U.S. Patent and Trademark Office
OFFICE OF THE CHIEF ECONOMIST
Economic Working Paper Series

Embracing Technological Similarity for the Measurement of Complexity and Patent Thickets

Charles deGrazia, *Economist*
Jesse P. Frumkin, *Data Scientist and Statistician*
Nicholas A. Pairolero, *Economist*

USPTO Economic Working Paper No. 2018-01
February 2018

The views expressed are those of the individual authors and do not necessarily reflect official positions of the Office of Chief Economist or the U. S. Patent and Trademark Office. USPTO Economic Working Papers are preliminary research being shared in a timely manner with the public in order to stimulate discussion, scholarly debate, and critical comment.

For more information about the USPTO's Office of Chief Economist, visit www.uspto.gov/economics.

UNITED STATES
PATENT AND TRADEMARK OFFICE
uspto

Embracing Technological Similarity for the Measurement of Complexity and Patent Thickets

Charles deGrazia, Jesse Frumkin, Nicholas A. Pairolero
United States Patent and Trademark Office

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, 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 (Shapiro 2000). Researchers have developed several measures to gauge the extent and impact of patent thickets. This paper contributes to that literature by proposing a new measure of patent thickets that incorporates patent claim similarity to more precisely identify technological similarity, which is shown to increase the information contained in the measurement of patent thickets. Further, the measure is universally computable for all patent systems. These advantages will enable more accurate measurement and allow for novel economic research on technological complexity, fragmentation in intellectual property, and patent thickets within and across all patent jurisdictions.

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, when a product depends upon intellectual property held across multiple firms, contracting failures may lead to suboptimal economic outcomes. Shapiro (2000) defines a patent thicket as, "a dense web of overlapping intellectual property rights that a company must hack its way through in order to actually commercialize new technology," and contends this fragmentation of patent rights has negative economic consequences. 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. Patent thickets 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, set forth by Galasso and Schankerman (2010), contends that patent thickets have a more ambiguous effect and may ease licensing negotiations and expedite court settlements. They maintain that fragmented patent rights reduce the value at stake in each individual license negotiation, easing bargaining tensions and facilitating deals. Therefore, in certain cases, there is a priori ambiguity regarding the impact of patent thickets that must be resolved empirically.

Researchers have developed several measures of fragmented patent rights to empirically study patent thickets. The Ziedonis (2004) fragmentation index (hereafter "fragmentation index") measures a firm's patent portfolio fragmentation based on the share of citations to competitors within a particular technology. von Graevenitz et al. (2011) develop a patent thicket measure (hereafter "VG measure") that relies on groups of three firms (called triples), each of which holds a patent used to block the patent applications of the other firms.[1] Fischer and Ringler (2015) extend the VG measure (hereafter "FR measure") to other patent jurisdictions by using all patent citations, relaxing the "blocking patent" requirement. While a variety of empirical studies on the effects of fragmented patent rights have used these pioneering methods, each measure suffers from limitations.

The fragmentation index and FR measure rely on citations alone and, therefore, fail to recognize variation in technological similarity[2] across patent citations. Further, with recent research showing that citations suffer from significant noise, measurement error and weakening informational content [3] (Gambardella et al. 2008, Roach and Cohen 2013, Kuhn et al. 2017), the accuracy of citation-based patent thicket measures should also be called into question.

The VG measure avoids this limitation by using triples of firms that hold mutually blocking patents. However, confining fragmented patent rights to only those patent applications used as prior art in a patent office rejection may ignore applications without patentability issues ex ante but overlapping patent claims ex post. Whether or not the applicant initially drafts allowable claims or modifies them during patent prosecution to overcome a rejection should not impact the measurement of patent thickets since all granted patents are a part of the thicket measure. Thus, while the citation-based measures may be too inclusive, the VG measure may be too restrictive and omit technologically similar patents with overlapping claims. An additional disadvantage of the VG measure is that it relies on European Patent Office (EPO) X and Y citations and, therefore, is not computable for other patent jurisdictions that do not maintain such systems.

In this paper, we propose a new measure of patent thickets based on the technological similarity of patent claim text. Our measure improves on the existing measures in the literature in three critical ways. First, we apply standard natural language processing to quantify the technological similarity of claims between citing and cited patents, better capturing the dispersion of technology 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 technological similarity without the limitation of only using blocking patents determined by the patent office.[4] We find significant overlap in the distribution of claim text technology similarity between blocking and non-blocking citations, suggesting that patent thicket measures derived from only blocking patents fail to capture all overlapping patent rights. Lastly, our measure is computable

[1] Dietmar Harhoff gave a presentation to the IP Statistics for Decision Makers Conference, suggesting the use of patent similarity to measure patent thickets. Our measure was developed independently around the same time. The presentation can be found at `https://www.oecd.org/site/stipatents/1_2_Harhoff.pdf`.

[2] We define technological similarity to be the similarity of two inventions based on a textual comparison of each invention's patent claims.

