described previously. We estimate whether examiner citations differ, on average, from inventor citations, and if so whether the difference is in the hypothesized direction in which "nearby" citations are generated by inventors and more distant citations by examiners. Finally, we then turn to multivariate logit regressions to estimate the odds of a given citation dyad as being generated by an examiner or an inventor, conditional on all of the dimensions measured in the univariate means tests. A key test is whether, controlling for all other factors, we can observe statistical differences in the generation of the two citation streams and whether the direction of the difference meets our simple assumptions about inventor knowledge.

***Patent citations as a measure of high-impact innovations***

A second group of studies using patent data, which we broadly term productivity studies, makes a much weaker assumption about the theoretical meaning of patent citations. Here, citation data are used as a measure of the impact of patents, as captured by forward citation counts. The core assumption is that forward citations that accrue to an issued patent are a good measure of the economic and social value of the cited invention. A number of studies have shown that forward citations correlate with non-patent measures of value, such as firm market value, litigation, and expert evaluation of technological impact, so there is good evidence that this assumption is correct (see Hall & Trajtenberg (2000) for a review). However, the process by which highly-cited patents are generated is still a matter of speculation. A highly-cited patent is assumed to attain high impact status through its diffusion to other inventors; as Henderson, Jaffe & Trajtenberg (1998) write: "Implicit in this approach is a view of technology as an evolutionary process, in which the significance of any particular invention is evidenced, at least partly, by its role in stimulating and facilitating future inventions. We assume that at least some of such future inventions will reference or cite the original invention in their patents, thereby making the number and character of citations received a valid indicator of the technological importance of an invention."

If, however, citations made by patent examiners are responsible for generating highly-cited patents, we may infer that important inventions emerge not so much through evolutionary processes but through ascription via an administrative process. Inventors may learn about prior art from patents that are frequently cited by examiners, indicating a hub-and-spoke process of knowledge transfer with examiners at the center, rather than a structure connecting inventors directly.

We analyze a cohort of very highly-cited patents and disaggregate their forward citations into those made by inventors and those made by examiners. We find that examiners and inventors do not overlap very much in their selection of important patents, but that these differences diminish over time.

**Data and variables**

Starting in January 2001, the USPTO has indicated on the front page of patent images which prior art citations were added by examiners. We collected the front page images of all utility patents granted between January 1st 2001 and August 26th 2003 from the USPTO. This yielded a group of 442,839 patents citing back to 5,434,883 patents in their prior art.[9] Tables 1 and 2 provide summary information for the dataset. On average patents cite 12.2 patents. Patent examiners are an important source of citations representing about 40 per cent of all citing-cited dyads in each of the three years. The magnitude of examiner citations is even higher when measured on a per patent basis: for the average patent in our dataset, examiners imposed 67 per cent of all prior art citations. The difference between the dyad and patent-level means derives from the fact that between 38 and 40 per cent of patents granted over the period have *all* citations imposed by examiners; in contrast, only 8 per cent of patents had no examiner-added citations. About 70 per cent of the patents have at least half or more of their prior art citations introduced by examiners.

Our analysis requires that we match individual elements of citing and cited patents, in particular the names of individuals and organizations; to reduce this to a manageable task, we create a sample of 1,500 citing patents from the full dataset. Since

[9] Consistent with other studies, we only analyze cited patents granted in the US. In our sample, 849 citing patents cite at least one patent filed in another system (mainly Japan and the European Union), generating 4,971 citing-cited pairs. Of these, roughly 90% are inventor citations and 10% citations imposed by examiners. However, according to our interviews, it is likely that a significant portion of foreign patents listed as inventor citations were in fact added by foreign examiners during application for international patents.

the USPTO often issues patents in batches, with seasonal and firm-level variation, we do not sample specific weeks or months but instead randomly select 500 patents from each of the three years in the data to achieve a random distribution of patents across time.

The 1,500 citing patents generate 17,866 prior art citations ("cited patents"); only 26 patents (2.4 per cent of the sample) have no citations at all. Detailed data are not available for patents granted before 1976, leading us to remove 1,767 of the cited patents granted before that date, yielding a final sample of 16,089 citing-cited patent pairs. Standard distributional tests (e.g. Mann-Whitney and Kolmogorov-Smirnov ) provide supporting evidence (bottom of table 2) that our sample is random and representative of the full dataset with respect to the following relevant variables: citations per patent, percentage of examiner citations per patent and application year. To explore whether the distribution across technology classes in our samples is similar to that of the full datasets, we compare the top 20 most frequent classes in both groups and find an overlap of over 80 per cent for each year. The results indicate that we have a representative sample from the population of patents issued in that time period.

Except for technology classifications, patent data are not standardized, resulting in a great deal of variation in data formats across common elements. To correct for this, we perform a number of operations on the data, in order to identify common assignees, geographic locations, and individuals. Changes in the data introduced by our cleaning operations reveals the extent to which failure to perform similar operations can produce errors in identifying matching elements contained in patent data.

For corporate ownership, we create two variables: *dif_company* and *dif_parent*, which take a value of one if the citing-cited dyad are each assigned to different firms or different corporate parents, respectively. To construct these variables we perform three steps. We first standardize names by correcting for differences in spelling and format (for example: Sam Sung Electronics/Samsung Electronics; Minnesota Mining and Manufacturing Co./3M). In the second step, we group assignees with different names (e.g., Nokia Finland and Nokia USA) that are subsidiaries of the same corporate parent. We identify the ultimate parent for each assignee using the Directory of Corporate Affiliations based on their parent in the year of patent application, going back to 1991. Assignees on patent applications before 1991(27 per cent) were matched to the 1991 directory. We further correct for mergers, acquisitions, and name changes since 1976. Taken together, these changes reduce the number of unique assignee names by 28 per cent, from 5,933 to 4,239: 1,002 assignee names are eliminated through corrections for name variations; 1,694 unique assignee names are removed in the second step accounting for corporation affiliations and mergers. These changes indicate that self-citation rates could be affected without accounting for mergers and corporate parents, specifically overestimation of "cross-firm" citation rates and underestimation of the rate of self-citation.

To control for heterogeneity in patent practices across different types of inventors, we identify whether the patent assignee on citing and cited patents was one of four possible types: *Government, Academia, Corporate, Other.*[10]

[10] Government includes US government agencies; "other" includes individual inventors, non-university research institutions, foreign government agencies, and unspecified assignees.

We assigned the geographic location of citing and cited patents based on the locations of inventors. Our sample generated 40,797 inventors (58 per cent located in the US) and 8,474 different locations (51 per cent in the US). We identify locations for all inventors listed on each patent, not only first inventors. Similar to assignee data, locations also present problems and require significant cleaning and checking. We first perform a manual cleaning of city, state, and country names[11]. Second, we identify longitude and latitude point data using the United States Postal Office for locations within the US and GEOnet name server of the National Imagery and Mapping Agency for all other locations. These steps allow us to identify 73 per cent of locations, leaving 2,301 locations (mostly in Asia) unidentified. Country natives checked each list to match place names to those given in the GEOnet and USPTO databases. As a result, we are able to identify at least one inventor location for all but four patents involved in six citing-cited pairs.

Prior studies have used both discrete geographic units as well as continuous distance in miles to measure geographic proximity. We adopt both methods, as each captures different aspects of the relationship between geography and knowledge flows. We first measure administrative boundaries at the country, city, state, economic area and county level[12]. We construct dummy variables that take a value of one if none of the inventors in the citing-cited pair share a common location for each of these units

[11] To our surprise, state and country data was far from perfect. Unfortunately, the USPTO uses the same abbreviations for countries and states. For example, CA can be California or Canada, IL can be Illinois or Israel. This problem is also present in the NBER dataset. Although this problem does not seem to affect a great number of patents, researchers should be aware of it.
[12] In an effort to identify geographic areas that mimic economic activity and not state or administrative boundaries, the Bureau of Economic Analysis (BEA) defined 171 economic areas that span the US. Each economic area consists of one or more nodes – metropolitan or similar areas that serve as centers of economic activity – and the surrounding counties that are economically related to the nodes. The main factor used in determining the economic relationships among counties is commuting patterns, so each economic area includes, as far as possible, the place of work and the place of residence of its labor force.

(*dif_country, dif_city,* etc.)[13]. In addition, we create a variable *distance* which measures the great circle distance in miles between citing and cited patents using latitude and longitude coordinates.

Identifying individual inventors presents a challenge since it is reasonable to expect that identical names can correspond to different individuals. We construct different rules with increasingly stringent criteria for matching inventors between citing and cited patents. First, we identify common full names, in which first name, middle, and last names must all be the same. The variable *same_inventor* takes a value of one if the citing-cited pair share a common inventor according to this matching principle. This still leaves the possibility that individuals may have exactly the same name and middle initials. To increase the hurdle for a match, we create two additional variables. The first, *same_inventor_company* identifies if the full names are the same *and* company assignee is the same. The same name/company increases the probability that the identified pair is indeed the same person. To allow for job mobility, we create a third variable, *same_inventor_city* in which full names are the same *and* the locations for citing and cited patent are no more than 100 miles apart. The first matching principle is the most flexible because it recognizes that inventors can move to other locations or companies. However, it is also the most likely to generate false matches leading to Type 2 measurement error (inferring a common link when no link exists). The last two matching rules are more restrictive, in that they are less likely to assume common inventors in cases where the same names belong to different people.

