Routledge
Taylor & Francis Group

Economics of Innovation and New Technology

ISSN: 1043-8599 (Print) 1476-8364 (Online) Journal homepage: https://www.tandfonline.com/loi/gein20

Embracing invention similarity for the measurement of vertically overlapping claims

Charles A. W. deGrazia, Jesse P. Frumkin & Nicholas A. Pairolero

ECONOMICS OF INNOVATION AND NEW TECHNOLOGY
2020, VOL. 29, NO. 2, 113–146
https://doi.org/10.1080/10438599.2019.1593035

Embracing invention similarity for the measurement of vertically overlapping claims

Charles A. W. deGrazia[a,b], Jesse P. Frumkin [a] and Nicholas A. Pairolero [a]

[a]United States Patent & Trademark Office, Alexandria, VA, USA; [b]Department of Economics, Royal Holloway, University of London, Egham, Surrey, UK

ABSTRACT

Clear and well-defined patent rights can incentivize innovation by granting monopoly rights to the inventor for a limited period of time in exchange for public disclosure of the invention. However, with cumulative innovation, when a product draws from intellectual property held across multiple firms (including fragmented intellectual property or patent thickets), contracting failures may lead to suboptimal economic outcomes. However, an alternative theory, developed by a variety of scholars, contends that patent thickets have a more ambiguous effect. Researchers have developed several measures to gauge the extent and impact of cumulative innovation and the various channels of patent thickets. This paper contends that mis-measurement may contribute to the incoherence and overall lack of consensus within the patent thickets literature. Specifically, the literature is missing a precise measure of vertically overlapping claims. We propose a new measure of vertically overlapping claims that incorporates invention similarity to more precisely identify inventive overlap. The measure defined in this paper will enable more accurate measurement, and allow for novel economic research on cumulative innovation, fragmentation in intellectual property, and patent thickets within and across all patent jurisdictions.

ARTICLE HISTORY
Received 23 October 2018
Accepted 22 February 2019

KEYWORDS
Innovation; intellectual property; cumulative innovation; invention similarity; vertically overlapping claims

JEL CLASSIFICATION
O30; O31; O32; O34; O38

1. Introduction

Clear and well-defined patent rights can incentivize innovation by providing the inventor monopoly rights over the invention for a limited period of time in exchange for public disclosure. However, with cumulative innovation, when a product depends upon intellectual property held across multiple firms, contracting failures may lead to sub-optimal economic outcomes. Shapiro (2000) contends this fragmentation of patent rights, called a patent thicket, has negative economic consequences. Patent thicket is defined by Shapiro (2000) as, ‘a dense web of overlapping intellectual property rights that a company must hack its way through in order to actually commercialize new technology’. Theoretically, the Cournot complements problem shows that when firms must license components from multiple patentees, individual firm license costs are increasing in the number of entities required for contracting. This reduces the return on research and development, impeding entry and resulting in a suboptimal level of follow-on innovation. Additionally, patent thickets may also increase the potential for hold-up, particularly when vague ex-ante licensing conditions enable patentees to demand royalties or threaten injunction after follow-on product development.

However, an alternative theory, developed by a variety of scholars, contends that patent thickets have a more ambiguous effect. For example, Galasso and Schankerman (2010) show that

CONTACT Nicholas A. Pairolero nicholas.a.pairolero@gmail.com United States Patent & Trademark Office, 600 Dulany Street, Alexandria, VA, USA

fragmentation may ease licensing negotiations and expedite court settlements. In particular, fragmented patent rights reduce the value at stake in each individual license negotiation, easing bargaining tensions and facilitating deals. Additionally, Spulber (2017) theoretically shows that the complements problem is less severe with bargaining, rather than the post-prices assumption of Cournot. Therefore, in certain cases, there is a priori ambiguity regarding the impact of fragmentation and patent thickets that must be resolved empirically.

In an appraisal of this debate, Egan and Teece (2015) contend that there is 'growing confusion' on patent thickets in the economics literature. In particular, the term patent thicket covers several economic issues, and definitions are both becoming more complex over time, and inconsistently used across researchers. Egan and Teece (2015) argue that the underlying economic issues labeled as patent thickets can be classified into several categories. Beyond this overall confusion, we contend that imprecise measurement of the economic channels of patent thickets may contribute to the overall incoherence of the literature, and suggest that refinements could allow researchers to more readily exploit variation within a given channel of patent thickets.

Our paper focuses on a particular channel of patent thickets: vertically overlapping claims (also referred to as overlapping patents), the measurement of which is missing from the literature. According to Egan and Teece (2015), overlapping claims can be divided into two groups, vertical and horizontal overlap, the former of which is the focus of our paper. In a cumulative innovation setting (vertical overlap), a follow-on patent overlaps with all related patents that precede it and, therefore, is susceptible to the complements problem. Horizontal claim overlap, or patents that claim overlapping inventive space through imperfectly defined intellectual property rights, results in 'wasteful duplication of resources' as well as the susceptibility to the complements problem (Egan and Teece 2015).

In this paper, we propose and validate a new measure of vertically overlapping claims (capturing cumulative inventions rather than imperfectly defined patent rights) based on the similarity of patent claim text, or what we call invention similarity. Our measure could be combined with a fragmentation-style and transaction cost indexes to formally test the complements problem and the various theoretical bargaining variants (Spulber 2017) in the literature. Leveraging invention similarity and the structure of patent citations, our overlapping claims measure uses a technique that is an improvement over those used to measure inventive relationships in the patent thicket measurement literature. First, we apply natural language processing based on word frequencies to quantify the invention similarity of claims between citing and cited patents (Younge and Kuhn 2016; Kuhn, Younge, and Marco 2018; Arts, Cassiman, and Gomez 2018), better capturing the dispersion of invention similarity across citations and minimizing noise from less relevant citations. Additionally, because patent claims precisely define the content and scope of the invention, claim text similarity captures inventive overlap more precisely than full patent text similarity. Second, our measure emphasizes invention similarity without the limitation of only using blocking patents used in rejections at the patent office.[1] We find significant overlap in the distribution of invention similarity between blocking and non-blocking citations, suggesting that the patent thicket channel measures derived from only blocking patents fail to capture all overlapping patent rights. Lastly, our measure is computable for all patent systems, allowing scholars to address cumulative innovation-related research questions consistently within and across all jurisdictions.

The paper proceeds in the following way. First, we describe background on the economics of cumulative innovation and categorize the measurement literature used to identify the channels of patent thickets, and show that a precise measurement of the vertically overlapping claims channel of patent thickets is generally missing from the literature. Second, we define and describe our vertically overlapping claims measure. We then validate our measure using methods previously applied in the patent thicket measurement literature (von Graevenitz, Wagner, and Harhoff 2011; Fischer and Ringler 2014). Results show that our overlapping claims measure is consistently higher in complex versus discrete technologies.[2] This difference persists after normalizing for patent volume and average citation counts, indicating that invention similarity of claims conveys additional information for distinguishing between complex and discrete technologies (Cohen, Nelson, and Walsh 2000).

We show our measure to be positively correlated with patent examiner search intensity, application pendency and USPTO patent examination complexity factors. Since patent examiners are given and/or expected to require more time to search and prosecute complex technologies, the overlapping claims measure should be positively correlated with these variables. Additionally, our results show that the overlapping claims measure is not positively correlated with whether or not a US patent application receives a rejection on the first office action[3] because of obviousness or lack of novelty. This result is consistent with our premise that reliance on blocking patents for measuring local vertical inventive overlap is overly restrictive. Using invention similarity based on patent claim text retains the information contained in each citation and offers a more precise method for measuring vertically overlapping claims.

Finally, to test whether our overlapping claims measure captures vertically overlapping claims, while excluding overlap derived from improperly defined patent rights, we correlate our measure with Patent Trial and Appeals Board (PTAB) institution outcomes. Patents are only instituted at PTAB if there 'is a reasonable likelihood that the petitioner would prevail with respect to at least 1 of the claims challenged in the petition'.[4] If the measure excludes overlap from improperly granted patent rights, then we expect our overlapping claims measure to be uncorrelated with these decisions at PTAB.

2. Background and literature review

2.1. Cumulative innovation, patents and patent thickets

In the absence of an appropriation mechanism, innovation may suffer from the public goods problem. Knowledge is both non-rival (one person's use of knowledge does not preclude someone else use of the same knowledge) and non-excludable (in the absence of any appropriation mechanism, it is challenging to exclude someone else from using the knowledge). For these reasons, the generation of knowledge creates positive externalities, and therefore the equilibrium quantity of knowledge is not socially optimal. From a practical perspective, firms may be reluctant to invest in costly innovation if once realized, other firms are able to easily assimilate the innovation (because of non-rivalry and non-excludability).

Patents provide the right to exclude others from using an invention, and therefore are one potential solution to the public goods problem of innovation. Increasing the incentive to innovate is not costless however. The right to exclude potentially allows the innovator to charge monopoly rents for use of the innovation (Nordhaus 1969), creating a deadweight loss. Monopoly rents and innovation incentives are a static trade-off, and when inventions are cumulative, patents introduce additional dynamic considerations. Cumulative innovation occurs when new inventions utilize ideas from earlier inventions. In this case, optimal patent policy must consider additional trade-offs beyond monopoly rights and the incentives to innovate. In particular, optimal patent policy must ensure that the rewards to innovation within firms is sufficient to cover their costs along the sequential chain of innovation (Scotchmer 2004).

There are three types of cumulative innovation: (1) basic/applied research; (2) research tools; and (3) quality ladders (Scotchmer 2004). In the basic and applied research case, the original innovator develops basic research that leads to additional applied research. Green and Scotchmer (1995) develop a licensing model for this type of cumulative innovation with two main results. First, the nature of licensing agreements matter. For example, if *ex ante* licensing agreements are unavailable (or not binding), then the applied innovator may not invest since the basic innovator will charge a higher licensing fee ex-post, effectively holding up the applied innovator. Second, ex-ante licensing agreements resolve this hold-up problem, but even if the overall private value of the basic/applied innovations is positive, ex-ante licensing terms may be insufficient to cover the basic innovator's cost. In this case, optimal patent policy is to induce investment by providing a longer patent term to increase the profit allocated to the basic innovator.

The research tools case of cumulative innovation occurs when a single invention relies upon several complementary input innovations. One concern is Cournot's complements problem, which arises when a firm must purchase inputs from several monopolists (Shapiro 2000). In the case of patents, the problem materializes when a firm must license components from multiple patent inputs. Under Cournot's assumptions (licensors post prices), licensing royalties are increasing in the number of input firms. This reduces research and development, deters entry by follow-on innovators, and hinders the development of products. Furthermore, private and consumer welfare is lower than if the patent-holders coordinated their licensing agreements. Further research shows, however, that the predictions of the complements problem are highly dependent on the theoretical bargaining assumptions. For example, Spulber (2017) modifies the bargaining assumptions of the complements problem and finds results that are far less severe.

The final type of cumulative innovation is the quality ladder. In this case, each successive innovation serves the same purpose; however, successive innovations improve on quality. The potential problem is that the effective patent term is far less than the actual patent term since successive innovation competes with previous innovation. Because of this competition, innovators may not be able to cover the costs of invention, and therefore are less likely to invest. O'donoghue and Scotchmer (1998) develop a theoretical model of quality ladders and show that patent breadth may be used to encourage an optimal amount of investment. The idea is that innovators are issued broader patents, and therefore are able to extract licensing royalties from higher quality successive innovation. This encourages innovation by increasing the returns to successful invention.

In addition to the theoretical literature on cumulative innovation, a second and related literature has developed broadly classified as 'patent thickets'. Despite the popularity of Shapiro (2000) definition, defined in the introduction, Egan and Teece (2015) find that at least 164 papers in the literature define patent thicket, and these definitions are both highly varied, and inconsistently used across and within authors. In particular, Egan and Teece (2015) find several economic mechanisms (or channels) of patent thickets contained in the literature; including, saturated invention spaces, diversely held complementary inputs, overlapping patents, gaming the patent system, transaction costs and probabilistic patents. The common characteristic of these mechanisms is that each has the potential to cause economic inefficiency.

