Semantically-Based Patent Thicket Identification

Mateusz Gątkowski[1], Marek Dietl[2], Lukasz Skrok[2], Ryan Whalen[3], Katharine Rockett[1]

Abstract

Patent thickets have been identified as a major stumbling block in the development of new technologies, creating the need to accurately identify thicket membership. Various citations-based methodologies (Graevenitz et al, 2011; Clarkson, 2005) have been proposed, which have relied on broad survey results (Cohen et al, 2000) for validation. Expert evaluation is an alternative direct method of judging thicket membership at the individual patent level. While this method potentially is robust to drafting and jurisdictional differences in patent design, it is also costly to use on a large scale. We employ a natural language processing technique, which does not carry these large costs, to proxy expert views closely. Furthermore, we investigate the relation between our semantic measure and citation based measures, finding them quite distinct. We then combine a variety of thicket indicators into a statistical model to assess the probability that a newly added patent belongs to a thicket. We also study the role each measure plays, as part of creating a prospective screening model that could improve efficiency of the patent system, in response to Lemley (2001).

Keywords: Patent Thicket, Intellectual Property, Semantic Distance, Latent Semantic Analysis, Natural Language Processing, Complexity
JEL Classification: L13, L20, O34

[1] Dept. of Economics, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ
mgatko@essex.ac.uk; kerock@essex.ac.uk
[2] Department of Business Economics, Collegium of World Economy, SGH Warsaw School of Economics, al. Niepodległości 162, 02-554 Warsaw
mdietl@sgh.waw.pl; lskrok@sgh.waw.pl
[3] University of Hong Kong, Faculty of Law, Cheng Yu Tung Tower, Pokfulam, Hong Kong
whalen@hku.hk

1. Introduction

Patent thickets—where multiple entities have intertwining intellectual property rights over related technologies—can act as stumbling blocks in the development of new technologies as they raise the spectre of increased licensing costs, increased innovation transaction costs, and ultimately suboptimal levels of innovation.[4] Meanwhile, drafting problems resulting from a "lack of resources and misaligned incentives at patent offices dealing with a flood of patents" (Hall et al., 2013)[5,6] can lead to "junk patents" that can in turn contribute to thickets (Holman, 2006). While some might be tempted to rely on litigation or the patent fee structure to weed out low-quality patents, recent papers by Schankerman and Schuett (2018) and Frakes and Wasserman (2019) cast doubt on the feasibility of these approaches. Given concerns about the consequences of thickets and their proliferation, methods to better understand them and identify the risk that a particular patent might become part of one is an important task for scholars of intellectual property and innovation policy, patent portfolio managers, and would-be inventors.

Many patent thicket detection methods have been suggested. By and large, these are based on citations as proxies for underlying linkages among patents.[7] Graevenitz et al. (2011) recognise the need to demonstrate the external validity of these proxies, and do so by noting the correspondence between the occurrence of triples—where three firms' patents have bilateral citations—and the complexity of the technology, identified by Cohen et al. (2000) in surveys of managers. Indeed,

---

[4] See Egan and Teece (2015), for a recent review of this extensive literature.

[5] See WIPO (2013, 2014, 2015) for recent details on this steep rise. See Barnett (2014) for further discussion of patent interference and "junk patents". Lemley (2001) finds that overall time spent per application at the United States Patent and Trademark Office (USPTO) is about 18 hours spread over the months of the granting process.

[6] Consistent with the interpretation that thickets result from low quality patent review, Lemley and Shapiro (2005) note that "when patents are granted covering technologies that were already known or were obvious, the resulting patents could cause social costs without offsetting benefits"; however, they also propose a more strategic interpretation, noting that patent thickets result when, "companies fil[e] numerous patent applications on related components that are integrated into a single functional product". This can create the opportunity for royalty requests or for outright blocking of technology development, as proposed by Heller and Eisenberg (1998).

[7] Established ways of measuring patent thickets have relied on qualitative methods such as interviews with executives on patenting strategies (Hall and Ziedonis 2001) or examining prior art citations and their fragmentation (concentration) as measured by a Herfindahl index, HHI (Ziedonis, 2004; Galasso and Schankerman, 2010). Others, such as Clarkson (2005) and Clarkson and De Korte (2006) suggest calculating measures based on citation network density, while Graevenitz et al. (2011) suggest identification using critical prior art references and calculating the density of "triples", which are specific subnetworks of these references. DeGrazia et al (2019) point out, however, that citations practices can introduce substantial "noise" into this measure, as they are also susceptible to drafting errors (excluding links across patents that have not been detected or including irrelevant links). In defence of citations, however, they do form real legal linkages among patents, so they have validity on their own as detection measures.

those surveys suggest that more complex technologies are more likely to give rise to thickets, which Graevenitz et al. detect using their methodology.

Expert opinion based on a careful reading of the patent and knowledge of the technology offers an alternative way to detect thickets and to validate other proposed thicket measures. Indeed, expert-curated patent thicket maps can provide an arguably more nuanced comparison than groupings based on technology characteristics such as complexity, as they provide tailored assessment at the individual patent level. At the same time, expert evaluation is costly and time consuming. In the first part of this paper, we evaluate natural language processing as a way to replicate expert views, finding that it proxies these evaluations closely and identifies the expert-identified thickets in a statistically significant way. This finding demonstrates a sophisticated semantic technique that allows us to capture the patent linkages that are revealed by experts but that may not have been previously detectable given budgetary and time constraints. It also provides us with a source of external validation for both semantic analysis and other methodologies of thicket detection.

In the second part of the paper, we compare alternative citation-based measures to the semantic approach. The definition of thickets guiding our experts, while standard, is broader than that underlying citation-based measures. As a result, we expect some variation across methodologies. We find that, while all measures are informative of expert views, statistical testing shows that the various proposed patent thicket measures are very distinct, suggesting that they capture very different things. We next investigate the prospective evaluation of the propensity of an individual patent to be part of a thicket, rather than identifying groups of technologies that are prone to thicketing, as in some earlier work. As such, the emphasis of our work is complementary to citation-based approaches to thicket detection. To implement our evaluation, we combine the different measures into a single framework for detecting thickets and generating predictions of the likelihood that a specific patent will belong to a thicket. We propose this latter model as an answer to Lemley's (2001) appeal for such an *ex ante* "screen" to weed out thickets before the patent is granted in the first place.

Previewing our results in slightly more detail, we find that semantic distance among patents changes systematically depending on the membership of any pair of patents in expert-identified

thickets. Patents in the same thicket tend to be semantically similar. Moreover, the average semantic distance between these combinations of thicket membership differ in a statistically significant way from one another, making it possible to identify thicket membership and furthermore supporting semantic analysis as an accurate way to replicate expert assessment. Finally, we find that the semantic distance between patents within the same thicket in discrete technological areas is shorter than it is for complex technologies, making the difference between thicketed and non-thicketed patents greater for discrete technologies. This suggests that semantic distance may be a more powerful tool to identify thickets in discrete technologies, and also accords with the intuition that in complex technologies patent thickets cover a wider range of patent claims and so are more diverse.

We then statistically compare the degree to which expert opinion is captured by semantic distance or, alternatively, citation-based measures. Doing so reveals that each of these measures is broadly informative of expert opinion, but our semantic implementation and the citation-based measures lead to quite distinct results. This suggests that semantic similarity may be a useful tool to identify patent thickets, but not necessarily the same thickets as those identified by citation measures. This introduces the possibility of combining the different measures to provide a more accurate evaluation of thicket membership. We propose a logit model incorporating all measures alongside controls to implement this idea. In addition to illustrating the power of citations and semantically-based thicket measures in this prediction, our implementation also captures many of the features found in earlier work on the roles of fragmentation, crowding, and complexity in thicket formation.

In the remainder of this paper we introduce our methodology in section 2. Section 3 establishes that our semantic analysis corresponds well with expert views. Subsections detail this and various extensions of the results. Section 4 moves on to a comparison of our experts' classification with the results of various citation-based methods. Because this analysis suggests substantial differences, we then combine citation and semantic-based methods in our final predictive model of section 5. Section 6 presents our conclusions from this set of exercises and outlines further work that could deepen our preliminary explorations.

2. Using Semantic Analysis to Represent Expert Views

2.1 Sample Selection and Expert Evaluation

To determine the way experts view patent thicket membership, we use data from the USPTO on 11,872 patents from 58 patent groups (subclasses within the United States Patent Classification – USPC – technology classification scheme), sampled as of the end of February 2015. The dataset contains the full text and bibliographic data of the patents, including data on the filing company, application and granting dates and the number of claims.

We selected a group of eight technical subject matter experts and asked them to review patents in the 58 patent groups under study.[8] We provided each expert with the subset of patents from our original sample that were related to his or her field of expertise. The experts subsequently examined all of their assigned patents and, where they deemed appropriate, identified thickets of patents falling within the provided definition.

Because of the cost and time required to have experts identify patent thickets, the set of patent groups we examined was not comprehensive. Rather, we conducted an initial selection of groups that included different forward citation network structures. In particular, we covered a spectrum of degrees of centrality, as citations network structures have been identified as important to thickets.[9] We thus selected our sample using groups with varying citation tendencies, but did not attempt a random sample. Instead, the advantage of our sample is the high-quality evaluation by experts in the groups we included. We anticipated that the evaluation process would generate a number of clarification questions, so we chose experts who could keep in good contact with our help desk. The result was a sample that spanned enough variety in citation structure and technological complexity that we are reassured that our results on semantic analysis as a method of capturing expert views does not depend upon any specific technological or citation feature.

One of the core difficulties involved in such an exercise is to give experts instructions that enable them to reliably detect what the researcher means by a "patent thicket." Here, the literature provides mixed guidance. Patent thickets have been variously referred to as "blocking patents",

[8] We used experts in the fields of electrical systems, chemical engineering, material engineering, electricity (measuring and testing), electrolytic coating, nanostructures, dentistry, drugs, medical chemistry, surgery, and image processing.
[9] Our initial investigation of centrality did not yield insights that added to the main points already made in this piece, so in the interest of brevity we exclude any further discussion of this issue. The point is that our groups, while not representing a random sample, do represent a variety of relevant characteristics.

"patent floods", or "patent clusters" (IPO, 2011). In a recent review of the literature, Egan and Teece (2015) detail the many different definitions that have been used and associate each definition with one or more of seven distinct policy concerns.

One common definition of a patent thicket, taken as a starting point in the Egan and Teece paper and corresponding to four of their seven policy concerns, is "an overlapping set of patent rights requiring that a company must hack its way through in order to actually commercialize new technology" (Shapiro, 2001). The popular citation-based measure demonstrated by Graevenitz et al. (2011) is based on a slightly broader view that "the combination of complex technology and high-volume patenting creates patent thickets, which can be defined as dense webs of overlapping patent rights."

The definition of patent thicket we provided to our experts was a modification of the Shapiro (2001) definition: "Patent thickets are dense webs of overlapping intellectual property rights owned by one or more different companies (patent owners), which create a potential high cost in commercializing a new technology, and this cost is difficult to assess upfront." This definition's incorporation of additional language into the Shapiro definition was inspired by feedback from our domain experts about the shorter original definition. It allows thicketed patents to be associated with a single firm, although it requires a high cost of commercialisation to result from the thicket. As a result, defensive as well as hold up reasons for thicket generation can potentially be included in our expert identified thickets. Hence, we allow for the full scope of issues that Egan and Teece flag as being evoked by this definition.[10]

Upon completing their review of the patents within each of their technical areas, the subject matter experts identified 303 patent thickets containing 2,615 patents. In our sample the density of patent thickets (defined as the percentage of patents belonging to thickets from the full sample of patents) is 22%. On average each thicket contains 8.9 patents.

2.2 Semantic Analysis

[10] See Table C.2, Appendix C for the questions posed to the experts.

With this expert-identified set of patent thickets in hand, we then set out to measure the semantic distance between patents that they had identified as thicketed and non-thicketed.[11] Document similarity can be measured with a wide variety of methods (Harispe, Ranwez, Janaqi, & Montmain, 2015). Many of the most common rely on document metadata to infer content and perform similarity comparisons. In the context of patents this is often done by categorizing inventions according to their technical classification (e.g. United States Patent Classification or International Patent Classification) and treating inventions of the same category as similar (see e.g., Fleming, 2001). However, classification-based approaches are by definition coarse and do not allow for precise similarity measurements between documents. To achieve more precise similarity measures, one must look at the contents of documents and compare them to one another.[12]

Latent Semantic Analysis (LSA)[13] offers a well-established method that uses document contents to detect latent similarities between texts and thereby allows pairwise similarity measurements between documents (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990; Landauer, Foltz, & Laham, 1998). It does so by leveraging word co-occurrence to infer "concepts" or "topics" that the documents discuss (e.g. "car" and "automobile" are different words but refer to the same concept and thus co-occur with a similar set of vocabulary). This allows documents to be represented as a mix of concepts, providing a reduced-dimensional vector representation of the documents which can more accurately and efficiently identify similar or dissimilar documents

[11] A limited number of other papers have performed semantic analysis on patents. After Yoon and Park's (2004) initial work on keywords, Gerken and Moehrle (2012) used semantic analysis to detect novelty, Preschitschek et al. (2013) used it to study technology convergence, Khun and Thompson (2017) used word counts to analyse patent scope, and Bergeaud et al. (2019) classified patent technologies using the semantic content of patent abstracts. Whalen (2018) uses semantic citation distance measures to illustrate an increase in "boundary spanning" inventions and identify the challenges they raise for patent offices. Closest to our work, DeGrazia et al. (2019) apply semantic analysis to weight "triples" in their study on "vertical overlap" across patents (i.e., overlap across cumulative innovation). Compared to their work, our study does not modify citations measures, but instead evaluates semantic distance as a proxy on its own for expert views and then compares and ultimately combines citations and semantic measures in a predictive model. Furthermore, while deGrazia et al. (2019) rely on the earlier survey results we have described above to validate their modified measure, our emphasis is on using bespoke expert opinion to validate semantic analysis on its own.

[12] This can be done very simply by using a relatively straightforward "bag-of-words" approach (see e.g., Lang, 1995) that treats each document as the set of the words it uses, or alternately the somewhat more nuanced term-frequency-inverse-document frequency ("TF-IDF") approach, which weights words based on both their importance to the document and their frequency within the entire corpus being analysed (Salton & McGill, 1986; DeGrazia et al., 2019). These methods are, however, hampered by their inability to detect latent similarities between documents that might contain similar content, but use different vocabulary to discuss it.

[13] As readers may not be familiar with this methodology, we provide a primer on LSA in Section A.1 of Appendix A. Detailed knowledge of LSA is not required to understand this paper's main points.

(Landauer et al. 2013). Despite its advantages, to our knowledge, a semantic similarity approach like this has yet to be used in thicket detection.

Our LSA model uses a multi-step procedure where we first process all USPTO utility patents granted between 1976 and 2015 by removing very common words like "the", which are uninformative about the content of the patent, as well as misspellings and other such rare but misleading occurrences. The remaining words are then awarded a score that reflects how often the word appears in a given patent document versus how often it occurs within the entire set of patents.[14] A high score reflects a word that is very common in a specific patent but rare overall. For example, a low score might be awarded to a word like "small" since it occurs often throughout the set of patents and so is not very informative about the unique aspects of any single patent; a high score might be awarded to "nanotechnology" since this would be less common overall but might occur frequently in a patent relevant to this field. The result of this "reweighting" exercise of the terms in the patent document is then used as the inputs to the LSA model. This model uses matrix decomposition to produce a 500-dimension document–concept matrix, wherein each patent is represented by a 500 dimensional weighted "list" of concepts. The distance between each pair of patents can then be measured by taking the cosine distance between their concept vectors. Patents with a low score are considered proximate to one another within the patent topic space, suggesting they contain text describing similar technical content, while patents with high distance scores have less in common with one another.

We hypothesize that the overlapping rights indicative of patent thickets will correspond with semantic similarity between patents occupying the same thicket. To confirm whether this is the case, we benchmark our semantic similarity measures against the set of patent thickets identified by our panel of experts. This provides a degree of external validity that is uncommon in the patent thicket literature.

Our method was to use the LSA concept vectors to calculate distance scores for each patent pair within the 58 patent groups assigned to our experts[15] and to compare the average distances for four different sets of the pairs using Welch's unequal variances t-test test for mean equality[16]

[14] A "TF-IDF" score. See Appendix A for detail.
[15] There were overall more than 3.7 million patent pairs
[16] This is a version of a Student-t test. It is more robust when samples have unequal variances and sizes.

(Welch, 1951). Pairs were divided in four sets:[17] I) Same thicket – where both patents belong to the same thicket; II) Different thickets - where both patents belong to a thicket but not the same one; III) Thicket/no thicket – where only one of the patents belongs to a thicket; IV) No thicket – where neither of the patents belong to a thicket. We then undertook a variety of statistical comparisons of these groups to show that semantic similarity captures expert classification well. We now turn to the results of this work.

3. Results on Similarity and Expert-Identified Thickets

3.1 Semantic distance between patents is the shortest in the same thicket

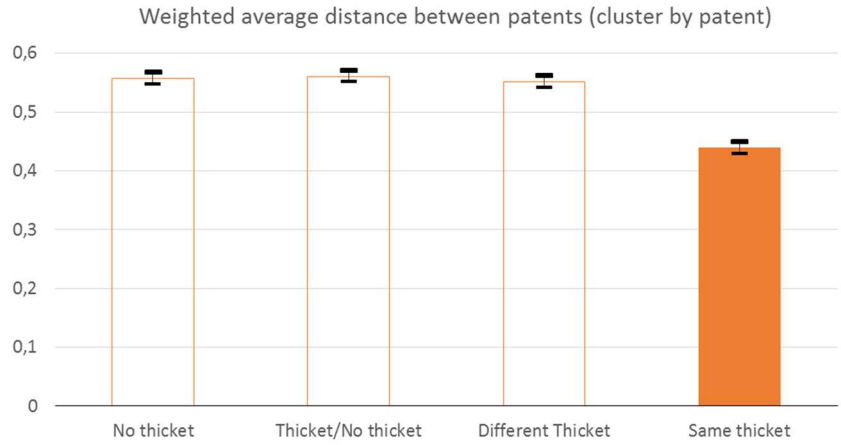
Our primary finding is that the average semantic distance between pairs of patents belonging to the same thicket is statistically different from other sets of pairs, and the result is strongly significant. This suggests that the semantic content of within-thicket patent pairs is more similar than pairs of patents that do not inhabit the same thicket, and that this difference is detectable using natural language processing techniques.

We illustrate this in Figure 1, showing the details of the calculated average semantic distance and the size of the sample. The figure presents average semantic distance between pairs of patents in each of the sets that we have just defined, calculated as the average of the distances in each of the 58 patent groups, weighted by the number of patents. We calculated the significance of differences using a linear OLS regression model with error clustering. The base scenario is set I (same thickets). Dummies were used for the remaining sets. We clustered errors by patents (the same patent could belong to more than one pair). All the coefficients were significant and positive with $p < 0.001$. In other words, the average distance in pairings outside the same thicket was significantly higher than within it. We have also tested the regression with clustering of errors by patent groups (58 clusters) and the results continued to hold.