[13] The last three variables are constructed only for citations where both patents have at least one American inventor.

We also create *same_examiner* and *same_lawyer* to identify common examiners and lawyers on citing and cited patents. We match examiners, lawyers, and law firms using rosters provided by the USPTO.[14] These variables control for self-citation patterns not just for inventors but for other individuals involved in drafting prior art citations.

We create a variable *dif_technology* to identify whether the citing-cited pairs belong to the same technology class. We use the International Patent Classification (IPC) instead of the United States Classification (USC) for this purpose, and match citing-cited patent pairs at the 4-digit level.[15] A number of reasons drive this choice. First, the IPC system follows a nested hierarchical structure, allowing us to look at different levels of aggregation in the technology domain. Second, the IPC system is more similar to a traditional industry end-use classification system than the US system, which classifies patents by function. One problem with the IPC is that older patents are not reclassified when classification codes change (which happens infrequently). This would make our matching test more conservative, insofar as patents that belong to the same class are coded differently because the more recent patent was subject to a newer classification code. To account for this, we update IPC codes based on USPTO-IPC concordance tables.

Table 3 shows definitions for the variables used to measure linkages between citing and cited patents for all of the elements discussed above.

---

[14] This roster not only allows us to match names of lawyers on citing-cited patent pairs but also provides information on whether the lawyer is an in-house counselor or not. Companies that introduce numerous applications per year, such as Intel, IBM, Procter & Gamble, have a group of internal lawyers that deal with the applications for that company. In some cases, in-house counselors and external law firms are both involved in a patent application.

[15] We also estimate our models at the 2- and 3-digit levels with similar results.

**Estimating the proportion of examiner-added citations**

We expect that because of the limited time for search allotted to examiners and the complementarity between examiner and inventor search processes, the proportion of examiner citations will be inversely proportional to the number of citations provided by inventors. We also control for characteristics of the assignee that would affect their skills in patenting; to the degree that inventors are skilled in patenting and perform comprehensive searches, examiners should add proportionately fewer citations.

At the citing patent level, we estimate the following specification:

$$Y_i = X_i\beta + \varepsilon \qquad (1)$$

Where $Y_i$ is the percentage of examiners imposed citations and $X_i$ is a vector of the following patent traits: the logarithm of number of inventor citations that were originally submitted in the patent application (*log_inventor_citations*), plus a set of control variables. Foreign companies may be less familiar with the American patent application process, thus we would expect examiners adding fewer citations to American-origin patents; we indicate whether the assignee is an American company (*american_company*). Academic institutions have become proactive in patenting in recent years, and could have developed advanced patent-writing skills; we also specify whether the assignee belonged to one of the following groups – government, academia, corporate, and other. Organizations that have abundant experience at patenting would be better prepared for the application process by offering a more complete list of prior art, minimizing the role of examiners; we include a variable indicating whether the assignee is among the top 200 owners of US patents, based on patents awarded over 1988 to 2003. Analogously, patents from assignees with legal counselors that help them to conduct more thorough prior art searches would be expected to receive a smaller portion of examiner citations; we code for whether or not there was a lawyer involved in drafting the patent application (*lawyer*).

Table 5.1 shows the summary statistics and table 6.1 correlation values for all variables. Table 7 shows these results for OLS estimation of equation (1) using fixed effects by assignee. For all models, the expected relationship between inventor and examiner citations holds: the larger the number of citations added by inventors, the lower the proportion of patents subsequently imposed by the examiner. The coefficient for lawyers is negative but insignificant. Regarding experience, we do not find any effect for firms that have patented exhaustively nor for foreign firms.

**Comparing examiner and inventor citations: univariate tests**

We first conduct univariate tests comparing means of inventor and examiner citing-cited dyads along the following dimensions: geographic co-location, both in terms of continuous distance and discrete measures of co-location; self-citation at the individual, firm and corporate level; time; and technology class. Table 4 shows the results from these tests.

Two surprising findings stand out. First, we find slightly more localization effect for inventor citations than examiner citations, but the magnitude of the difference is so small for most measures as to be economically insignificant. The greatest difference in inventor and examiner proportions occurs for citations to patents with locations outside the US. Overall, however, we do not see large differences in means that would indicate "noisy signal" or "gap filling" patterns by examiners. To further show this, we plot the distance of examiner citations against inventor citations (figure 2). The distributions track one another closely, suggesting a pattern consistent with Figure 1C.

The second unexpected finding concerns self-citation patterns. While we find that patent applicants are more likely to cite themselves at the firm level, they have a lower rate of self-citation at the individual level. For our least restricted inventor match, where only the names are matched, we find a higher (but not significantly so) rate of self-citation among inventors than examiners. However, for our restrictive inventor measures, in which we have higher confidence of identifying the same individuals (same inventor/same company; same inventor/same city), examiners are *more* likely to include self-citations than inventors themselves. Since we have difficulty accepting that inventors forget about their own past patents, we assume that they omit self-citations because they lack patenting skills, or for strategic or legal reasons. Examiners subsequently add back what inventors "should have" included in their original lists. Our discussions with attorneys and patent professionals do not suggest any strong theoretical reasons for this pattern, e.g., in which inventors would gain by avoiding self-citation, so we interpret this pattern as evidence of poor patent practice on the part of inventors.

Regarding the other variables, we find that inventors are more likely to add prior art from different technological classes than examiners, and here the difference is large (49 per cent and 38 percent, respectively). Two possible explanations present themselves. The first is that inventors have a greater breadth of knowledge about patented technologies than examiners, who are narrowly specialized by technological field. The second (and we feel, more plausible) explanation stems from the patenting process itself. Patents are not classified until they go through the examination process, so inventors are adding citations without knowledge of the ultimate detailed classification code of the invention. In the process of examining patents, examiners develop classification codes based on individual claims, at the same time searching for prior art that is relevant to those claims. This endogenous process in which classifications and prior art are simultaneously generated by examiners would be consistent with a pattern in which examiners match citing and cited patents on technology to a greater degree than inventors.

The results on vintage effects indicate that examiners are more likely to add recent citations than inventors, with a mean difference in years of 7 versus 9.8 years. The long time lags for both inventors and examiners (7 and 10 years) mean the difference is probably not due to administrative delays which would allow examiners to know about new patents sooner than inventors. While we do not have a strong theoretical explanation for this finding, we note that it is consistent with a pattern in which inventors and lawyers may choose to cite older patents whose owners are less likely to litigate than owners of recently issued patents.

**Comparing inventor and examiner citations: Multivariate analysis**

To analyze differences between examiner and inventor citations streams, we also estimate models at the dyad level for each citing-cited pair in our sample. We estimate the following empirical specification:

$$Prob(examiner\ citation_{ij}=1 \mid X_{ij})=F(\beta X_{ij}) + u_{ij} \qquad (2)$$

where $F(Z)=e^{z}/(1+e^{z})$ is the cumulative logistic distribution, our dependent variable is binary and equal to 1 if the citation was imposed by the examiner and 0 if it comes from the inventor, with $X_{ij}$ a set of variables that indicate similarities between citing patent $i$ and cited patent $j$ along the dimensions shown in table 3: self-citation by individuals, assignees, corporate parents, lawyers and examiners; geographic location; technology classes and time. Since citation pairs are not necessarily independent of citing patents, estimating equation (2) without paying further attention to error terms could generate biased estimates. We explore three alternatives to deal with this problem: using fixed effects per citing patent; correcting standard errors for heteroskedasticity by clustering on citing patents; and random-effects model on a panel data structure. We adopt the latter for a number of reasons. First, with a fixed-effects model, all citing patents that have all or zero citations added by examiners would drop from our sample, resulting in a loss of 48 per cent of citing-cited pairs. Second, a Hausman test comparing fixed and random effects specifications favor the latter. Third, by explicitly modeling the individual component (citing patent) that is common across cited patents, a random effects model offers an extra advantage over heteroskedasticity correction of standard errors. Fourth, tests for all models show that the panel-level variance component (within citing patent variation) is important and the panel estimator is different from the pooled estimator. Thus, our empirical approach is to estimate the following equation

$$\text{Prob}(\textit{examiner citation}_{ij}=1 \mid X_{ij}, v_i)=\text{F}(\beta X_{ij} + v_i) + \varepsilon_{ij} \qquad (3)$$

where $u_{ij=} v_i + \varepsilon_{ij}$ , $v_i$ is the unobserved heterogeneity for the $i$th citing patent with mean zero and variance $\sigma_v^2$ and

$$\text{examiner citation}=1 \Leftrightarrow \beta X_{ij} + v_i + \varepsilon_{ij} > 0$$

for $\varepsilon_{ij}$ iid logistic distributed with mean zero and variance $\sigma_\varepsilon^2$ so that $\varepsilon_{ij}$ is also orthogonal to $v_i$ (Greene 2003).