[3] Kuhn, et al. (2017) suggest that the average similarity between citing and cited patents should increase given the ever-expanding set of prior art. However, the paper finds that the "technological relationship reflected by the average patent citation has weakened over time, and that the weakening is continuing to accelerate."

[4] All patents that were a basis for a rejection(s) should be included in the list of backward citations. See 37 CFR 1.104 and the USPTO's MPEP 707

for all patent systems, allowing scholars to address patent thicket-related research questions within and across all jurisdictions.

The paper proceeds in the following way. First, we describe the theoretical and empirical literature on patent thickets. Second, we define and describe our proposed patent thicket measure. We then validate our measure using methods previously applied in the patent thicket measurement literature (von Graevenitz et al. 2011, Fischer and Ringler 2010). Results show that our patent thicket measure is consistently higher in complex versus discrete technologies. This difference persists after normalizing for patent volume and average citations, indicating that technological similarity of claims conveys additional information for distinguishing between complex and discrete technologies (Cohen et al. 2000).

As an additional validation, 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 thicket measure should be positively correlated with these variables. Finally, we show that our thicket measure is not predictive of whether or not a US patent application receives a prior art rejection on the first office action.[5] This result is consistent with our premise that reliance on blocking patents for measuring patent thickets omits technological similar citations for which the applicant considered but initially drafted claims to avoid rejection. Using technological similarity based on patent claim text retains the information contained in each citation and offers a more precise method for measuring technological complexity and fragmentation in patent rights.

2 Literature Review

2.1 Theoretical Literature

Shapiro (2000) establishes the theoretical framework regarding the negative economic consequences of patent thickets. The first consequence mirrors the Cournot complements problem, that is, issues arising when a firm must purchase inputs from several monopolists. In the case of intellectual property, this problem materializes when a firm must license components from multiple patentees, leading to individual license costs that are increasing in the number of entities required for contracting. As a result, licensing costs reduce the return on research and development, deterring entry by follow-on innovators and hindering the development of certain products. Further, patentee profits and consumer welfare is lower than if the patent-holders coordinated their licensing agreements. The complements problem is particularly acute where patent rights are fragmented. Both product complexity and overlapping patent claims tend to increase the number of licenses needed for a given product and therefore further exacerbates the complements problem.

[5] 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." (USPTO 2018)

A second channel through which patent thickets have economic consequences is patent hold-up. The hold-up problem occurs after product development (or after fixed costs have been incurred) when a patentee demands payment for infringement and threatens injunction in the courts. The potential for hold-up is increasing in patent rights fragmentation for a variety of reasons. First, patent thickets increase the likelihood that a firm is unaware of potentially infringing third party patent rights. Second, mechanisms that may counter the complements problem of patent thickets may actually exacerbate hold-up. As Shapiro (2000) notes, patent pools[6] are one solution to the complements problem in patent thickets by allowing coordination that increases overall patent owner profit, lowers prices and increases output. However, since patent collusion might inspire anti-trust regulation, patentees are reluctant to specify precise terms in patent pool licensing agreements. Vague ex-ante licensing conditions may allow patentees to hold-up follow-on innovators post product development. Shapiro (2000) also notes that cross-licenses are an effective way for large companies to solve the complements problem amongst themselves, at least for two competitors. However, defensive patenting intensifies patent thickets as firms attempt to increase their cross licensing bargaining positions for potential hold-up situations (Hall et al. 2013). This intensification of patent thickets negatively affects small firms not able to engage in cross licensing through both the complements problem and the potential for patent hold-up.

Some more recent literature proposes a more ambiguous effect of patent thickets on economic outcomes. Galasso and Schankerman (2011) show that, under certain conditions, overlapping patent rights reduce time to settlement in patent disputes. Even though there are more parties to negotiate with, fragmentation lowers the value at stake in each individual negotiation, easing bargaining tensions and lowering costs for each license. Therefore, in certain cases, there is ambiguity regarding the impact of patent thickets that must be resolved empirically.

2.2 Empirical Literature

Hall et al. (2013) surveys the empirical literature on the economic consequences of patent thickets, finding mixed evidence for the established theoretical results. Hall and Ziedonis (2001) show that the number of patents per research and development dollar for semiconductors increased between 1982 and 2003, while maintaining relatively constant for manufacturing companies. The result supports the assertion that strategic patenting in semiconductors was relatively defensive during that period (Hall et al. (2013) further cite Federal Trade Commission (2003) and Somaya (2003) for additional support to this claim). Cockburn et al. (2010) find that firms with increasing patent citation fragmentation (across companies) introduce relatively more products if licensing is not required, and less if licensing of inputs is required. Cockburn et al. (2011) study the impact of patent thickets on entry. They find that a 1 percent increase in the number of patents in a thicket reduces new firm entry by .8 percent. Further, entry is less likely for firms that hold less patents.