Several of the economic channels of patent thickets relate directly to cumulative innovation. The diversely held complements problem is the Cournot complements problem, relevant to the research tools form of cumulative innovation. Overlapping patents may be both 'vertical' or 'horizontal'. Vertical overlapping claims occur directly in cumulative innovation when a successive innovation utilizes the previous invention. 'Horizontally' overlapping patents are groups of patents that might overlap because of improperly defined intellectual property rights. Horizontally overlapping patents may be complementary, substitutes, or cumulative in nature. Several of the economic channels in this literature represent potential strategic responses to potential inefficiencies in the cumulative innovation setting. For example, firms may patent strategically to increase their bargaining position in licensing negotiations and litigation (called defensive patenting).

Because of the complexity of this literature, and perhaps incoherence (Egan and Teece 2015), careful examination and further development of empirical research are crucial. Additionally, this is especially important since theoretical predictions in the literature are highly dependent on assumptions. For example, as noted earlier, both the complements problem in research tool cumulative innovation, and inefficiencies in basic/applied research cumulative innovation rely heavily on the structure of licensing negotiations. For these reasons, along with the importance of understanding cumulative innovation and patents, the next section surveys the empirical measurement literature on these issues. Our purpose is to show that vertically overlapping claims (a crucial component of cumulative innovation) is improperly measured in the empirical literature.

2.2. Measurement literature

This paper contends that the set of measurements used to identify the channels of patent thickets are incomplete, and this may be contribute to the overall incoherence of the patent thicket literature (Egan and Teece 2015). In this section, we classify several of the popular measures within their appropriate patent thicket channel (or channels), and show that a precise measure for vertically overlapping claims is missing from the literature.

Ziedonis (2004) develop the fragmentation index, which is particularly useful in the patent thicket literature for identifying the conditions necessary for the complements problem. The measure is defined in the following way

$$Fragmentation_{kat} = 1 - \sum_{j=1}^{n} s_{kjat}^{2}$$

for firm k, technology a, and time t. The summation is over the set of firms n that k cites within technology a where s_{kjat} is the share of k's citations to j within technology a in period t. Generally, as ownership rights to a firm's complementary patents become more dispersed, the fragmentation index will increase (Ziedonis 2004). Increased fragmentation leaves firms more vulnerable to the complements problem. The measure has been used in a variety of empirical studies on the effects of market fragmentation, although the measure does not capture overlapping claims. Inventive similarity is important for the complements problem, since fragmentation is only relevant in the event of licensing.

In von Graevenitz, Wagner, and Harhoff (2011), the authors focus on triads of firms that each hold patents previously used to block the patent applications of the other firms (hereinafter the 'VG' measure). Their measure does incorporate elements of overlapping claims, but also measures the transaction costs channel of patent thickets. Because of this, researchers using this method are unable to disentangle the overlapping claims and transaction cost channels of patent thickets. The VG measure captures inventive overlap by using X and Y references from European Patent Office data. An X reference calls into question the novelty or inventive step of an application. A Y reference does the same regarding the inventive step, but in conjunction with other documents. A patent i is said to block patent application j if i is used in an X or Y citation during European Patent Office examination of application j. A triad is defined to be three firms, each of which holds a blocking citation on the others. The VG measure at the firm level is the sum of all triads in which the firm is a member. The measure at the technology level is the sum of all triads where all the blocking patents are in the same technology.

Overall, the VG measure reasonably focuses on blocking patents; however, neglects all similar citation pairs that are not cited in a rejection by the patent office. For example, consider a patent application z and the following two scenarios. In the first scenario, the applicant submits application z and receives a rejection based upon a prior patent. The applicant then modifies the claims of the patent application to z'. The examiner then allows the patent application. In the second scenario, the applicant preemptively recognizes the prior patent and initially submits claims z'. The examiner allows the patent application with no rejections based on the prior patent. The relationship between the resulting patent and the prior patent is the same in both scenarios and a license may need to be extended in each case; however, the VG measure only captures the citation pair in the first scenario. Additionally, inventions may share components of other patented inventions but still receive a patent. For example, an invention may use several components from other patented inventions, but be a non-obvious combination and therefore receive a patent without unnecessary rejections on the record involving the other patents.[5] Finally, the examiner may have ample citations to reject the claim and therefore might omit some or have insufficient time for a complete search (Frakes and Wasserman 2017, Lei and Wright 2017). Therefore, the use of blocking citations in the VG measure may be too restrictive, at least for the measurement of overlapping claims.

Gaessler, Harhoff, and Sorg (2017) uses a measure of patent fencing containing a count of semantically similar patents (similarity to the focal patent greater than the 95th percentile of patent similarities) that are contained in the focal patent holder's portfolio. This measure is primarily focused on the gaming the patent system channel of patent thickets; however, incorporates elements of the overlapping claims and saturation channels of patent thickets. Again, as with the VG measure, researchers are unable to distinguish between these channels with the Gaessler, Harhoff, and Sorg (2017) measure alone. Although this measure uses a version of technological similarity (captured by a patent's title, abstract, claims, and description), this measure was generated independently of our own measure (and vice versa).[6] Finally, Fischer and Ringler (2014) extend the VG measure by broadening the attention beyond blocking patents; however, by using all patent citations to construct the triads, the measure does not capture all inventive overlap. Furthermore, as with the VG measure, the FR measure confounds overlapping claims with the transaction costs channel of patent thickets.

To summarize, the measurement literature on the channels of patent thickets does not contain precise measures of vertical and horizontal overlapping claims, and therefore is currently incomplete. Researchers using these measures are therefore unable to disentangle the various channels of patent thickets. Furthermore, we argue in this paper that pure citation and blocking citation-based measures omit information contained in the citations regarding overlapping patent rights. The measure developed in this paper uses invention similarity to capture vertically overlapping claims. Vertical overlap, as opposed to horizontal overlap, is important for identifying the complements problem (diversely held complementary inputs). Our measure could be combined with a fragmentation-style and transaction cost indexes to formally test the complements problem and the various theoretical bargaining variants (Spulber 2017) in the literature.

We clarify this with an example from the literature, namely the research question and empirical analysis in Cockburn, MacGarvie, and Mueller (2010). Although we describe this paper in detail, we do so simply to clarity of our argument. Cockburn, MacGarvie, and Mueller (2010) is an early and influential paper on patent thickets, and greatly contributes to our overall understanding of this topic. To test the complements problem, Cockburn, MacGarvie, and Mueller (2010) empirically test the following hypothesis: 'The more fragmented the ownership of patents that read on a firm's product, the higher are the licensing costs associated with commercializing that product' (Cockburn, MacGarvie, and Mueller 2010).

If the complements problem exists in the studied setting, then Cockburn, MacGarvie, and Mueller (2010) expect to find empirical evidence for this hypothesis. Empirically, the authors analyze the relationship between fragmentation and both the probability of in-licensing, and licensing expenditure per unit sales. The fundamental problem here is that the degree of backward inventive overlap is not included in the empirical specification, and therefore leads to omitted variable bias. In particular, if the degree of inventive overlap is positively correlated with both fragmentation and licensing costs, then the impact of fragmentation will be biased upward. Although the authors control for industry, this does not eliminate unobserved variation in backward inventive overlap at the firm level over time. Therefore, Cockburn, MacGarvie, and Mueller (2010) may in fact overestimate the impact of fragmentation on licensing costs. Since other forms of cumulative innovation (basic and applied research, and quality ladders) may also lead to economic inefficiencies, controlling for cumulative innovation is crucial for understanding the complements problem (an inefficiency exclusive to research tools cumulative innovation).

2.3. Institutional background on patents

Although the focus of our paper is on measuring vertically overlapping claims, we draw on knowledge of patent law and patent examination procedure to develop and validate the measure. For this reason, in this section we provide a brief discussion of the institutional background on patents. We only provide the details relevant for our analysis; more thorough discussions are available in the literature (Marco et al. 2017; Graham, Marco, and Miller 2018; Lu, Myers, and Beliveau 2017).

Patents contain a variety of information about the invention disclosure including, but not limited to, a precise definition of the invention, patent technology classification, and information on the inventors. The claims of the patent precisely define the content and scope of the invention, while the specification describes the invention, but may also discuss the background of the invention and define the terms and structures recited in the claims. A patent also contains a list of citations to earlier patents, patent application publications, and non-patent literature. Patent citations may be added by either the applicant, the examiner, or third parties throughout the course of patent prosecution.

Patent examiners are organized by technology at the United States Patent and Trademark Office (USPTO). At the highest level of technology aggregation, examiners are placed within a technology center. Within technology centers, examiners are placed within work groups, and then more narrowly art units. Within an art unit, examiners are responsible for certain technology class/sub-class combinations of the United States Patent Classification (USPC).

Patents are examined under a set of statutory criteria including, but not limited to, novelty (35 USC § 102), non-obviousness (35 USC § 103), subject matter eligibility (35 USC § 101) and various clarity and enablement issues (35 USC § 112).[7] Rejections for lack of novelty and obviousness are often called 'prior art' rejections since they cite evidence of earlier disclosures (called 'prior art') including public documents and other forms of evidence. Generally, an examiner initially assesses the validity of a patent application and issues a first-action decision. Two common first-action decisions for new applications are non-final rejections and allowances. Included with the first-action decision is a list of references cited, documenting the prior art that the examiner considered when deciding the first-action. An applicant receiving a non-final rejection has the opportunity to revise the patent application for further consideration. The patent application is terminally disposed by either an abandonment or an allowance. At the USPTO, examiners are unable to terminally dispose of the application by rejecting the application. The amount of time between application filing and the terminal disposal is called pendency.

USPTO patent examiners are guided by a system of incentives (Marco et al. 2017). Examiners are provided different amounts of time to evaluate patent applications based upon their seniority and the underlying complexity of the technology being examined. Complexity factors are scalars that reflect the underlying complexity level of a particular U.S. Patent Classification (USPC) class-subclass combination. A higher complexity factor indicates that an examiner should be given more time. Seniority factors are scalars that vary based upon the seniority of the examiner (determined by the Federal General Schedule, with some exceptions).

There are many types of relationships between patent applications filed at the USPTO and foreign patent applications. For example, a USPTO application may claim priority to a foreign patent application for the purpose of obtaining an earlier filing date. Within the USPTO, patent applications can be a continuation, continuations-in-part, or divisionals of an earlier patent application. The continuation requires the new patent application to have the same specification, while the continuation-in-part allows the applicant to add information to the specification of the preceding application. The relationships between patent applications at the USPTO are more formally described in Graham, Marco, and Miller (2018).

The validity of patents may be re-assessed post-grant at the USPTO's Patent Trial and Appeals Board (PTAB). The re-assessment of validity at PTAB follows a sequential procedure. First, patents are petitioned. Petition requires a detailed description of the petioners' issues with the granted patent. Upon reviewing the petition, if PTAB finds 'reasonable likelihood that the petitioner would prevail with respect to at least 1 of the claims challenged in the patent', then the Board will institute the patent for further review. Following institution, the Board makes a final written decision on the petition.

Finally, as a result of the patent examination process, the patent claims precisely define the content and scope of the claimed invention, so that similarity of claims more precisely measures inventive overlap compared to using other portions of the disclosure. For example, two patents may overlap in their detailed descriptions by sharing a similar background, motivation, general-purpose elements, and boilerplate legal disclaimers but assert completely different inventions as formally defined in the claims. Hence, for our overlapping claims measure, we will focus on the similarity of the patent claims rather than other parts of the patent disclosure.

3. Methodology

3.1. Overlapping claims measure

The core building blocks for our overlapping claims measure are patent citations with each citation weighted by the invention similarity of the patent claims. A triad is defined to be three distinctly-owned patents i, j, and k such that j cites i, and k cites i and j. See Figure 1 for an example. Each citation carries a weight determined by the similarity of the patent claims in the citing and cited patent. The citation weights are combined to form an overall triad weight. The sum of the weighted triads containing distinct patent owners emphasizes overlapping claims across firms, although the measure may be adjusted to include self-citations if relevant for particular applications.[8] Weighting the triads by patent claim similarity sharpens the measure by emphasizing technologically similar citations. The first definition formalizes the notion of a triad.