Table 5.2 shows the summary statistics and table 6.2 the correlation values for all variables. Table 8 contains results in two sets. The first estimates the model for all dyads (models 1-4), the second set excludes patents with all citations imposed by examiner (model 5); then excludes patents with no patent added by examiner (model 6) and finally excludes both sub-samples (model 7). We include the latter three models as robustness checks of our full-sample models. Coefficients for all variables are expressed as odds ratios. An odds ratio greater than one indicates that examiners are more likely to have added the citation than inventors; odds ratios less than one indicate that examiners are less likely than inventors to have added a citation. Statistical significance for a given coefficient indicates that examiners and inventors differ in the propensity to add a citation.

Table 8 provides two panels at the bottom. The first panel presents the number of citing-cited pairs, number of citing patents, and minimum, average and maximum number of cited patents by citing patent. The second panel offers two tests to evaluate the models. The Wald Chi-square test provides evidence of model fit (similar to an F test), the Chi bar-square tests whether the pooled estimator is equal to the panel estimator. For all models, tests indicate that the panel data estimators are preferred over the pooled estimators.

Now we turn to discuss our results. We first consider self-citation at the firm level. In model 1, self-citation is measured for the company level: a positive, and statistically significant coefficient on *dif_company* suggests that a citation across different firms is 78% more likely to be an examiner citation. For self-citation at the corporate parent level (*dif_parent,* model 2), the coefficient is again positive and significant, indicating that self-citations at the corporate level are more likely to be generated by inventors. This corresponds to our simple expectation that inventors are more likely to generate self-citations than examiners.

The coefficient for self-citation at the individual inventor level in models 1 and 2 is not statistically significant. In models 3 and 4 we include more restricted definitions for inventor self-citation. Here, the coefficients reinforce the findings of the univariate analysis: self-citations to patents with at least one of the same inventors as the citing patent are more likely to come from examiners! The magnitude of this unpredicted effect is striking. Odds ratios of 2.52 and 1.92 for same inventor/ same city (model 3) and same inventor/same company (model 4) respectively indicate that links between patents with same inventors are at least twice more likely to come from examiners than from inventors. This clearly violates the assumption that self-citations are more likely to be generated by inventors than examiners.

Self-citation at the examiner level is also significant: citing-cited pairs that share the same examiner are more likely to be added by the examiner ($P<0.01$). The magnitude of the effect is high: a link between two patents is at least 85% more likely by an examiner if she reviews the application of both patents. In other words, examiners add citations to patents with which they had previous experience; examiners assume a role of linking patents based on their own examination practices histories. The specialization of examiner by technology also increases the likelihood that examiner self-citation would be high, since in some art units examiners have examined much of the relevant prior art (Cockburn, Kortum and Stern 2004).

Technology and vintage effects are also consistent with the univariate tests, and are not in line with our simple expectations about inventor citations more likely cluster in "nearby" technology classes or years. We estimate models with technology specified at the 4-digit classification level. Examiners are less likely to add citations when the cited and citing patents differ in technological class ($p<0.01$, across all models). As shown in model 1 an inventor is 43% (1/0.70) more likely to cite patents from different 4-digit technological classes than from the same classes of the citing patent, suggesting that inventors are adding more breadth of prior art than examiners. We also find a similar pattern for the variable *years,* which indicates that examiners are less likely to cite older patents than inventors, however, the magnitude of the difference is small. Figure 4 shows graphically these differences in time for examiner and inventor citations.

Finally, the coefficient on distance is equal to one and highly significant, indicating an equal probability of a citation being generated by inventors or examiners across distance. We examine the effects of geography in more detail in models that follow (table 9). This result does not indicate the presence of greater localization for inventor than for examiner citations.

Patents for which the examiner imposes either all citations or zero citations could have special characteristics that may affect our findings. To verify that our results are robust to these potentially problematic patents, we replicate model 1 for three sub samples: excluding all patents where all citations are examiner imposed (model 5), excluding all patents with zero citation added by examiner (model 6), and excluding both groups (model 7). Note that the number of observations changes for these models since 580 citing patents have all their citations imposed by examiner and 114 have zero patents added. All coefficients in models 5, 6 and 7 are similar in magnitude and significance to those discussed previously, indicating that our findings are robust to inclusion of these groups.

We also test for whether results change when lawyers are involved in the process. We re-run all models for patents with and without lawyers, but do not find significant differences in the estimated coefficients from the full model; we do not report those results (available from authors). One concern is that the sample of patents in which lawyers are not involved is so small – only 5 per cent of dyads – that this approach is not an adequate test for the true effect of lawyers on inventor prior art. We suspect that they are important, and partially responsible for the closeness between examiner and inventor citation means.

We explore further the role of geographic localization in table 9. We re-estimate model 1 from table 8, but instead of measuring distance in continuous miles between citing and cited patent we define distance with binary variables that indicate whether there is at least one pair of inventors in the citing-cited pair that is in the same country, state, county, economic area, or city. Note that only 7,632 observations are used to estimate models 2 through 5 since the geographic definitions used in these models require that at least one inventor on both citing and cited patents be located in the US. The drop in sample size underscores the very high proportion of dyads (52%) that include at least one inventor located outside the United States.

We focus our attention on the geographic component (results for the other independent variables remain similar to those in table 8). Examiners are 25% more likely than inventors to generate citations to patents that are in other countries than inventors, confirming their role as connectors of knowledge across national boundaries. This lends strong support for Jaffe and Trajtenberg's (1998) prior finding that knowledge spillovers as evidenced by patent citations are strongly national in character. Within the United States, we find more localization for inventor than examiner citations: examiners are more likely to cite patents originating in different states, counties, and economic areas. However, the coefficient for city is not statistically significant. City may be too small to be an economically meaningful unit in measuring knowledge flows; particularly knowledge flows related to employment communication, and transaction patterns in a local area. The economic area is designed to overcome these limitations, and is statistically significant. Overall, then, we find evidence of more localization for inventor citations than examiner citations, which is congruent with our expectations about what would occur if citations indicate inventor knowledge. At the same time, the magnitude of the difference between examiners and inventors in geographic citing patterns is very low, as shown in figures 2 and 3. In other words, while the difference is in the expected direction, the magnitude of the difference is small. This raises the possibility that examiner citations are potentially inflating localization patterns that are being attributed to knowledge spillovers.

**Analysis of highly-cited patents**

We now turn to our analysis of highly-cited patents. Our objective is to learn whether highly-cited patents – which are associated with patents of unusually high economic and technological value – "earn" their status through citation by examiners (an administrative process) or citation by inventors (an evolutionary, or diffusion, process). We identify all patents granted in 1998 (119,852 patents) that are in the top 1% according to the total number of forward citations they receive between January 1 2001 and August 30 2003 – the years for which we are able to distinguish between examiner and inventor citations. This yields a group of 1,175 highly-cited patents. The correlation between forward citations received from Jan 2001 to August 2003 and those received from grant date to August 2003 is 0.915, so we are confident that our group represents the cohort of highly-cited patents from 1998. For each highly-cited patent, we identify if it also belongs to the top 1% according to citations made only by examiners (*top1_examiner*) and citations made only by inventors (*top1_inventor)* between January 1, 2001 and August 30 2003. Since forward citations peak at about four years post-issue, we are fairly confident that even though our counts are both left- and right-truncated, we have a representative picture of the total flow of inventor and examiner citations to the 1998 cohort.

The correlation between *top1_examiner* and *top1_inventor* is –0.39 ($p<0.01$) suggesting that inventors and examiners select different patents as being important. Table 10 provides a more detailed view of the differences between examiners and inventors. Each cell presents the number of patents in a category and its percentage share from the group of 1,175 highly-cited patents. We see a fairly high level of separation between patents selected by examiners and those selected by inventors, such that one or the other determines entry into the group of highly-cited patents. Only 17% (204 patents) of highly cited patents are in the top 1% as selected by both examiners and inventors. Only 12 percent are neither in the top 1% of inventors or examiners; for these patents, it is the addition of examiner and inventor citations that earns entry into the top-cited group. Most patents are selected by either inventors or examiners, but not both: 47 per cent are highly-cited by inventors but not examiners; 24 percent are highly-cited by examiners but not inventors. Overall, inventors are more important than examiners in determining top-cited patents, with some 65% of the group in the top-cited by inventor category as against only 41 per cent in the top-cited by examiner category. Across all patents, the correlation between the number of inventor and examiner citations is equal to -0.13 ($p<0.01$), further indicating different selection processes by these two groups.

To further explore differences between inventors and examiners in allocating citations to patents, we estimate correlations between examiner and inventor forward citations received between 2001 and August 2003 for all patents granted since 1998. We are interested in whether the differences we found in the highly-cited group between inventor and examiner citing patterns hold for all patents, and whether there are time effects that change the ways in which patents are cited by examiners and inventors. We found in our logit models that inventors are more likely to cite older patents than examiners. It is possible that patents are initially cited by examiners, who are most aware of recently-issued patents, and those same patents are subsequently cited by inventors – who learn of the prior art from examiners. Here, knowledge flows occur not between inventors directly, but indirectly with the examiner acting as intermediary.