Figure 1 illustrates these results including confidence intervals with significance level of 95% at the top of each bar.

[17] We use “set” to describe groups of pairs of patents – depending whether patents belong to a thicket or not; we use "patent group", when we refer to the USPTO patent classification.

Figure 1. Weighted average distance between patents shown by large vertical bars (errors clustered by patent), for each set. Small bars at top indicate confidence intervals (α=95%).



	Set (IV) No thicket	Set (III) Thicket/ No thicket	Set (II) Different thickets	Set (I) Same thicket
Average distance	0.558	0.560	0.552	0.439
Standard error	0.0054	0.0050	0.0051	0.0052
No of pairs in the sample	2 425 272	1 100 243	162 910	30 420

Source: Own calculations

While Figure 1 strongly suggests that patents identified as in the same thicket are semantically more similar, we confirmed that these differences were statistically significant by testing for the mean equality between sets of pairs of patents using the Welch test, which revealed that the average semantic distance between patents in set I —when both patents are from the same thicket—is significantly lower than for other sets.[18]

In order to investigate the overall results of Figure 1 in more detail, Figure 2 illustrates the average semantic differences for all patents, now broken down by patent group. This gives us an idea of the stability of the results in Figure 1 across technology groups and also allows us to track why certain groups might not exhibit a statistically significant difference in semantic distance across patents within the same thicket compared to others. Figure 2 also illustrates information that is closer to what a user would need to know: if semantic analysis were used to identify patents at risk of belonging to thickets in specific patent portfolios, then this likely would be performed for certain technology groups only.

[18] For results, see Table B.1 and accompanying discussion in Appendix B.

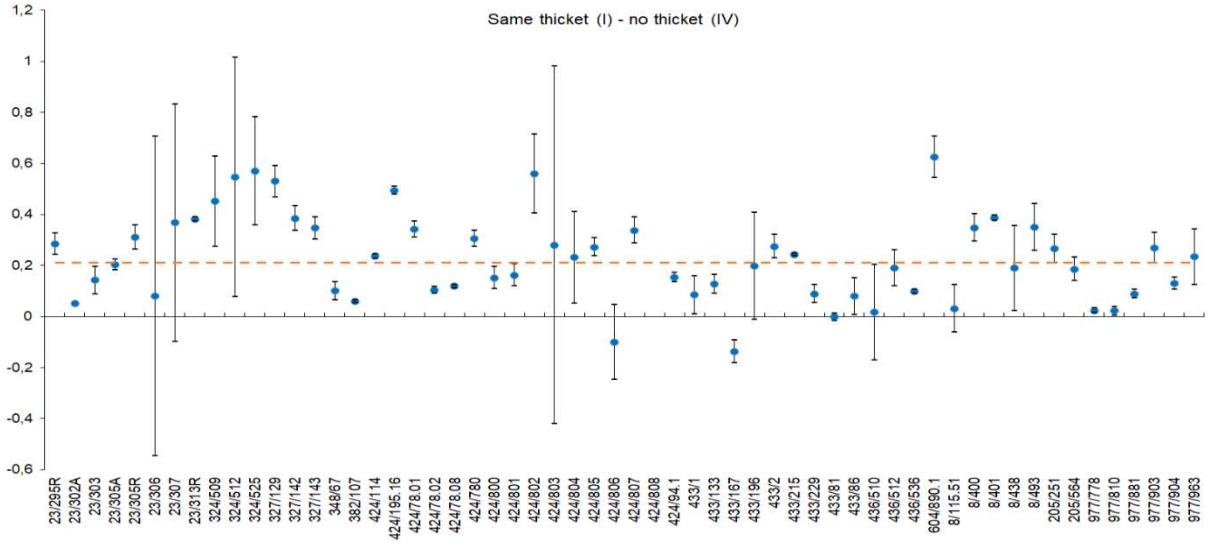
Figure 2 presents two charts showing differences in average semantic distances for different patent groups (dots) with 95% confidence intervals (whiskers). For brevity, a set with the strongest differences (top panel) and another exhibiting weaker differences (lower panel) were included in the text, with the remaining charts relegated to the appendix.[19] The first chart shows the difference between "same thicket" and "no thicket" sets; the second chart shows "different thickets" and "thicket/no thicket" sets broken down by group. As is evident from Figure 2, the first case exhibits larger differences in average semantic distance than the second. It also shows that there are groups where the semantic distance in set I ("same thicket") is not the shortest (the dots with a negative value). This is a rare event and correspond to groups with relatively few observations. Similarly, groups where the result was not statistically significant (the "whiskers" around the dots overlap zero) also corresponds with few observations within the group. Thickets are rare events in any case, so when we break down our dataset by group, small numbers of data points can generate quite misleading results; however, the general tendency for set I to have the shortest semantic distance is clearly visible.[20]

The horizontal dotted line on each chart of Figure 2 illustrates the median difference in semantic distances between sets. A comparison of the medians (0.203 for the chart with "same thicket" and "no thicket" difference and 0.022 for "different thickets" and "thicket/no thicket") reveals that the sets on the first chart are more distant from one another than those of the second chart. This underlines the role of semantic distance as a potential tool for distinguishing between patents belonging to a single thicket or not. For example, the lower portion of the chart shows that the technique did not find a significant difference between patents belonging to different thickets compared to those where only one of the patents belonged to a thicket. We can say, then, that the tool we propose allows us to identify not just whether patents are in a thicket but are in the same thicket.

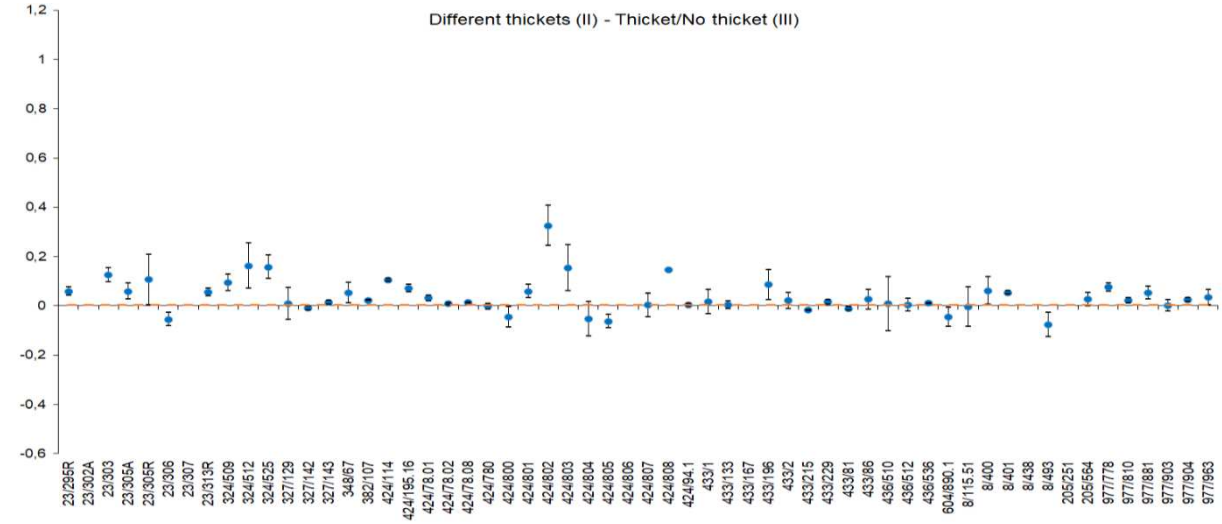
[19] The remaining four tests are included in figure B.1 of Appendix B. Note that sets III and IV exhibit slightly weaker links than II-III, included here. The lesson, however, is that the full set of charts allows us to rank distances among sets, which allows us to conclude that patents in the same thicket are closer than those of any other combination of sets.

[20] The number of observations and characteristics of each group is presented in table C.1 of Appendix C.

Figure 2. Average semantic distances (dots) between chosen sets with confidence intervals ("whiskers" above and below dots). Where the confidence interval overlaps with 0, the result is statistically insignificant (at 95% level of confidence). A horizontal dotted line indicates median for each panel. Notice the different median for each panel.



Source: Own calculations



Finally, comparing semantic distances between different sets of patents involves multiple comparisons, because the average semantic distance for one set needs to be compared simultaneously with the results for three other sets. A Bonferroni correction allows us to perform this simultaneous multiple comparison.[21] We find that semantic distance does perform well under

[21] The Bonferroni correction requires that in order to reach 95% statistical significance for a difference, each of the three tests for the equality of mean semantic distances between pairs of sets must have the p-value lower than 0.05/3

this correction with 72.5% of groups identifying patents in the "same thicket" set at a high significance level.[22]

The main finding from this analysis is that patents that experts identified as belonging to the same thicket are semantically more similar to one another than other patents. Indeed, one can augment the results from Figure 2 with the full set of comparisons[23] which, when sequenced, show that the distance between two patents in a single thicket is less than when the two patents belong to different thickets, which in turn is less than that observed when there is no thicket or just one of the patents belongs to a thicket. Shapiro's (2001) definition of patent thickets as "dense webs of overlapping intellectual property rights" might lead one to expect patents within the same patent thicket to share semantic similarity. Our results confirm that semantic similarity is a good proxy for expert identification. The results also suggest that we can use semantic similarity to identify potential patent thickets, taking expert opinion as reflecting a valid definition of a thicket. We explore the implications of this further in Section 4, but for the remainder of section 3 we identify additional characteristics of our semantic groupings.

3.2 Semantic distance is greater in discrete than in complex technology areas

In addition to comparing the semantic distance between patents inside and outside of thickets, we can also explore how this distance measure relates to the complexity of the technology field in question. To do so we first divide technology areas in accordance with the discrete and complex technology definitions presented in Cohen et al. (2000) and used by Graevenitz et al. (2011). The main difference between a complex and a discrete technology lies in how many separate patentable elements are incorporated in market-ready products. Where there are few elements, the technology is assessed as discrete. On the other hand, products requiring many unique patentable elements are considered complex[24].

= 0.01667. For example, to confirm that set I, our set of interest, is semantically distinguishable from sets II, III and IV simultaneously at a p-value lower than 0.05, we would require that set I pass a much higher bar of a p-value of 0.01667 when compared to each individual set.

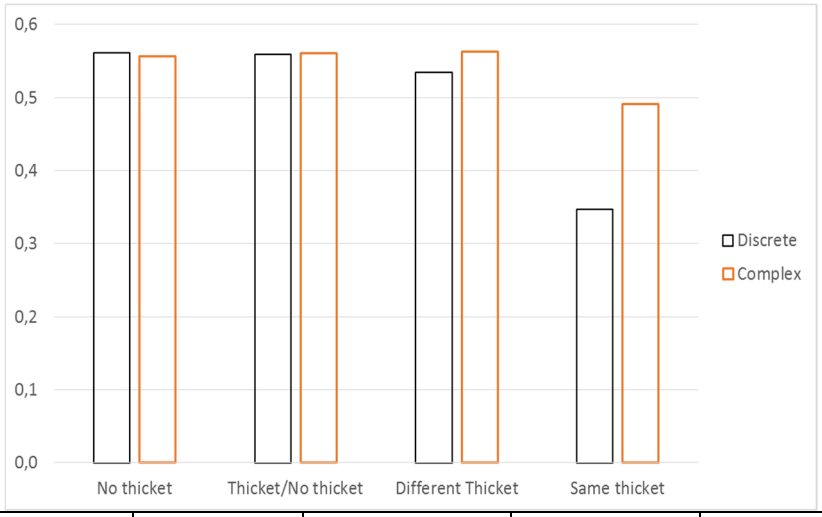
[22] Details of how many groups were significant under Bonferroni correction are listed in Table B.3 of Appendix B.

[23] See Appendix B, Figure B.1.

[24] The list of patent groups and their membership in complex or discrete technology type can be found in Table C.1 of Appendix C.

We find that the average semantic distance we observe between patents in the same thicket (set I) is shorter when those patents are in discrete technology areas and longer in complex ones. Furthermore, the difference between set I and other sets is much greater in discrete cases than it is in complex ones. Figure 3 depicts these differences in bar graphs, with standard errors below and extremely small confidence intervals. Interestingly, patents that do not belong to thickets have a larger average semantic difference in discrete technologies than in complex technologies, perhaps reflecting the wider ranging nature of claims in complex technologies.

Figure 3. Weighted average distance between patents for discrete and complex technologies



Standard error	Set (IV) No thicket	Set (III) Thicket/ No thicket	Set (II) Different thickets	Set (I) Same thicket
Discrete	0.00006	0.00007	0.00021	0.00086
Complex	0.00005	0.00006	0.00017	0.00084

Note: Confidence intervals are of a magnitude 10^{-6} and can't be visibly reported on the graph. Source: Own calculations

Further investigation of the percentage of the groups where the differences between average semantic distances of sets are statistically significant confirms that analysing discrete and complex technologies separately does not change our overall conclusions from the full sample: semantic distance isolates expert-identified thickets well. The overall tendency, however, is that complex technology areas possess a higher percentage of significantly different groups than discrete areas.[25]

The above results suggest that semantic distance as an indicator of potential patent thickets is likely to be more powerful when assessing discrete rather than complex technologies. Equally,

[25] See Appendix B, Table B.2, for details.

knowing whether the underlying technology area tends to be complex or discrete can aid in calibrating the method: if the difference in semantic distance between those patents sharing membership in a thicket and those outside the thicket is smaller for complex technologies, it will be more difficult to distinguish between what is, and what is not, in the thicket using semantic distance. Given the fact that complex technology areas tend to have a greater number of patents within the technology class, and that these patents are more semantically similar, this would lead one to expect more detected thicketing in complex areas, all else equal. This supports the findings of Graevenitz et al. (2011), who detect more thickets in complex technologies than in discrete ones. Furthermore, the greater semantic distance within thickets in complex technology areas suggests that patents belonging to thickets in these areas are more diverse, i.e. these thickets are also more complex, covering a wider variety of rights.

3.3 The results hold if we control for experts

Table 1 below breaks down our tests by expert. This control for expert identity is both interesting in itself and a way for us to ensure that errors in individual judgement were not driving our overall results.

Table 1 indicates that there is no difference in the main conclusions presented in the previous subsections: for each expert the average semantic distance for patents in the same thicket is the shortest; the results for most of the groups are statistically significant; and for the majority of the groups the average semantic distance for patents in different thickets is also statistically significant, apart from expert C, who assessed only one group. This reassures us that our source of external validation is not being driven by errors in judgement.

Table 1. Results of the tests for difference of mean semantic distance between sets, given as percentage of the groups where the difference was significant at 95%. Each row represents an expert.

Expert	No. groups	(1)	(2)	(3)	(4)	(5)	(6)	Set I	Set II	Set III	Set IV
A	**2**	**100%**	**100%**	**50.0%**	100%	100%	100%	0.535	0.593	0.574	0.551
B	**2**	**100%**	**100%**	**100%**	100%	100%	100%	0.166	0.499	0.505	0.525
C	**1**	**100%**	**100%**	**100%**	0%	0%	100%	0.042	0.572	0.581	0.571
D	**13**	**69.2%**	**61.5%**	**61.5%**	69.2%	69.2%	69.2%	0.288	0.622	0.626	0.574

E	**3**	**100%**	**66.7%**	**66.7%**	100%	100%	100%	0.155	0.660	0.598	0.468
F	**8**	**100%**	**87.5%**	**62.5%**	62.5%	87.5%	75.0%	0.625	0.694	0.680	0.642
G	**9**	**77.8%**	**77.8%**	**55.6%**	44.4%	66.7%	77.8%	0.422	0.603	0.598	0.626
H	**20**	**80.0%**	**80.0%**	**80.0%**	60.0%	60.0%	55.0%	0.387	0.540	0.541	0.527

Note: Bold columns show results for the differences of mean semantic distance between "same thicket" and other sets. Grey column is average semantic distance within "same thicket". (1): I – IV (Same thicket and No thicket); (2): I – III (Same thicket and Thicket/No thicket); (3): I – II (Same thicket and Different thickets); (4): II – III (Different thickets and Thicket/No thicket); (5): II – IV: (Different thickets and No thicket); (6): III – IV (Thicket/No thicket and No thicket).
Number of groups assigned to an expert and mean semantic distance between patents in each set are shown for each expert, listed in the left column.
Source: Own calculations;

4. Comparison of Expert-Based, Triples, and Network Density Methods of Patent Thicket Identification

In this section we compare the sample of USPTO patents examined by experts against two thicket measures described in literature – triples introduced in Graevenitz et al. (2011) and weighted average patent network density presented in Clarkson (2005).

The Graevenitz et al. (2011) "triples" patent thicket identification method forms triads of firms' portfolios of critical patents within a technology group, where there are bilateral citations between the portfolios of three different firms. This corresponds to the idea that, where three firms have overlapping portfolios, the negotiation process between them or with another entity is more costly. The idea of triples, used as a proxy measure for patent thicket density, has been used recently to investigate competition (Graevenitz et al., 2013), new entries into technological areas (Hall et al., 2015), and patent opposition (Harhoff et al., 2016).

We compare the results obtained with the triples method with expert patent thicket identification by comparing the share of patents that experts identify as belonging to thickets with the share of patents that belong to triples within given technology groups[26]. We do not necessarily expect our measure to be closely tied to triples due to both implementation issues and our thicket definition. Triples are calculated on much larger sample of patents than the sample our experts examined. Furthermore, the triples methodology places much more prominence on fragmentation of rights than the definition that we provided to the experts, as pointed out recently by DeGrazia et al. (2019). Still, the question we address here is not whether but how different the methods are.

The answer is that they are very different methods indeed. The comparison shows that only 3.7% percent of patents in expert-identified thickets belong to the triples. This is barely higher

[26] Appendix A, Section A.2 outlines our implementation of triples on our database.

than the baseline 3.2% thicket membership we observe when we look at all of the patents from our USPTO sample that were mapped to European Patent Office (EPO) patents. This small increase in the percentage of patents that belong to triples, when moving from the whole sample to patents that our experts identified as within thickets, suggests that the triples methodology and the experts identified very distinct groups of patents. A simple regression run on the data shows little overlap between the two sets with R^2=0.049.

The above results do not mean that the triples method is not good as a proxy for identifying density of thickets in a technology area at the aggregate level, but it does suggest that it may not closely agree with expert judgement on existence of thickets amongst specific patents, using a standard definition. In any event, the two methods have identified quite different sets. That said, the USPTO sample we used was comparatively small (and non-random), so our findings should be interpreted with caution.