Definition 3.1: Let $\{i, j, k\}$ be patents, each with a different patent owner. Suppose that k cites j and i, and j cites i. In this case, patents $\{i, j, k\}$ are said to form a triad. i is said to be associated with triad $\{s, q, t\}$ if $i \in \{s, q, t\}$ where the set $\{s, q, t\}$ forms a triad.

The next definition formalizes the notion of a triad weighting function.

Definition 3.2: Let $S = [0, 1]^3$. A triad weighting function f is a mapping from S to R^+ such that for each $(w_1, w_2, w_3) \in S$

$$\frac{\partial f}{\partial w_i} \geq 0$$

for all $i \in \{1, 2, 3\}$.

A relatively simple triad weighting function adopted for the empirical section of this paper is

$$f(w_{ij}, w_{ik}, w_{jk}) = (w_{ij} + w_{ik} + w_{jk})$$

The linearity of the weighting function imposes a very particular assumption about the way the measure captures information from invention similarity. For example, rather than applying linear

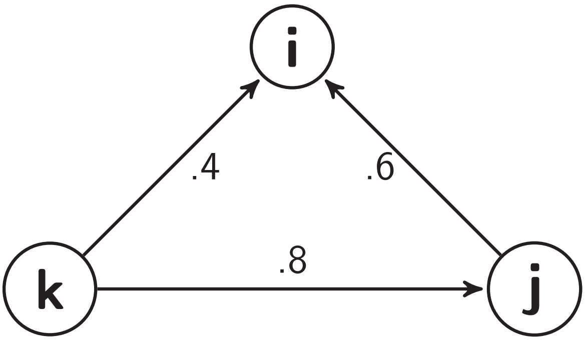

Figure 1. Example of a triad.

weights, quadratic weights could be used. This would provide relatively more weight to more similar citations, and further marginalize less similar citations. Alternatively, one could provide a threshold such that a citation similarity is only used if it is above the threshold. We use the linear weighting function in the main empirical section of this paper to abstract from these complexities, and show that the measure satisfies a variety of validation tests with this assumption. Despite this, researchers may choose to modify the weighting function to emphasize information differently in applications. In the appendix, we further explain the theoretical differences between these weighting functions (linear, truncated and quadratic), and examine the consistency of our empirical results across these functions.

We define the local overlapping claims measure for a given patent as the sum of similarity weighted triads for which the patent is a member.

Definition 3.3: Suppose that i is a patent and f is a triad weighting function. Let T_i be the set of all patent pairs $\{j,k\}$ for which $\{i,j,k\}$ form a triad. Define the local overlapping claims measure as

$$\sum_{\{j,k\}\in T_i} f(w_{ij}, w_{ik}, w_{jk})$$

This measure is local since it relies exclusively on all triads associated with i.

Next, we define the overlapping claims measure at the global technology level as the weighted triads in the technology summed and normalized.

Definition 3.4: Let $G_t = \{(i,j,k) \mid (i,j,k) \text{ form a triad}\}$ be the set of triads contained in technology t. The global overlapping claims measure for technology t is defined as

$$\sum_{(i,j,k)\in G_t} f(w_{ij}, w_{ik}, w_{jk})/n$$

where n is some normalizing constant relevant to technology t.

We leave the normalization procedure general since it may depend on the situation. In later sections of this paper, we use the number of patents and average number of citations to normalize the global overlapping claims measure. Generally, the global overlapping claims measure should be normalized by the number of patents in the technology space since some categories might simply be larger than others. However, by additionally normalizing by the average number of citations, we are able to isolate the impact of invention similarity for measuring overlapping claims.

Finally, the global overlapping claims measure may be used to measure the overall inventive overlap of a patent portfolio, or a set of patents for a particular inventor, by changing G_t to the relevant set of patents (either firm portfolio, or set of patents by a particular inventor).

Before formalizing invention similarity, it is useful to carefully describe how the structure of our overlapping claims measure maps to identifying the complements problem. At the time of grant, the local overlapping claims measure captures claim overlap between an invention and the prior inventions upon which it relies. However, as the local invention space evolves, the overlapping claims measure begins to capture claim overlap in forward inventions. Therefore, the local overlapping claims measure represents aspects of the complements problem in the local inventive area, which may translate into the direct risk to the patent owner for utilizing the invention in a specific product. In particular, it is possible that the patent owner may want to utilize inventive aspects of the inventions contained in the forward citations. In this case, capturing inventive overlap in forward citations may more accurately reflect the complementary input risk for a particular invention. Despite this, a straightforward modification could limit the local measure to only backward citations from patent i over time.

3.2. Invention similarity

Citations are listed on the face of a patent for a variety of reasons; for example, if the cited documents provide the definition of a term-of-art, supply clarification that is somewhat pertinent to the claims, describe an element used in the invention, or were extraneously noted by the applicant on their Information Disclosure Statement (IDS) but were still reviewed by the examiner.[10] Hence, the content and importance of cited patents will vary in their degree of similarity to the claimed invention. To account for this variation, we use natural language processing techniques based on word frequency to quantify the invention similarity between citing and cited patents.

Two textual documents that are semantically similar will often have similar word frequency distributions. Therefore, we algorithmically compare the word-frequencies (i.e. the 'bag-of-words' or 'multisets') from patent claims text to compute the invention similarity between two patent documents. To compute the word frequencies, we first pre-process the patent claim text by removing claim numbers, dropping punctuation characters, and converting the mixed-caps words to lowercase. To de-emphasize the common words across many patent documents (e.g. 'a', 'the', etc.), we use a 'Term-Frequency-Inverse-Document-Frequency' (TF-IDF) approach (Salton and Mcgill 1986; Younge and Kuhn 2016; Kuhn, Younge, and Marco 2018). This approach adjusts the term-frequency scores by scaling down the scores of the common words the most. To scale down the scores, we precompute the number of patent documents that include a given word, divided by the total number of patents documents. We then take the base-two-logarithm of the inverse of this fraction to compute the 'Inverse-Document-Frequency' (IDF). To finish computing the TF-IDF, we multiply the IDF scaling factor by the raw term frequencies in each patent document. We quantify the invention similarity between two patent documents as the cosine similarity between the TF-IDF vectors derived from their patent claims.[11]

Before moving to the empirical methodology, it is important to discuss what invention similarity does and does not measure. Precisely, invention similarity measures the degree to which two inventions share underlying technology. The vertical relationship may be complementary (one invention uses the second invention), or the two inventions share some underlying third invention. Additionally, invention similarity could capture imprecisely defined patent rights; however, since our measure is based on citations reviewed by examiners for patented inventions, it is unlikely that our measure reflects overlap from imprecisely defined patent rights. In Section 5.4, we test the relationship between our overlapping claims measure and imprecisely defined patent rights. Specifically, we correlate our overlapping claims measure with the probability of being instituted at PTAB.

3.3. Empirical methodology

This section describes the empirical methodology used to validate the overlapping claims measure and to highlight the additional information the similarity-weighted citations provide for the measurement of overlapping claims. We use the linear weighting function for all of the empirical results described in the main portion of the paper, and reproduce these regressions using variants of the weighting function in the Robustness Checks section.

We first validate our overlapping claims measure using methods familiar to the literature (von Graevenitz, Wagner, and Harhoff 2011). According to Cohen, Nelson, and Walsh (2000), products in complex technologies are comprised of many separate components (or inventions) while products in discrete technologies contain very few. For example, smart phones are comprised of thousands of patents on the various inputs used to produce a single smart phone and are classified as complex technologies. On the other hand, pharmaceutical drugs contain relatively few patentable inventions and therefore are discrete. Complex technologies are characterized by a high degree of cumulative innovation and technological interoperability, and therefore we expect a larger degree of technical overlap (or specifically, overlapping claims) across inventions in complex technologies relative to discrete technologies (Hall et al. 2013). In this first validation, we compare the global overlapping claims measure across discrete and complex technologies using the Cohen, Nelson, and Walsh (2000) classification. To do so, we first normalize our global overlapping claims measure by the number of patents to mitigate variation in the overall size of each technology space. Second, we normalize by the average number of citations within the technology. Any difference in our measurement of overlapping claims remaining between discrete and complex technologies can be attributed to differences in the invention similarity of patent claims among citations within those technologies.

Next, we compare our overlapping claims measure to a variety of USPTO patent examination characteristics. Specifically, we estimate the following regression model

$$y_i = \gamma_a + \gamma_t + x_i\beta + \epsilon_i \tag{1}$$

where γ_a are technology fixed effects, γ_t are year fixed effects and x_i is a set of application-level characteristics. We use a variety of dependent variables y_i; including, post-first-action patent application pendency, examiner search intensity and USPTO examination complexity factors. Post-first-action patent application pendency is the length of time between the initial response from the examiner (called the first action) and terminal disposal. Examiner search intensity is the amount of time the examiner spent searching for prior patents and other documents during examination and is proxied by the number of search pages in the first action. For the final validation, we compare our overlapping claims measure to USPTO examination complexity factors. These are established factors that reflect the expected level of complexity for patent applications examined in a particular technology classification. A higher complexity factor indicates that an examiner is allotted more time to complete an examination (Marco et al. 2017). Additionally, we include a set of incoming application characteristics x_i because of possible relationships with both the outcome variable and the overlapping claims measure. In particular, we include incoming patent application claim scope, and parent type (for example, continuation, continuation-in-part, divisional etc.). To capture incoming claim scope, we use both the number of independent claims (ICC) and the length of the shortest independent claim (ICL) at pre-grant publication (Marco, Sarnoff, and deGrazia 2017). These controls are important since application characteristics are likely related to aspects of patent prosecution (for example, broader claims may be more likely to face resistance at the patent office and therefore have longer pendency) and perhaps the overlapping claims measure (very narrow claims may rely heavily on prior art).

Our overlapping claims measure should be positively correlated with variables related to technological complexity, including the amount of time patent applications are in the patent office, the intensity of examination search, and the amount of time provided to the examiner for prosecution.[12] To estimate these correlations, we run ordinary least squares on Equation (1) with technology center/action year fixed effects. USPTO patent examiners are organized into technology centers based on the technologies they are assigned to examine. In the robustness checks section, we run additional regressions with more granular technological fixed effects (USPC).

Next, we assess the informational content of invention similarity for measuring overlapping claims. Recall, that one disadvantage of only using blocking citations to measure inventive overlap is that initially allowable patent applications will not receive a blocking rejection, therefore similar yet unblocking patent citations will not be included in the computation. Whether or not the patent application received the blocking rejection and then an amendment, or submitted allowable claims initially should not impact the measurement of inventive overlap. It is important to recognize that whether or not an application receives a rejection is endogenous. It is reasonable to assume that applicants adjust their behavior in more cumulative technologies by searching more prior to filing. Therefore, one cannot a priori assume that blocking rejections are more likely in cumulative or complex technologies. Additionally, as discussed in the literature review section, inventions may share components of other patented inventions but still receive a patent without citing the other patents.

To assess the informational coverage of invention similarity for measuring overlapping claims, we compare the distributions of invention similarity for blocking versus non-blocking citations. The degree of overlap between these two distributions indicates the amount of information lost by *only* using prior art rejections to measure overlapping rights. In particular, non-blocking citations that are just as technologically similar as blocking citations are not used in pure blocking citation measures (for example, the VG measure). Second, we run ordinary least squares regressions to estimate the impact of overlapping claims on the probability of receiving a prior art rejection on the first office action at the USPTO.[13,14] Specifically, we estimate specification (1) with y_i equal to one if the application received a prior art rejection on the first office action, and zero otherwise. An insignificant or negative estimate on the marginal effect of the overlapping claims measure would indicate that our measure contains additional information beyond that captured in blocking patents.