We construct two measures of forward citation for patents issued since 1998. We first count the number of forward citations made by examiners and inventors. We then rank patents within a given year into deciles, according to citations received by inventors and examiners. We correlate inventor and examiner citations streams for each of these two measures. Table 11 shows the results, which reveal a pattern that varies over time. For recently issued patents, correlations are quite high and negative, suggesting a negative relationship between inventor and examiner selection of important patents.   Over time, these difference steadily attenuate, with correlations becoming positive, such that for the oldest cohort (patents issued in 1998) the correlation is 0.38 – weak, but still indicative of some overlap between inventor and examiner choice of important patent. The data show a gradual pattern of convergence between inventor and examiner citations, which is congruent with a story of examiner-mediated learning sketched above. We intend to explore these patterns further with additional statistical analysis as well as interviews with examiners and other patent professionals.

**Discussion and conclusion**

Knowledge is difficult to measure, and researchers have understandably been eager to apply patent citation data to test theories of knowledge creation and diffusion by organizations and individuals. However, the question has always remained as to the extent to which these data actually do measure knowledge and track knowledge flows. In particular, the addition of examiner citations and the process by which firms and attorneys craft patents would seem to add significant noise – and possibly distortions to –assumed patterns of knowledge flows. Apart from a few studies that show the potential weakness in the knowledge transfer assumption (Jaffe, Trajtenberg and Fogarty 2000; Meyer 2000; Michel and Bettels 2001; Thompson 2003; Breschi and Lissoni 2004; Cockburn, Kortum and Stern 2004) ours is among the first to compare the generation of inventor citations to examiner citations in the aggregated data. We also provide the first analysis of how inventors and examiners differ in selecting highly-cited patents, which have been associated with high economic and technological importance. Our study is also important in showing the degree to which careful attention to cleaning and matching names, places and organizations is needed to avoid over- or under-estimating true rates of matching in patent data.

One methodological problem we face is that we are not able to separate citations specified by inventors from those added by inventors' lawyers. Insofar as lawyers are likely to be cognitively and behaviorally much closer to examiners than to inventors this is a limitation of our analysis. We attempted to show whether lawyers affect citing patterns at the dyad level; however, so few dyads did not involve lawyers (less than 5 per cent of dyads) that we cannot produce results that would tease apart the effects these actors have on inventor-only citations. Another methodological drawback is that we do not analyze the generation of non-patent references, which form an important body of prior art, particularly for emerging technologies. We also do not consider citations to patents issued outside the United States. However, our analysis of patent citations should help inform many studies of knowledge flows which only consider prior art captured in US patents.

Our simple expectation of what inventor citations "ought" to reveal – and the corresponding patterns of examiner citations – most approximates a world in which inventors reveal what is in their heads, and examiners fill in the missing pieces. We make a simple assumption that inventors have the greatest knowledge of proximate technologies, and measure proximity along a number of dimensions. We find the greatest statistical support for this scenario in our analysis of localization of citations and self-citation at the firm level. However, regarding geography, the closeness with which examiner and inventor citations track each other raises the question of the economic significance of these differences, and raises the possibility of Type 2 error for inferring localization effects from aggregate data. We find that examiner citations differ from inventor citations in unexpected directions (examiners cite more proximate patents) along the dimensions of technological distance and time. The most unexpected finding is for self-citation at the individual level. The fact that self-citations by individuals (using the most stringent matching rules) are more likely to come from examiners than inventors indicates that citations are not straightforward codification of inventor knowledge. We believe that inventors know of their prior patents, but they omit many of those from their lists of prior art, which are subsequently added back in by examiners. We do not have a strong theoretical explanation for these patterns, which are even more perplexing given high inventor self-citation at the firm level: we can however, state that the results show that examiners are not adding noise to the data, but are including citations we would naturally attribute to inventors.

Overall, our results do not change the presumption that patents trace out knowledge flows: inventors face strong legal pressures to reveal all they know, and our results do show that inventor citations follow a pattern we would associate with inventor knowledge. However, while researchers have argued that aggregate citations are a noisy signal of inventor knowledge, our analysis indicates that changes introduced by patent examiners are both high and non-random. Instead, we find that the "invisible hand" of administration and the legal system is strong in generating aggregate citation streams. Indeed, we suspect that even inventor citations taken on their own are an imperfect measure of inventor knowledge, given the active role that attorneys and patent searchers play in the process. Our analysis should help indicate the direction of the bias that might occur for a variety of theoretical hypotheses that make the assumption that aggregate citations measure knowledge transfer.

Our results point to interesting processes by which citations are generated that have not received as much attention in the literature. We show that examiners and inventors adhere to different processes of selection of important patents. It may be that inventors and examiners develop “favorite” patents for citation that are guided by different criteria: in the case of inventors, these might be self citations or older patents with less risk of litigation, and for examiners, patents they know well or “thick” patents that encapsulate a great deal of prior art. We intend to explore these different selection processes with further analysis of highly-cited patents.

The finding that inventor and examiner forward citations slowly converge over time is potentially indicative of a learning process between examiners and inventors that has not been shown before in the literature and would further add complexity to the picture of citations as measuring knowledge flows. At the turn of the last century, patent agents – the forerunners of modern patent attorneys and examiners – were positioned as information hubs about patents, becoming important intermediaries in emerging markets for technology(Lamoreaux and Sokoloff 2003). We believe there are parallels in the current institutional environment that have been overlooked: firm-level learning from examiners is likely important, both through citations and personnel movements as examiners leave the US Patent Office to become attorneys and patent professionals working for inventors.

Our paper suggests that prior art citations are not only codified lists of knowledge held by inventors, but legal and strategic tool in which the interests of inventors, attorneys, examiners, and competitors come into play. Our results show that aggregate citations should not be viewed as a noisy signal of knowledge flows, but as a multi-dimensional signal involving heterogeneous processes and actors – knowledge flows among inventors, learning between inventors and examiners, and a complex administrative process of codification by lawyers and examiners– that intersect to create and shape technology fields.

## References


Allison, J., M. Lemley, K. Moore and D. Trunkey (2003). "Valuable Patents." UC Berkeley School of Law **Research Paper 133**.

Breschi, S. and F. Lissoni (2004). Knowledge Networks from Patent Data: Methodological Issues and Research Targets. Handbook of Quantitative S&T. M. H. Glänzel W., Schmoch U., Kluwer Academic Publishers.

Cockburn, I., S. Kortum and S. Stern (2004). Are All Patent Examiners Equal? Examiners, Patent Characteristics, and Litigatioin Outcomes. Patents in the Knowledge-based Economy. W. M. Cohen and S. Merrill. Washington D.C., National Academy Press.

Gittelman, M. and B. Kogut (2003). "Does Good Science Lead to Valuable Knowledge? Biotechnology Firms and the Evolutionary Logic of Citation Patterns." Management Science **49**(4): 366-382.

Greene, W. H. (2003). Econometric Analysis. New Jersey, Prentice Hall.

Hall, B., A. and M. T. Jaffe (2000). "Market value and patent citations: A first look." NBER Working Paper 7741.

Hall, B. H., A. B. Jaffe and M. Trajtenberg (2001). "The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools." NBER Working Paper **8498**.

Henderson, R., A. B. Jaffe and M. Trajtenberg (1998). Universities as a source of commercial technology: A detailed analysis of university patenting, 1965-1988. The Review of Economics and Statistics. Cambridge**:** 119-127.

Jaffe, A. and M. Trajtenberg (1998). "International Knowledge Flows: Evidence from Patent Citations." NBER Working Paper 6507.

Jaffe, A., M. Trajtenberg and M. S. Fogarty (2000). "The Meaning of Patent Citations: Report on the NBER/Case Western Research Survey of Patentees." NBER Working Paper **7631**.

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.

Lamoreaux, N. and K. Sokoloff (2003). Intermediaries in the U.S. Market for Technology, 1870-1920. Finance, Intermediaries, and Economic Development. e. a. Stanley L. Engerman. New York, Cambridge Univerasity Press**:** 209-246.

Levin, R. C., A. K. Klevorick, R. R. Nelson, S. G. Winter, R. Gilbert and Z. Griliches (1987). Appropriating the Returns from Industrial Research and Development; Comments and Discussion. Brookings Papers on Economic Activity. Washington**:** 783 49 pages.

Meyer, M. (2000). "Does science push technology? Patents citing scientific literature." Research Policy **29**(3): 409.

Michel, J. and B. Bettels (2001). "Patent citation analysis: A closer look at the basic input data from patent search reports." Scientometrics **52**(1): 185-20.

Song, J., P. Almeida and G. Wu (2003). "Learning-by-hiring: When is mobility more likely to facilitate interfirm knowledge transfer?" Management Science **49**(4): 351.

Thompson, P. (2003). "Patent Citations and the Geography of Knowledge Spillovers: What do Patent Examiners Know?" Working Paper **Carnegie Mellon University**.