[6] "Under a patent pool, an entire group of patents is licensed in a package, either by one of the patent holders or by a new entity established for this purpose, usually to anyone willing to pay the associated royalties," (Shapiro 2000).

Both findings in Cockburn et al. (2010) and Cockburn et al. (2011) supports the theoretical consequences of the complements problem discussed above.

A variety of papers show that rather than solving the complements problem, patent pools actually reduce innovation and patenting by members of the patent pool (Lampe et al. 2010, Joshi et al. 2011, Lampe et al. 2012). These papers suggest that the theoretical benefits of patent pools as discussed by Shapiro (2000) may not occur in practice. Galetovic et al. (2015) investigate patent hold-up with standard essential patents (SEPs) and find that contrary to the discussion on standards in Shapiro (2000), quality adjusted prices fall more slowly in SEP industries rather than non-SEP industries. Additionally, changes in patent law reducing the power of standard essential patents to extract hold-up did not affect innovation in SEP industries.

Finally, Galasso and Schankerman (2010) show that patent litigation takes less time to settle when the alleged infringer is using technology where patent rights are more fragmented. They find that greater patent rights fragmentation reduces settlement time per dispute, and that this effect was much larger before the establishment of the Court of Appeal of the Federal Circuit introduced more certainty over patent rights enforcement.

2.3 Measurement Literature

Ziedonis et al. (2004) develop the fragmentation index, 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.[7] The measure has been used in a variety of empirical studies on the effects of market fragmentation, although it suffers from the assumption that all citations represent the same technological similarity.

von Graeventiz et al. (2011) improve upon the fragmentation index by focusing on triads of firms that each hold patents previously used to block the patent applications of the other firms. The measure uses X and Y references from EPO data. An X reference is prior art that solely calls into question the novelty or inventive step of an application. A Y reference does the same 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 EPO patent 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 for patent thickets 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 applied as blocking prior art 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

[7] 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.

receives a rejection based upon prior art. 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 recognizes the prior art and initially submits claims z'. The examiner allows the patent application with no prior art rejection. The relationship between the resulting patent and the prior art is the same in both scenarios; however, the VG measure only captures the citation pair in the first scenario.

Gaessler, et al. (2017) uses a measure of patent thickets 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. Though 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). To the best of our knowledge, the measure used in Gaessler, et al. was first proposed by Dietmar Harhoff in a presentation at the IP Statistics for Decision Makers Conference [8], but a working paper for the measure is not yet available. Finally, Fischer and Ringler (2015) extend the VG measure by broadening the attention beyond blocking patents; however, by using all patent citations to construct the triads, the measure suffers from the same criticism as the fragmentation index. In the next section, we describe the methodology we apply for constructing a new measure of patent rights fragmentation that alleviates the disadvantages of the existing measures in the literature that do not account for technological similarity.

3 Methodology

3.1 Patent Thicket Measure

The core building blocks for our patent thicket measure are patent citations with each citation weighted by the technological 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 gives an indicator of the likely contracting inefficiencies associated with the patent thicket. 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 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.

von Graevenitz et al (2011) discusses the importance of only including triads with distinct patent owners. The complements problem and the potential for patent hold-up are more severe

[8] (`https://www.oecd.org/site/stipatents/1_2_Harhoff.pdf`)

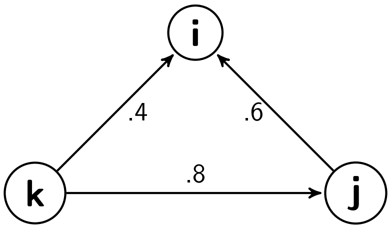

Figure 1: Example of a Triad.

with multiple patent owners since it is more difficult to negotiate multi-lateral than bi-lateral licensing agreements. The next definition formalizes the notion of a triad weighting function.

Definition 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

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

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

Definition 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 thicket 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 thicket measure at the global technology level as the weighted triads in the technology summed and normalized.

Definition 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 thicket 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 technology thicket measure. Generally, the global technology thicket 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 technological similarity for measuring patent thickets.

Finally, any moment from the distribution of the local patent thicket measures for each patent owner or inventor could be used to define a thicket measure at the patent owner or inventor level. We leave these notions without precise definitions since the form may depend on the particular study.

3.2 Technological Similarity

Citations are listed on the face of a patent if they provide the definition of a term-of-art, supply clarification that is somewhat pertinent to the claims, or were extraneously noted by the applicant on their Information Disclosure Statement (IDS) but were still reviewed by the examiner [9]. 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 standard natural language processing techniques to quantify the technological similarity between citing and cited patents.

Specifically, we algorithmically compare the word-frequencies (i.e., “bag-of-words” or “multisets”) from patent claims text to compute the technological similarity between two patent documents. We calculate word-frequencies to include the words in each of the independent and dependent claims. The patent claims precisely define the content and scope of the invention, so that similarity of claims more precisely measures inventive overlap than using other portions of the disclosure. For example, two patents may overlap in their detailed descriptions by sharing a similar background but assert completely different inventions as formally defined in the claims.