Finally, we test whether our overlapping claims measure emphasizes vertically overlapping claims, while excluding overlap derived from improperly defined patent rights. To do this, we correlate our overlapping claims measure with Patent Trial and Appeals Board (PTAB) Inter Partes Review (IPR) institution decisions. We use the empirical model specified in (1), with $y_i = 1$ if at least one patent claim is instituted at PTAB, and $y_i = 0$ otherwise (conditional on petition). IPR cases at PTAB are prior art validity proceedings at the USPTO. These cases follow a sequential procedure. First, the grounds for invalidity are initially inspected for petitioned patents at PTAB. If the grounds for invalidity are such that there is a 'reasonable likelihood that the petitioner would prevail with respect to at least 1 of the claims challenged in the patent', then the PTAB institutes the claims for further review.[15] If instituted, the PTAB fully evaluates the claims and writes a final written decision that determines the validity of each instituted claim. Since instituted claims are likely to have at least one invalid claim, we correlate the overlapping claims measure at patent grant with the probability of being instituted at PTAB, with technology (Technology Center) fixed effects, institution year fixed effects and additional covariates x_i specified above. If the overlapping claim measure excludes overlap derived from improperly granted patent rights, then we expect the measure to be generally uncorrelated with the PTAB institution outcome.

4. Data

To compute our overlapping claims measure, we rely on citations, issue and expiration dates, technology classifications and claim text for each patent. All of these data are contained in publicly-available datasets from the USPTO's Office of the Chief Economist (OCE). We use data on patent application scope from the Patent Claims Research Dataset (Marco, Sarnoff, and deGrazia 2017), citations and patent owners from PatentsView,[16] and patent issue/expiration dates from the Historical Patent Data Files (Marco et al. 2015). We remove self-citations by excluding citing/cited pairs with the same patent owner.

Post-first-action application pendency and parent type[17] are available in the OCE's PatEx dataset (Graham, Marco, and Miller 2018). However, we extract the variables from the USPTO Patent Application Location Monitoring (PALM) database. PatEx is derived from Public Pair, which is derived from PALM. Thus, post-first-action application pendency should be the same as if it were extracted from PatEx. We extract the number of search pages in the first office action from the USPTO's Image File Wrapper (IFW). Patent examination complexity factors are not publicly available, therefore we extract those data from PALM. We utilize data on blocking patents from the OCE's Office Actions Dataset (Lu, Myers, and Beliveau 2017). Specifically, we extract an indicator from the Office Actions Dataset for whether the application received a prior art rejection on the first action. Lastly, the PTAB data was obtained from internal USPTO sources, but the data is generally available publicly via USPTO systems.[18]

We compute our local overlapping claims measure for each unexpired patent between the years 2000 and 2014. Note that the overlapping claims measure changes year over year since the network of citations evolves with new forward citations and expiring patents. Since the prior art rejection data on the first action is only available post-2008 (from the Office Actions Dataset), for consistency across regressions we limit the data to every observation with the first action including and after 2008. With the full set of variables x_i, this includes 1,259,086 observations for post-first-action pendency, 1,246,302 for complexity, 1,282,299 for search, 1,251,199 for the probability of receiving a prior art rejection on the first action, and 2258 for PTAB institutions. Additionally, since the overlapping claims measure, post-first-action pendency and search pages variables are skewed, we log these variables. Finally, since a number of applications have overlapping claims measure of zero (because the applications are not a member of any triad), we add 0.001 before logging the variable. We follow Acemoglu and Finkelstein (2008) and include several additional robustness checks to ensure our results are insensitive to the weight from the overlapping claims measure at zero. In particular, we run additional specifications that include either a dummy variable for overlapping claims measure greater than zero, and additional specifications that only include the logged overlapping claims measure if the underlying overlapping claims measure is strictly positive. These results are reported in the robustness checks. Descriptive statistics for each of the regressions are contained in the appendix. The complexity factor, search pages, post-first-action pendency, prior art rejection and PTAB descriptive statistics are contained in Tables A13, A14, A15, A16 and A17 respectively.

5. Results

5.1. Discrete vs complex technologies

This section describes the results of the discrete v. complex technology validation tests for our overlapping claims measure. Figure 2 compares our overlapping claims measure across discrete and complex technologies, with and without various normalizations. Recall that our global overlapping claims measure is increasing in the average number of citations, the number of patents in the

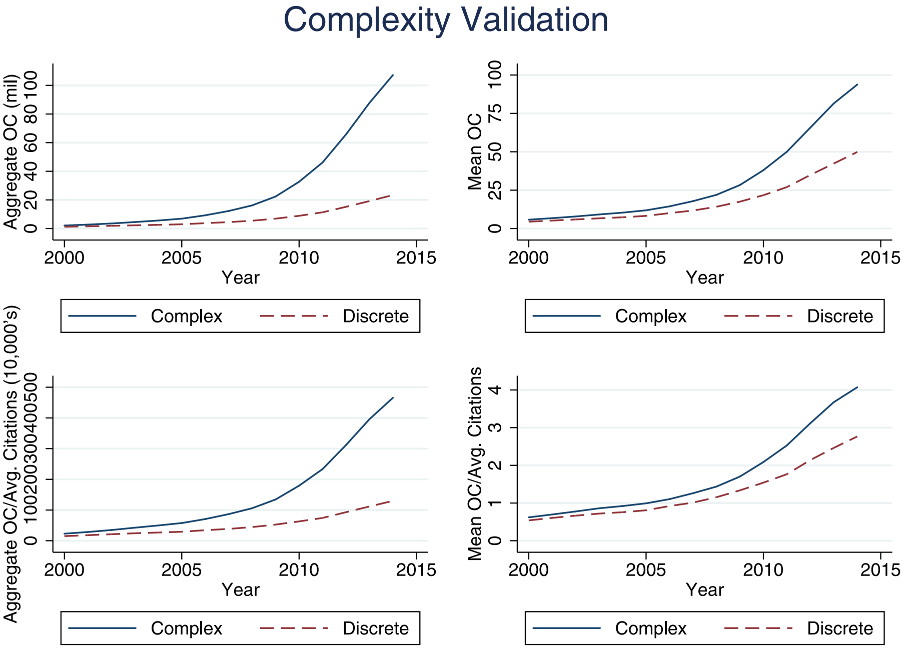

Figure 2. Complex vs discrete technologies: the aggregated (upper-left) and mean (upper-right) overlapping claims (OC) measure is plotted by year for both discrete and complex technologies. The aggregated (lower-left) and mean (lower-right) overlapping claims measure divided by the mean number of citations in either discrete or complex technologies is plotted by year for both discrete and complex technologies. Empirically, after the aforementioned normalizations, the overlapping claims measure in complex technologies is higher than discrete technologies. Furthermore, invention similarity contains information for distinguishing complex vs discrete (bottom-right).

technology space, and the invention similarity between patents. The top left panel in Figure 2 shows that the non-normalized global overlapping claims measure[19] for complex technologies is always higher than for discrete technologies and grows at a much faster rate. The top right pane shows that the global overlapping claims measure normalized by patent volume displays a smaller yet still growing gap between complex and discrete technologies over time. This is our preferred measure of overlapping claims since it controls for the size of the technology space.

In order to identify the degree to which invention similarity of patent claim text is driving the difference in our measure between discrete and complex technologies, we normalize by both the number of patents and the average number of citations per patent in the technology area. The bottom left pane shows a similar trend when normalizing by the average number of citations per patent. The bottom right pane normalizes the overlapping claims measure by both patent volume and average number of citations. The gap between discrete and complex technologies persists. Recall that, after controlling for patent volume and average citations, any difference in our measurement of overlapping claims remaining between discrete and complex technologies can be attributed to differences in the invention similarity of patent claims. This is especially true since, because patent volume and average number of citations are most likely positively correlated, the dual normalization may understate the invention similarity effect.

Overall, initial results support the validity of our overlapping claims measure for distinguishing between complex versus discrete technologies and our assertion that invention similarity, as captured by patent claim similarity, drives some of the persistent and growing gap in overlapping claims between these two sets of technologies.

5.2. USPTO patent examination

This section describes the results of several regressions used to estimate the relationship between the overlapping claims measure and USPTO application pendency, examiner search intensity, and USPTO performance measurement complexity factors. Recall that in more complex technologies, the USPTO provides examiners more time to prosecute a given patent application, and therefore the examiner should have higher search intensity and the application should take longer on average to prosecute. Positive correlations between the overlapping claims measure and these variables then would further validate our measure.[20]

Table 1 reports the regression results. The overlapping claims measure is statistically significant and positive in all regressions. This confirms our expectation that the overlapping claims measure is positively correlated with examination complexity factors, the number of search pages on the first office action, and post-first-action pendency. Although we only report these estimates as correlations, since the coefficients are relatively small and some of the variables are logged, we further explain the estimates. First, in the post-first-action pendency regression with the full set of covariates x_i, a 10% increase in the overlapping claims measure leads to a 0.14% change in post-first-action pendency. Since the overlapping claims measure is highly skewed, we evaluate a change of the magnitude between the inter-quartile range. An increase in overlapping claims from the 25th to 75th percentile leads to an increase in post-first-action pendency by 2 months, or 12.91%. Second, an increase in overlapping claims by 10% increases the number of search pages by 0.11%. Furthermore, an increase in overlapping claims from the 25th to 75th percentile increases the number of search pages by one-fourth of a page, or 10.4%. Finally, an increase overlapping claims by 1% increases the complexity factor by 0.0002. If the overlapping claims measure increases from 25th to the 75th percentile, the complexity factor increases by 0.017, or 0.08%.

5.3. Informational content of patent citations

This section further explores the informational coverage of invention similarity for the measurement of overlapping claims. As discussed earlier, one disadvantage of relying exclusively on blocking

Table 1. Complexity factor, search intensity and pendency regression results.

	(1)	(2)	(3)	(4)	(5)	(6)
Variables	Ex. comp	Ex. comp	Search	Search	Pendency	Pendency
Overlap. claims (Log)	0.0184***	0.00206***	0.0118***	0.0117***	0.0138***	0.0146***
	(0.000604)	(0.000654)	(0.000172)	(0.000187)	(0.000122)	(0.000131)
ICC		0.119***		0.00344***		0.0145***
		(0.00444)		(0.000420)		(0.000562)
ICL		−0.000380***		0.000263***		−0.000845***
		(3.48e−05)		(1.76e−05)		(3.62e−05)
Parent – Con.		0.551***		−0.0450***		−0.0312***
		(0.0183)		(0.00496)		(0.00340)
Parent – Div.		0.410***		1.93e−06		−0.110***
		(0.0193)		(0.00547)		(0.00374)
Parent – For.		0.0990***		−0.0114**		−0.00665**
		(0.0175)		(0.00469)		(0.00336)
Parent – N/A		0.298***		0.0134***		0.0422***
		(0.0175)		(0.00466)		(0.00336)
Parent – NST		−0.203***		−0.0312***		0.0695***
		(0.0177)		(0.00488)		(0.00340)
Parent – Prov.		−0.0107		0.0275***		0.0760***
		(0.0182)		(0.00493)		(0.00349)
Constant	21.67***	21.06***	0.937***	0.922***	6.241***	6.274***
	(0.0128)	(0.0275)	(0.00456)	(0.00665)	(0.00249)	(0.00609)
Observations	1,340,111	1,246,302	1,382,648	1,282,299	1,354,042	1,259,086
R-squared	0.556	0.561	0.013	0.014	0.164	0.191
Action Year FE	Yes	Yes	Yes	Yes	Yes	Yes
TC FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Robust standard errors in parentheses. Parent-type fixed effects are relative to continuations-in-part.
Overlap. claims (Log) = log(Overlap. claims + 0.001).
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

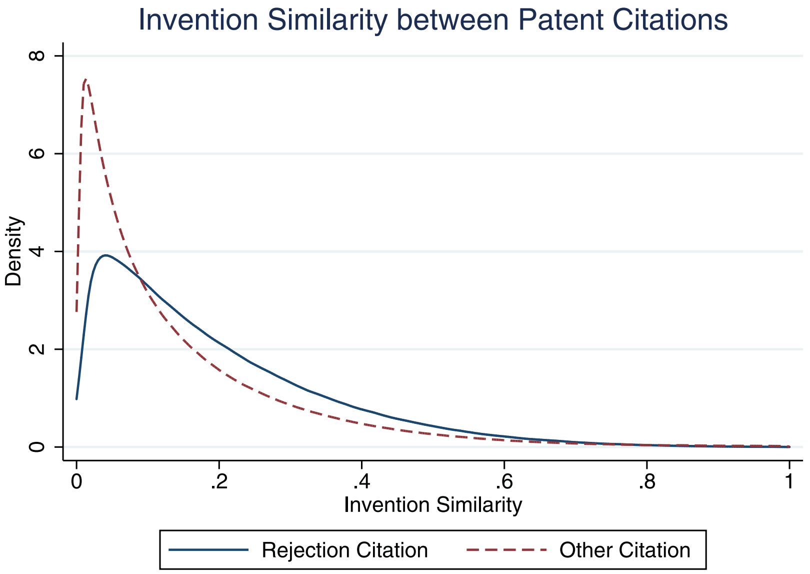

Figure 3. The informational content of patent citations: rejection citations are those used in prior art rejections at the USPTO. Other citations are those citations listed on the face of the patent which were not used in prior art rejections. Although the distribution of invention similarity for the other citations is less similar than the prior art rejection citations, there is significant overlap. This overlap signifies information lost when only considering blocking patents for the measurement of inventive overlap.

patents to measure inventive overlap is that once an applicant receives a prior art rejection, she must modify her claims in order to receive a patent grant later on in prosecution. Alternatively, if she had originally submitted claims to the USPTO that were an appropriate distance from the prior art, then the application would not receive a prior art rejection. Furthermore, patented inventions may share elements of other patented inventions, and still never receive a prior art rejection in the patent prosecution process. Finally, examiners may not record all blocking patents in the rejection for a variety of reasons.