Table 1
Summary statistics of full dataset

| | 2001 | 2002 | 2003 * | Total |
|---|---|---|---|---|
| Citing patents | 166,064 | 167,424 | 109,351 | 442,839 |
| Cited patents | 1,960,448 | 2,040,345 | 1,433,690 | 5,434,483 |
| Inventor | 57% | 59% | 60% | 59% |
| Examiner | 43% | 42% | 40% | 42% |
| % examiner citation x patent | | | | |
| Average | 63% | 63% | 63% | |
| 0% | 8% | 8% | 7% | |
| (0%, 10%] | 5% | 5% | 6% | |
| (10%, 20%] | 6% | 6% | 7% | |
| (20%, 30%] | 6% | 6% | 6% | |
| (30%, 40%] | 5% | 5% | 6% | |
| (40%, 50%] | 8% | 8% | 9% | |
| (50%, 60%] | 3% | 3% | 3% | |
| (60%, 70%] | 6% | 6% | 7% | |
| (70%, 80%] | 4% | 4% | 5% | |
| (80%, 90%] | 6% | 6% | 6% | |
| (90%, 100%) | 42% | 41% | 39% | |
| 100% | 40% | 39% | 38% | |

* From January 1 to August 26 2003

Table 2. Comparison of full dataset and 3 year sample

| | 2001 Universe | 2001 Sample | 2002 Universe | 2002 Sample | 2003 Universe | 2003 Sample |
|---|---|---|---|---|---|---|
| Citing patents | | | | | | |
| Total | 166,064 | 500 | 167,424 | 500 | 109,351 | 500 |
| With no citations | 2.4% | 2.0% | 2.3% | 2.4% | 1.9% | 2.8% |
| Cited patents | | | | | | |
| Total | 1,960,448 | 5,668 | 2,040,345 | 5,902 | 1,433,690 | 6,296 |
| Citation/patent | | | | | | |
| Mean | 11.84 | 11.33 | 12.23 | 11.80 | 13.18 | 12.62 |
| Std. Dev | 17.74 | 13.63 | 18.47 | 21.00 | 20.75 | 18.41 |
| % Examiner citations | | | | | | |
| Mean | 0.63 | 0.63 | 0.63 | 0.62 | 0.63 | 0.64 |
| Std Dev | 0.37 | 0.37 | 0.37 | 0.37 | 0.37 | 0.38 |
| Application year | | | | | | |
| Mean | 1,998.69 | 1,998.68 | 1999.699 | 1999.7 | 2,000.40 | 2,000.32 |
| Std. Dev | 1.25 | 1.19 | 1.22 | 1.18 | 1.24 | 1.13 |
| Mann-Whitney Test | z | Prob >\|z\| | z | Prob >\|z\| | z | Prob >\|z\| |
| Citation/patent | 0.606 | 0.5442 | 0.861 | 0.389 | 0.549 | 0.5828 |
| % Examiner citations | 0.38 | 0.70 | 0.746 | 0.455 | (0.85) | 0.40 |
| Application year | 0.317 | 0.751 | 0.045 | 0.9369 | 0.932 | 0.3514 |
| Kolmogorov-Smirnov Test | D | p-value | D | p-value | D | p-value |
| Citation/patent | 0.0301 | 0.769 | 0.0305 | 0.755 | 0.0313 | 0.73 |
| % Examiner citations | 0.0246 | 0.929 | 0.0305 | 0.755 | -0.0377 | 0.499 |
| Application year | 0.0099 | 1 | 0.007 | 1 | 0.0178 | 0.998 |

Mann-Whitney Test: H0: Sample= Universe, H1: Sample~= Universe
Kolmogorov-Smirnov Test: H0: Distribution of Sample= Distribution of Universe, H1: Distribution of Sample~= Distribution of Universe

| **IPC code** | **Description** | **Rank in universe** | **Rank in sample** | **% of sample** | **% of universe** |
|---|---|---|---|---|---|
| G06F | Electric Digital Data Processing | 1 | 2 | 6.1% | 6.3% |
| H01L | Semiconductor Devices | 2 | 1 | 6.4% | 5.8% |
| A61K | Preparations For Medical, Dental, or Toilet Purposes | 3 | 3 | 3.6% | 3.6% |
| A61B | Diagnosis; Surgery; Identification | 4 | 5 | 2.2% | 2.0% |
| G02B | Optical Elements, Systems, or Apparatus | 5 | 4 | 2.3% | 1.8% |
| G01N | Investigating or Analysing Materials By Determining Their Chemical or Physical Properties | 6 | 7 | 2.0% | 1.5% |
| B32B | Layered Products | 7 | 10 | 1.7% | 1.5% |
| G11B | Information Storage Based On Relative Movement Between Record Carrier And Transducer | 8 | 11 | 1.6% | 1.5% |
| H01R | Electrically-Conductive Connections | 9 | 18 | 0.9% | 1.3% |
| H04N | Pictorial Communication, e.g. Television | 10 | 6 | 2.1% | 1.3% |
| G11C | Static Stores | 11 | 14 | 1.2% | 1.3% |
| H04B | Transmission | 12 | 9 | 1.7% | 1.2% |
| B41J | Typewriters; Selective Printing Mechanisms | 13 | 12 | 1.3% | 1.2% |
| G06K | Recognitionof Data; Presentationof Data; Record Carriers; Handling Record Carriers | 14 | 24 | 0.8% | 1.2% |
| A61F | Prostheses; Orthopaedic, Nursing or Contraceptive Devices; Treatment or Protectionof Eyes or Ears | 15 | 16 | 1.0% | 1.1% |
| H04L | Transmissionof Digital Information | 16 | 26 | 0.8% | 1.0% |
| B65D | Containers For Storage or Transportof Articles or Materials | 17 | 34 | 0.7% | 1.0% |
| C07C | Organic Chemistry | 18 | 8 | 1.9% | 0.9% |
| C07D | Heterocyclic Compounds | 19 | 15 | 1.1% | 0.9% |
| B01D | Separation | 20 | 17 | 1.0% | 0.9% |
| | Total top 20 classes | | | 40.5% | 37.5% |

Table 3
Variable definitions and measurement

| Dimension | Variable | Definition |
|---|---|---|
| Dependent variable | Examiner | 1 if examiner citation, 0 if inventor citation |
| **Self Citation, by:** | | |
| *Inventors* | same_inventor | 1if citing and cited patents have the same first inventor, 0 otherwise |
| | same_inventor_all | 1 if citing and cited patents have at least 1 inventor in common, 0 otherwise |
| *Firms* | dif_company1 | 0 if citing and cited patents have the same first assignee, 1 otherwise |
| | dif_company_all | 0 if citing and cited patents have at least 1 assignee in common, 1 otherwise |
| *Corporate parent* | dif_parent1 | 0 if citing and cited patents have the same first ultimate parent 1 otherwise |
| | dif_parent_all | 0 if citing and cited patents have at least 1 ultimate parent in common, 1 otherwise |
| *Lawyers* | same_law_firm | 1 if citing and cited patents have the law firm, 1 otherwise |
| | same_law_firm_all | 1 if citing and cited patents have either at least 1 law firm in common, 0 otherwise |
| *Examiners* | same_examiner | 1 if citing and cited patents have the same primary examiner, 0 otherwise |
| | same_examiner_all | 1 if citing and cited patents have either the same primary or assitant examiner, 0 otherwise |
| **Location, by:** | | |
| *Country* | dif_country1 | 0 if the first inventors in the citing and cited patents are in the same country, 1 otherwise |
| | dif_county_all | 0 if at least 1 inventor in the citing and cited patents is in the same country, 1 otherwise |
| *State (US)* | dif_state1 | 0 if the first inventors in the citing and cited patents are in the same state, 1 otherwise |
| | dif_state_all | 0 if at least 1 inventor in the citing and cited patents is in the same state, 1 otherwise |
| *Economic area (US)* | dif_ea1 | 0 if the first inventors in the citing and cited patents are in the same economic area, 1 otherwise |
| | dif_ea_all | 0 if at least 1 inventor in the citing and cited patents is in the same economic area 1 otherwise |
| *County (US)* | dif_county1 | 0 if the first inventors in the citing and cited patents are in the same county, 1 otherwise |
| | dif_county1_all | 0 if at least 1 inventor in the citing and cited patents is in the same county 1 otherwise |
| *City (all)* | dif_city1 | 0 if the first inventors in the citing and cited patents are in the same city, 1 otherwise |
| | dif_city1_all | 0 if at least 1 inventor in the citing and cited patents is in the same city 1 otherwise |
| *Miles* | distance1 | Distance in miles between the location of first inventors for citing and cited patents |
| **Technology** | dif_technology4 | 0 if citing and cited patents have same primary IPC technology classficiation, 1 otherwise |
| | dif_technology4_all | 0 if citing and cited patents have at least 1 IPC technology classficiation in common, 1 otherwise |
| **Time** | years | application year citing-application year cited |

Table 4.
Comparison of Means, Inventor and Examiner Citations

| | Inventor Citations (n=9370) | Examiner Citations (n=6725) | T test |
|---|---|---|---|
| **Self Citation:** | | | |
| Same inventor | 0.063 | 0.062 | 0.2 |
| Same inventor, same company | 0.039 | 0.046 | -4.12** |
| Same inventor, same city | 0.022 | 0.033 | -2.29** |
| Different company | 0.89 | 0.9 | -1.55 |
| Different parent | 0.87 | 0.88 | -3.2** |
| Same law firm [a] | 0.09 | 0.08 | 1.5 |
| **Technology Class:** | | | |
| Different Technology, 4 digit IPC code | 0.49 | 0.38 | 13.1** |
| **Geographic Distance:** | | | |
| Different country, all inventors | 0.34 | 0.47 | -0.17** |
| Different state, all inventors, US only[b] | 0.7 | 0.73 | -3.04** |
| Different city, all inventors | 0.9 | 0.91 | -1.05 |
| Different economic area, all inventors, US only[b] | 0.74 | 0.77 | -3.02** |
| Distance, miles | 2197 | 2605 | -11.0** |
| **Vintage:** | | | |
| Years | 9.8 | 7.1 | 28.3** |