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 common words in many patent documents, we use a standard “Term-Frequency-Inverse-Document-Frequency” (tfidf) approach (Salton and McGill 1986). Specifically, we pre-compute 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 tfidf, we multiply the idf factor by the term frequencies in each patent document.

We quantify the technological similarity between two patent documents as the cosine similarity between the tfidf vectors derived from their patent claims[10]. These technological similarities are then incorporated into the measurement of patent thickets through the weighting of triads of

[9] See the USPTO’s MPEP 1302

[10] Formally, Let t be a particular term in the corpus and D be the set of documents in the corpus. Define the term frequency of term t in document d ($tf(t,d)$) to be the number of times t occurs in d. The inverse document frequency of term t in D is

citations. The next section describes the empirical methodology used to validate the thicket measure and to highlight the additional information the technological similarity weighted citations provides for the measurement of complexity and patent thickets.

3.3 Empirical Methodology

We first validate our patent thicket measure using methods familiar to the literature (von Graevenitz et al. 2011). According to Cohen et al (2000), products in complex technologies are comprised of many patents while products in discrete technologies contain very few. For example, smart phones are comprised of thousands of patents and are classified as complex technologies. On the other hand, pharmaceutical drugs contain relatively few patents and therefore are discrete. Complex technologies are more susceptible to patent thickets than discrete technologies, and increased complexity over time has contributed to the growth of patent thickets (Hall et al 2013). In this first validation, we compare the patent thicket measure across discrete and complex technologies using the Cohen et al. (2000) classification. To do so, we first normalize our global technology thicket 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 in the technology. Any difference in our measurement of patent thickets remaining between discrete and complex technologies can be attributed to differences in the technological similarity of patent claims among citations within those technologies.

Next, we compare our patent thicket measure to a variety of USPTO patent examination characteristics; specifically, post-first action[11] patent application pendency, examiner search intensity and USPTO examination complexity factors. Post-first action patent application pendency is the length of time an application spends in the patent office after the initial response from the examiner (called the first action). Examiner search intensity is the amount of time the examiner spent searching prior art during examination and is proxied by the number of search pages in the first action. For the final validation, we compare our patent thicket 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 [12] (Marco et al.

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

Finally, the term frequency inverse document frequency of term t in document d given D is

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

Let a and b be the term frequency inverse document frequency vectors for document A and document B given corpus D. The term frequency inverse document frequency cosine similarity between document A and B is given by

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

[11] The term first action always refers to the first office action as defined in the footnote above

[12] Complexity factors are scalars that reflect the underlying level of complexity for all technology examined in a particular U.S. Patent Classification (USPC) class-subclass combination. A higher complexity factor indicates that

2017). Our local patent thicket measure should be positively correlated with greater technological complexity (Hall, et al. 2013). Therefore, our measure should also be 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.[13] To estimate these correlations, we run ordinary least squares on the local patent thicket measure with and without technology center/action year fixed effects. USPTO patent examiners are organized into technology centers based on the technologies they are assigned to exam. We run additional regressions with more granular technological fixed effects (USPC, etc.), however these are left unreported since the results were consistent with technology center estimations.

Finally, we assess the inventive content of technological similarity for measuring complexity and patent thickets. Recall that one disadvantage of only using blocking citations to construct a thicket measure is that patent applications initially written properly will not receive a block, therefore similar yet unblocking art will not be included in the computation. Whether or not the patent application received the blocking rejection and then modified the claims away from the art, or submitted appropriate claims initially should not impact the measurement of technological complexity and patent thickets. 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 complex technologies by searching more prior to filing. Therefore, one cannot a priori assume that blocking rejections are more likely in complex technologies. Finally, the examiner may have a variety of art to reject the claim and therefore might omit some or have insufficient time to fully search all prior art (Frakes and Wasserman 2017, Lei and Wright 2017).

To assess the informational coverage of technological similarity for measuring patent thickets, we compare the distributions of patent claim technology similarity for blocking versus non-blocking citations. The degree of overlap between these two distributions indicates the amount of information lost by *only* using blocking patents to measure overlapping rights. In particular, non-blocking citations that are just as technologically similar as blocking citations are not included in the VG measure. Secondly, we run ordinary least squares regressions to estimate the impact of the patent thicket measure on the probability of receiving a prior art rejection on the first office action at the USPTO.[14] An insignificant estimate on the marginal effect of the patent thicket measure would

an examiner should be given more time to balanced disposal (USPTO term for application completion]). The main portion of an examiner's production (or output) goal calculation is

$$\frac{\text{Number of Examining Hours} \cdot \text{Seniority Factor}}{\text{Complexity Factor}}$$

This formula provides the number of counts an examiner must complete in the pay period to meet the production quota. Counts are given for specific milestones in the patent examination process. The seniority factor adjusts the production quota so that more experienced examiners are required to do more work. For more information see Marco et al. 2017.