Weighted average patent network density (Clarkson, 2005) is a measure calculated as a proportion of directed (in or out) citations in patent networks to all possible (in or out) citations, with the network defined on a patent group. Clarkson (2005) suggests that where the density is higher than the surrounding set of patents a patent thicket can be identified. The measure is based on the idea that patents in a potential patent thicket should cite one another more frequently than patents not belonging to the thicket. Because citations are more frequently made to closely-related inventions, substitute technologies are more likely to be subsumed by this thicket definition than complementary inventions, even though both types could potentially result in the sort of hold-up that has been associated with thicket "problems." For example, Clarkson presents calculations for two patent pools MPEG-3 (a video compression technology) and PRK (a medical technology) and obtains results 0.029 and 0.203 respectively. The MPEG-3 technology is a pool of complementary patents essential to a standard, while PRK contains substitute patents, describing similar approaches to the same technology.[27]

To compare Clarkson's density with our expert-based method we calculate Clarkson's measure[28] on the USPTO classification groups included in our expert thicket identification patent

[27] Régibeau et al. (2012) support the view that Clarkson density is a noisy measure, its value depending strongly on how broadly the patent network, i.e. technology, is defined.
[28] We use the weighted average patent network density described by formula (6) in Clarkson (2005).

set and compare the results with the expert identified thickets. Similarly to the triples comparison, a simple regression shows little overlap between the two measures, with R^2=0.037.[29]

To summarise, our findings are that the expert judgement, derived from a standard thicket definition and well correlated with the semantic similarity of the whole body of the patent texts, is not well correlated with two citation-based measures at the individual patent level. In turn, this suggests that semantic similarity may be a useful tool in identifying patent thickets, but not necessarily the same thickets identified by existing citation measures. The fact that the sets are so distinct suggests that the information in them could possibly be combined to provide a better aggregate evaluation of thicket membership. We turn to this exercise in the next section.

5. A Semantic Network Model for Thicket Recognition

One way to incorporate the divergence of these methods into an overall approach to thicket identification is to propose a classification model based on the network of pairwise semantic distances and drawing from information contained in other methods, specifically triples and Clarkson's weighted average patent network density. We do so here. The model below is aimed both at illustrating individual measures of patent overlap as indicators of thicket membership and also at predicting a newly added patent's probability of membership in (any) existing thicket within a given patent group. The results on individual measures helps align our work with earlier results on the roles of citations structure, fragmentation, crowding, and technological complexity. The results on prediction point to a method for implementing the Lemley (2001) screening suggestion mentioned in the introduction as a step in the patent application or examination processes.

Logit regression holds several advantages as a classification method for tasks such as the one we propose here. In addition to allowing for inference from coefficients, it can be used as a predictive tool by allowing us to calculate theoretical probabilities of the dependent variable's taking a unit value (i.e., the probability that a patent belongs to a thicket). Logit regression also allows us to adopt a critical value for this probability to exceed a threshold. This, in turn, allows us to classify our data. For example, we could specify the lowest probability that would classify a patent as belonging to a thicket with some confidence level.

[29] For details on both correlations, see Appendix A, Sections A.2 and A.3 and Figure A.1.

To provide an illustration of its potential as a forecasting tool, we have estimated our model on an "in-sample" dataset and tested on "out-of-sample" data. This allows us to check how well the model performs if estimated at one point in time whilst making a prediction at another point of time. As a measure of performance, we assess the quality of the model by forecasting the theoretical probability of being in a thicket given specific characteristics of a patent application in question using in-sample estimates, but an out-of-sample testing period. The year 2001 was chosen as a break point between in-sample and out-of-sample portions of the data.[30] More precisely, the estimation sample consists of patents applied for between 1976 and 2000 (5,482 patents of which 1,088 are in thickets[31]), while the testing sample contains patents awarded from the period 2001-2010 (3,089 patents of which 467 belong to thickets).[32] We have included results from all our work below using the full 1976-2010 sample and have reported the results in Appendix D. With one exception (see Appendix D), the results do not differ substantially.

The logit modelling results are presented in Table 2, below.[33] The dependent variable is the membership of a patent in an expert-identified thicket.[34] The independent variables of interest are: minimal semantic distance – distance to the most similar earlier patent; Clarkson's ratios for a group (at the moment of filing), calculated as described in Section 4 as the ratio of existing pairs of patents, in which one cites the other, over the maximal potential number of such pairs, which depends only on the number of patents in the group; the triples ratio for a group, in other words, the share of patents belonging to triples (at the moment of filing) where triples are identified using

[30] The rule for selecting the break year was that it was the first year for which two thresholds for our sample were surpassed – firstly, 60% of all patents in the full dataset were filed for and, secondly, 70% of all patents in thickets were filed for in the estimation dataset.

[31] This allows us to have about ten times as many positive observations as explanatory variables in the most extended model specification, per standard practice (means n = 10k, where n represents observations with dependent variable = 1 and k = number of explanatory variables in the regression). See, however, Vittinghoff and McCullogh (2007) for discussion.

[32] The full sample running through 2015 was not used due to a likely sample selection. Namely, inclusion of the latest available years could have affected assessment of the predictive power of the model, since patents applied for before the cut-off date for our whole sample, but granted afterwards, would not be included. This could change the structure of the patents in the last period covered substantially, by removing from the sample applications subject to particularly long deliberation. In our truncated sample (i.e., 1976-2010), 90% patents were granted within 4.6 years and 95% within 5.6 years. The LSA analysis uses all years to 2015.

[33] Our logit model was estimated using the generalized linear model and its implementation in R.

[34] In both estimation and test samples the earliest patents in thickets were not counted as "in a thicket", because the model takes into account time-varying structure of patents groups. This means that the first patent does not belong to any thicket at the moment of its filing.

Graevenitz et al's (2011) method translated to our dataset.[35] We also included controls for the number of backward citations; number of claims; number of patent groups to which a patent under consideration belongs to (a measure of interdisciplinary character of a patent); thicket ratio for a group – share of patents belonging to thickets in a group of application (at the moment of filing); complex group dummy variable – group from complex or discrete technology area; HHI calculated for patents for a given group (at the moment of filing) as a measure of ownership concentration,[36] based on filing dates of eventually successful patents); number of prior (eventually successful)[37] filings by assignee;[38] total number of applications and of patents granted in a given group at the moment of, respectively, filing or granting a given patent; dummies for class (or group) of patent and the application year.

We built up a model using various specifications, but as these generally had the same substantive results we present only our preferred specification in the text, whilst relegating other specifications to Appendix D for brevity.[39]

To facilitate interpretation of the coefficients' magnitudes, we report not only estimates (along with standard errors), but also odds ratios and "incremental steps". The latter correspond to either one standard deviation in our sample (for continuous variables) or one (for dummy variables). The "Odds ratio" column in Table 2 shows how the change by one incremental step in each control variable changes those odds. For example, we can see that an increase in semantic distance by one standard deviation would lead to a fall in odds of a patent belonging to a thicket by 42%.

---

[35] One should exercise caution in drawing strong conclusions from this analysis: our measure underestimates the number of triples – and, therefore, number of patents in triples. This is because the method requires us to restrict attention to those patents that could be mapped to the EPO database. It, therefore, omits any triples created by US-only patents, not to mention more complex relations between different national and international patent systems. Nevertheless, our view is that it is indicative, as the direction of bias – including over time and across groups – is not clear.

[36] Other measures of fragmentation are possible. We have chosen on that is readily computable from our dataset. We note that our results are consistent with the rest of the literature, so we have not investigated alternative measures more thoroughly at this point.

[37] Patents for which we did not have data on the assignee where omitted while calculating the HHI index.

[38] Filings based on known assignees for the patents included in our sample. We have made an effort to match names containing obvious typos and differences in abbreviations or other conventions. The R package by van der Loo (2014) was utilised.

[39] The Preferred Model in the text is listed as Model 9 in all Appendices. Section D.1 of Appendix D lists the results for all models and data years 1976-2000, Section D.2 lists the results for the full sample, 1976-2010, for all models, and Section D.3 lists the results for data years 1976-2000 and includes the coefficient values for all groups and years.

Table 2. Results for Preferred Model

Variable	Logit Estimate	Odds ratio	CI low (2.5 %)	CI high (97.5 %)	Incremental step
Semantic distance	-3,434*** (0.291)	0.579	0.528	0.634	0.159
Number of backward citations	0.067*** (0.014)	1.198	1.114	1.291	2.683
Number of claims	0.001 (0.003)	1.016	0.942	1.094	13.979
Number of groups	0.617*** (0.142)	1.156	1.082	1.233	0.234
Thicket ratio for a group (in %)	4.291*** (0.295)	2.183	1.968	2.429	0.182
Clarkson ratio for a group	0.593 (1.252)	1.020	0.935	1.103	0.033
Complex group (average)	2.247 (3.328)	9.463	0.042	2,269,290.0	1
Triples incidence in group	-1.534 (2.255)	0.963	0.862	1.074	0.025
HHI	-0.900* (0.489)	0.906	0.813	1.003	0.110
Prior applications of assignee	0.014*** (0.004)	1.142	1.060	1.228	9.477
Prior applications in the group	0.002* (0.001)	1.494	0.981	2.328	190.862
Prior patents in the group	-0.004*** (0.001)	0.470	0.299	0.717	192.733
Patent class dummies	yes				
Observations	5,482				
Log Likelihood	-2.184.413				
Akaike Inf. Crit.	4,412.825				

Note: Dependent variable for the logit regressions is belonging to a thicket at the moment of application. Class dummies are included in the Preferred Model specification, but group and year dummies are not. In odds calculations, one incremental step is equal to the standard deviation of each variable or 1 for dummies and 'complex group' (which is a dummy variable averaged over all the groups for a given patent – in almost all cases it is either 0 or 1). "CI" refers to "confidence interval".
$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.
Source: Own calculations

These results suggest several conclusions. First, the model suggests that (eventually successful) patent applications belonging to groups that include patents belonging to thickets, are substantially more likely to be in a thicket as well. This suggests both that thickets are a characteristic of a patent group and that a larger pool of thicket patents breeds a higher likelihood that further research will overlap with those existing thickets.

Second, patent applications that are semantically closer to an earlier most similar patent have a greater probability of belonging to a thicket. This is similar to our earlier discussion: semantic distance predicts the evaluation of our experts well.

Third, patents belonging to many technology groups are more likely to be in a thicket. Complexity is unlikely to underlie this, as the relation holds when we control for complexity.[40] A similar relationship can be found for the number of backward citations. The positive correlation with backwards citations suggests that crowding in a group is associated with thicket emergence.

Fourth, the number of claims in a patent application is not particularly relevant to the probability of the patent belonging to a thicket once one allows for various group characteristics.

Fifth, Clarkson's density ratio and the triples ratio are not significant. Furthermore, the negative coefficient on triples (significant in one of the alternative specifications reported in the Appendix D) indicates that the presence of existing triples in the group is actually negatively correlated with further thicket membership. As a high triples ratio indicates a relatively well defined set of patent holders, this may reduce the complexity of the patent examiner's as well as the assignee's task in creating new and distinct patents.

Sixth, concentration of patent ownership (measured by HHI) lowers the probability of occurrence of a new thicket (or to increase of size of the preexisting one) even though the historic propensity of a group to include thickets is controlled for. Hence fragmentation at the group patent level is positively related to the prediction that a patent will fall in a thicket, as suggested by previous studies focusing on hold-up. Furthermore, the magnitude of the coefficient suggests a relatively strong effect.

---

[40] One could speculate why this would occur, but more thorough investigation would be required to support any specific interpretation. The result is intriguing, however, in the light of Noel and Schankerman's (2006) model of enforcement costs related to the number of points of conflict in a patent. While points of conflict may be related to fragmentation, as in their work, it could also (perhaps additionally) be related to large applicability, which could be indicated by membership in a large number of groups.

Seventh, there is a greater chance that a thicket will be created (or joined by a further patent) when the assignee has filed for a greater number of patents in the past. This suggests the possibility of defensive or strategic patenting driving some of the results, but is not definitive: the result could also suggest that patents resulting from a single research trajectory, as might be pursued by a single researcher, are more likely to interfere with each other because the underlying subject matter will tend to overlap.

Finally, the opposite signs of the total number of prior filings and prior positive decisions in a given group suggest, taken together, that: a) patents that were granted after longer deliberation had a lower probability of belonging to a thicket, while the ones that were granted relatively quickly had a greater probability of being in a thicket. This last result is particularly intriguing, as it could be interpreted as suggesting that there may be a link between the quality of patent review and the likelihood of thicket membership. It is not definitive, however, as this quick review could also be associated with the familiarity of the patent examiner with the technology. Hence, learning effects could also be driving this result without any link to lower quality[41].

To further assess model performance we consider two ratios: (1) a "false positive" ratio - which shows how many patents would be unnecessarily identified, i.e. how many patents flagged by the model as in thickets are actually not in a thicket; and (2) a “false negative” ratio – which shows how many patents in thickets would be wrongly omitted, i.e. how many patents are flagged by the model as not in a thicket, when they actually belong to a thicket. The “false positive” ratio can be regarded as an indicator of type I error, whereas the “false negative” ratio of type II error. The magnitude of the ratio will depend on a theoretical probability threshold of the assessment “not in a thicket” or “in a thicket” as an outcome of the model. We call this probability a critical value and present ratios for a range of critical values in Figure 4[42] for the Preferred Model[43]. Selected

[41] We have investigated this in another working paper, Dietl et al. (2017), finding that even when we account for learning effects, thickets seem to be associated with shorter delay.

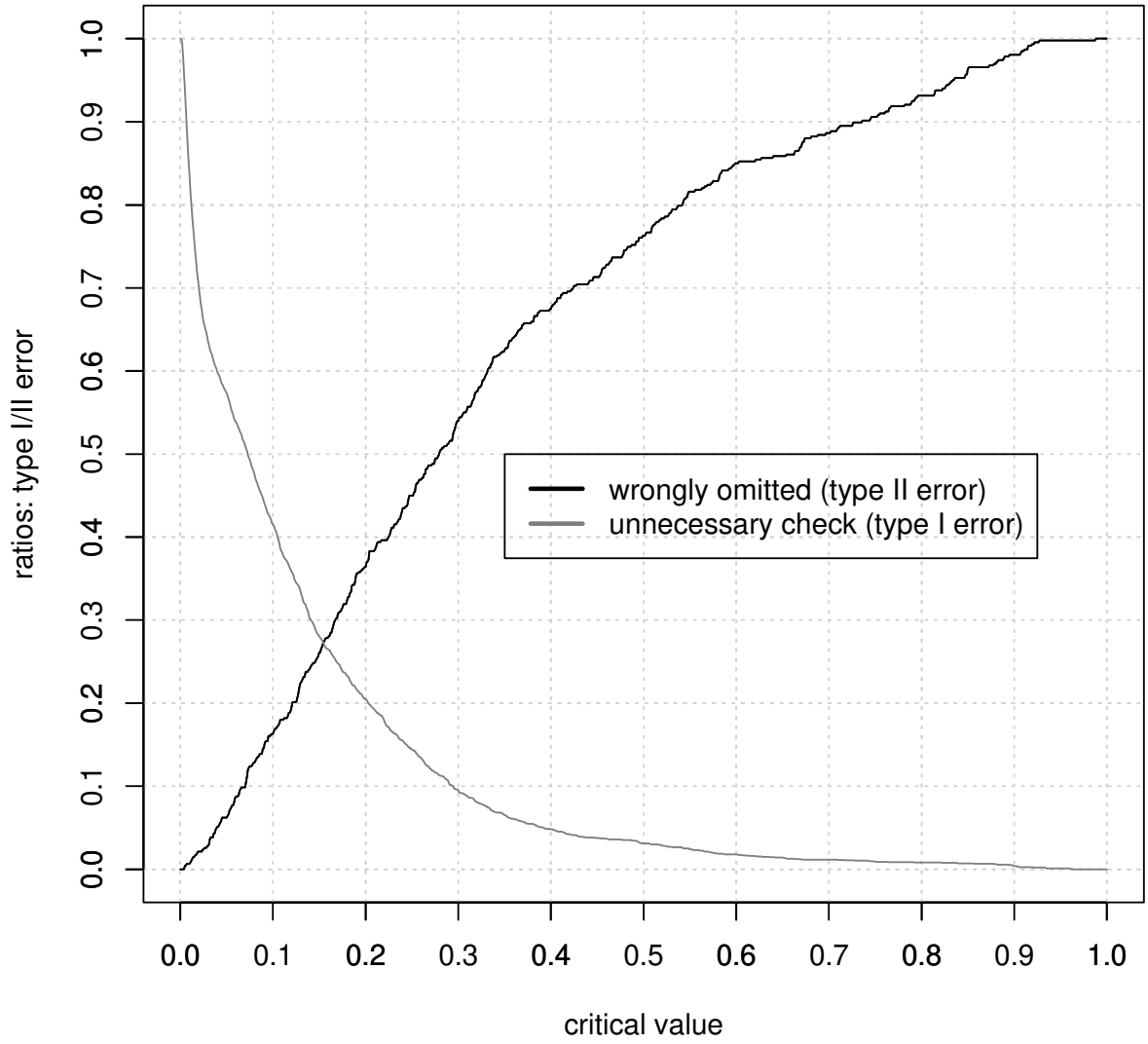
[42] For example, Figure 4 suggests that when the critical value is 0.1, around 15% of patents that are members of thickets are wrongly classified by the model as not belonging to a thicket; whereas 40% of patents that do not belong to a thicket would be wrongly classified as belonging to it. Said differently, 60% could be subject to a quick check if the model was used to screen for thicket membership. A critical value of 0.2 would result in error values of 35% and 20%, respectively.

[43] For analogous charts for alternative specifications see Appendix E. The same appendix contains the performance of all versions of the logit model, listed in Table E.1 and Figure E.9.

specifications are compared on Figures 5 and 6, which illustrate the performance of the full model compared to versions with certain right hand variables excluded. Performance improves to the south-west of Figures 5 and 6.