Figure 3 displays the distributions of invention similarity, as captured by patent claim similarity, between citations applied in a prior art rejection and all other citations from patent applications. The distribution of invention similarity between non-blocking citations is generally less similar than the distribution of rejection citations; however, there is significant overlap. Crucially, a large volume of citations not used in rejections are just as technologically similar as those citations used in blocking prior art rejections. Therefore, a measure for overlapping claims that only uses blocking citations loses all of this additional information.

Table 2 displays the regression results for the probability of receiving a 102/103 prior art rejection at the USPTO. If the coefficient on the overlapping claims variable is positive and significant, then a densely-cumulative technology space is more prone to 102/103 rejections, implying that much of the information captured in our overlapping claims measure would already be captured by blocking citations. Alternatively, a negative and significant or insignificant estimate would indicate that invention similarity contains additional information beyond that captured by an overlapping claims measure relying on blocking patents alone.

The coefficient on the overlapping claims measure is negative and statistically significant in both regressions. Since the estimates are not significantly positive, the regression results verify that invention similarity contains information on inventive overlap not captured by blocking patents alone. Given that the estimates are small, we provide further interpretation. First, if the overlapping claims measure increases by 10%, then the probability of receiving a prior art rejection decreases by 0.01 percentage points. An increase in the overlapping claims measure from the 25th percentile to the 75th percentile reduces the probability of receiving a prior art rejection by 1 percentage point, or 1.83%.

Table 2. 102/103 rejection regression results.

Variables	(1) 102/103 Rejection	(2) 102/103 Rejection
Overlap. claims (Log)	−0.00128***	−0.00112***
	(8.49e−05)	(9.21e−05)
ICC		−0.00657***
		(0.000346)
ICL		−0.000488***
		(2.00e−05)
Parent – Con.		0.00120
		(0.00236)
Parent – Div.		0.0239***
		(0.00258)
Parent – For.		0.0274***
		(0.00225)
Parent – N/A		0.0360***
		(0.00222)
Parent – NST		−0.00556**
		(0.00230)
Parent – Prov.		0.0566***
		(0.00232)
Constant	0.525***	0.591***
	(0.00173)	(0.00394)
Observations	1,345,077	1,251,199
R-squared	0.219	0.233
Action Year FE	Yes	Yes
TC FE	Yes	Yes

Notes: Robust standard errors in parentheses. Parent-type fixed effects are relative to continuations-in-part.
Overlap. claims (Log) = log(Overlap. claims + 0.001).
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

5.4. PTAB institution decisions

This section reports the PTAB institution regressions results, which tests whether the overlapping claims measure emphasizes cumulative innovation (vertically overlapping claims), while generally excluding overlap from improperly defined patent rights. This distinction is important because overlapping, but valid claims and improperly granted claims have different economic implications in patent thickets (Egan and Teece 2015). We contend that our measure contains the cumulative portion of overlapping claims but does not capture improperly granted claims. This assertion can be directly tested using results from Inter Partes Review (IPR) from the Patent Trial and Appeals Board (PTAB). If our measure contains elements related to improperly granted claims, then we would expect a significant relationship between our measure and the probability of invalidity at PTAB. Specifically, a higher degree of overlapping claims would capture improperly granted claims that should have been rejected by the USPTO under 35 USC § 102 or 35 USC § 103. Recall that since institutions at PTAB indicate that at least one claim is likely to be invalid for novelty or non-obviousness (prior art rejections at the USPTO), institution decisions can be used as a proxy for validity relative to the prior art.[21]

Table 3 reports the results of linear probability models that correlate the overlapping claims measure to the probability of being instituted at PTAB, conditional on petition. Notably, the overlapping claims measure is insignificant in both regression specifications. This result, coupled with the previous validations, implies that our measure captures vertically overlapping claims and neglects invention overlap from improperly defined patent rights.

5.5. Robustness checks

In this section, we report the results of several robustness checks to the regression analysis. The first robustness check re-estimates the regression specifications with more granular technology fixed

Table 3. PTAB institution results.

Variables	(1) Institution	(2) Institution
Overlap. claims (Log)	−0.000250	0.000316
	(0.00198)	(0.00257)
ICC		−0.00270
		(0.00273)
ICL		−0.000121
		(0.000103)
Parent – Con.		0.00937
		(0.0384)
Parent – Div.		−0.0239
		(0.0467)
Parent – For.		0.0140
		(0.0595)
Parent – N/A		−0.0242
		(0.0443)
Parent – NST		0.0443
		(0.0585)
Parent – Prov.		−0.0835*
		(0.0449)
Constant	0.634***	0.961***
	(0.0519)	(0.0656)
Observations	3,033	2,258
R-squared	0.019	0.023
Action Year FE	Yes	Yes
TC FE	Yes	Yes

Notes: Robust standard errors in parentheses. Parent-type fixed effects are relative to continuations-in-part.
Overlap. claims (Log) = log(Overlap. claims + 0.001).
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

effects. In particular, rather than using broad USPTO Technology Centers, we use USPC class fixed effects. The results are reported in Table A18 in the appendix. The results for the search, pendency, rejection and institution regressions are consistent in sign, significance and magnitude. We do not estimate the complexity factor regression with USPC fixed effects since the complexity factors have very little variation within the USPC class.

The second set of robustness tests checks the sensitivity of our results to the mass of overlapping claims at zero. With this type of distribution for the independent variable of interest, Acemoglu and Finkelstein (2008) suggests the following robustness checks. In addition to the regressions reported earlier in the paper, we re-estimate the models with both (1) a dummy variable for a strictly positive overlapping claims measure, and (2) the logged overlapping claims measure restricted to strictly positive values of overlapping claims. The dummy variable regressions are shown in Tables A8, A7 and A9. The signs and statistical significance of the results are completely consistent except for column (2) of Table A8. These regressions generally confirm our earlier results. The results from the regressions restricting to observations with strictly positive overlapping claims are provided in Tables A11, A10 and A12. The signs and statistical significance are completely consistent with the results presented earlier and the magnitudes are generally larger.

In the final set of robustness checks, we recalculate and compare the overlapping claims measure with various weighting functions and re-estimate the validation regressions with these modified measures. In particular, we compare the linear weighting function, quadratic weighting function and truncated weighting functions described in the measure methodology section. The theoretical differences are described in the appendix, though both the truncated and quadratic weighting functions provide relatively less weight to lower similarity citations. Figure A1 in the appendix shows that marginalizing lower weight citations with the quadratic and truncated weighting functions does decorrelate the overlapping claims measure from the linear weighting function. In particular, the quadratic weighting function is less correlated with the linear weighting function than the truncated weighting function. This is appealing since the general purpose of our measure is to reduce the contribution of weakly related citations to the measurement of overlapping claims. Finally, as shown in the appendix, our regression results are generally insensitive to the weighting function used.

6. Conclusion

This paper contributes to the cumulative innovation and patent thicket literature by developing a precise measure of vertically overlapping claims that exploits invention similarity to capture inventive overlap. As stated in the literature review, a measure for this particular channel of patent thickets is generally missing from the literature. Generally speaking, empirical studies of patent thickets need to be cognizant of each related economic channel and how the absence of a particular measure may lead to omitted variable bias. For example, our measure alone cannot identify the complements problem, but if it were combined with a fragmentation-style index and transaction cost measure, we could potentially identify the complements problem, and other bargaining variants in the literature (Spulber 2017). Furthermore, our measure isolates the vertical channel of overlapping claims (cumulative innovation or complementary inputs), while excluding inventive overlap from improperly defined patent rights. This is important since these two separate notions of overlapping claims have different economics implications (Egan and Teece 2015). In particular, vertical overlap is important for the complements problem, while horizontal overlap is important for the duplication of resources channel of patent thickets. We validated our measure using a variety of tests, including those from the existent literature.

Finally, our measure is universally computable for all patent systems and will enable novel empirical research regarding cumulative innovation and patent thickets within and across all patent jurisdictions. Further research will use the overlapping claims measure defined here, and weighted patent similarities more broadly, to address these questions.

Notes

1. We follow von Graevenitz, Wagner, and Harhoff (2011) and define a blocking patent as a patent used in a rejection at the patent office.
2. Complex technologies are characterized by a high degree of cumulative innovation and technological interoperability, and therefore we expect a larger degree of technical overlap (or specifically, overlapping claims) across inventions in complex technologies relative to discrete technologies (Hall et al. 2013).
3. As defined by the USPTO, 'An Office action is a document written by a patent examiner in the course of examination of a patent application. The Office action may cite prior art and gives reasons why the examiner has allowed (approved) the applicant's claims, and/or rejected the claims. A first Office action on the merits (FAOM) is typically the first substantive examination of the application' (https://www.uspto.gov/learning-and-resources/statistics/first-office-action-estimator).
4. 35 USC § 314a
5. https://www.uspto.gov/web/offices/pac/mpep/s2158.html
6. To the best of our knowledge, the measure used in Gaessler, Harhoff, and Sorg (2017) was first proposed by Dietmar Harhoff in a presentation at the IP Statistics for Decision Makers Conference (https://www.oecd.org/site/stipatents/1_2_Harhoff.pdf), but a working paper for the measure is not yet available.
7. https://www.uspto.gov/web/offices/pac/mpep/s706.html
8. The condition necessary for the complements problem requires licensing across firms.
9. See the USPTO's MPEP 1302.
10. A formal definition of TF-IDF cosine similarity is in the appendix.
11. There may exist simultaneity in these simple models. For example, an examiner receives more time to examine an application in a more technologically complex field, giving the examiner additional time to perform a more thorough search and cite additional relevant literature. Therefore, all else equal, the increase in examination time could lead to an increase in the overlapping claims measure.
12. A blocking patent is the basis for a prior art rejection. For example, if patent application 1 is rejected under 35 USC § 102 (novelty) based on patent 2, then patent 2 is the blocking patent.
13. 102 rejections are given for lack of novelty and 103 rejections are given for obviousness.

14. 35 USC § 114a
15. www.patentsview.org
16. Values for application type include continuation, continuation-in-part, divisional, foreign priority, national stage entry, provisional or new application.
17. For example, the PTAB API; https://developer.uspto.gov/api-catalog/ptab-api.
18. The aggregate of all weighted triads in the technology.
19. Recall from earlier that complex technologies are characterized by a high degree of cumulative innovation and technological interoperability, and therefore we expect a larger degree of technical overlap (or specifically, overlapping claims) across inventions in complex technologies relative to discrete technologies (Hall et al. 2013).
20. Validity for IPRs are only based on patentability requirements set forth in 35 USC § 102 and 35 USC § 103.