[13] There may exist simultaneity in these simple models. For example, an examiner receives more time to examine an application in a more technologically complex art, giving the examiner additional time to perform a more thorough prior art search and cite additional relevant literature. Therefore, all else equal, the increase in examination time could lead to an increase in the patent thicket measure.

[14] 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.

indicate that our measure contains additional information beyond that captured in blocking patents.

4 Data

To compute our patent thicket 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 claim text from the Patent Claims Research Dataset (Marco et al. 2017), citations and patent owners from PatentsView[15], 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. We compute our local thickets measure for each unexpired patent between the years 2000 and 2014. Note that the local thicket measure changes year over year since the network of citations evolves with new forward citations and expiring patents.

Post-first action application pendency is available in the OCE's PatEx dataset (Graham et al. 2015). However, we extract the variable from the USPTO PALM (Patent Application Location Monitoring) 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. Lastly, we utilize data on blocking patents from the OCE's Office Actions dataset (Lu et al. 2017).

5 Results

5.1 Discrete v. Complex Technologies

This section describes the results of the three validation tests for our patent thicket measure. Figure 2 compares our patent thicket measure across discrete and complex technologies, with and without various normalizations. Recall that our global technology thicket measure is increasing in the average number of citations, the number of patents in force[16] and the similarity between patents. The top left panel in figure 2 shows that the non-normalized global technology patent thicket measure[17] 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 thicket measure normalized by patent volume displays a smaller yet still growing gap between complex and discrete technologies over

[15] www.patentsview.org

[16] A patent is in force if the patent owner has not let the patent expire by paying all required maintenance payments or has not exceeded the statutory term of a patent. We retrieved patent expiration data provided from the Historical Patent Data Files (Marco, et al. 2015). The Marco, et al. (2015) expiration date calculations are somewhat incomplete as the authors do not account for all possibilities related to the transition from in force to expired patent (e.g. invalidation, etc.). However, these inconsistencies are relatively few in number.

[17] The aggregate of all weighted triads in the technology.

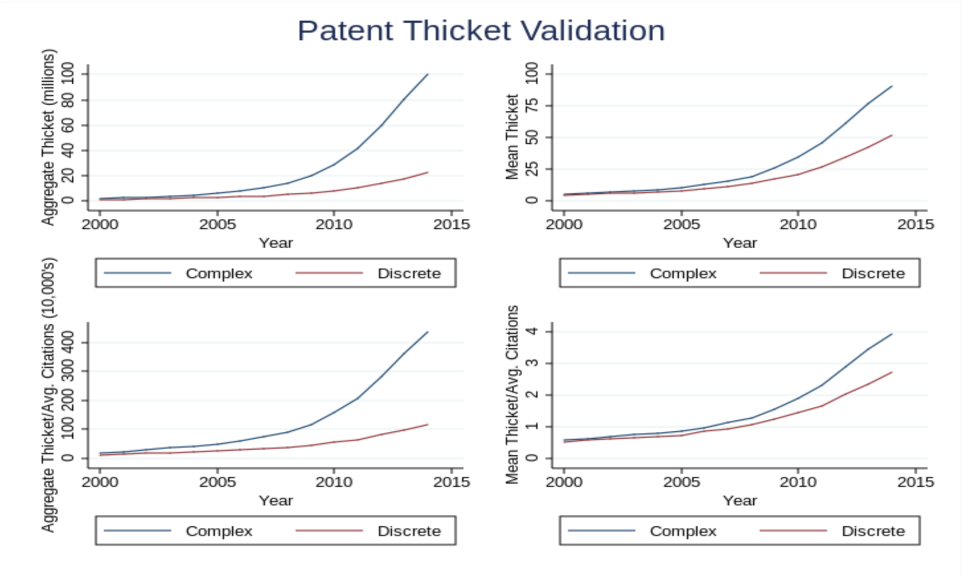

Figure 2: Complex v. Discrete Technologies: Patent thickets are more prevalent in complex technologies. The aggregated (upper-left) and mean (upper-right) patent thicket measure is plotted by year for both discrete and complex technologies. The aggregated (lower-left) and mean (lower-right) patent thicket 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 thicket measure in complex technologies is higher than discrete technologies. Further, technological similarity contains information for distinguishing complex v. discrete. (bottom-right)

time. This is our preferred measure of patent thickets since it controls for the size of the technology space.

In order to identify the degree to which technological 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 in force and the average number of citations per patent. The bottom left pane shows a similar trend when normalizing by the average number of citations per patent. The bottom right pane normalizes the patent thickets measure by both patent volume and average number of citations. The gap between discrete and complex technologies mostly persists[18]. Recall that, after controlling for patent volume and average citations, any difference in our measurement of patent thickets remaining between discrete and complex technologies can be attributed to differences in the technological 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 technological similarity effect.

Overall, initial results support the validity of our measure for assessing patent thickets in complex versus discrete technologies and our assertion that technological similarity, as captured by patent claim similarity, drives some of the persistent and growing gap in fragmentation between

[18] The patent thicket measure is higher in complex technologies over discrete technologies in every year but the first.