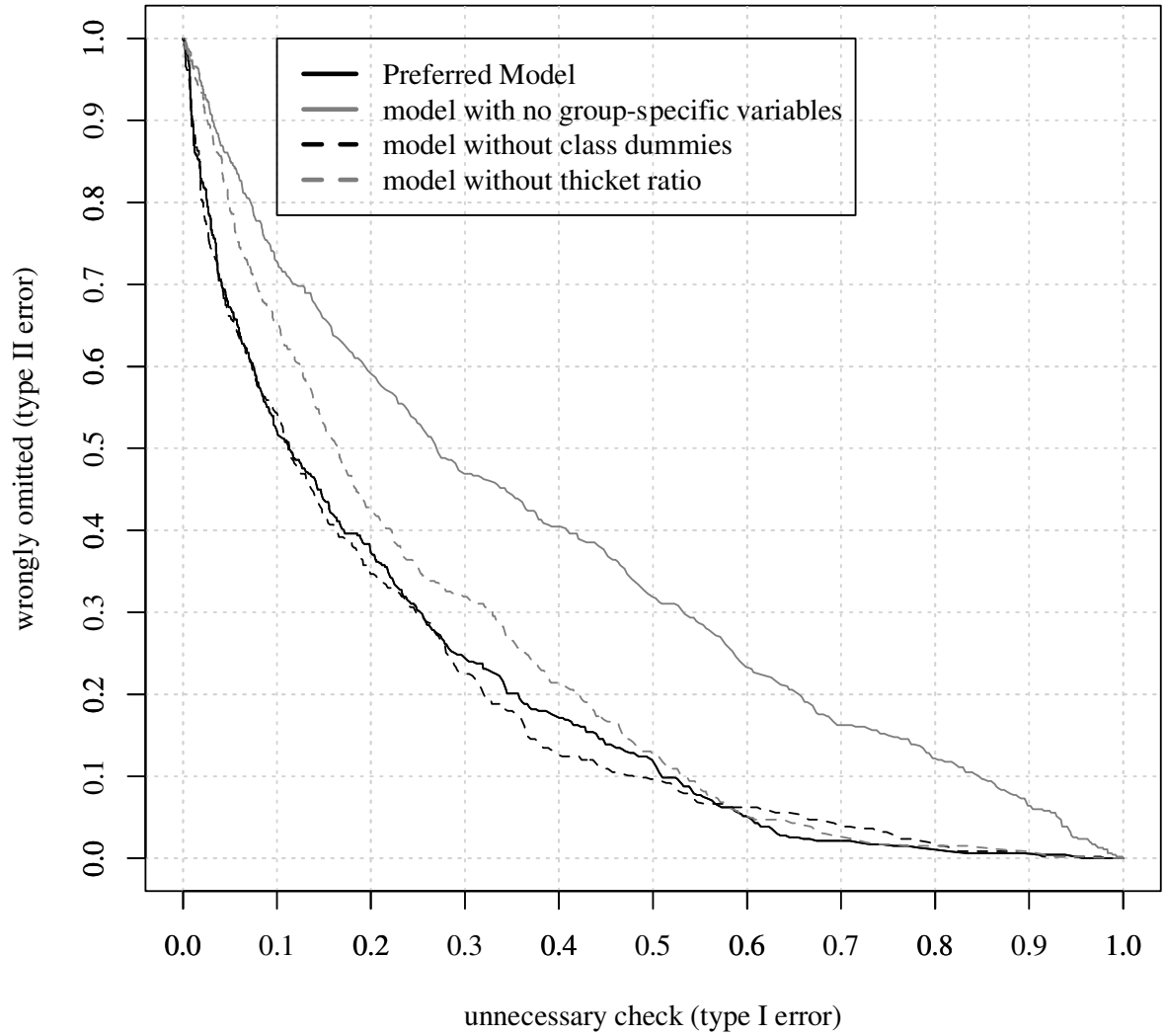
The conclusion from this exercise and accompanying figures is that the different thicket measures can be combined meaningfully into a screening model that captures many of the relationships that have been previously identified between thickets and underlying characteristics of patents and technologies as well as reflecting expert view as a form of external validation, exactly how that is best done depends on how one trades off errors of omission (type II error) or errors of inclusion (type I error). While the Preferred Model works relatively well compared to the others, it is not completely dominant. In particular, Figure 5 shows that exclusion of the class dummies can improve or worsen the specification in some cases. On the other hand, the same figure illustrates that omission of all group-specific variables simultaneously (thicket ratio, Clarkson ratio, triples ratio, HHI ratio, number of past applications and granted patents, class dummies) significantly worsens the performance. Figure 6 illustrates, however, that the same cannot be said for each of these ratios excluded in isolation. Omission of semantic distance, however, tends to worsen the predictive capability of the model in Figure 6. Hence, we view the measures used as complementary to the type of semantic analysis we conduct as building blocks in a well-functioning predictive model.

Figure 4. False positive/negative ratios as functions of the critical value for the Preferred Model.



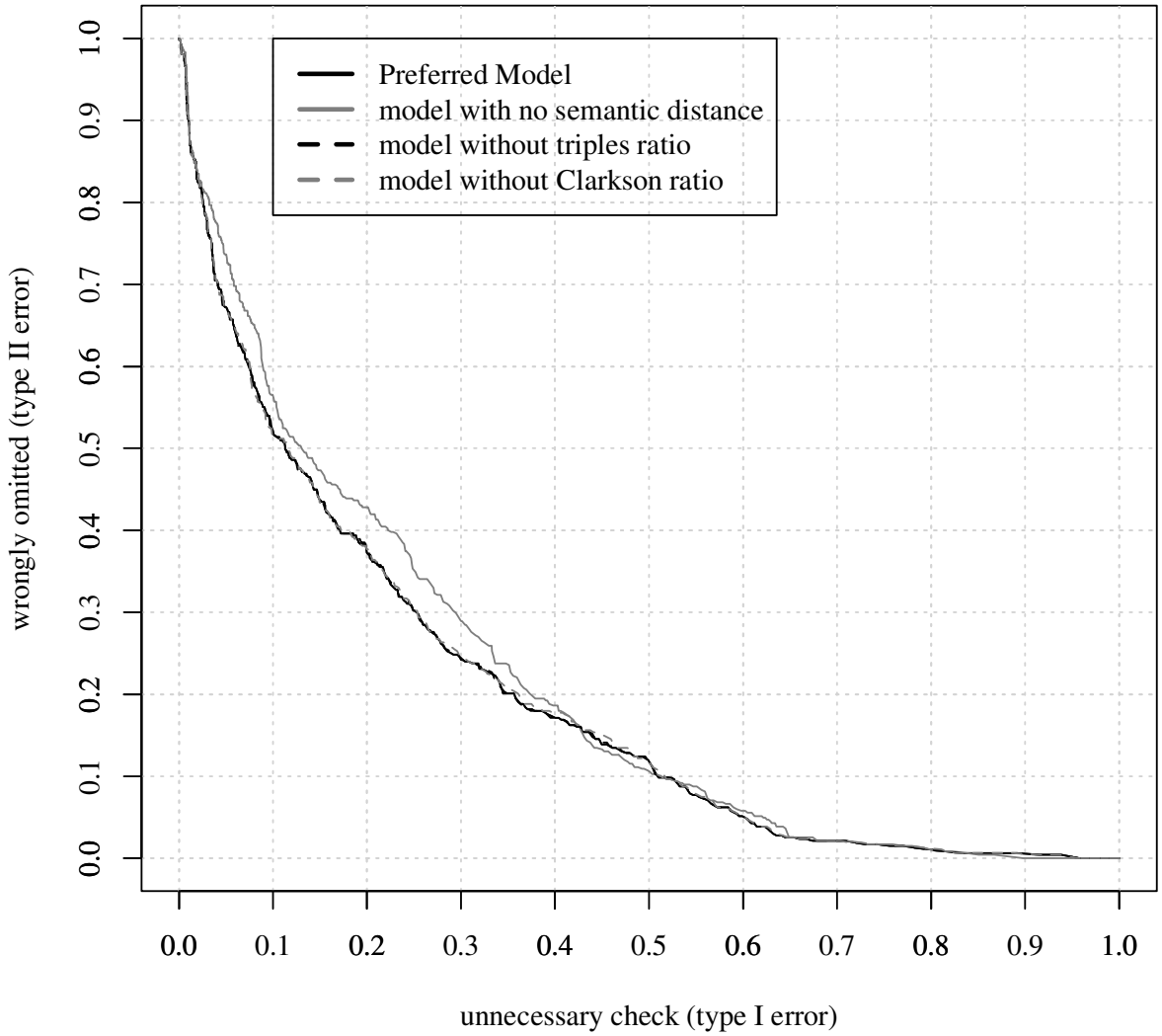
Note: The lines are not smooth as they are derived from the tests on out-of-sample datasets. The ratios are: number of applications that would not be in thickets flagged as in-thicket patents (i.e. selected for an unnecessary check) and number of applications that would be in thickets flagged as not-in-thicket patents (i.e. wrongly omitted from selection for a check).
Source: Own calculations

Figure 5. False positive/negative ratios tradeoff for the Preferred Model and alternative specifications (I).



Source: Own calculations

Figure 6. False positive/negative ratios tradeoff for the Preferred Model and alternative specifications (II).



Source: Own calculations

6. Conclusions

Because patent rights depend ultimately on interpreting patent language, expert reading and opinion would appear to provide a way to measure the substantive technical content claimed, and relatedly, thicketing. While in itself this is an unwieldy technique, advances in natural language

processing may allow expert views to be efficiently replicated. To test this, we first gather expert views on thickets and then compare them to semantic networks that we create via latent semantic analysis. We validate our distance scores by comparing them to expert opinion, applied to a standard definition of patent thickets. This illustrates how semantic distance corresponds with expert views. To our knowledge, this is the first such comparison as well as a first application of expert classification at the individual patent level as external validation for patent thicket identification. Semantic analysis and the underlying similarity that it captures has the advantage of being much more finely-grained than the survey results that have been used to date.

Our key conclusion is that patents belonging to the same expert-identified thicket are closer semantically than other pairs of patents and this holds across diverse technology fields spanning differing underlying citations structures and technological complexity. While this result is dependent upon the definition provided to the experts and is a costly method of detection in itself, semantic distance does appear to be a promising method of proxying the view that would be obtained by a careful reading of patent documents. The recent availability of computing power and natural language processing tools allows the ready implementation of this proxy of expert view. Indeed, now that we have shown the link to expert view, applying semantic distance measures to a wider set of patents would be a natural extension.

We use our thicket measure to investigate earlier results. We find that the semantic distance between patents belonging to thickets in discrete technology areas is shorter than for those in complex areas, which confirms the intuition that patents for complex technologies cover a more diverse set of rights. It also suggests that it is easier for thickets to arise in complex technology areas, where there are more patents and those patents are more semantically similar, confirming the findings of Graevenitz et al. (2011). These findings hold when controlling for the experts used to identify patent thickets and thus are not influenced by expert idiosyncrasies.

We also check whether existing citation-based methods of identification perform well compared to expert classification, our source of external validation. We find that there is little overlap between individual patents indicated by experts as belonging to a thicket and patents belonging to Graevinitz et al. (2011) triples-defined thickets. Similarly, the patent network density measure introduced by Clarkson (2005) shows no significant relation to the share of individual

patents in these thickets. This is not entirely unexpected: the definition provided to our experts, while standard, is broader than that captured by triples. Furthermore, our implementation requires us to apply an EPO-based methodology to USPTO patents. Still, the differences are significant in our view. Furthermore, the fact that these measures do capture such distinct sets suggests that they may work well together to identify thickets.

In a third step we then combine the various measures into a single predictive model of thicket identification, and evaluate its performance as a potential "screen" in terms of its identification of false positives (membership of a thicket where this is not actually the case) and false negatives (lack of membership, when membership in the thicket actually is the case). Most significantly, the model shows that semantic distance combined with other information including citation-based measures and controls for fragmentation can be helpful in assessing a newly-filed application for its "risk" of thicket membership. It also generates results on the predictive contribution of individual measures within this combined framework. For example, crowding and technology group complexity play a role in the likelihood that a patent will belong to a thicket and also enter into our specification. A key indicator of a patent belonging to a thicket is the previous density of thickets within a patent group. Finally, fragmentation is also an indicator of thicket formation, which has been emphasised consistently in the literature.

In addition to delineating thicket membership, the different types of relationships amongst thicket members are important as well. A standard set of possible relationship types would include blocking, complementary, independent, or substitute patents (Clarkson, 2005). While most citations-based measures focus on blocking relationships the method presented in this paper, based on semantic networks, could potentially cover any relationship including substitution or complementarity and so provides a more comprehensive measure of the source of linkage via overlapping content only. Our final screening model, nesting various thicket measures and so capturing various relationships, includes a measure of fragmentation, which is a proxy for the portion of patent thickets reflecting hold-up rather than defensive patenting concerns (Noel and Schankerman, 2006, Galasso and Schankerman, 2010).[44]

[44] This measure turns out to be significant. See discussion around Table 2.

The predictive model we present can provide support for those interested in identifying patent thickets prospectively as a means of anticipating thicket-based strategic issues that may arise later. This includes identification at early stages where the text of the patent is still being drafted, as discussed with an emphasis on measurement by Hall et al., (2013) and with an emphasis on the theory of cumulative innovation by Gallini (2017). At the same time, this remains one proposal only: we do not optimise this combination of measurement tools in our logit approach. Instead, our contribution should be interpreted as making the point that semantic distance adds valuable information to the analysis of thickets, and information that can be obtained early – at the point the patent is created – so that it can have value as a prospective measure. Many methods exist to combine different measures together for a more accurate view of patent thickets and exploring these is left to future work.

Our method exploits expert opinion to identify thickets, and this method has some weaknesses. First among these is individual expert error. We have investigated the role this might play in driving our results and have found some evidence that it is not. Still, our aim is not to develop an infallible tool for thicket detection but rather a method of delineating a set of patents with a high probability of becoming members of a thicket. Indeed, it would be interesting to repeat the approach used above with a broader group of experts and technology areas. It would also be interesting to perform similar analyses on EPO data or data from other patent offices or with different definitions for experts to facilitate comparison with the triples and Clarkson methodologies. Equally, alternative methods of detection exist, including court cases. While court cases are based on many factors that may not be present in all thickets, these nonetheless may do a good job of isolating those thickets that are likely to be the most troublesome in their consequences.

Acknowledgement

Part of the research was supported by the National Science Centre (NCN) [grant No. DEC-2013/11/B/HS4/00682 “A new method for identification of patent thickets”]

We want to thank Dr Michał Rudolf for his very helpful assistance with data and fruitful discussions with Vanessa Behrens and other discussants at the 7th ZEW/MaCCI Conference on the Economics of Innovation and Patenting. Pierre Régibeau has provided invaluable ideas and

advice. We would also like to thank the editor of this journal and two anonymous referees for their valuable comments.

7. References

1. Barnett J. M., 2014. From Patent Thickets to Patent Networks: The Legal Infrastructure of the Digital Economy. Jurimetrics J. 55, 1–53.
2. Bergeaud A., Potiron Y., Raimbault J., 2017. Classifying Patents Based on Their Semantic Content. PLoS ONE 12(4). https://doi.org/10.1371/journal.pone.0176310
3. Bessen J., 2004. Patent Thickets: Strategic Patenting of Complex Technologies. Working paper 0401. Research on Innovation.
4. Bradford, R. B., 2008. An Empirical Study of Required Dimensionality for Large-scale Latent Semantic Indexing Applications. Proceedings of the 17th ACM Conference on Information and Knowledge Management, 153–162. New York, NY, USA: ACM. doi:10.1145/1458082.1458105
5. Clarkson G., 2005. Patent Informatics for Patent Thicket Detection: a Network Analytic Approach for Measuring the Density of Patent Space. Acad. Manag. Honolulu.
6. Cohen, W., Nelson, R., Walsh, J., 2000. Protecting their Intellectual Assets: Appropriability Conditions and Why U.S. Manufacturing Firms Patent (or Not). National Bureau of Economic Research Working Paper 7552.
7. Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A., 1990. Indexing by Latent Semantic Analysis. JAsIs 41(6), 391–407.
8. DeGrazia, C. A., Frumkin, J. P, & Pairolero, N. A., 2019. Embracing Invention Similarity for the Measurement of Vertically Overlapping Claims. Economics of Innovation and New Technology. DOI: 10.1080/10438599.2019.1593035
9. Dietl, M., Skrok, L, Benalcazar, P., Gatkowski, M., Rockett, K., 2017. Pendency and Thickets. Economics Discussion Papers 19979, University of Essex, Department of Economics.
10. Egan, E., Teece, D., 2015. Untangling the Patent Thicket Literature. Tusher Center for the Management of Intellectual Capital Working Paper 7. March.

11. Fleming, L., 2001. Recombinant Uncertainty in Technological Search. Management Science 47(1), 117–132. doi:10.1287/mnsc.47.1.117.10671.
12. Frakes, M., Wasserman, M., 2019. Irrational Ignorance at the Patent Office. Vanderbilt Law Review. 72. Forthcoming. Available at SSRN: https://ssrn.com/abstract=3284109.
13. Galasso, A., Schankerman, M., 2010. Patent Thickets, Courts, and the Market for Innovation. Rand Journal of Economics 41(3), 472-503.
14. Gallini, N., 2017. Do Patents Work? Thickets, Trolls, and Antibiotic Resistance. Canadian Journal of Economics 50(4), 893-926.
15. Gerken, JM., Moehrle MG., 2012. A New Instrument for Technology Monitoring: Novelty in Patents Measured by Semantic Patent Analysis. Scientometrics 91(3), 645–670.
16. Graevenitz, von G., Wagner S., Harhoff, D., 2011. How to Measure Patent Thickets-A Novel Approach. Econ. Lett. 111(1), 6–9.
17. Graevenitz, von G., Wagner, S., Harhoff, D., 2013. Incidence and Growth of Patent Thickets: The Impact of Technological Opportunities and Complexity. The Journal of Industrial Economics 61, 521–563.
18. Hall B., Helmers C., Graevenitz von G., Rosazza-Bondibene C., 2013. A Study of Patent Thickets. Intellect. Prop. Off. 401, 7–76.
19. Hall, B. H., Helmers, C., & Graevenitz, von G., 2015. Technology Entry in the Presence of Patent Thickets. National Bureau of Economic Research Working Paper 21455, Revised 2017.
20. Hall, B. H., Ziedonis, R.H., 2001. The Patent Paradox Revisited: An Empirical Study of Patenting in the US Semiconductor Industry, 1979-1995. RAND Journal of Economics 32(1), 101–128.
21. Harhoff D., Graevenitz von G., Wagner S., 2016. Conflict Resolution, Public Goods and Patent Thickets. Management Science INFORMS 62(3), 704-721, March.
22. Harispe, S., Ranwez, S., Janaqi, S., & Montmain, J., 2015. Semantic Similarity from Natural Language and Ontology Analysis. Synthesis Lectures on Human Language Technologies 8(1) 1–254. doi:10.2200/S00639ED1V01Y201504HLT027.