* $p<0.05$ ** $p<0.01$

a. $N_{inventors}=6986$; $N_{examiners}=4988$

b. $N_{inventors}=5253$; $N_{examiners}=2379$

Table 5.1
Summary statistics for variables in regressions at the patent level

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| perimposed | 1,456 | 0.63 | 0.4 | 0 | 1 |
| log_inventor_citations | 1,456 | 1.55 | 1.0 | 0 | 5.43 |
| lawyer | 1,456 | 0.95 | 0.2 | 0 | 1 |
| american_company | 1,456 | 0.52 | 0.5 | 0 | 1 |
| top_200 | 1,456 | 0.36 | 0.5 | 0 | 1 |
| citing_type_academia | 1,456 | 0.02 | 0.1 | 0 | 1 |
| citing_type_industry | 1,456 | 0.95 | 0.2 | 0 | 1 |
| citing_type_govt | 1,456 | 0.01 | 0.1 | 0 | 1 |

Table 5.2
Summary statistics for variables in regressions at the citing-cited pair level

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| distance | 16,089 | 2,368 | 2,322 | 0 | 10,781 |
| dif_country_all | 16,089 | 0.40 | 0.49 | 0 | 1 |
| difstateall | 7,632 | 0.71 | 0.45 | 0 | 1 |
| dif_county_all | 7,632 | 0.81 | 0.39 | 0 | 1 |
| dif_city_all | 16,089 | 0.90 | 0.30 | 0 | 1 |
| dif_company_all | 16,089 | 0.89 | 0.31 | 0 | 1 |
| dif_parent_all | 16,089 | 0.87 | 0.33 | 0 | 1 |
| same_inventor_all | 16,089 | 0.06 | 0.24 | 0 | 1 |
| same_inventor_city_all | 16,089 | 0.03 | 0.16 | 0 | 1 |
| same_inventor_company_all | 16,089 | 0.04 | 0.20 | 0 | 1 |
| same_examiner_all | 16,089 | 0.07 | 0.25 | 0 | 1 |
| same_lawyer | 1,998 | 0.12 | 0.32 | 0 | 1 |
| same_law_firm | 11,974 | 0.09 | 0.28 | 0 | 1 |
| years | 16,089 | 8.70 | 6.16 | 0 | 27 |
| dif_technology4 _all | 16,089 | 0.45 | 0.50 | 0 | 1 |

Table 6.1
Correlation table for regressions at patent level

| | perimposed | log_original_citations | american_company | top_200 | citing_type_academic | citing_type_industry | citing_typ_govt |
|---|---|---|---|---|---|---|---|
| perimposed | 1 | | | | | | |
| log_original_citations | -0.6807 | 1 | | | | | |
| | (0.00) | | | | | | |
| american_company | -0.3522 | 0.2721 | 1 | | | | |
| | (0.00) | (0.00) | | | | | |
| top_200 | 0.0606 | -0.0007 | -0.1706 | 1 | | | |
| | (0.00) | (0.93) | (0.00) | | | | |
| citing_type_academic | -0.0125 | -0.0168 | 0.0089 | -0.0688 | 1 | | |
| | (0.11) | (0.03) | (0.26) | (0.00) | | | |
| citing_type_industry | -0.0314 | 0.0743 | 0.0174 | 0.0784 | -0.6512 | 1 | |
| | (0.00) | (0.00) | (0.03) | (0.00) | (0.00) | | |
| citing_typ_govt | 0.0306 | -0.0458 | 0.0176 | 0.0314 | -0.0101 | -0.3731 | 1 |
| | (0.00) | (0.00) | (0.03) | (0.00) | (0.20) | (0.00) | |

Table 6.2
Correlation table for regressions at patent level

| | distance | dif_country_all | dif_state_all | dif_county_all | dif_city_all | dif_company_all | dif_parent_all | same_inventor_all | same_inventor_city_all | same_inventor_company_all | same_examiner_all | same_layer_all | same_firm_all | dif_technology4 _all | years |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| distance | 1 | | | | | | | | | | | | | | |
| dif_country_all | 0.787 | 1 | | | | | | | | | | | | | |
| | (0.00) | | | | | | | | | | | | | | |
| dif_state_all | 0.5637 | . | 1 | | | | | | | | | | | | |
| | (0.00) | (1.00) | | | | | | | | | | | | | |
| dif_county_all | 0.4596 | . | 0.7439 | 1 | | | | | | | | | | | |
| | (0.00) | (1.00) | (0.00) | | | | | | | | | | | | |
| dif_city_all | 0.2968 | 0.2655 | 0.5761 | 0.7755 | 1 | | | | | | | | | | |
| | (0.00) | (0.00) | (0.00) | (0.00) | | | | | | | | | | | |
| dif_company_all | 0.2908 | 0.2519 | 0.4526 | 0.5429 | 0.5515 | 1 | | | | | | | | | |
| | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | | | | | | | | | | |
| dif_parent_all | 0.3084 | 0.2636 | 0.4859 | 0.5738 | 0.5521 | 0.9133 | 1 | | | | | | | | |
| | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | | | | | | | | | |
| same_inventor_all | -0.2292 | -0.2025 | -0.4196 | -0.5403 | -0.6154 | -0.4743 | -0.478 | 1 | | | | | | | |
| | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | | | | | | | | |
| same_inventor_city_all | -0.1613 | -0.1351 | -0.2884 | -0.388 | -0.5089 | -0.3472 | -0.336 | 0.6488 | 1 | | | | | | |
| | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | | | | | | | |
| same_inventor_company_all | -0.192 | -0.1659 | -0.3363 | -0.4389 | -0.5381 | -0.6046 | -0.552 | 0.8147 | 0.5894 | 1 | | | | | |
| | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | | | | | | |
| same_examiner_all | -0.0296 | -0.0145 | -0.0264 | -0.033 | -0.054 | -0.0671 | -0.052 | 0.0677 | 0.0802 | 0.0783 | 1 | | | | |
| | (0.00) | (0.07) | (0.02) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | | | | | |
| same_layer_all | -0.2042 | -0.1618 | -0.4053 | -0.426 | -0.4121 | -0.4886 | -0.48 | 0.4435 | 0.2872 | 0.3965 | 0.1082 | 1 | | | |
| | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | | | | |
| same_firm_all | -0.262 | -0.2313 | -0.4257 | -0.4945 | -0.4792 | -0.6004 | -0.626 | 0.4674 | 0.3324 | 0.4496 | 0.05 | 0.7049 | 1 | | |
| | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | | | |
| dif_technology4 _all | 0.0372 | 0.0509 | 0.1277 | 0.1439 | 0.134 | 0.1572 | 0.1466 | -0.1347 | -0.113 | -0.1487 | -0.1405 | -0.2471 | -0.1464 | 1 | |
| | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | | |
| years | 0.0041 | 0.0061 | 0.0278 | 0.0378 | 0.0445 | 0.0584 | 0.0518 | -0.0645 | -0.0446 | -0.0589 | -0.127 | -0.0569 | -0.0512 | 0.0878 | 1 |
| | (0.60) | (0.44) | (0.02) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.01) | (0.00) | (0.00) | |

Table 7
Results of regressions at patent level

Dependent variable is % of examiner citations
Fixed effects for assignee

| | (1) | (2) | (3) |
|---|---|---|---|
| | perimposed | perimposec | perimposed |
| log_inventor_citations | -0.11 | -0.109 | -0.109 |
| | [0.015]** | [0.015]** | [0.015]** |
| lawyer | -0.065 | -0.066 | -0.066 |
| | [0.066] | [0.066] | [0.066] |
| american_company | -0.082 | | |
| | [0.133] | | |
| top_200 | | 0.027 | |
| | | [0.365] | |
| citing_type_academia | | | 0.351 |
| | | | [0.442] |
| citing_type_industry | | | 0.591 |
| | | | [0.360] |
| citing_type_govt | | | 0.936 |
| | | | [0.440]* |
| Constant | 1.054 | 0.973 | 0.381 |
| | [0.233]** | [0.192]** | [0.321] |
| F-test for fixed effects | 1.19 | 1.24 | 1.24 |
| Prob > F | 0.0121 | 0.0029 | 0.0025 |
| Observations | 1456 | 1456 | 1456 |
| R-squared | 0.75 | 0.75 | 0.75 |

Standard errors in brackets
* significant at 5%; ** significant at 1%

Table 8
Results of logit regressions

Dependent variable is equal to 1 if citation comes from examiner, 0 otherwise
Results assume random effects structure for error term
Models 1, 2, 3, and 4 include all citing patents
Model 5 excludes 580 citing patents where all citations come from examiner
Model 6 excludes 114 citing patents where zero citations come from examiner
Model 7 excludes citing patents where both all (580) or zero (114) citations come from examiner