VARIABLES	(1) Complexity	(2) Complexity	(3) Search	(4) Search	(5) Pendency	(6) Pendency
Thicket Measure	-4.09e-05***	2.41e-05***	2.73e-05***	2.66e-05***	2.06e-05***	2.02e-05***
	(2.70e-06)	(2.32e-06)	(1.59e-06)	(1.57e-06)	(9.22e-07)	(8.12e-07)
Constant	22.57***	21.42***	1.113***	0.908***	5.782***	6.210***
	(0.00271)	(0.294)	(0.000747)	(0.00453)	(0.000595)	(0.00249)
Observations	2,752,440	2,752,440	1,383,730	1,383,730	1,354,069	1,354,069
R-squared	0.000	0.563	0.001	0.011	0.001	0.156
Action Year FE	No	Yes	No	Yes	No	Yes
TC FE	No	Yes	No	Yes	No	Yes

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 1: Patent Examination Regression Results. Complexity is for complexity factor, search is for examiner search intensity, and pendency is for post first action pendency.

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 patent thicket measure and USPTO application pendency, examiner search intensity, and USPTO performance measurement complexity factors.[19] Recall that through technological complexity, we expect the patent thicket measure to be positively correlated with all three dependent variables. Specifically, 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 patent thicket measure and these variables then would further validate our thicket measure.

Table 1 reports the regression results. For the application pendency and examiner search intensity regressions, the coefficient on the local patent thicket measure is positive and significantly different than zero at the one percent level, both with and without technology and first action year fixed effects.[20] For the USPTO complexity factors, the coefficient on the patent thicket measure is negative and statistically different than zero at the one percent level without fixed effects. However, it is positive and significant with technology and first action fixed effects. Given the potential for varying citation tendencies over time and across technologies, we prefer the regressions that control for technology and first action year. On the whole, the positive correlations between the patent examination variables and the patent thicket measure provide further support for the validity of our measure.

[19] For each regression, we only include applications with a "first action on the merits" date between 2008 and 2014.

[20] First action year fixed effects were chosen since the search report used to proxy for examination search intensity was from the first office action, and the pendency variable is overall pendency post first action.

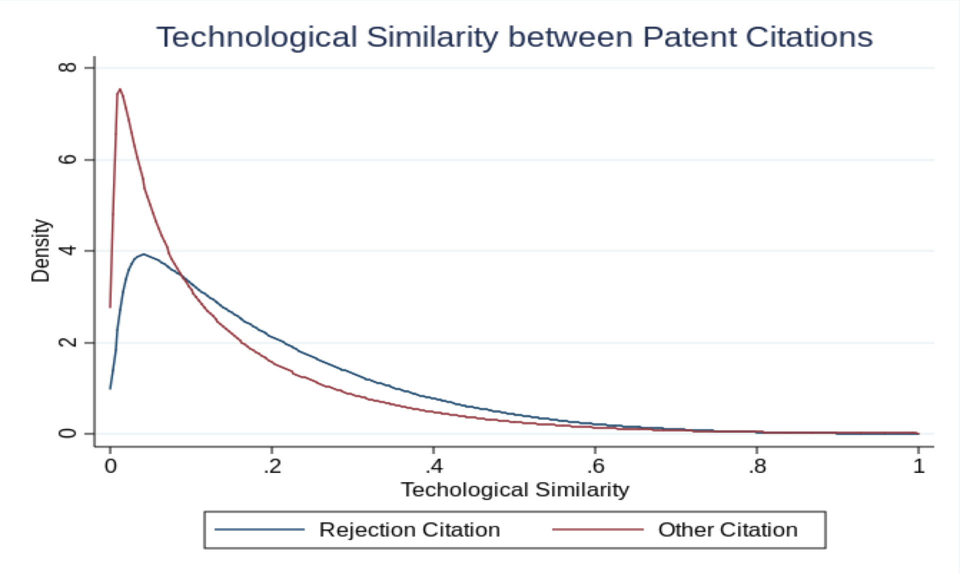

Figure 3: The Informational Content of Patent Citations: Rejection citations are those used in prior art rejections at the USPTO. Other citations are those citations listed on the face of the patent that were not used in prior art rejections. Although the distribution of technological similarity for 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 in the measurement of patent thickets.

5.3 Informational Content of Patent Citations

This section further explores the informational coverage of technological similarity for the measurement of patent thickets. As discussed earlier, one disadvantage of relying exclusively on blocking patents to measure fragmentation 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 would have modified the claims an appropriate distance from the prior art upon initial submission to the USPTO, then the application would not receive a prior art rejection. The patent thicket measure by von Graevenitz et al. (2011) does not count the latter citation as a blocking patent, even though the application in both scenarios result in the same patent. Additionally, examiners may not record all blocking patents in the rejection for a variety of reasons.