23. Holman, C., 2006. Clearing a Path through the Patent Thicket. Cell 125(4), 629–33.
24. Intellectual Property Office Informatics Team, 2011. Patent Thickets. Intellectual Property Office, Newport, UK.
25. Kuhn, J. M., Thompson, N., 2017. The Ways We've Been Measuring Patent Scope are Wrong: How to Measure and Draw Causal Inferences with Patent Scope. Available at SSRN: https://ssrn.com/abstract=2977273 or http://dx.doi.org/10.2139/ssrn.2977273.
26. Lemley, M., 2001. Rational Ignorance at the Patent Office. University of California at Berkeley School of Law UC Berkeley Public Law & Legal Theory Research Paper Series 95(4).
27. Landauer, T. K., Foltz, P. W., & Laham, D., 1998. An Introduction to Latent Semantic Analysis. Discourse Processes 25(2–3), 259–284. doi:10.1080/01638539809545028.
28. Landauer, T. K., Laham, D., & Foltz, P., 1998. Learning Human-like Knowledge by Singular Value Decomposition: A Progress Report. Advances in Neural Information Processing Systems 45–51.
29. Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds), 2013. Handbook of Latent Semantic Analysis. Routledge, New York.
30. Lang, K., 1995. Newsweeder: Learning to Filter Netnews, in: Proceedings of the 12th International Conference on Machine Learning, pp. 331–339.
31. Noel, M., Schankerman, M., 2006. Strategic Patenting and Software Innovation. Journal of Industrial Economics 61(3), 481-520.
32. OECD, 1994. The Measurement of Scientific and Technological Activities Using Patent Data as Science and Technological Indicators: Patent Manual OECD Publishing, Paris. https://doi.org/10.1787/9789264065574-en.
33. Preschitschek N, Niemann H, Leker J, Moehrle MG., 2013. Anticipating Industry Convergence: Semantic Analyses vs IPC Co-classification Analyses of Patents. Foresight 15(6), 446–464.
34. R Core Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

35. Régibeau P, Rockett, K., Mariam, S., 2012. Patent Pendency, Learning Effects, and Innovation Importance at the US Patent Office. Economics Discussion Papers 709. University of Essex, Department of Economics.
36. Řehůřek, R., & Sojka, P., 2010. Software Framework for Topic Modelling with Large Corpora, in: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. ELRA, Valletta, Malta, pp 45-50.
37. Rijsbergen, C. J. V., 1979. Information Retrieval (2nd ed.). Butterworth-Neinemann. Newton, MA, USA.
38. Salton, G., & McGill, M. J., 1986. Introduction to Modern Information Retrieval. McGraw-Hill, Inc., New York, NY, USA.
39. Schankerman, M. and Schuett, F., 2018. Screening for Patent Quality. CEPR Discussion Paper 11688 (updated 2018).
40. Shapiro C., 2001. Navigating the Patent Thicket: Cross Licenses, Patent Pools, and Standard Setting, in: Jaffe, A., Lerner, J., and Stern, S. (Eds) Innovation Policy and the Economy, v1. National Bureau of Economic Research, Cambridge, MA. 119-150.
41. Turney, P. D., & Pantel, P., 2010. From Frequency to Meaning: Vector Space Models of Semantics. Journal of Artificial Intelligence Research 37(1), 141–188.
42. Van der Loo, M., 2014. The Stringdist Package for Approximate String Matching. The R Journal 6, pp. 111-122. <URL: https://CRAN.R-project.org/package=stringdist>.
43. Vittinghoff, E, & McCullogh, C., 2007. Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regressions. American Journal of Epidemiology 165(6), 710-719.
44. Welch, B. L., 1951. On the Comparison of Several Mean Values: An Alternative Approach. Biometrika 38 (3-4), 330–336.
45. Whalen, R., 2018. Boundary Spanning Innovation and the Patent System: Interdisciplinary Challenges for a Specialized Examination System. Research Policy 47 (7), 1334–1343.
46. World Intellectual Property Organization, 2013. WIPO IP Facts and Figures. WIPO Publication No. 943E/13, p. 44.

47. World Intellectual Property Organization, 2014. WIPO IP Facts and Figures. WIPO Publication No. 943E/14, p. 6.
48. World Intellectual Property Organization, 2015. WIPO IP Facts and Figures. WIPO Publication No. 943E/15, p. 8.
49. Yoon, B., Park Y., 2004. A Text-mining-based Patent Network: Analytical Tool for High-technology Trend. The Journal of High Technology Management Research 15(1), 37–50
50. Ziedonis, R.H., 2004. Don't Fence Me In: Fragmented Markets for Technology and the Patent Acquisition Strategies of Firms. Management Science 50, 804–820

Appendix A – Latent Semantic Analysis, Triples, and Network Density

A.1. Latent Semantic Analysis – A Primer

We use LSA to calculate pair-wise patent semantic distance, creating a semantic patent network, where links between patents are weighted by the semantic distance between them. We hypothesize that the overlapping rights indicative of patent thickets will correspond to semantic similarity between patents occupying the same thicket. We benchmark semantic similarity measures against a set of expert-identified patent thickets to incorporate external validity.

We use the entire corpus of patents published by the USPTO between 1976 and 2015 to calculate our LSA model. These documents were downloaded from the public data dumps made available by the USPTO. We then take the full text of each granted patent - comprising the abstract, the description, and the claims - and use that as the terms representing each document. LSA takes as its starting point a document-term matrix, which is then transformed using singular value decomposition. We begin the creation of our matrix by generating a term-document matrix with a row for each granted patent (our input documents), and a column for each unique term (i.e. word) used across the corpus. The matrix values are the frequency of that term within each row's relevant patent document. Because very common and very rare words provide little in the way of insight we remove all the words from a common set of stop words (Rijsbergen, 1979), as well as terms from the corpus that occur in more than 50% of all documents or fewer than 5 of the documents. This removes very common words like 'the' or 'claim' or 'and' as well as highly unusual terms that are often typos or spelling errors.

Once these low-information terms have been removed from the matrix, we then subject the corpus to a term frequency–inverse document frequency (TF-IDF) transformation to further improve the semantic signal (Salton & McGill, 1986). We use a standard TF-IDF transformation, which multiplies term i's frequency in the given document j ($f_{i,j}$) by the logarithmically-scaled inverse document frequency—that is the number of documents in the corpus (D) divided by the number of documents where the term appears (d_i).

$$weight_{i,j} = f_{i,j} * log_2 D/d_i$$

A high TF-IDF score for a particular term demonstrates that it occurs frequently within the given document, but rarely across the corpus, suggesting that it provides a strong signal as to the

document's topical focus. Essentially, this re-weights terms based on the degree of insight they provide into a document's topics. For instance, common words like "small" will occur frequently across the corpus, and will thus be discounted by the TF-IDF transformation, which is appropriate as they are likely to provide a weak signal as to the document's focus, whereas less frequently occurring terms such as "nanotechnology" will have their signal amplified from their raw frequency counts by the TF-IDF transformation. The resulting document/TF-IDF matrix is used as the input matrix for our LSA model.

Once the input matrix has been assembled, we use the Gensim Python library to perform the dimensional reduction (Řehůřek & Sojka, 2010). Gensim takes the input matrix X and performs a rank-reduced singular value decomposition on it, allowing the creation of a k-dimensional document-concept matrix. The k-dimensional document-concept matrix X_k is the matrix of rank-k that best approximates the original TF-IDF transformed document-term matrix **X.** The document-concept matrix is the output of primary interest when attempting to determine the similarity of documents within the corpus. The literature on determining the appropriate value of k generally recommends a value between 300–500 (Bradford, 2008) for larger sets of documents. Because our corpus of documents is quite large—approximately 5.5 million granted patents—and because patents cover a wide-variety of technical areas, we opt for 500 dimensions. The result is a 500-dimension vector for each patent, representing its semantic content as "weights" within each of the 500 topics identified by the LSA process.

Once the document-concept matrix has been computed, we can use vector-space distance measures to measure how distant documents are from one another in the reduced-dimensional space. We rely on the commonly-used cosine distance (Landauer, Laham, & Foltz, 1998) to calculate pairwise distance for the patents in our study.

$$distance(a, b) = 1 - \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \sqrt{\sum_{i=1}^{n} b_i^2}}$$

The cosine distance is calculated as above, where a_i and b_i correspond to the *i*th dimension of each patent's 500-dimension LSA vector.

Patents with a high cosine distance have concept vectors with dissimilar weightings, demonstrating that they cover unrelated technical topics. On the other hand, patents that have low cosine distance have similar concept vector weightings, suggesting that they are more similar.

If we imagine that technical knowledge exists as a multidimensional space with some types of knowledge being "closer" together while others are more distantly-related, the entire process can be conceptualized using a spatial metaphor. For instance, the knowledge required to build an axe is quite similar to the knowledge required to build a hammer, and they are thus close to one another in technical space. On the other hand, the knowledge required to build an axe is very dissimilar from the knowledge required to develop a complex tax minimization strategy and they are thus distant from one another in technical space. The LSA process essentially locates each of the patents in our corpus within a 500-dimensional technical space, while the cosine distance calculation measures how closely (or distantly) related the information within each patent document is.

A.2. Triples Methodology Applied to USPTO data

Triples may be calculated only on the EPO database as they require cited patents to be assessed as to whether they constitute a critical innovation. As our experts evaluated USPTO data, we reproduced the triples thicket identification method on patents granted by the EPO and mapped these, where possible, to patents from our USPTO sample, using the PATSTAT database. Out of 11,872 patents, 5,912 (44%) had matches in the PATSTAT database. Of 2,615 patents identified as belonging to thickets, for 1,233 (47%) PATSTAT matches were found.

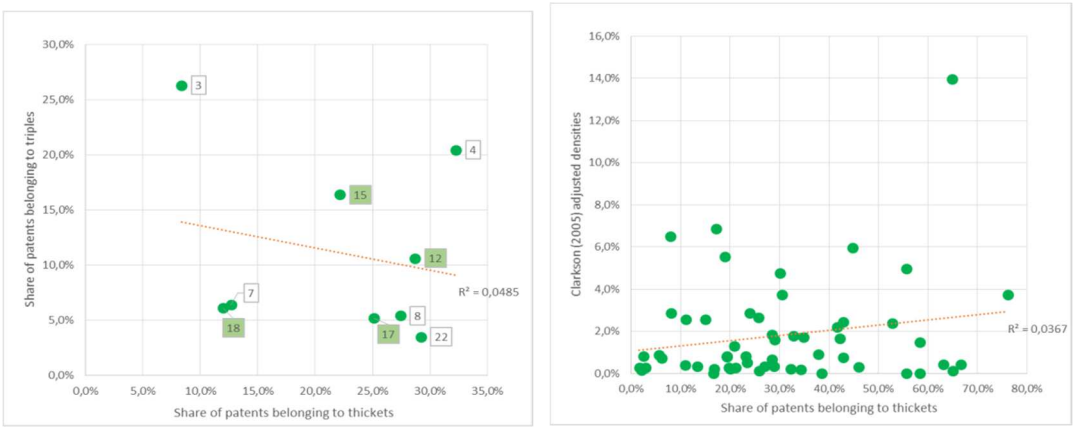
Triples were calculated in the OST-INPI/FhG-ISI technology areas in order to retain comparability with the original Graevenitz et al. (2011) paper and because the measure requires broad samples. Subsequently, we compared patents that belong to triples with patents that were identified as belonging to thickets by field experts.

A.3. Triples, Network Density, and Expert Evaluation Correlations

The left panel of Figure A.1 plots the share of patents in expert-identified thickets against the share of patents belonging to triples identified using the Graevenitz et al. (2011) method for individual technology areas. A simple regression run on the data shows little overlap between the

two with R^2=0.049. The right panel of Figure A.1 plots the Clarkson network density measure against the share of patents in expert-identified thickets. Similarly to the triples, the simple regression shows little overlap between the two measures, with R^2=0.037. In order to account for the different number of patents within groups we have estimated an OLS regression with dummy variables for small groups and an outlier with density of 13%. None of the coefficients was significant, nor was the F-test of the regression model. The robustness of the above findings was checked by calculating Clarkson's measure on patent classes and on the OST-INPI/FhG-ISI technology areas. In none of the cases could the share of patents belonging to the expert-identified thickets be related to Clarkson's density in a statistically significant manner.

Figure A.1 Dots indicate the share of patents belonging to expert-identified thickets vs share of patents belonging to triples in different technologies (left panel) or Clarkson's (2005) adjusted densities (right panel). Discrete technology areas are shaded on the left pane.



Note: Three outliers were removed from the chart on the left due to very small number of patents in our USPTO sample. Out of 58 USPTO patent groups two were removed from the chart on the right, because of no internal citations
The labels on the left panel indicate the technology area OST-INPI/FhG-ISI technology nomenclature (OECD, 1994). 22 – Environment; 12- Pharmaceuticals/Cosmetics; 15 - Petrol Chem./Materials Chem; 17- Materials; 18- Chemical Engineering; 3 - Telecommunications; 4 - IT; 7- Analysis/Measurement/Control Technology; 8 - Medical Technology
Source: Own calculations

Appendix B: Confirming Differences in Average Semantic Distance

B.1. Welch Test confirming statistical significance among groups

We performed a Welch test on six different combinations starting with a mean equality test between set I and set II, (i.e. between pairs of patents belonging to the same thicket and pairs of patents belonging to different thickets), and then for each combination of sets I-IV. The same six tests were repeated for 58 patent groups.

Table B.1 presents these results, showing the percentage of the groups for which the Welch test confirmed the statistical significance of the difference between means with various p-value thresholds. We use a 95% significance level as a cut-off value for the test[45].

Table B.1 Results of the test for mean equality of semantic distance: the percentage of the number of patent groups for which the hypothesis of equality is rejected for a given significance level (p-value).

I – IV (Same thicket and No thicket)	**I – III (Same thicket and Thicket/No thicket)**	**I – II (Same thicket and Different thickets)**	II – III (Different thickets and Thicket/No thicket)	II – IV (Different thickets and No thicket)	III – IV (Thicket/No thicket and No thicket)	p-value
66.1%	**62.5%**	**62.7%**	41.2%	54.9%	48.3%	<=0.0001
73.2%	**64.3%**	**64.7%**	49.0%	66.7%	53.4%	<=0.001
75.0%	**73.2%**	**72.5%**	60.8%	68.6%	56.9%	<=0.01
85.7%	**80.4%**	**78.4%**	72.5%	80.4%	70.7%	<=0.05
14.3%	**19.6%**	**21.6%**	27.5%	19.6%	29.3%	>0.05
56	**56**	**51**	51	51	58	No. groups

Note: The grey cells contain cases, where hypothesis of equality cannot be rejected with more than 95% significance. "No. groups" indicates number of groups for which tests could be performed. Bold columns show results for the differences of mean semantic distance between "same thicket" and other sets.
Source: Own calculations

These tests confirm that the average semantic distance between patents in set I —when both patents are from the same thicket—is significantly lower than for other sets. Depending on the

[45] In some cases there was only one thicket in a patent group, which did not allow for a comparison between thickets, or this single thicket was smaller than 3 patents, which did not allow for comparison within a thicket. We have excluded such groups, and have reported instead the results for cases where the test could be performed. The percentage of the total dataset for which the test was not possible was relatively small, however, standing at 3% of the sample for most columns. We report the number of groups for which the test could be performed in the last row of the table. As the percentage of excluded groups varies depending on how we split the data, these numbers vary.

setup, 85.7% (when testing for difference between averages in sets I and IV, i.e. patents belonging to the same thicket and patent outside any thicket) to 78.4% (I-II, i.e. patents from the same thicket compared with patents from different thicket) of groups have passed the test for difference in average semantic distance with at $p < 0.05$. The differences in average semantic distances between other sets are also evident; however, the difference is significant least often for the sets II – III, that is between different thickets and thicket/no thicket sets. Nevertheless, tests show that for the majority of groups, all four sets are distinguishable.

B.2. Discrete and Complex Technology Differences

Table B.2 shows the percentage of the groups where the differences between average semantic distances of sets are statistically significant and confirms that analysing discrete and complex technologies separately does not change our overall conclusions from the full sample presented in Table B.1 of this appendix: semantic distance isolates expert-identified thickets well. The overall tendency, however, is that a higher percentage of groups possess statistically significant differences at 95% for complex than for discrete areas, with an exception of difference between set I and set II (i.e. same versus different thickets). This may be explained by the smaller average number of patents per group in discrete technologies.

Table B.2. Results of the test for mean equality for discrete and complex areas: the percentage of the patent groups, for which the hypothesis of equality is rejected with a given significance level (p-value).

I – IV (Same thicket and No thicket)	**I – III (Same thicket and Thicket/No thicket)**	**I – II (Same thicket and Different thickets)**	II – III (Different thickets and Thicket/No thicket)	II – IV (Different thickets and No thicket)	III – IV (Thicket/No thicket and No thicket)	p-value
Discrete						
73.3%	**70.0%**	**69.2%**	46.2%	57.7%	40.6%	<=0.0001
76.7%	**70.0%**	**73.1%**	53.8%	69.2%	46.9%	<=0.001
76.7%	**73.3%**	**76.9%**	61.5%	69.2%	50.0%	<=0.01
83.3%	**80.0%**	**88.5%**	76.9%	76.9%	62.5%	<=0.05
16.7%	**20.0%**	**11.5%**	23.1%	23.1%	37.5%	>0.05
30	**30**	**26**	26	26	32	No. groups
Complex						

57.7%	**53.8%**	**56.0%**	36.0%	52.0%	57.7%	<=0.0001
69.2%	**57.7%**	**56.0%**	44.0%	64.0%	61.5%	<=0.001
73.1%	**73.1%**	**68.0%**	60.0%	68.0%	65.4%	<=0.01
88.5%	**80.8%**	**68.0%**	68.0%	84.0%	80.8%	<=0.05
11.5%	**19.2%**	**32.0%**	32.0%	16.0%	19.2%	>0.05
26	**26**	**25**	25	25	26	No. groups

Note: The top grey cells contain cases where test's result was not significant at more than 95% level of confidence. "No. groups" indicates number of groups for which tests could be performed. Please see footnote 10 for details. Bold columns show results for the differences of mean semantic distance between "same thicket" and other sets.
Source: Own calculations;

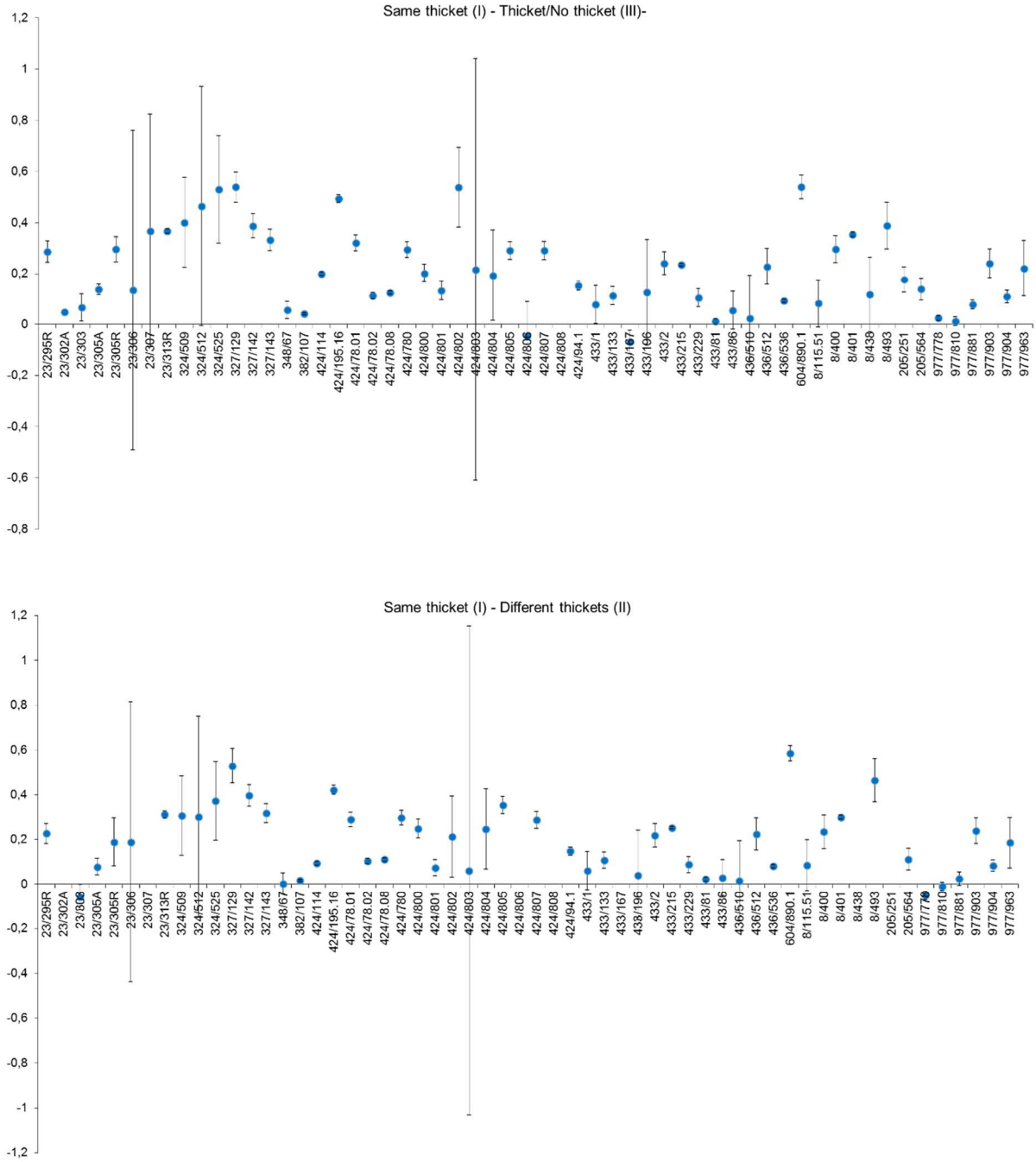
B.3. Bonferroni Correction and Confidence Intervals