Acknowledgments

The authors would like to thank Andrew Toole, Amanda Myers, Samantha Zyontz, Daniel Spulber, Ralph Siebert, Diego de Morales Silva, Frank Van Rijnsoever, Juan Pablo Rud, Maris Goldmanis, three anonymous referees, participants at the Searle Center Sixth Annual Roundtable on Standard Setting Organizations and Patents, and others at the DRUID18 Conference in Copenhagen for their helpful feedback. An earlier version of this paper appeared in Nicholas A. Pairolero's doctoral dissertation (Pairolero 2016). The views expressed are those of the individual authors and do not necessarily reflect official positions of the United States Government, the United States Patent and Trademark Office (USPTO), or the Office of the Chief Economist at the USPTO.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Jesse P. Frumkin http://orcid.org/0000-0002-9601-8648
Nicholas A. Pairolero http://orcid.org/0000-0003-0628-0536

References

Acemoglu, Daron, and Amy Finkelstein. 2008. "Input and Technology Choices in Regulated Industries: Evidence from the Health Care Sector." *Journal of Political Economy* 116 (5): 837–880.

Arts, Sam, Bruno Cassiman, and Juan Carlos Gomez. 2018. "Text Matching to Measure Patent Similarity." *Strategic Management Journal* 39 (1): 62–84.

Cockburn, Iain M., Megan J. MacGarvie, and Elisabeth Mueller. 2010. "Patent Thickets, Licensing and Innovative Performance." *Industrial and Corporate Change* 19 (3): 899–925.

Cohen, Wesley M., Richard R. Nelson, and John P Walsh. 2000. "Protecting their Intellectual Assets: Appropriability Conditions and Why US Manufacturing Firms Patent (or not)." *National Bureau of Economic Research*.

Egan, Edward J., and David J Teece. 2015. "Untangling the Patent Thicket Literature.".

Fischer, Timo, and Philipp Ringler. 2014. "The Coincidence of Patent Thickets – A Comparative Analysis." *Technovation* 38: 42–49.

Frakes, Michael D., and Melissa F Wasserman. 2017. "Is the Time Allocated to Review Patent Applications Inducing Examiners to Grant Invalid Patents? Evidence from Microlevel Application Data." *The Review of Economics and Statistics* 99 (3): 550–563.

Gaessler, Fabian, Dietmar Harhoff, and Stefan Sorg. 2017. "Patents and Cumulative Innovation – Evidence from Post-Grant Patent Oppositions." *Academy of Management Proceedings* (Vol. 2017, No. 1, p. 12800). Briarcliff Manor, NY: Academy of Management.

Galasso, Alberto, and Mark Schankerman. 2010. "Patent Thickets, Courts, and the Market for Innovation." *The RAND Journal of Economics* 41 (3): 472–503.

Graham, Stuart J. H., Alan C. Marco, and Richard Miller. 2018. "The USPTO Patent Examination Research Dataset: A Window on Patent Processing." *Journal of Economics & Management Strategy* 27 (3): 554–578.

Green, Jerry R., and Suzanne Scotchmer. 1995. "On the Division of Profit in Sequential Innovation." *The RAND Journal of Economics* 26: 20–33.

Hall, Bronwyn, Christian Helmers, C Rosazza-Bondibene, and Georg Von Graevenitz. 2013. "A Study of Patent Thickets." *UKIPO*.

Kuhn, Jeffrey M., Kenneth A. Younge, and Alan C Marco. 2018. "Patent Citations Reexamined." *SSRN*.

Lei, Zhen, and Brian D Wright. 2017. “Why Weak Patents? Testing the Examiner Ignorance Hypothesis.” *Journal of Public Economics* 148: 43–56.

Lu, Qiang, Amanda Myers, and Scott Beliveau. 2017. “USPTO Patent Prosecution Research Data: Unlocking Office Action Traits.” *SSRN*.

Marco, Alan C., Michael Carley, Steven Jackson, and Amanda Myers. 2015. “The USPTO Historical Patent Data Files Two Centuries of Innovation.” *SSRN*.

Marco, Alan C., Josh Sarnoff, and Charles deGrazia. 2017. “Patent Claims and Patent Scope.” *SSRN*.

Marco, Alan C., Andrew Toole, Richard Miller, and Jesse Frumkin. 2017. “USPTO Patent Prosecution and Examiner Performance Appraisal.” *SSRN*.

Nordhaus, William. 1969. *Invention, Growth, and Welfare*. Cambridge: The MIT Press.

O’donoghue, Ted, Suzanne Scotchmer, and Jacques-François Thisse. 1998. “Patent Breadth, Patent Life, and the Pace of Technological Progress.” *Journal of Economics & Management Strategy* 7 (1): 1–32.

Pairolero, Nicholas. 2016. “Pricing in Complex Networks.” *ProQuest*.

Salton, Gerard, and Michael Mcgill. 1986. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill Inc.

Scotchmer, Susan. 2004. *Innovation and Incentives*. Cambridge: MIT Press.

Shapiro, Carl. 2000. “Navigating the Patent Thicket: Cross Licenses, Patent Pools, and Standard Setting.” *Innovation Policy and the Economy* 1: 119–150.

Spulber, Daniel F. 2017. “Complementary Monopolies and Bargaining.” *The Journal of Law and Economics* 60 (1): 29–74.

von Graevenitz, Georg, Stefan Wagner, and Dietmar Harhoff. 2011. “How to Measure Patent Thickets – A Novel Approach.” *Economics Letters* 111 (1): 6–9.

Younge, Kenneth, and Jeffrey Kuhn. 2016. “Patent-to-patent Similarity: a Vector Space Model.” *SSRN*.

Ziedonis, Rosemarie Ham. 2004. “Don’t Fence Me in: Fragmented Markets for Technology and the Patent Acquisition Strategies of Firms.” *Management Science* 50 (6): 804–820.

Appendix 1. Definitions of term frequency, inverse document frequency, and cosine similarity

To quantify the similarity between two textual claims, we compare the frequencies of the claim terms after reducing the impact of the terms that are used across many documents (Salton and Mcgill 1986). Formally, let t be a particular term (e.g. a particular word) and D be the set of N documents (e.g. making up the corpus of N patent documents). Define the term frequency of term t in document d to be the number of times the term t occurs in document d and referred to as $tf(t,d)$. The inverse document frequency of term t in D is

$$idf(t,D) = \log_2\left(\frac{N}{|\{d \in D|t \in d\}|}\right)$$

Finally, the Term Frequency Inverse Document Frequency (TF-IDF) of term t in document d given D is

$$tfidf(t,d) = tf(t,d) \cdot idf(t,D)$$

Using *tfidf* scores, a particular document is represented as a vector, where each dimension of the vector is a *tfidf* score for a particular term. To quantify the similarity between the two vectors (e.g. from two claims), we use the cosine similarity of the vectors. Specifically, let a and b be the term-frequency-inverse-document-frequency vectors for document A and document B in the corpus D. The cosine similarity of between document A and B is given by

$$cos(\theta) = \frac{a \cdot b}{||a||\,||b||}$$

The cosine similarity ranges from 0 to 1, where the more positive scores indicate more textual similarity between the documents (e.g. between the claims).

Appendix 2. Weighting functions

In this appendix, we further explore the weighting function used to construct the overlapping claims measure. The weighting function is used to combine the citation similarity weights into an overall triad weight. The precise form of this function modifies the emphasis placed on higher and lower similarity values. Although for simplicity we use the linear weighting function in the empirical portion of the paper, in this appendix we explain the implications of several weighting function variants (truncated and quadratic) on the local overlapping claims measure (that is, the patent level measure). First we define and explore the weighting functions theoretically. Second, we empirically compare the weighting functions, and ensure the validation results of this paper are not sensitive to the particular weighting function.

Recall, the linear weighting function is defined in the following way

$$f_l(w_{ij}, w_{ik}, w_{jk}) = (w_{ij} + w_{ik} + w_{jk})$$

This function simply takes the similarity weights between each pair of patents in the triad and sums them up. Information from less relevant patent citations is reduced through the similarity weights. In particular, less similar patents will have a lower overall contribution to the triad weight since $\partial f / \partial w_{ij} > 0$ (higher w_{ij} is more similar). Recall Figure 3. From the figure, it is clear that although many non rejection citations contain relevant information, a large portion of them contain very little information (the sizable mass towards zero). The triad weighting function can be further modified to reduce the noise from these less similar citations. We provide two examples. The truncated weighting function is defined in the following way:

$$f_t(w_{ij}, w_{ik}, w_{jk}) = 1\{w_{ij} > 0.10\} \cdot w_{ij} + 1\{w_{ik} > 0.10\} \cdot w_{ik} + 1\{w_{jk} > 0.10\} \cdot w_{jk}$$

The truncated weighting function explicitly removes the similarity weights below the threshold of 0.10. We chose 0.10 since it's the crossing point of the distributions in Figure 3. Although arbitrary, we do so for illustrative purposes. The truncation further reduces the noise from less relevant citations by explicitly removing them from the triad weighting function.

We introduce one more weighting function that further reduces the noise from less relevant citations. The quadratic weighting function is defined in the following way:

$$f_q(w_{ij}, w_{ik}, w_{jk}) = (w_{ij}^2 + w_{ik}^2 + w_{jk}^2)$$

The quadratic weighting function provides relatively higher weight to more similar citations since if $x' > x$,

$$\frac{x'^2}{x^2} > \frac{x'}{x}$$

Figure A1 shows that marginalizing lower weight citations with the quadratic and truncated weighting functions does de-correlate the overlapping claims measure from the linear weighting function. In particular, the quadratic weighting function is less correlated with the linear weighting function than the truncated weighting function. This is appealing since the general purpose of our measure is to reduce the contribution of weakly related citations to the measurement of overlapping claims. Finally, as shown in the following tables, our regression results are consistent across the type of weighting function used.