Figure 3 displays the distributions of technological 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 technological 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. A measure for technological complexity and patent thickets that only uses blocking citations loses all of this additional information.

Table 2 displays the regressions results for the probability of receiving a 102/103 prior art

	(1)	(2)	(3)	(4)
VARIABLES	102/103 Rejection	102/103 Rejection	102/103 Rejection	102/103 Rejection
Thicket Measure	2.87e-06***	-2.73e-08	2.04e-06***	-6.77e-07
	(4.75e-07)	(4.39e-07)	(4.75e-07)	(4.41e-07)
Constant	0.421***	0.0188***	0.348***	-0.0600***
	(0.000427)	(0.000313)	(0.00152)	(0.00139)
Observations	1,346,077	1,346,077	1,346,077	1,346,077
R-squared	0.000	0.213	0.005	0.219
Action Year FE	No	Yes	No	Yes
TC FE	No	No	Yes	Yes

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 2: 102/103 Rejection Regression Results.

rejection at the USPTO [21] for all combinations of technology and first action year fixed effects. If the coefficient on the patent thicket variable is significant, then a densely-thicketed technology space is more prone to 102/103 rejections, implying that much of the information captured in our thicket measure is already captured by blocking citations. Alternatively, an insignificant estimate would indicate that technological similarity contains additional information beyond that captured by thicket measures that rely on blocking patents alone. The value of this additional information is motivated by the theoretical arguments pertaining to the endogeneity of patent examination. However, we empirically validate the differences in informational content but do not directly assess the value of the this additional information.[22]

The coefficient on the local thicket measure is positive and significant in the two regressions that do not control for first action year, and insignificant when first action year fixed effects are included. Again, we prefer the regressions that include a full set of fixed effects given the potential for citation variation unrelated to technological complexity and patent thickets across technologies and time. For this reason, the regression results enforce the notion that technological similarity contains information on inventive overlap not captured by thicket measures computed on blocking patents. This result is also consistent with our argument regarding the endogeneity of the patent examination process. Conditioning patent thickets to only rejection citations omits technological similar citations the applicant fully considered and appropriately distanced their claims from prior to filing at the USPTO. For the purpose of measuring fragmentation, such technological similar citations are equivalent to the blocking citations and therefore should be included in the measurement of patent thickets. The measure defined in this paper fully incorporates this additional information.

[21] 102 rejections are given for non-novelty and 103 rejections are given for non-obviousness. See `https://www.uspto.gov/web/offices/pac/mpep/mpep-9015-appx-l.html#al_d1fbe1_234ed_52` for more details.

[22] If one were to calculate the patent thicket measure based only on blocking patents, this patent thicket measure would be positively related to the probability of receiving a 102 or 103 rejection (by construction). Any patent without a 102 or 103 rejection would have a patent thicket measure of zero. Further, if a patent only had one 102 or 103 rejection, the patent thicket measure would be zero as well.

6 Conclusion

This paper defines a measure for technological complexity and patent thickets that improves upon existing measures in the literature by incorporating the technological similarity, as captured by patent claim similarity, of each citation. Various tests support the validity of our measure for assessing patent thickets in complex versus discrete technologies and the notion that technological similarity captures relevant information omitted by blocking-based thicket measures.

Additionally, our measure is universally computable for all patent systems and will enable novel empirical research regarding intellectual property fragmentation. A consistent measure computable across jurisdictions will enable cross-country comparisons in technological complexity and patent thickets. Further research will use the thicket measure defined here, and weighted patent similarities more broadly, to address these questions.

References

[1] Bessen J., Maskin E., Sequential Innovation, Patents, and Imitation, *The RAND Journal of Economics*, 2009

[2] Borgatti S., Mehra A., Brass D., Labianca G., Network Analysis in the Social Sciences, *Science*, 2009

[3] Cohen W., Nelson R., Walsh J., Protecting Their Intellectual Assets: Appropriability Conditions and why U.S. Manufacturing Firms Patent (or not), 2000

[4] Clarkson G., Objective Identification of Patent Thickets: A Network Analytic Approach, 2004

[5] Clarkson G., DeKorte D., The Problem of Patent Thickets in Convergent Technologies, *The Problem of Patent Thickets in Convergent Technologies*, 2006

[6] Cockburn I., MacGarvie J., Patents, Thickets and the Financing of Early - Stage Firms: Evidence from the Software Industry, *Journal of Economics and Management Strategy*, 2009

[7] Cockburn I, MacGarvie M., Muller E., Patent Thickets, Licensing and Innovative Performance, *Industrial and Corporate Change*, 2010

[8] Fischer T., Ringler P., The Coincidence of Patent Thickets - A Comparative Analysis, *Technovation*, 2015

[9] Frakes M., Wasserman M., ”Is the Time Allocated to Review Patent Applications Inducing Examiners to Grant Invalid Patents? Evidence from Microlevel Application Data”, Review of Economics and Statistics, 2017