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| dif_company | 1.785 | | 1.966 | 2.069 | 1.672 | 1.843 | 1.758 |
| | [0.185]** | | [0.192]** | [0.239]** | [0.175]** | [0.189]** | [0.186]** |
| dif_parent | | 1.822 | | | | | |
| | | [0.181]** | | | | | |
| same_inventor | 1.245 | 1.286 | | | 1.282 | 1.54 | 1.343 |
| | [0.159] | [0.167] | | | [0.160]* | [0.190]** | [0.172]* |
| same_inventor_city | | | 2.521 | | | | |
| | | | [0.407]** | | | | |
| same_inventor_company | | | | 1.918 | | | |
| | | | | [0.398]** | | | |
| same_examiner_all | 1.941 | 1.919 | 1.856 | 1.899 | 1.899 | 1.98 | 1.894 |
| | [0.205]** | [0.203]** | [0.198]** | [0.202]** | [0.200]** | [0.216]** | [0.201]** |
| dif_technology_4_all | 0.701 | 0.696 | 0.698 | 0.697 | 0.705 | 0.694 | 0.705 |
| | [0.045]** | [0.044]** | [0.044]** | [0.046]** | [0.043]** | [0.043]** | [0.044]** |
| years | 0.916 | 0.916 | 0.917 | 0.917 | 0.92 | 0.92 | 0.921 |
| | [0.005]** | [0.005]** | [0.005]** | [0.005]** | [0.005]** | [0.005]** | [0.005]** |
| distance | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | [0.000]* | [0.000]* | [0.000]* | [0.000]* | [0.000]* | [0.000]** | [0.000] |
| Observations | 16089 | 16089 | 16089 | 16089 | 12977 | 14809 | 11697 |
| Number of group(citing) | 1,456 | 1,456 | 1,456 | 1,456 | 876 | 1,342 | 762 |
| Min cited per citing | 1 | 1 | 1 | 1 | 1 | 1 | 2 |
| Avg cited per citing | 11.05 | 11.05 | 11.05 | 11.05 | 14.814 | 11.035 | 15.35 |
| Max cited per citing | 234 | 234 | 234 | 234 | 234 | 234 | 234 |
| Wald Chi 2 | 422.56 | 431.836 | 448.923 | 420.883 | 397.437 | 430.397 | 392.806 |
| Degrees of freedom | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| Rho | 0.62 | 0.618 | 0.62 | 0.616 | 0.415 | 0.583 | 0.347 |
| Chi bar2 | 7,258.57 | 7,255.77 | 7,276.21 | 7,264.78 | 2,918.24 | 6,455.35 | 2,505.46 |

Standard errors in brackets
* significant at 5%; ** significant at 1%

Table 9
Results of logit regressions for different geographci units

Dependent variable is equal to 1 if citation comes from examiner, 0 otherwise
Results assume random effects structure for error term
Models 1, 2 include all citing patents
Model 3, 4, 5, and 6 include only those dyads where both citing and cited patents have all inventors in the US

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| distance | 1 | | | | | |
| | [0.000]* | | | | | |
| dif_country_all | | 1.254 | | | | |
| | | [0.091]** | | | | |
| dif_state_all | | | 1.355 | | | |
| | | | [0.143]** | | | |
| dif_county_all | | | | 1.55 | | |
| | | | | [0.214]** | | |
| dif_ea_all | | | | | 1.524 | |
| | | | | | [0.176]** | |
| dif_city_all | | | | | | 1.27 |
| | | | | | | [0.213] |
| dif_company | 1.785 | 1.78 | 1.453 | 1.388 | 1.355 | 1.56 |
| | [0.185]** | [0.184]** | [0.208]** | [0.204]* | [0.199]* | [0.219]** |
| same_inventor | 1.245 | 1.255 | 1.275 | 1.413 | 1.352 | 1.286 |
| | [0.159] | [0.161] | [0.206] | [0.244]* | [0.222] | [0.237] |
| same_examiner_all | 1.941 | 1.908 | 2.363 | 2.363 | 2.365 | 2.392 |
| | [0.205]** | [0.200]** | [0.354]** | [0.358]** | [0.356]** | [0.360]** |
| years | 0.916 | 0.916 | 0.904 | 0.905 | 0.904 | 0.905 |
| | [0.005]** | [0.005]** | [0.007]** | [0.007]** | [0.007]** | [0.007]** |
| dif_technology_4_all | 0.701 | 0.696 | 0.677 | 0.68 | 0.674 | 0.685 |
| | [0.045]** | [0.045]** | [0.064]** | [0.065]** | [0.065]** | [0.064]** |
| Observations | 16089 | 16095 | 7632 | 7632 | 7632 | 7632 |
| Number of group(citing) | 1,456 | 1,456 | 715 | 715 | 715 | 715 |
| Min cited per citing | 1 | 1 | 1 | 1 | 1 | 1 |
| Avg cited per citing | 11.05 | 11.054 | 10.674 | 10.674 | 10.674 | 10.674 |
| Max cited per citing | 234 | 234 | 129 | 129 | 129 | 129 |
| Wald Chi 2 | 422.562 | 420.481 | 253.133 | 252.349 | 256.307 | 248.088 |
| Degrees of freedom | 6 | 6 | 6 | 6 | 6 | 6 |
| Rho | 0.617 | 0.614 | 0.605 | 0.605 | 0.606 | 0.605 |
| Chi bar2 | 7258.568 | 7058.75 | 2517.611 | 2524.263 | 2521.894 | 2530.132 |

Standard errors in brackets
* significant at 5%; ** significant at 1%

Table 10

Comparisons of highly-cited patents according to inventors and examiners

| Is in top 1% according to inventors? | | Is in top 1% according to examiners? No | Yes | Total |
|---|---|---|---|---|
| | No | 137 (12%) | 278 (24%) | 415 (35%) |
| | Yes | 556 (47%) | 204 (17%) | 760 (65%) |
| | Total | 693 (59%) | 482 (41%) | 1,175 |

Top 1% for patents granted in 1998 based on forward citations received from Jan 2001 to August 2003

Correlation between forward citations received from Jan 2001 to August 2003 and those received from grant date to August 2003 is 0.915

Table 11

Correlations of inventor and examiner forward citations by grant year of all cited patents

Correlation between examiner and inventor citations:

| | By # of forward citations | By decile of # forward citations |
|---|---|---|
| 1998 | 0.381 | 0.119 |
| 1999 | 0.316 | 0.107 |
| 2000 | 0.223 | 0.081 |
| 2001 | 0.083 | 0.012 |
| 2002 | -0.112 | -0.098 |
| 2003 | -0.714 | -0.343 |