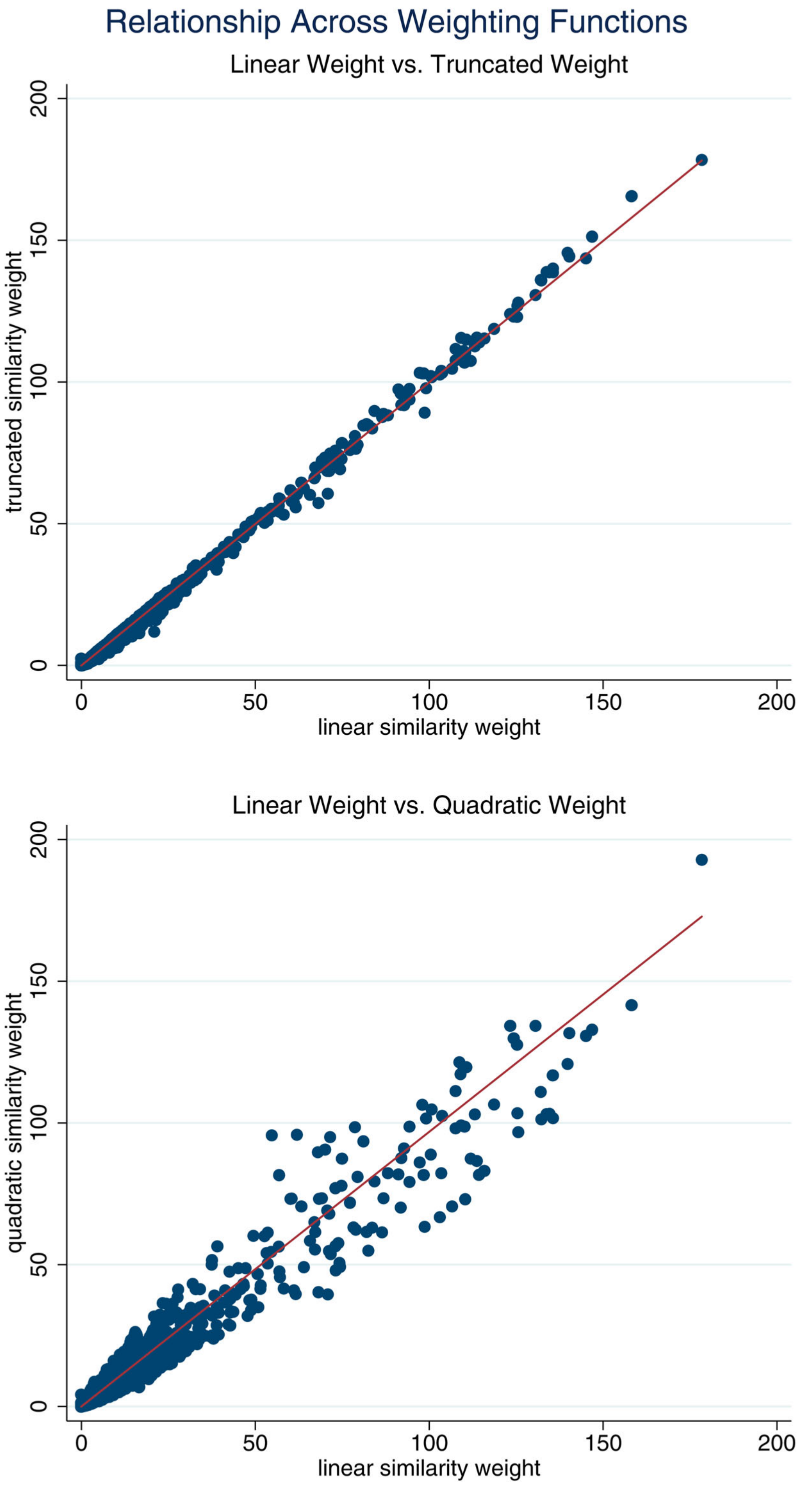

Figure A1. Comparing the overlapping claims measure derived from the linear, truncated and quadratic weighting functions.

Table A1. Robustness check: truncated weighting function. Complexity factor, search intensity and post-first-action pendency regression results.

	(1)	(2)	(3)	(4)	(5)	(6)
Variables	Ex. comp	Ex. comp	Search	Search	Pendency	Pendency
Overlap. claims (Log)	0.0177***	0.00137**	0.0117***	.0116***	0.0137***	0.0145***
	(0.000608)	(0.000658)	(0.000172)	(0.000187)	(0.000123)	(0.000132)
ICC		0.119***		0.00360***		0.0145***
		(0.00444)		(0.000420)		(0.000563)
ICL		−0.000380***		0.000250***		−0.000846***
		(3.48e−05)		(1.69e−05)		(3.62e−05)
Parent – Con.		0.552***		−0.0423***		−0.0314***
		(0.0183)		(0.00497)		(0.00340)
Parent – Div.		0.410***		−0.000359		−0.111***
		(0.0193)		(0.00547)		(0.00374)
Parent – For.		0.0964***		−0.0111**		−0.00776**
		(0.0175)		(0.00469)		(0.00336)
Parent – N/A		0.297***		0.0161***		0.0419***
		(0.0175)		(0.00466)		(0.00336)
Parent – NST		−0.205***		−0.0310***		0.0688***
		(0.0177)		(0.00489)		(0.00340)
Parent – Prov.		−0.0116		0.0262***		0.0759***
		(0.0182)		(0.00493)		(0.00349)
Constant	21.67***	21.06***	0.932***	0.917***	6.242***	6.276***
	(0.0128)	(0.0275)	(0.00455)	(0.00662)	(0.00249)	(0.00610)
Observations	1,340,111	1,246,302	1,382,648	1,282,299	1,354,042	1,259,086
R-squared	0.556	0.561	0.014	0.015	0.163	0.190
Action Year FE	Yes	Yes	Yes	Yes	Yes	Yes
TC FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Robust standard errors in parentheses. Parent-type fixed effects are relative to Continuations-in-part.
Overlap. claims (Log) = log(Overlap. claims + 0.001).
***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

Table A2. Robustness check: truncated weighting function. 102/103 rejection regression results.

Variables	(1) 102/103 Rejection	(2) 102/103 Rejection
Overlap. claims (Log)	−0.00133***	−0.00114***
	(8.54e−05)	(9.26e−05)
ICC		−0.00657***
		(0.000346)
ICL		−0.000488***
		(2.00e−05)
Parent – Con.		0.00124
		(0.00236)
Parent – Div.		0.0239***
		(0.00258)
Parent – For.		0.0273***
		(0.00225)
Parent – N/A		0.0359***
		(0.00222)
Parent – NST		−0.00562**
		(0.00230)
Parent – Prov.		0.0566***
		(0.00232)
Constant	0.524***	0.591***
	(0.00173)	(0.00394)
Observations	1,345,077	1,251,199
R-squared	0.219	0.233
Action Year FE	Yes	Yes
TC FE	Yes	Yes

Notes: Robust standard errors in parentheses. Parent-type fixed effects are relative to continuations-in-part.
Overlap. claims (Log) = log(Overlap. Claims + 0.001).
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A3. Robustness check: truncated weighting function. PTAB institution results.

Variables	(1) Institution	(2) Institution
Overlap. claims (Log)	−0.000388	0.000172
	(0.00198)	(0.00257)
ICC		−0.00268
		(0.00272)
ICL		−0.000120
		(0.000103)
Parent – Con.		0.00949
		(0.0384)
Parent – Div.		−0.0238
		(0.0467)
Parent – For.		0.0134
		(0.0594)
Parent – N/A		−0.0245
		(0.0443)
Parent – NST		0.0439
		(0.0584)
Parent – Prov.		−0.0836*
		(0.0449)
Constant	0.634***	0.960***
	(0.0520)	(0.0654)
Observations	3,033	2,258
R-squared	0.019	0.023
Action Year FE	Yes	Yes
TC FE	Yes	Yes

Notes: Robust standard errors in parentheses. Parent-type fixed effects are relative to continuations-in-part.
Overlap. claims (Log) = log(Overlap. Claims + 0.001).
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A4. Robustness check: quadratic weighting function. Complexity factor, search intensity and post-first-action pendency regression results.

	(1)	(2)	(3)	(4)	(5)	(6)
Variables	Ex. comp	Ex. comp	Search	Search	Pendency	Pendency
Overlap. claims (Log)	0.0213***	0.00267***	0.0127***	0.0127***	0.0150***	0.0161***
	(0.000671)	(0.000729)	(0.000191)	(0.000209)	(0.000135)	(0.000146)
ICC		0.119***		0.00346***		0.0144***
		(0.00444)		(0.000420)		(0.000562)
ICL		−0.000381***		0.000251***		−0.000847***
		(3.49e−05)		(1.72e−05)		(3.63e−05)
Parent – Con.		0.551***		−0.0453***		−0.0330***
		(0.0183)		(0.00497)		(0.00340)
Parent – Div.		0.410***		−0.00210		−0.111***
		(0.0193)		(0.00547)		(0.00374)
Parent – For.		0.100***		−0.0116**		−0.00603*
		(0.0175)		(0.00470)		(0.00336)
Parent – N/A		0.299***		0.0161***		0.0438***
		(0.0175)		(0.00466)		(0.00336)
Parent – NST		−0.201***		−0.0311***		0.0707***
		(0.0177)		(0.00489)		(0.00340)
Parent – Prov.		−0.00999		0.0267***		0.0774***
		(0.0182)		(0.00493)		(0.00349)
Constant	21.69***	21.06***	0.942***	0.929***	6.251***	6.286***
	(0.0128)	(0.0275)	(0.00456)	(0.00664)	(0.00251)	(0.00611)
Observations	1,340,111	1,246,302	1,382,648	1,282,299	1,354,042	1,259,086
R-squared	0.556	0.561	0.014	0.015	0.163	0.191
Action Year FE	Yes	Yes	Yes	Yes	Yes	Yes
TC FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Robust standard errors in parentheses. Parent-type fixed effects are relative to Continuations-in-part.
Overlap. claims (Log) = log(Overlap. claims + 0.001).
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A5. Robustness check: quadratic weighting function. 102/103 rejection regression results.

Variables	(1) 102/103 Rejection	(2) 102/103 Rejection
Overlap. claims (Log)	−0.00179***	−0.00155***
	(9.43e−05)	(0.000103)
ICC		−0.00652***
		(0.000345)
ICL		−0.000487***
		(2.00e−05)
Parent – Con.		0.00151
		(0.00236)
Parent – Div.		0.0238***
		(0.00258)
Parent – For.		0.0264***
		(0.00225)
Parent – N/A		0.0353***
		(0.00222)
Parent – NST		−0.00659***
		(0.00230)
Parent – Prov.		0.0562***
		(0.00232)
Constant	0.522***	0.590***
	(0.00174)	(0.00395)
Observations	1,345,077	1,251,199
R-squared	0.219	0.233
Action Year FE	Yes	Yes
TC FE	Yes	Yes

Notes: Robust standard errors in parentheses. Parent-type fixed effects are relative to continuations-in-part.
Overlap. claims (Log) = log(Overlap. claims + 0.001).
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A6. Robustness check: quadratic weighting function. PTAB institution results.

Variables	(1) Institution	(2) Institution
Overlap. claims (Log)	−0.000396	0.000133
	(0.00213)	(0.00276)
ICC		−0.00268
		(0.00273)
ICL		−0.000120
		(0.000103)
Parent – Con.		0.00952
		(0.0384)
Parent – Div.		−0.0238
		(0.0467)
Parent – For.		0.0133
		(0.0595)
Parent – N/A		−0.0246
		(0.0443)
Parent – NST		0.0437
		(0.0585)
Parent – Prov.		−0.0836*
		(0.0449)
Constant	0.634***	0.960***
	(0.0522)	(0.0656)
Observations	3,033	2,258
R-squared	0.019	0.023
Action Year FE	Yes	Yes
TC FE	Yes	Yes

Notes: Robust standard errors in parentheses. Parent-type fixed effects are relative to continuations-in-part.
Overlap. claims (Log) = log(Overlap. claims + 0.001).
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Appendix 3. Robustness check: dummy variable overlapping claims

Table A7. Robustness check: dummy variable overlapping claims. 102/103 rejection regression results.

	(1)	(2)
Variables	102/103 Rejection	102/103 Rejection
Overlap. claims (Dummy)	-0.00522^{***}	-0.00412^{***}
	(0.000770)	(0.000817)
ICC		-0.00668^{***}
		(0.000347)
ICL		-0.000488^{***}
		(2.00e−05)
Parent – Con.		0.000790
		(0.00236)
Parent – Div.		0.0242^{***}
		(0.00258)
Parent – For.		0.0300^{***}
		(0.00224)
Parent – N/A		0.0375^{***}
		(0.00222)
Parent – NST		−0.00310
		(0.00229)
Parent – Prov.		0.0576^{***}
		(0.00232)
Constant	0.531^{***}	0.596^{***}
	(0.00175)	(0.00394)
Observations	1,345,077	1,251,199
R-squared	0.219	0.233
Action Year FE	Yes	Yes
TC FE	Yes	Yes

Notes: Robust standard errors in parentheses. Parent-type fixed effects are relative to continuations-in-part.
Overlap. claims (Dummy) =1[Overlap. claims >0].
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A8. Robustness check: dummy variable overlapping claims. Complexity factor, search intensity and post-first-action pendency regression results.