[10] Federal Trade Commission, ”To Promote Innovation - The Proper Balance of Competition and Patent Law and Policy”, Washington, DC., 2003

[11] Galasso A., Schankerman M., Patent Thickets, Courts, and the Market for Innovation, *RAND Journal of Economics*, 2010

[12] Gambardella A., Harhoff D., Verspagen B., "The Value of European Patents", European Management Review, 2008

[13] Graham S., Marco A., Miller R., "The USPTO Patent Examination Research Dataset: A Window on the Process of Patent Examination", SSRN, 2015

[14] Hall B., Exploring the Patent Explosion, *The Journal of Technology Transfer*, 2004

[15] Hall B., Helmers C., Rosazza-Bondibene C., A Study of Patent Thickets, *UKIPO*, 2013

[16] Hegde D., Mowery D., Graham S., Pioneering Inventors or Thicket Builders: Which US Firms Use Continuations in Patenting?, *Management Science*, 2009

[17] Holland P., Leinhardt S., Local Structure in Social Networks, *Sociological Methodology*, 1976

[18] von Graevenitz G., Wagner S., Harhoff D., How to Measure Patent Thickets - A Novel Approach, *Economic Letters*, 2011

[19] von Graevenitz G., Wagner S., Harhoff D., Incidence and Growth of Patent Thickets: The Impact of Technological Opportunities and Complexity, *Journal of Industrial Economics*, 2013

[20] Jackson M., Social and Economic Networks, 2010

[21] Kortun S., Lerner J., "Stronger Protection or Technological Revolution: What is Behind the Recent Surge in Patenting?", *Carnegie-Rochester Conference Series on Public Policy*, 1998

[22] Kuhn J., Younge K., Marco A., "Patent Citations Reexamined: New Data and Methods", SSRN, 2017

[23] Lei Z., Wright B., "Why Weak Patents? Testing the Examiner Ignorance Hypothesis", Journal of Public Economics, 2017

[24] Lu Q., Myers A., Beliveau S., "USPTO Patent Prosecution Research Data: Unlocking Office Action Traits", SSRN, 2017

[25] Marco A., Carley M., Jackson S., Myers A., "The USPTO Historical Patent Data Files: Two Centuries of Innovation", SSRN, 2015

[26] Marco A., Sarnoff J., deGrazia C., "Patent Claims and Patent Scope", SSRN, 2017

[27] Marco A., Toole A., Miller R., Frumkin J., "USPTO Patent Prosecution and Examiner Performance Appraisal", SSRN, 2017

[28] Meyer M., Patent Citations in a Novel Field of Technology - What can they tell about Interactions between Emerging Communities of Science and Technology?, 2000

[29] Mossoff A., A Stitch in Time: The Rise and Fall of the Sewing Machine Patent Thicket, 2009

[30] Mossoff A., The Rise and Fall of the First American Patent Thicket: The Sewing Machine War of the 1850's, *Arizona Law Review*, 2011

[31] Pairolero N., "Identifying Innovation in Thickets: A Network Approach", *SSRN*, 2016

[32] Pairolero N., "Pricing in Complex Networks", *ProQuest*, 2016

[33] Roach M., Cohen W., "Lens or Prism? Patent Citations as a Measure of Knowledge Flows from Public Research", Management Science, 2013

[34] Salton G., Mcgill M., "Introduction to Modern Information Retrieval", McGraw-Hill Inc., 1986

[35] Noel M., Schankerman M., Strategic Patenting and Software Innovation, *The Journal of Industrial Economics*, 2013

[36] Shapiro C., Navigating the Patent Thicket: Cross Licenses, Patent Pools, and Standard Setting, *Innovation Policy and the Economy*, 2001

[37] Somaya D., "Strategic Determinants of Decisions not to Settle Patent Litigation", Strategic Management Journal 24, 2003

[38] Sternitzke C., Bartkowski A., Schramm R., Visualizing Patent Statistics by Means of Social Network Analysis Tools, *World Patent Information*

[39] USPTO, "First Office Action Estimator", `https://www.uspto.gov/learning-and-resources/statistics/first-office-action-estimator`, 2018

[40] USPTO, "Manual for Patent Examining Procedure", `https://www.uspto.gov/web/offices/pac/mpep/index.html`, 2015

[41] USPTO, "37 CFR 1.104 Nature of Examination", `https://www.uspto.gov/web/offices/pac/mpep/s707.html`, 2018

[42] Watts D., The New Science of Networks, *Annual Review of Sociology*, 2004

[43] Webb C., Dernis H., Harhoff D., Hoisl K., Analysing European and International Patent Citations: A Set of EPO Patent Database Building Blocks, 2005

[44] Younge K., Kuhn J., "Patent-to-Patent Similarity: A Vector Space Model", SSRN, 2016

[45] Ziedonis R., Don't Fence me in: Fragmented Markets for Technology and the Patent Acquisition Strategies for Firms, *Management Science*, 2004