Forward citations received from Jan 2001 to August 2003

All correlations are siginificant at 1%

Figure 1

Distributional Assumptions for aggregate, inventor and examiner citations

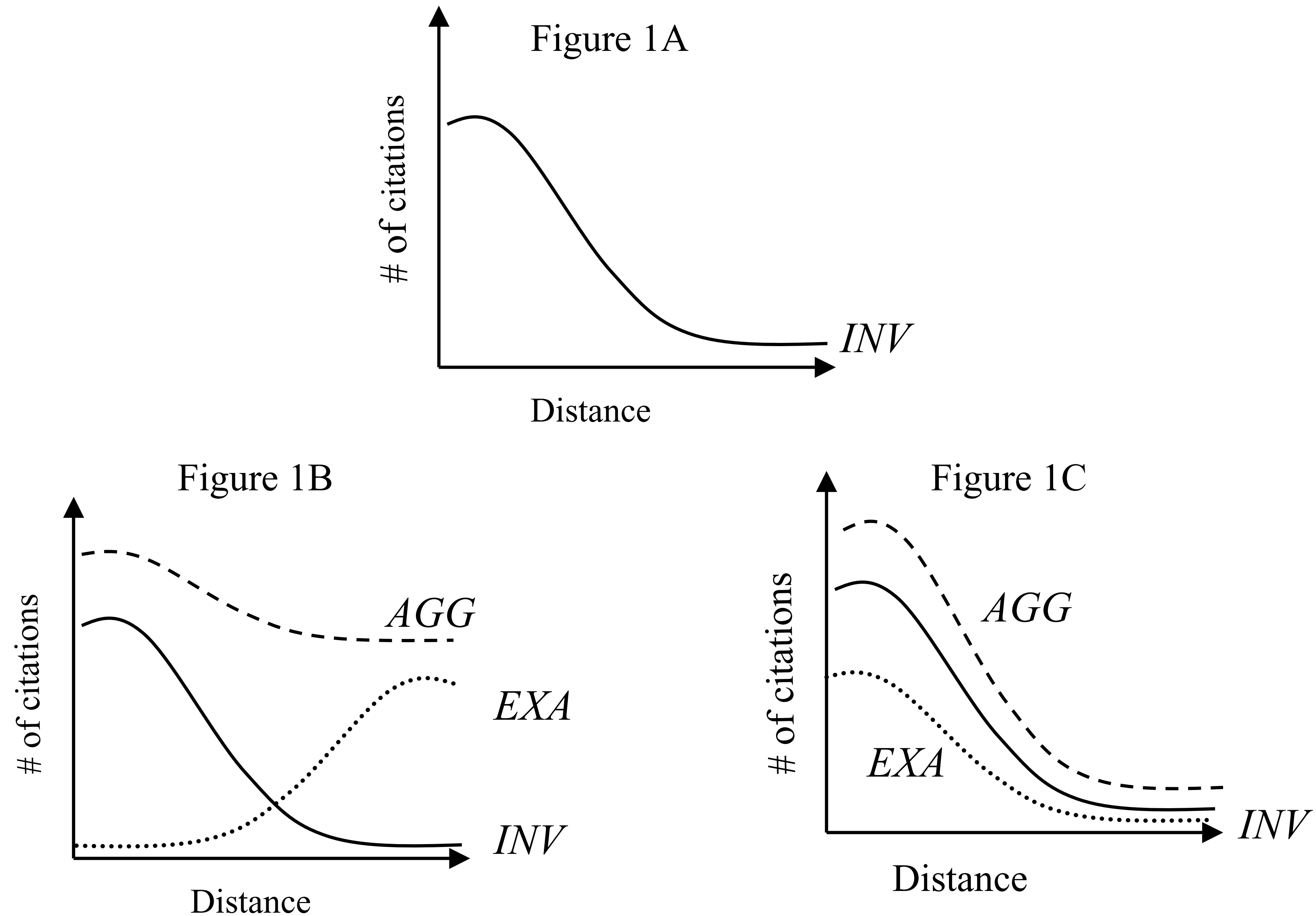

Figure 2

Kernel density graph for distance in miles

Examiner vs. Inventor citations

Citing and cited patents in Continental USA

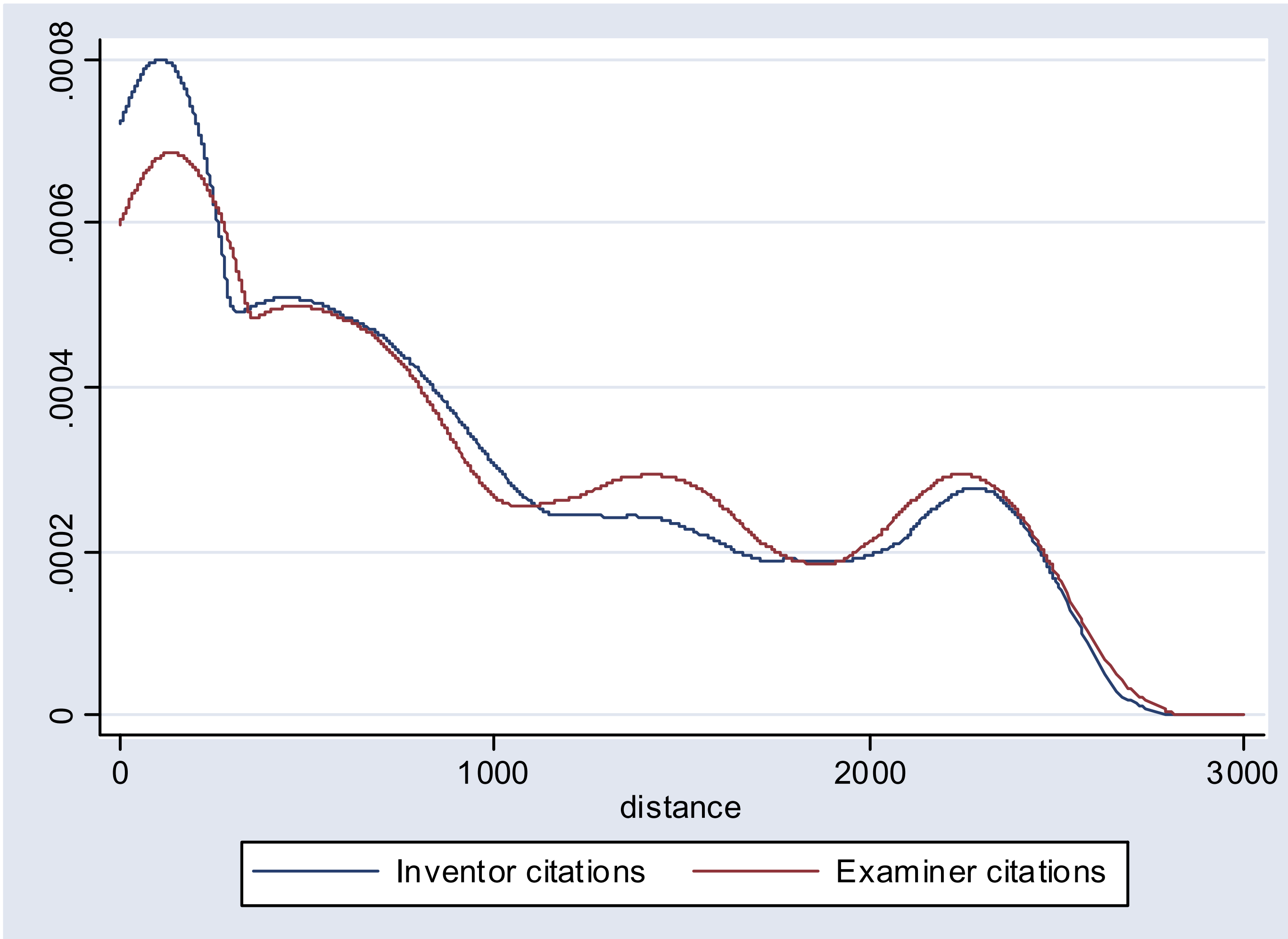

Figure 3

K-density graph for distance in miles
Examiner vs. Inventor citations
Citing patents in Continental USA, foreign cited patent

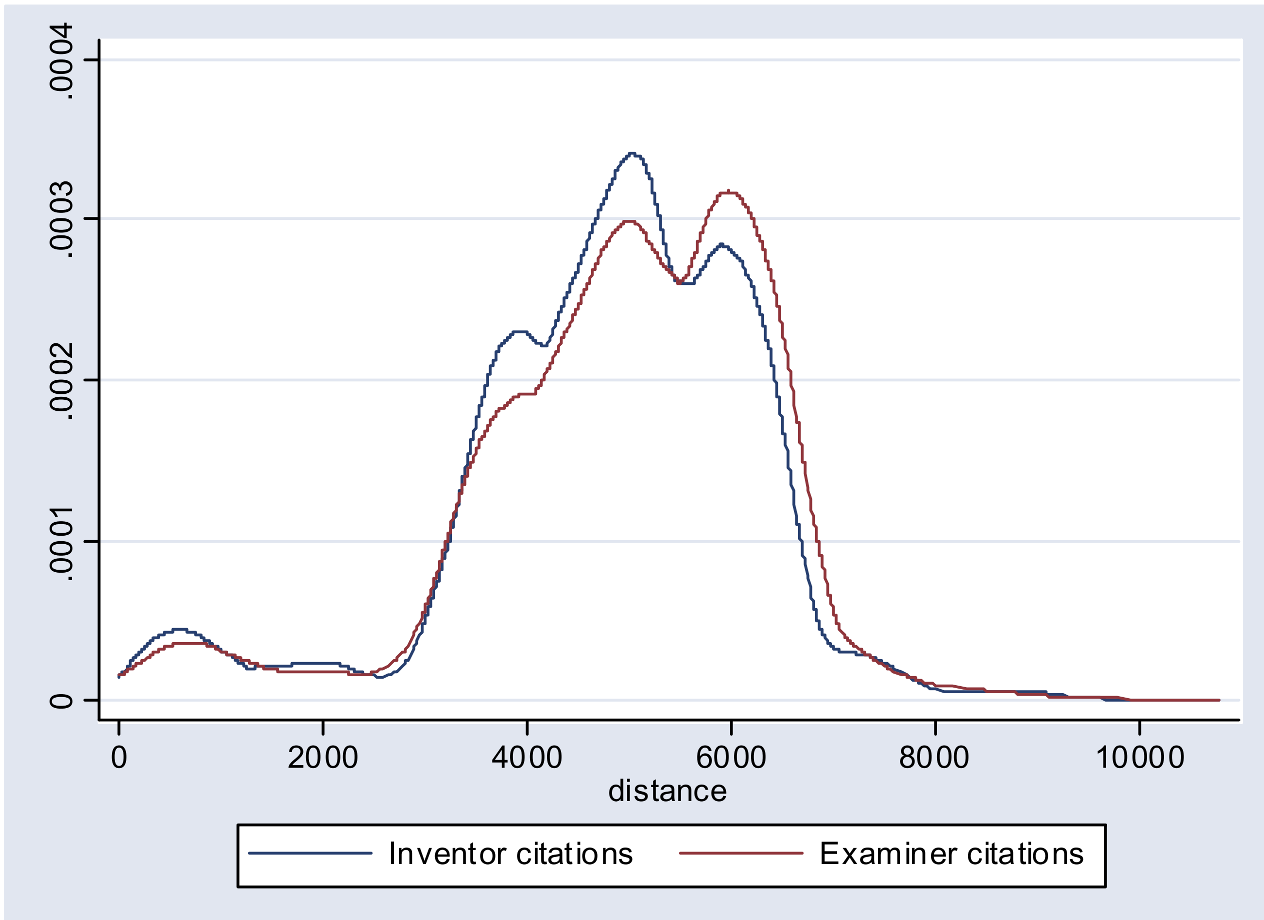

Figure 4
K-density graph for Time in years
Examiner vs. Inventor citations