	(1)	(2)	(3)	(4)	(5)	(6)
Variables	Ex. comp	Ex. comp	Search	Search	Pendency	Pendency
Overlap. claims (Dummy)	0.0880***	−0.0268***	0.0839***	0.0786***	0.102***	0.103***
	(0.00542)	(0.00573)	(0.00151)	(0.00162)	(0.00111)	(0.00116)
ICC		0.120***		0.00421***		0.0154***
		(0.00445)		(0.000425)		(0.000575)
ICL		−0.000377***		0.000266***		−0.000841***
		(3.48e−05)		(1.77e−05)		(3.60e−05)
Parent – Con.		0.553***		−0.0424***		−0.0270***
		(0.0183)		(0.00497)		(0.00342)
Parent – Div.		0.409***		−0.00288		−0.113***
		(0.0193)		(0.00548)		(0.00376)
Parent – For.		0.0844***		−0.0288***		−0.0261***
		(0.0175)		(0.00468)		(0.00336)
Parent – N/A		0.291***		0.00183		0.0288***
		(0.0175)		(0.00466)		(0.00337)
Parent – NST		−0.216***		−0.0482***		0.0503***
		(0.0177)		(0.00487)		(0.00340)
Parent – Prov.		−0.0153		0.0205***		0.0681***
		(0.0182)		(0.00493)		(0.00350)
Constant	21.59***	21.07***	0.867***	0.861***	6.157***	6.194***
	(0.0130)	(0.0273)	(0.00463)	(0.00673)	(0.00255)	(0.00611)
Observations	1,340,111	1,246,302	1,382,648	1,282,299	1,354,042	1,259,086
R-squared	0.555	0.561	0.012	0.013	0.161	0.188
Action Year FE	Yes	Yes	Yes	Yes	Yes	Yes
TC FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Robust standard errors in parentheses. Parent-type fixed effects are relative to continuations-in-part.
Overlap. claims (Dummy) = 1[Overlap. claims >0].
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A9. Robustness check: dummy variable overlapping claims. PTAB institution results.

	(1)	(2)
VARIABLES	Institution	Institution
Overlap. claims (Dummy)	0.00741	0.0216
	(0.0210)	(0.0280)
ICC		−0.00283
		(0.00273)
ICL		−0.000121
		(0.000103)
Parent – Con.		0.00882
		(0.0383)
Parent – Div.		−0.0248
		(0.0467)
Parent – For.		0.0185
		(0.0589)
Parent – N/A		−0.0220
		(0.0440)
Parent – NST		0.0493
		(0.0582)
Parent – Prov.		−0.0825*
		(0.0448)
Constant	0.631***	0.953***
	(0.0532)	(0.0686)
Observations	3,033	2,258
R-squared	0.019	0.024
Action Year FE	Yes	Yes
TC FE	Yes	Yes

Notes: Robust standard errors in parentheses. Parent-type fixed effects are relative to continuations-in-part.
Overlap. claims (Dummy) =1[Overlap. claims >0].
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Appendix 4. Robustness check: strictly positive overlapping claims

Table A10. Robustness check: strictly positive overlapping claim. 102/103 Rejection regression results.

	(1)	(2)
Variables	102/103 Rejection	102/103 Rejection
Overlap. claims (Log)	−0.00619***	−0.00503***
	(0.000246)	(0.000269)
ICC		−0.00620***
		(0.000461)
ICL		−0.000414***
		(2.57e−05)
Parent – Con.		−0.000979
		(0.00271)
Parent – Div.		0.0292***
		(0.00300)
Parent – For.		0.0320***
		(0.00271)
Parent – N/A		0.0355***
		(0.00258)
Parent – NST		−0.00320
		(0.00279)
Parent – Prov.		0.0605***
		(0.00269)
Constant	0.516***	0.568***
	(0.00245)	(0.00506)
Observations	779,434	727,197
R-squared	0.226	0.238
Action Year FE	Yes	Yes
TC FE	Yes	Yes

Notes: Robust standard errors in parentheses. Parent-type fixed effects are relative to continuations-in-part.
Overlap. claims (Log) = log(Overlap. claims + 0.001).
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.
Sample only includes patents if the overlapping claims value is strictly positive.

Table A11. Robustness check: strictly positive overlapping claims. Complexity factor, search intensity and post-first-action pendency regression results.

	(1)	(2)	(3)	(4)	(5)	(6)
Variables	Ex. comp	Ex. comp	Search	Search	Pendency	Pendency
Overlap. claims (Log)	0.0769***	0.0361***	0.0256***	0.0280***	0.0283***	0.0332***
	(0.00183)	(0.00199)	(0.000515)	(0.000565)	(0.000356)	(0.000388)
ICC		0.0994***		0.00283***		0.0133***
		(0.00569)		(0.000512)		(0.000739)
ICL		−0.000359***		0.000212***		−0.000774***
		(4.67e−05)		(2.21e−05)		(5.07e−05)
Parent – Con.		0.553***		−0.0572***		−0.0525***
		(0.0215)		(0.00573)		(0.00393)
Parent – Div.		0.395***		−0.00194		−0.110***
		(0.0229)		(0.00636)		(0.00437)
Parent – For.		0.0315		−3.84e−05		0.0239***
		(0.0213)		(0.00560)		(0.00408)
Parent – N/A		0.287***		0.0230***		0.0647***
		(0.0208)		(0.00543)		(0.00391)
Parent – NST		−0.191***		−0.0118**		0.101***
		(0.0217)		(0.00593)		(0.00413)
Parent – Prov.		−0.0163		0.0388***		0.0941***
		(0.0216)		(0.00574)		(0.00406)
Constant	21.56***	21.08***	0.848***	0.823***	6.218***	6.239***
	(0.0181)	(0.0351)	(0.00643)	(0.00856)	(0.00346)	(0.00804)
Observations	775,618	723,570	816,958	759,327	784,150	731,373
R-squared	0.548	0.554	0.014	0.016	0.173	0.201
Action Year FE	Yes	Yes	Yes	Yes	Yes	Yes
TC FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Robust standard errors in parentheses. Parent-type fixed effects are relative to Continuations-in-part.
Overlap. claims (Log) = log(Overlap. claims + 0.001).
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.
Sample only includes patents if the overlapping claims value is strictly positive.

Table A12. Robustness check: strictly positive overlapping claims. PTAB institution results.

	(1)	(2)
VARIABLES	Institution	Institution
Overlap. claims (Log)	−0.00443	−0.00840
	(0.00438)	(0.00520)
ICC		−0.00372
		(0.00284)
ICL		−0.000129
		(0.000103)
Parent – Con.		0.000208
		(0.0400)
Parent – Div.		−0.0456
		(0.0485)
Parent – For.		−0.0913
		(0.0747)
Parent – N/A		−0.0650
		(0.0479)
Parent – NST		−0.0118
		(0.0711)
Parent – Prov.		−0.132***
		(0.0478)
Constant	0.651***	1.002***
	(0.0629)	(0.0874)
Observations	2,432	1,883
R-squared	0.021	0.032
Action Year FE	Yes	Yes
TC FE	Yes	Yes

Notes: Robust standard errors in parentheses. Parent-type fixed effects are relative to continuations-in-part.

Overlap. claims (Log) = log(Overlap. claims + 0.001).

***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

Sample only includes patents if the overlapping claims value is strictly positive.

Appendix 5. Descriptive tables

Table A13. Descriptive statistics: complexity factor regression.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables	*N*	mean	sd	p25	p50	p75	min	max
Complexity factor	2.752e+06	22.57	4.495	19	21.30	25.90	11.60	123
Overlap. claims	2.752e+06	41.47	664.2	0	0.618	3.937	0	115,742
Overlap. claims (Log)	2.752e+06	−2.149	4.353	−6.908	−0.480	1.371	−6.908	11.66
ICC	2.089e+06	3.064	4.040	2	3	4	0	565
ICL	2.089e+06	110.3	96.77	59	91	137	0	14,547

Table A14. Descriptive statistics: search page regression.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables	*N*	mean	sd	p25	p50	p75	min	max
Search pages	1.408e+06	5.417	26.57	2	3	5	1	7,854
Search pages (Log)	1.408e+06	1.110	0.878	0.693	1.099	1.609	0	8.969
Overlap. claims	1.408e+06	66.39	886.8	0	0.756	5.304	0	115,742
Overlap. claims (Log)	1.408e+06	−1.851	4.487	−6.908	−0.278	1.669	−6.908	11.66
ICC	1.305e+06	2.873	2.510	2	3	3	0	564
ICL	1.304e+06	113.3	101.7	62	94	140	0	14,547

Table A15. Descriptive statistics: post-first-action pendency regression.

Variables	(1) *N*	(2) mean	(3) sd	(4) p25	(5) p50	(6) p75	(7) min	(8) max
Post-first-action pendency	2.865e+06	468.2	405.3	238	343	561	−633	13,865
PFA pend. (Log)	2.865e+06	5.887	0.706	5.472	5.838	6.330	3.497	9.537
Overlap. claims	2.865e+06	40.30	652.4	0	0.612	3.880	0	115,742
Overlap. claims (Log)	2.865e+06	−2.160	4.345	−6.908	−0.489	1.356	−6.908	11.66
ICC	2.114e+06	3.069	4.030	2	3	4	0	565
ICL	2.114e+06	110.3	98.94	59	91	137	0	14,547

Table A16. Descriptive statistics: prior art rejection regression.

Variables	(1) *N*	(2) mean	(3) sd	(4) p25	(5) p50	(6) p75	(7) min	(8) max
102/103 rejection	1.586e+06	0.357	0.479	0	0	1	0	1
Overlap. claims	1.586e+06	64.36	867.6	0	0.762	5.357	0	115,742
Overlap. claims (Log)	1.586e+06	−1.843	4.484	−6.908	−0.270	1.679	−6.908	11.66
ICC	1.469e+06	2.879	2.371	2	3	3	0	531
ICL	1.469e+06	113.1	102.7	61	93	140	0	14,547

Table A17. Descriptive statistics: PTAB institution regression.

Variables	(1) *N*	(2) mean	(3) sd	(4) p25	(5) p50	(6) p75	(7) min	(8) max
Overlap. claims	3033	195.4	1491	0.669	7.038	55.67	0	43,013
Overlap. claims (Log)	3033	0.918	4.399	−0.400	1.951	4.019	−6.908	10.67
Institution	3033	0.710	0.454	0	1	1	0	1
ICC	2258	3.732	3.496	2	3	4	1	54
ICL	2258	102.7	84.55	55	85	130	0	2306

Appendix 6. Robustness check: USPC fixed effects

Table A18. Robustness check: USPC fixed effects. Complexity factor, search intensity and post-first-action pendency regression results.

Variables	(1) Search	(2) Pendency	(3) Rejection	(4) Institution
Overlap. claims (Log)	0.0115***	0.0141***	−0.00130***	−0.000936
	(0.000696)	(0.000671)	(0.000260)	(0.00295)
ICC	0.00386***	0.0119***	−0.00579***	−0.00358
	(0.000877)	(0.00150)	(0.00107)	(0.00314)
ICL	0.000103	−0.000823***	−0.000482***	−0.000177*
	(0.000111)	(0.000232)	(0.000119)	(0.000107)
Parent – Con.	−0.0649***	−0.0411***	0.00509	0.0475
	(0.0114)	(0.0119)	(0.00582)	(0.0431)
Parent – Div.	−0.0282**	−0.115***	0.0303***	0.0290
	(0.0121)	(0.00895)	(0.00776)	(0.0452)
Parent – For.	−0.0228***	0.0206**	0.0337***	0.0385
	(0.00839)	(0.00882)	(0.00424)	(0.0604)
Parent – N/A	0.00901	0.0574***	0.0423***	−0.00115
	(0.00752)	(0.00814)	(0.00372)	(0.0406)
Parent – NST	−0.0514***	0.0970***	−0.00157	0.104
	(0.0106)	(0.00701)	(0.00335)	(0.0660)
Parent – Prov.	0.0108	0.0852***	0.0599***	−0.0653
	(0.0102)	(0.00666)	(0.00295)	(0.0436)
Constant	1.091***	6.136***	0.659***	1.023***
	(0.0143)	(0.0275)	(0.0178)	(0.0815)
Observations	1,156,677	1,259,086	1,251,199	2,258
R-squared	0.008	0.150	0.231	0.014
Number of USPCs	430	432	432	231
Action Year FE	Yes	Yes	Yes	Yes
USPC FE	Yes	Yes	Yes	Yes

Notes: Robust standard errors in parentheses. Parent-type fixed effects are relative to continuations-in-part.
Overlap. claims (Log) = log(Overlap. claims + 0.001).
***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.