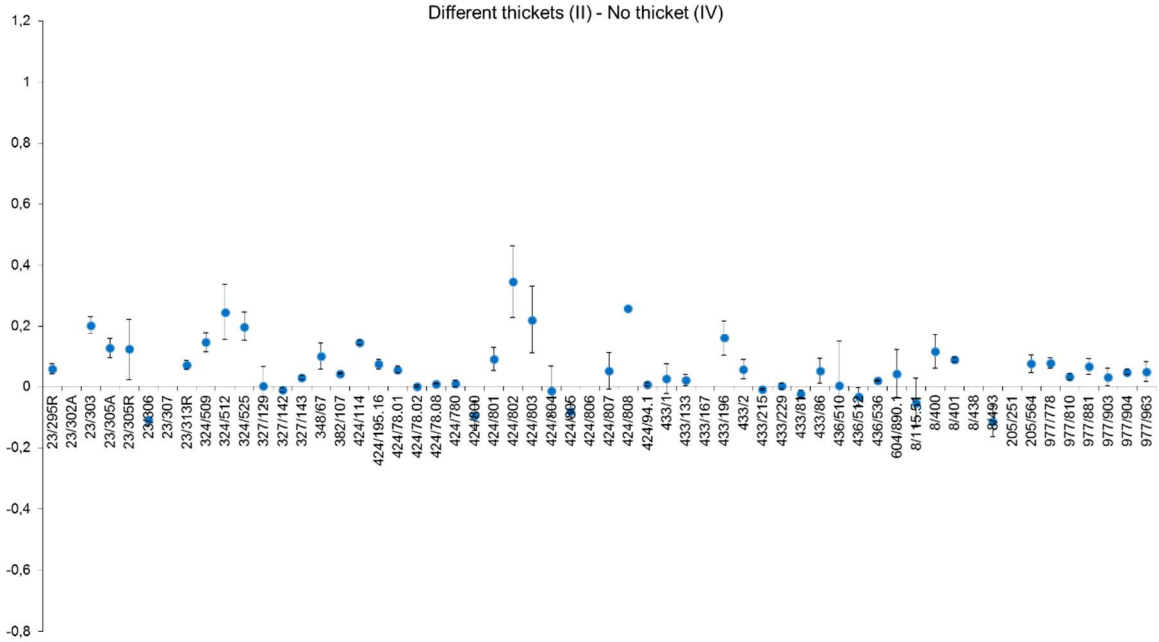
Table B.3. Results of tests with Bonferroni correction – number and percentage of patent groups where semantic distance remains significant at 95% level.

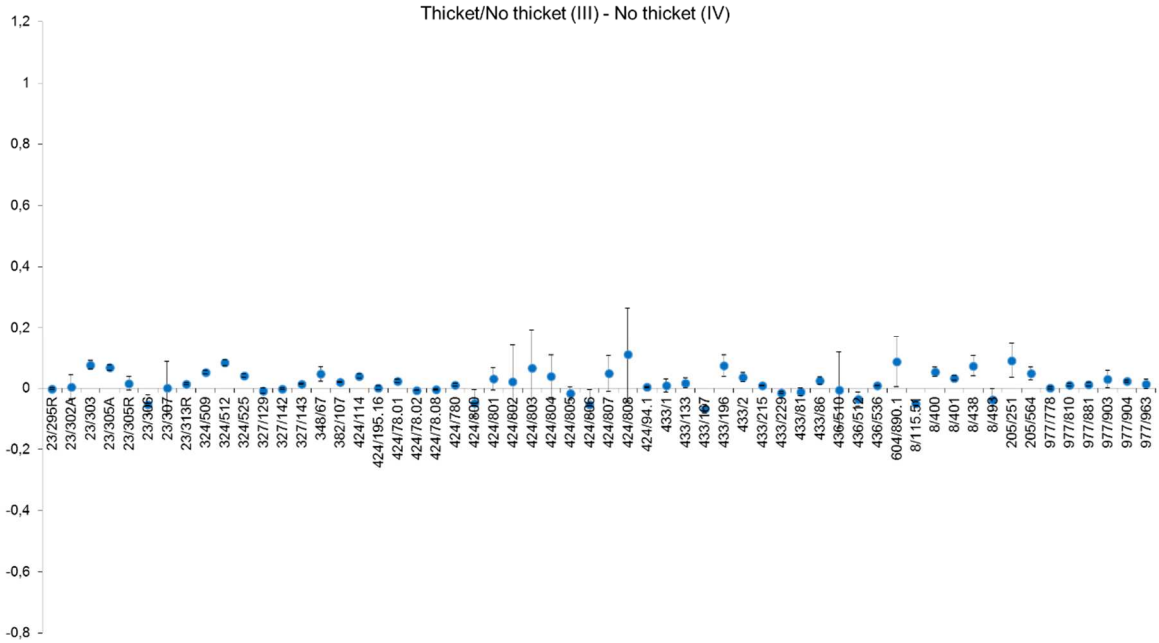
	Set I (Same thicket)	Set II (Different thickets)	Set III (Thicket/No thicket)	Set IV (No thicket)
Number of significant groups	37	26	21	26
Share of significant groups	72.5%	51%	41.2%	51%

Note: There were 51 groups where all three tests could be performed. When interpreting the results with a Bonferroni correction it is important to remember that this correction creates a more conservative test, lowering the probability of returning false positives
Source: Own calculations

Figure B.1. Average semantic distance between chosen sets with confidence intervals. Where the confidence interval overlaps with 0 line, the result is statistically insignificant (1-α=95%).




Different thickets (II) - No thicket (IV)
1,2
1
0,8
0,6
0,4
0,2
0
-0,2
-0,4
-0,6
-0,8


Thicket/No thicket (III) - No thicket (IV)
1,2
1
0,8
0,6
0,4
0,2
0
-0,2
-0,4
-0,6
-0,8

Appendix C: Included USPC group names and Survey Questions

Table C.1. Names of USPC groups used in the analysis with the number of patents.

Class / group	Name	Classification	Number of patents
327/129	Converting input frequency to output current or voltage. Generating sinusoidal output	Complex	143
23/295R	Chemistry: physical processes. Crystallization	Discrete	266
23/302A	Chemistry: physical processes. Crystallization. Alkali method and ammonium compounds. Ammonium compounds	Discrete	21
23/303	Chemistry: physical processes. Crystallization. Alkali method and ammonium compounds. Common salt	Discrete	60
23/305A	Chemistry: physical processes. Crystallization. Heavy metal or aluminium compounds. Aluminium compounds	Discrete	88
23/305R	Chemistry: physical processes. Crystallization. Heavy metal or aluminium compounds	Discrete	27
23/306	Chemistry: physical processes. Concentration of liquids in liquids	Discrete	22
23/307	Chemistry: physical processes. Concentration of liquids in liquids. With direct heating	Discrete	7
23/313R	Chemistry: physical processes. Agglomerating	Discrete	272
8/115.51	Bleaching and dyeing. Chemical modification of textiles or fibres or products thereof	Discrete	350
8/400	Bleaching and dyeing. Measuring, testing or inspecting dye process	Discrete	70
8/401	Bleaching and dyeing. Using enzymes, dye process, composition, or product of dyeing	Discrete	179
8/438	Bleaching and dyeing. Process of extracting or purifying of natural dye	Discrete	36
8/493	Bleaching and dyeing. Overall dimensional modification or stabilization. Modification of molecular structure of substrate by chemical means	Discrete	34
324/509	Electricity: measuring and testing. Fault detecting in electric circuits and of electric components of ground fault indication	Complex	427
324/512	Electricity: measuring and testing. Fault detecting in electric circuits and of electric components for fault location	Complex	215
324/525	Electricity: measuring and testing. Fault detecting in electric circuits and of electric components for fault location by resistance or impedance measuring	Complex	331
205/251	Electrolytic coating (process, composition and method of preparing composition). Depositing predominantly alloy coating. Gold is predominant constituent. Including arsenic, indium or thallium.	Discrete	26
205/564	Electrolytic coating (process, composition and method of preparing composition). Preparing single metal. Gallium, germanium, indium, vanadium or molybdenum produced.	Discrete	48
977/778	Nanostructure. Within specified host or matrix material (e.g., nanocomposite films, etc.)	Complex	235
977/810	Nanostructure. Of specified metal or metal alloy composition	Complex	177
977/881	Manufacture, treatment or detection of nanostructure. With arrangement, process, or apparatus for testing. With arrangement, process, or apparatus for testing	Complex	147
977/903	Specified use of nanostructure. For conversion, containment, or destruction of hazardous material	Complex	40

977/904	Specified use of nanostructure. For medical, immunological, body treatment, or diagnosis	Complex	243
977/963	Specified use of nanostructure. For medical, immunological, body treatment, or diagnosis. Specially adapted for travel through blood circulatory system	Complex	37
433/1	Dentistry. Veterinary dentistry	Complex	45
433/133	Dentistry. Apparatus. Having motor or means to transmit motion from motor to tool. Hand-held tool or handpiece. Contra angled handpiece	Complex	63
433/167	Dentistry. Prosthodontics	Complex	99
433/196	Dentistry. Prosthodontics. Orienting or positioning teeth	Complex	26
433/2	Dentistry. Orthodontics	Complex	66
433/215	Dentistry. Method or material for testing, treating, restoring, or removing natural teeth	Complex	1013
433/229	Dentistry. Miscellaneous	Complex	166
433/81	Dentistry. Apparatus. Having intra-oral dispensing means. Endodontic	Complex	135
433/86	Dentistry. Apparatus. Having intra-oral dispensing means. Endodontic. Ultrasonic tool	Complex	69
424/114	Drug, bio-affecting and body treating compositions. Plural fermentates of different origin	Discrete	229
424/195.16	Drug, bio-affecting and body treating compositions. Extract or material containing or obtained from a unicellular fungus as active ingredient	Discrete	163
424/78.01	Drug, bio-affecting and body treating compositions. Digestive system regulator containing solid synthetic organic polymer	Discrete	175
424/78.02	Drug, bio-affecting and body treating compositions. Topical body preparation containing solid synthetic organic polymer	Discrete	854
424/78.08	Drug, bio-affecting and body treating compositions. Solid synthetic organic polymer	Discrete	693
424/780	Drug, bio-affecting and body treating compositions. Extract or material containing or obtained from a micro-organism as active ingredient	Discrete	196
424/800	Drug, bio-affecting and body treating compositions. Antibody or fragment thereof whose amino acid sequence is disclosed in whole or in part	Discrete	51
424/801	Drug, bio-affecting and body treating compositions. Involving antibody or fragment thereof produced by recombinant DNA technology	Discrete	23
424/802	Drug, bio-affecting and body treating compositions. Antibody or antigen-binding fragment thereof that binds gram-positive bacteria	Discrete	9
424/803	Drug, bio-affecting and body treating compositions. Antibody or antigen-binding fragment thereof that binds gram-negative bacteria	Discrete	9
424/804	Drug, bio-affecting and body treating compositions. Involving IGG3, IGG4, IGA, or IGY	Discrete	16
424/805	Drug, bio-affecting and body treating compositions. Involving IGE or IGD	Discrete	66
424/806	Drug, bio-affecting and body treating compositions. Involving IGM	Discrete	13
424/807	Drug, bio-affecting and body treating compositions. Involving IGM. Monoclonal	Discrete	17
424/808	Drug, bio-affecting and body treating compositions. Involving IGM. Human	Discrete	10
424/94.1	Drug, bio-affecting and body treating compositions. Enzyme or coenzyme containing	Discrete	534
436/510	Chemistry: analytical and immunological testing. Immunochemical pregnancy determination	Complex	12

436/512	Chemistry: analytical and immunological testing. Involving antibody fragments	Complex	49
436/536	Chemistry: analytical and immunological testing. Involving immune complex formed in liquid phase	Complex	1261
604/890.1	Surgery. Controlled release therapeutic device or system	Discrete	22
348/67	Television. Improving the 3D impression of a displayed stereoscopic image	Complex	36
382/107	Image analysis. Applications. Motion or velocity measuring	Complex	953
327/142	Converting input frequency to output current or voltage. Synchronizing. Reset (e.g., initializing, starting, stopping, etc.)	Complex	523
327/143	Converting input frequency to output current or voltage. Synchronizing. Reset (e.g., initializing, starting, stopping, etc.). Responsive to power supply	Complex	849

Table C.2. Survey questions for the field experts

Question	Range of answers
Does given patent belong to a patent thicket?	Yes/No
To which patent thicket within a patent group does the patent belong?	Name of a thicket (like 'thicket_A', 'thicket_B')
What is the innovation level of the patent?	Choice of one of the five innovativeness levels: Very high, High, Average, Low, Very low

Appendix D: Logit Results

D.1. Logit Results for 10 Comparator Models

Compared to model (1), the first four models listed in Table D.1, models (2)-(5), differ from the first model by one variable (or one group of dummies) only. Respectively, these additions are: dummies for patent class (2), thicket ratio (3), Clarkson's ratio (4) or triples ratio (5). Model (6) contains dummies for patent groups instead of group-specific variables. Model (7) consists of patent-specific variables only. Model (8) is the same as model (1) but without semantic distance. Model (9) is the same as (1) but with no year dummies and is the Preferred Model in the text. Model (10) is a simplified version of (9) without information on number of prior applications and grants in a given patent group.

A few observations are due. Number of claims remains insignificant even if semantic distance is omitted (model 8), suggesting that additional factors on top of drafting are relevant. Number of Claims becomes significant, however, when either class dummies or group-specific thicket ratio

is omitted (models 3 and 7). This suggests that number of claims is indeed relevant, but it aligns with technology-specific and time-specific propensities for thickets to arise.

Table D.1. Estimates for different logit models (1-10) of the probability of the membership in an existing thicket for a new patent application

	Dependent variable:									
	Belonging to a thicket (at the moment of applying)									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Semantic distance	-3.425***	-3.240***	-3.837***	-3.418***	-3.425***	-3.870***	-3.348***		-3.434***	-3.425***
	(0.294)	(0.290)	(0.288)	(0.294)	(0.294)	(0.319)	(0.270)		(0.291)	(0.287)
Number of backward citations	0.071***	0.075***	0.081***	0.072***	0.071***	0.077***	0.046***	0.097***	0.067***	0.048***
	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.012)	(0.015)	(0.014)	(0.013)
Number of claims	0.002	0.001	0.005*	0.002	0.002	0.0002	0.008***	0.001	0.001	-0.001
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)
Number of groups	0.646***	0.706***	0.482***	0.652***	0.647***	0.562*	0.575***	0.709***	0.617***	0.691***
	(0.144)	(0.140)	(0.141)	(0.144)	(0.144)	(0.299)	(0.131)	(0.142)	(0.142)	(0.141)
Thicket ratio for a group (%)	4.216***	5.442***		4.218***	4.219***			4.484***	4.291***	4.582***
	(0.300)	(0.271)		(0.300)	(0.300)			(0.293)	(0.295)	(0.293)
Clarkson ratio for a group	1.237	1.677	1.037		1.211			0.890	0.593	1.571
	(1.210)	(1.297)	(0.990)		(1.213)			(1.204)	(1.252)	(1.191)
Complex group	2.351	0.095	3.216	2.272	2.337			1.417	2.247	1.800
	(3.371)	(0.090)	(3.059)	(3.327)	(3.365)			(3.246)	(3.328)	(3.395)
Triples ratio	0.916	-5.562***	1.007	0.772				0.992	-1.534	-3.556
	(2.321)	(1.936)	(2.125)	(2.318)				(2.264)	(2.255)	(2.217)
HHI for group	-1.436***	-1.991***	0.363	-1.374**	-1.495***			-1.991***	-0.900*	-0.496
	(0.546)	(0.536)	(0.392)	(0.540)	(0.528)			(0.538)	(0.489)	(0.480)
Prior appls of assignee	0.014***	0.012***	0.016***	0.013***	0.014***	0.012***	0.014***	0.018***	0.014***	0.011***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Prior appls in the group	0.003**	0.003**	0.002**	0.003**	0.003**			0.004***	0.002*	
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)			(0.001)	(0.001)	

Prior patents in the group	-0.004*** (0.001)	-0.004*** (0.001)	-0.004*** (0.001)	-0.004*** (0.001)	-0.004*** (0.001)			-0.004*** (0.001)	-0.004*** (0.001)	
Class dummies	yes	no	yes	yes	yes	no	no	yes	yes	yes
Group dummies	no	no	no	no	no	yes	no	no	no	no
Year dummies	yes	yes	yes	yes	yes	yes	yes	yes	no	no
Observations	5,482	5,482	5,482	5,482	5,482	5,482	5,482	5,482	5,482	5,482
Log Likelihood	-2,149.847	-2,201.675	-2,267.570	-2,150.346	-2,149.925	-2,103.826	-2,527.223	-2,224.488	-2,184.413	-2,207.123
Akaike Inf. Crit.	4,391.695	4,477.350	4,625.140	4,390.691	4,389.850	4,375.653	5,114.445	4,538.976	4,412.825	4,454.246
Note:	*p<0.1, **p <0.05, ***p<0.01									

Source: Own calculations

D.2. Results for All Comparator Models using Full sample

The results for models estimated for the full sample do not differ qualitatively from the ones reported in the article – in particular, with regards to semantic distance. The one difference, however, that emerged in the later years in the sample, is the positive, significant impact of patent belonging to one of the complex patent groups.

Table D.2. Estimates for different logit models (1-10) of the probability of the membership in an existing thicket for a new patent application – estimation using the full 1976-2010 sample.

	Dependent variable:									
	Belonging to a thicket (at the moment of applying)									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Semantic distance	-3.598*** (0.247)	-3.469*** (0.243)	-3.783*** (0.239)	-3.595*** (0.247)	-3.598*** (0.247)	-4.158*** (0.265)	-3.142*** (0.222)		-3.571*** (0.245)	-3.417*** (0.242)
Number of backward citations	0.033*** (0.007)	0.037*** (0.007)	0.038*** (0.007)	0.033*** (0.007)	0.033*** (0.007)	0.036*** (0.007)	0.025*** (0.006)	0.047*** (0.007)	0.032*** (0.007)	0.022*** (0.007)
Number of claims	0.002 (0.002)	0.003 (0.002)	0.004** (0.002)	0.002 (0.002)	0.002 (0.002)	0.001 (0.002)	0.008*** (0.002)	0.002 (0.002)	0.001 (0.002)	0.0001 (0.002)

Number of groups	0.937***	0.877***	0.824***	0.946***	0.931***	0.555**	0.643***	0.999***	0.928***	0.941***
	(0.107)	(0.103)	(0.106)	(0.107)	(0.107)	(0.268)	(0.094)	(0.105)	(0.106)	(0.106)
Thicket ratio for a group (%)	4.903***	5.992***		4.916***	4.896***			4.985***	4.939***	5.136***
	(0.265)	(0.232)		(0.265)	(0.264)			(0.258)	(0.262)	(0.260)
Clarkson ratio for a group	1.409	0.794	1.926*		1.413			1.167	0.385	1.580
	(1.194)	(1.217)	(0.987)		(1.186)			(1.143)	(1.208)	(1.171)
Complex group	7.731*	0.232***	6.347	7.660*	7.723*			8.501**	7.762*	7.605*
	(4.530)	(0.076)	(4.045)	(4.473)	(4.553)			(4.202)	(4.503)	(4.445)
Triples ratio	-1.879	-3.358**	-1.725	-1.897				-1.824	-2.691	-4.419***
	(1.789)	(1.587)	(1.628)	(1.789)				(1.746)	(1.679)	(1.660)
HHI for group	-2.261***	-2.325***	0.397	-2.155***	-2.142***			-2.763***	-1.318***	-0.869*
	(0.542)	(0.513)	(0.368)	(0.531)	(0.526)			(0.522)	(0.460)	(0.452)
Prior appls of assignee	0.004*	0.004	0.006**	0.004*	0.004*	0.003	0.005**	0.007***	0.004	0.002
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Prior appls in the group	0.004***	0.003***	0.003***	0.004***	0.004***			0.005***	0.003***	
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)			(0.001)	(0.001)	
Prior patents in the group	-0.004***	-0.005***	-0.004***	-0.004***	-0.004***			-0.005***	-0.004***	
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)			(0.001)	(0.001)	
Class dummies	yes	no	yes	yes	yes	no	no	yes	yes	yes
Group dummies	no	no	no	no	no	yes	no	no	no	no
Year dummies	yes	Yes	yes	yes	yes	yes	yes	yes	no	no
Observations	8,571	8,571	8,571	8,571	8,571	8,571	8,571	8,571	8,571	8,571
Log Likelihood	-3,161.077	-3,230.791	-3,372.763	-3,161.752	-3,161.631	-3,138.564	-3,792.093	-3,278.685	-3,204.201	-3,233.007
Akaike Inf. Crit.	6,434.154	6,555.582	6,855.525	6,433.504	6,433.262	6,465.128	7,664.187	6,667.371	6,452.403	6,506.015

Note: *p<0.1, **p <0.05, ***p<0.01

Source: Own Calculations

D.3. Full Set of Coefficients of Logit Models (1-10) for 1976-2000 Sub-sample

Table D.3. Estimates for different logit models (1-10) of the probability of the membership in an existing thicket for a new patent application - full version for the 1976-2000 subsample.

	Dependent variable:									
	Belonging to a thicket (at the moment of applying)									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	
Semantic distance	-3.425***	-3.240***	-3.837***	-3.418***	-3.425***	-3.870***	-3.348***		-3.434***	-3.425***
	(0.294)	(0.290)	(0.288)	(0.294)	(0.294)	(0.319)	(0.270)		(0.291)	(0.287)
Number of backward citations	0.071***	0.075***	0.081***	0.072***	0.071***	0.077***	0.046***	0.097***	0.067***	0.048***
	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.012)	(0.015)	(0.014)	(0.013)
Number of claims	0.002	0.001	0.005*	0.002	0.002	0.0002	0.008***	0.001	0.001	-0.001
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)
Number of groups	0.646***	0.706***	0.482***	0.652***	0.647***	0.562*	0.575***	0.709***	0.617***	0.691***
	(0.144)	(0.140)	(0.141)	(0.144)	(0.144)	(0.299)	(0.131)	(0.142)	(0.142)	(0.141)
Thicket ratio for a group (%)	4.216***	5.442***		4.218***	4.219***			4.484***	4.291***	4.582***
	(0.300)	(0.271)		(0.300)	(0.300)			(0.293)	(0.295)	(0.293)
Clarkson ratio for a group	1.237	1.677	1.037		1.211			0.890	0.593	1.571
	(1.210)	(1.297)	(0.990)		(1.213)			(1.204)	(1.252)	(1.191)
Complex group	2.351	0.095	3.216	2.272	2.337			1.417	2.247	1.800
	(3.371)	(0.090)	(3.059)	(3.327)	(3.365)			(3.246)	(3.328)	(3.395)
Triples ratio	0.916	-5.562***	1.007	0.772				0.992	-1.534	-3.556
	(2.321)	(1.936)	(2.125)	(2.318)				(2.264)	(2.255)	(2.217)
HHI for group	-1.436***	-1.991***	0.363	-1.374**	-1.495***			-1.991***	-0.900*	-0.496
	(0.546)	(0.536)	(0.392)	(0.540)	(0.528)			(0.538)	(0.489)	(0.480)
Prior appls of assignee	0.014***	0.012***	0.016***	0.013***	0.014***	0.012***	0.014***	0.018***	0.014***	0.011***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Prior appls in the group	0.003**	0.003**	0.002**	0.003**	0.003**			0.004***	0.002*	
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)			(0.001)	(0.001)	
Prior patents in the group	-0.004***	-0.004***	-0.004***	-0.004***	-0.004***			-0.004***	-0.004***	
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)			(0.001)	(0.001)	
Class 23	-0.108		-0.330	-0.135	-0.128			-0.246	0.142	0.092
	(0.223)		(0.207)	(0.221)	(0.217)			(0.216)	(0.212)	(0.212)

Class 324	-4.535	-6.680**	-4.481	-4.546		-3.806	-4.334	-3.862
	(3.412)	(3.104)	(3.369)	(3.406)		(3.289)	(3.369)	(3.435)
Class 327	-3.944	-5.937*	-3.878	-3.923		-2.945	-3.610	-3.202
	(3.372)	(3.059)	(3.328)	(3.365)		(3.246)	(3.327)	(3.395)
Class 348	-1.352	-0.436	-1.325	-1.358		-1.107	-1.606	-1.361
	(3.406)	(3.100)	(3.363)	(3.400)		(3.280)	(3.363)	(3.430)
Class 424	-0.196	-0.043	-0.231	-0.199		-0.081	-0.117	-0.264
	(0.178)	(0.164)	(0.174)	(0.177)		(0.173)	(0.178)	(0.177)
Class 433	-1.943	-2.876	-1.879	-1.956		-1.156	-1.744	-1.521
	(3.360)	(3.049)	(3.316)	(3.354)		(3.235)	(3.316)	(3.384)
Class 436	-2.630	-3.477	-2.563	-2.631		-1.645	-2.118	-2.191
	(3.364)	(3.052)	(3.319)	(3.357)		(3.238)	(3.317)	(3.386)
Class 604	1.129	0.507	1.133	1.112		1.485**	1.045	1.071
	(0.714)	(0.573)	(0.701)	(0.712)		(0.630)	(0.739)	(0.763)
Class 977	-2.131	-2.702	-2.085	-2.101		-1.495	-2.124	-1.655
	(3.350)	(3.038)	(3.306)	(3.343)		(3.225)	(3.307)	(3.375)
Group 23/302A					-16.329			
					(1,091.683)			
Group 23/303					0.216			
					(0.512)			
Group 23/305A					0.008			
					(0.382)			
Group 23/305R					-0.781			
					(0.849)			
Group 23/306					0.063			
					(0.859)			
Group 23/307					-0.021			

	(0.934)
Group 23/313R	-0.309
	(0.317)
Group 324/509	-16.055
	(279.642)
Group 324/512	-15.231
	(424.771)
Group 324/525	-1.617**
	(0.642)
Group 327/129	-2.530***
	(0.760)
Group 327/142	-1.745***
	(0.460)
Group 327/143	-2.722***
	(0.421)
Group 348/67	3.187***
	(0.633)
Group 424/114	1.250***
	(0.310)
Group 424/195.16	-2.482**
	(1.074)
Group 424/78.01	0.073
	(0.389)
Group 424/78.02	-0.105
	(0.284)
Group 424/78.08	0.820***
	(0.278)

Group 424/780	1.067**
	(0.439)
Group 424/800	0.409
	(1.051)
Group 424/801	-0.572
	(1.023)
Group 424/802	3.088**
	(1.439)
Group 424/803	-0.752
	(1.381)
Group 424/804	-0.232
	(0.743)
Group 424/805	-0.247
	(0.609)
Group 424/806	-0.615
	(0.831)
Group 424/807	-0.249
	(0.659)
Group 424/808	-1.712
	(1.503)
Group 424/94.1	-0.443
	(0.289)
Group 433/1	0.642
	(0.530)
Group 433/133	1.333***
	(0.427)
Group 433/167	-0.660

	(0.502)
Group 433/196	0.174
	(0.892)
Group 433/2	0.456
	(0.483)
Group 433/215	0.294
	(0.270)
Group 433/229	-0.232
	(0.333)
Group 433/81	2.985***
	(0.375)
Group 433/86	-0.384
	(0.563)
Group 436/510	1.258*
	(0.735)
Group 436/512	-0.127
	(0.437)
Group 436/536	-0.392
	(0.266)
Group 604/890.1	1.040*
	(0.612)
Group 8/115.51	-2.114***
	(0.581)
Group 8/400	0.562
	(0.441)
Group 8/401	2.210***
	(0.351)

Group 8/438						-1.364		
						(0.861)		
Group 8/493						0.112		
						(0.604)		
Group 977/778						2.319***		
						(0.708)		
Group 977/810						1.661		
						(1.037)		
Group 977/881						1.093***		
						(0.343)		
Group 977/903						2.006*		
						(1.166)		
Group 977/904						0.596		
						(0.433)		
Group 977/963						-13.966		
						(1,723.167)		
Year applied 1977	-0.104	-0.067	-0.401	-0.082	-0.101	-0.210	-0.289	-0.258
	(0.432)	(0.440)	(0.405)	(0.431)	(0.432)	(0.434)	(0.386)	(0.415)
Year applied 1978	0.038	0.057	-0.159	0.047	0.040	-0.103	-0.166	0.065
	(0.409)	(0.418)	(0.375)	(0.409)	(0.409)	(0.408)	(0.352)	(0.392)
Year applied 1979	0.247	0.223	0.038	0.263	0.249	0.057	-0.143	0.192
	(0.397)	(0.400)	(0.369)	(0.397)	(0.397)	(0.399)	(0.345)	(0.383)
Year applied 1980	0.321	0.351	0.222	0.345	0.324	0.394	0.109	0.267
	(0.400)	(0.403)	(0.370)	(0.399)	(0.400)	(0.407)	(0.345)	(0.386)
Year applied 1981	-0.423	-0.441	-0.391	-0.398	-0.416	-0.515	-0.347	-0.434
	(0.399)	(0.406)	(0.369)	(0.399)	(0.399)	(0.403)	(0.347)	(0.385)
Year applied 1982	-0.692*	-0.661	-0.830**	-0.668	-0.683*	-0.733*	-0.877**	-0.795**

	(0.410)	(0.413)	(0.385)	(0.409)	(0.409)	(0.413)	(0.366)	(0.398)
Year applied 1983	-0.926**	-0.775*	-1.123***	-0.905**	-0.918**	-0.982**	-1.037***	-0.933**
	(0.401)	(0.406)	(0.380)	(0.401)	(0.401)	(0.407)	(0.363)	(0.389)
Year applied 1984	0.354	0.520	-0.006	0.379	0.363	0.176	-0.067	0.224
	(0.381)	(0.383)	(0.359)	(0.381)	(0.381)	(0.386)	(0.339)	(0.371)
Year applied 1985	-0.472	-0.268	-0.840**	-0.443	-0.465	-0.577	-0.776**	-0.525
	(0.387)	(0.389)	(0.369)	(0.386)	(0.387)	(0.394)	(0.350)	(0.377)
Year applied 1986	-0.714*	-0.462	-1.105***	-0.686*	-0.704*	-0.878**	-1.100***	-0.782**
	(0.399)	(0.398)	(0.382)	(0.398)	(0.398)	(0.403)	(0.357)	(0.388)
Year applied 1987	-0.160	0.075	-0.609*	-0.130	-0.154	-0.253	-0.666**	-0.254
	(0.375)	(0.373)	(0.357)	(0.374)	(0.375)	(0.380)	(0.333)	(0.365)
Year applied 1988	-1.168***	-0.911**	-1.509***	-1.140***	-1.159***	-1.404***	-1.495***	-1.233***
	(0.405)	(0.404)	(0.387)	(0.404)	(0.405)	(0.416)	(0.365)	(0.397)
Year applied 1989	-0.477	-0.171	-0.883**	-0.451	-0.470	-0.606	-0.845***	-0.573
	(0.370)	(0.368)	(0.350)	(0.369)	(0.370)	(0.371)	(0.328)	(0.359)
Year applied 1990	-0.481	-0.164	-0.897**	-0.450	-0.473	-0.747**	-0.862***	-0.627*
	(0.371)	(0.366)	(0.352)	(0.370)	(0.370)	(0.371)	(0.326)	(0.361)
Year applied 1991	-0.315	-0.002	-0.673**	-0.284	-0.306	-0.588*	-0.662**	-0.383
	(0.356)	(0.350)	(0.336)	(0.355)	(0.356)	(0.357)	(0.310)	(0.346)
Year applied 1992	-0.281	0.024	-0.614*	-0.249	-0.272	-0.573	-0.716**	-0.398
	(0.358)	(0.350)	(0.339)	(0.356)	(0.357)	(0.356)	(0.309)	(0.348)
Year applied 1993	-1.006***	-0.657*	-1.407***	-0.972***	-0.996***	-1.363***	-1.409***	-1.116***
	(0.374)	(0.364)	(0.354)	(0.372)	(0.373)	(0.370)	(0.324)	(0.365)
Year applied 1994	-0.603*	-0.248	-1.069***	-0.566	-0.590	-1.092***	-1.271***	-0.673*
	(0.363)	(0.353)	(0.343)	(0.361)	(0.362)	(0.357)	(0.309)	(0.354)
Year applied 1995	-0.512	-0.143	-0.969***	-0.466	-0.497	-1.076***	-1.144***	-0.549

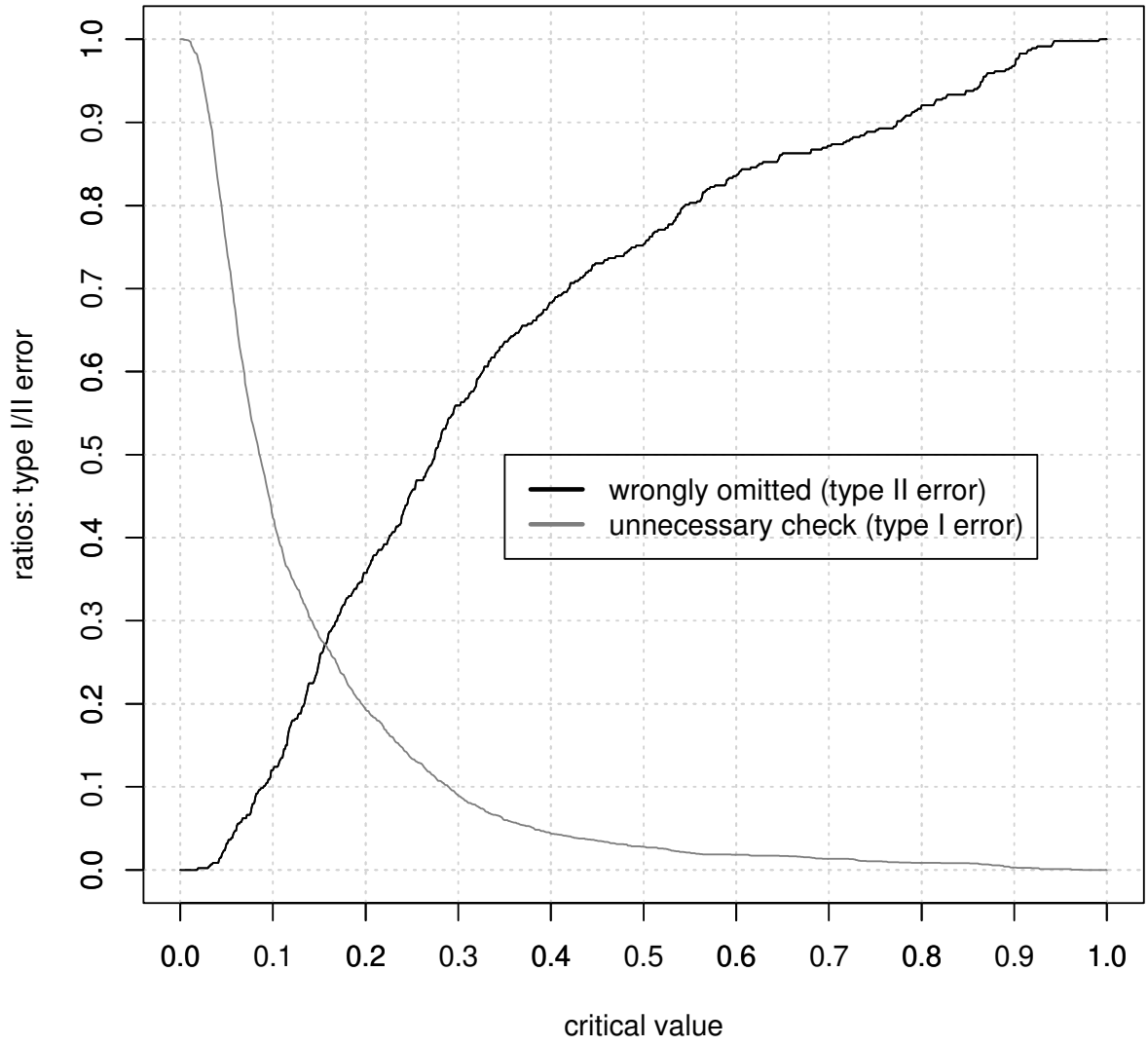
	(0.358)	(0.347)	(0.338)	(0.355)	(0.356)	(0.347)	(0.300)	(0.348)		
Year applied 1996	-0.842**	-0.375	-1.289***	-0.799**	-0.825**	-1.272***	-1.346***	-0.868**		
	(0.371)	(0.355)	(0.349)	(0.368)	(0.368)	(0.358)	(0.308)	(0.360)		
Year applied 1997	-1.047***	-0.636*	-1.476***	-1.004***	-1.029***	-1.677***	-1.718***	-1.156***		
	(0.376)	(0.361)	(0.356)	(0.373)	(0.373)	(0.363)	(0.313)	(0.366)		
Year applied 1998	-0.795**	-0.333	-1.213***	-0.753**	-0.777**	-1.430***	-1.439***	-0.827**		
	(0.373)	(0.354)	(0.352)	(0.370)	(0.370)	(0.356)	(0.304)	(0.362)		
Year applied 1999	-0.953**	-0.461	-1.429***	-0.908**	-0.935**	-1.651***	-1.582***	-1.068***		
	(0.379)	(0.359)	(0.360)	(0.376)	(0.376)	(0.359)	(0.308)	(0.370)		
Year applied 2000	-0.801**	-0.328	-1.334***	-0.755**	-0.785**	-1.514***	-1.540***	-0.949**		
	(0.379)	(0.356)	(0.358)	(0.376)	(0.377)	(0.356)	(0.304)	(0.369)		
Constant	-1.927***	-2.691***	-0.043	-1.916***	-1.907***	-0.410	-0.576*	-2.745***	-2.467***	-2.726***
	(0.416)	(0.365)	(0.369)	(0.416)	(0.413)	(0.373)	(0.308)	(0.399)	(0.271)	(0.268)
Observations	5,482	5,482	5,482	5,482	5,482	5,482	5,482	5,482	5,482	5,482
Log Likelihood	-2,149.847	-2,201.675	-2,267.570	-2,150.346	-2,149.925	-2,103.826	-2,527.223	-2,224.488	-2,184.413	-2,207.123
Akaike Inf. Crit.	4,391.695	4,477.350	4,625.140	4,390.691	4,389.850	4,375.653	5,114.445	4,538.976	4,412.825	4,454.246

*Note: *p < 0.1, **p<0.05, ***p<0.01*
Source: Own Calculations

Appendix E: Predictive Model Performance for Varying Specifications

A series of figures, presented below, allows us to examine the performance of the particular specifications of the logit predictive model. Since yearly dummies are not useable for forecasting, model (9), corresponding to model (1) without them, has been chosen as a Preferred Specification for the text. To facilitate comparison, the results of models (2)-(8) without yearly dummies and model (10) are presented in Table E.1, and Figures E.1-E.8 of Appendix E.

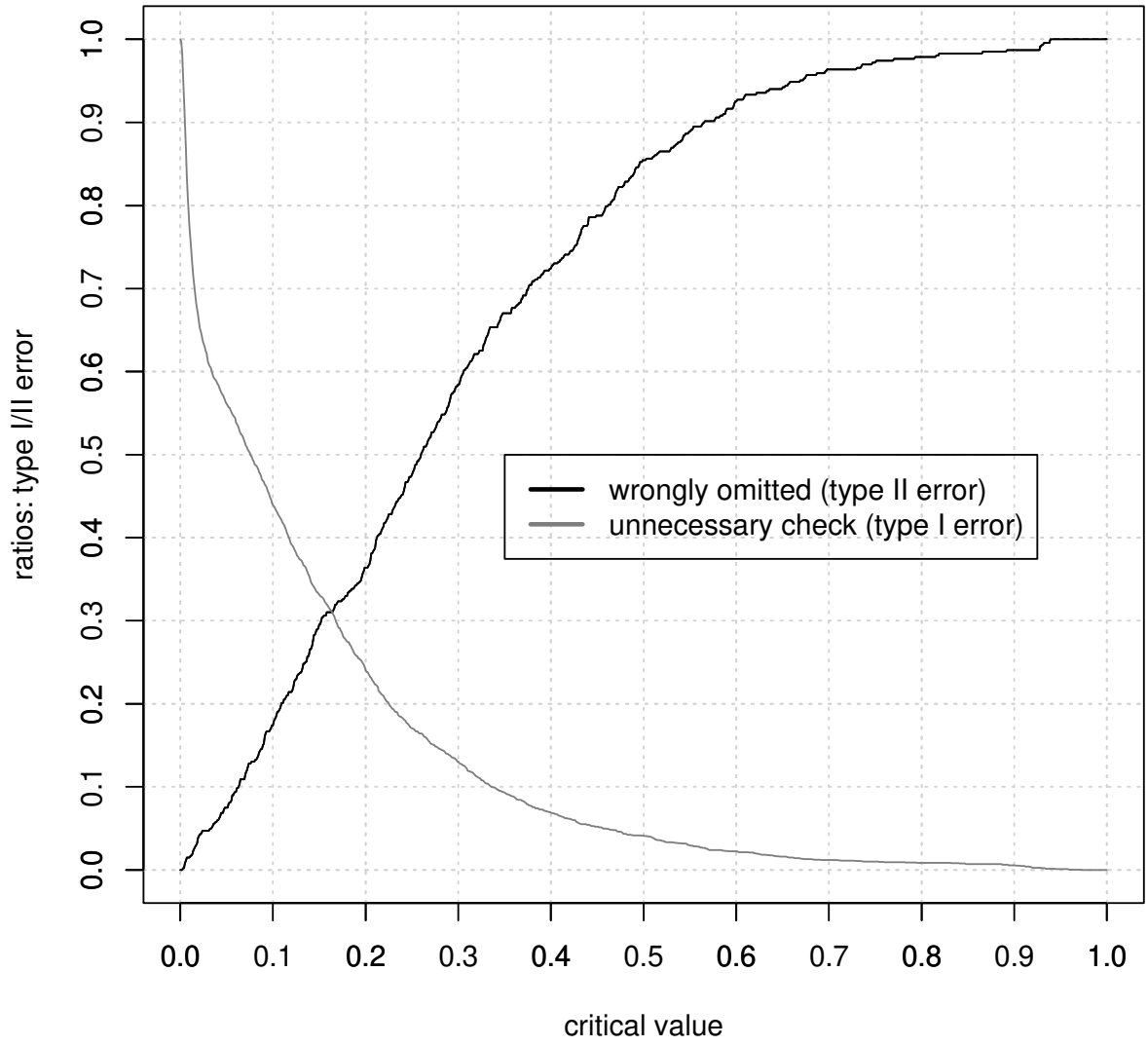
Figure E.1. False positive/negative ratios as functions of the critical value for the baseline model without class dummies (2').



Source: Own calculations

While to opposite is true for the model that omits group- and time- specific patents-in-thickets to patents ratio:

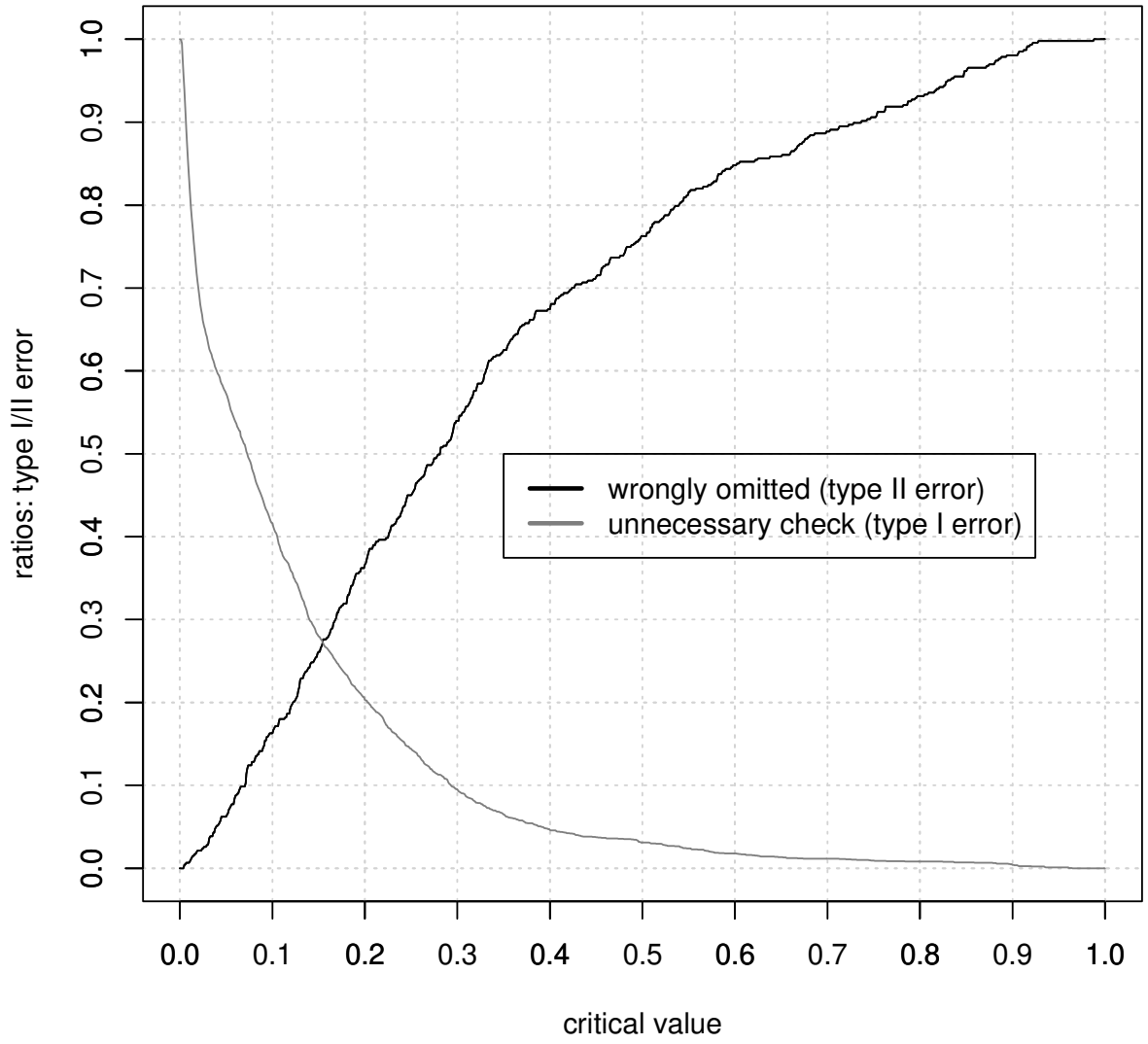
Figure E.2. False positive/negative ratios as functions of the critical value for the model without thicket ratio (3').



Source: Own calculations

Omitting Clarkson ratio does not seem to matter much:

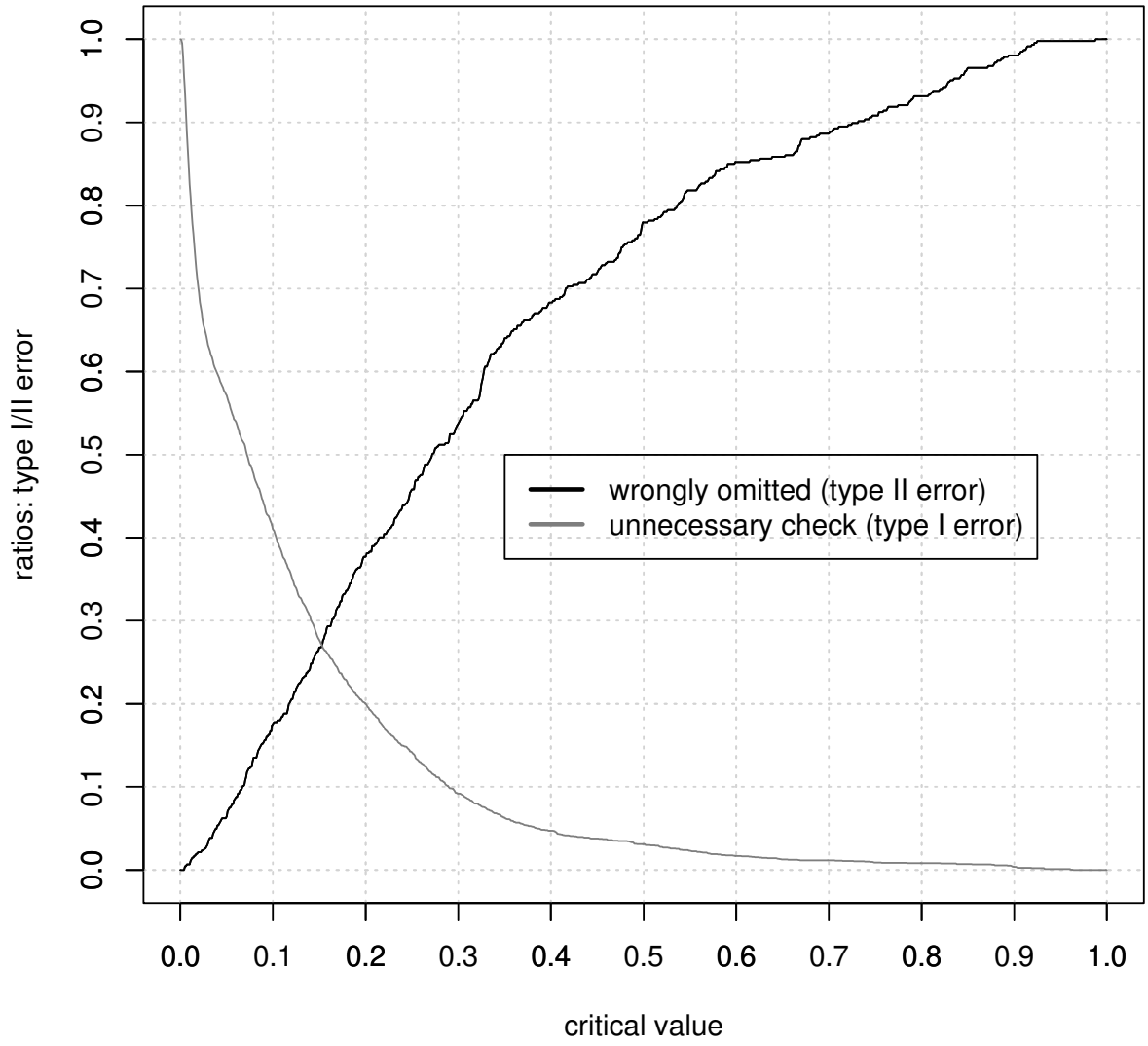
Figure E.3. False positive/negative ratios as functions of the critical value for the model without Clarkson ratio (4').



Source: Own calculations

And the results are similar for triples in the group:

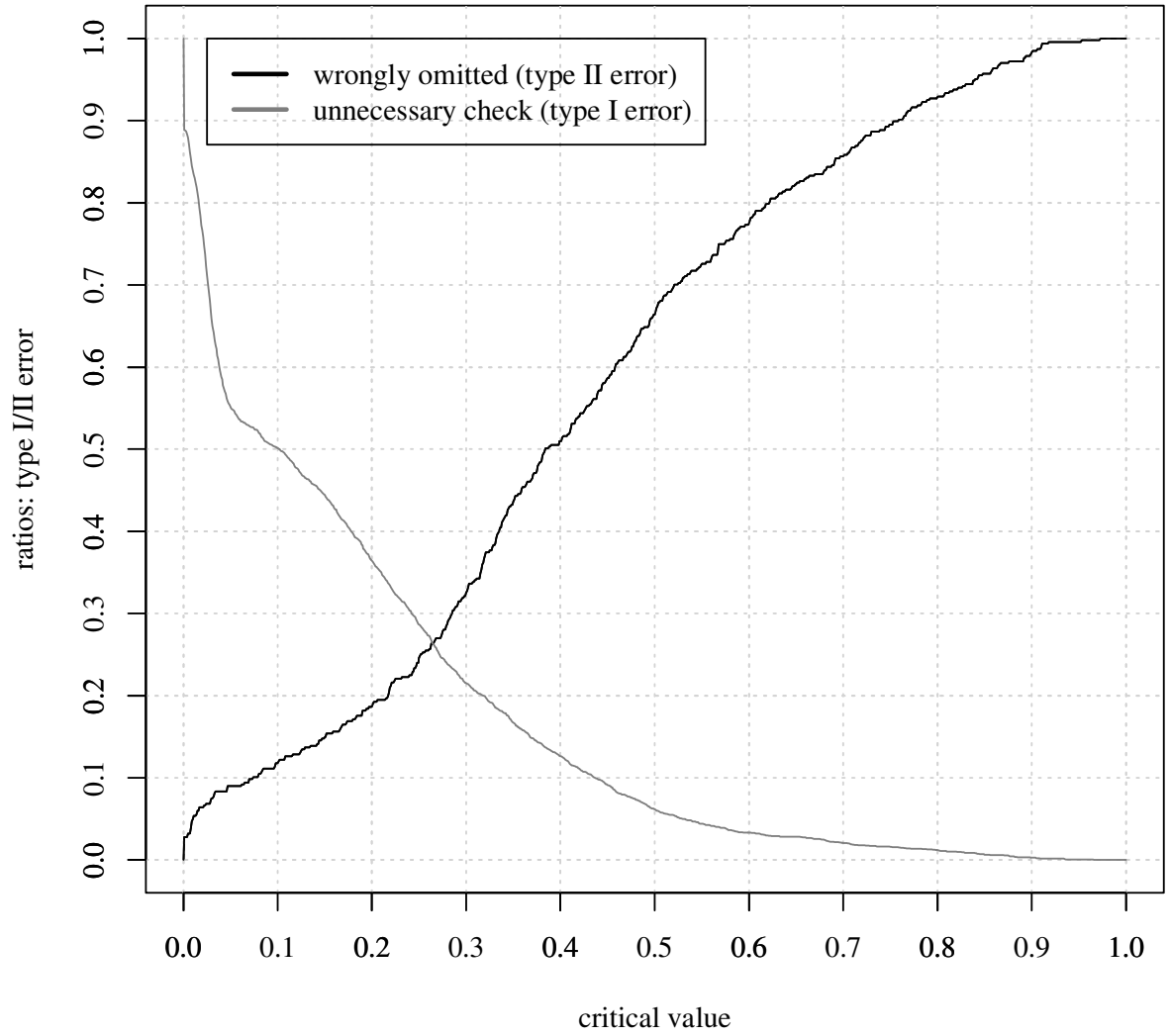
Figure E.4. False positive/negative ratios as functions of the critical value for the model without triples ratio (5').



Source: Own calculations

Replacing all group-specific with time-constant group dummies does not seem to improve the model much (or worsen it):

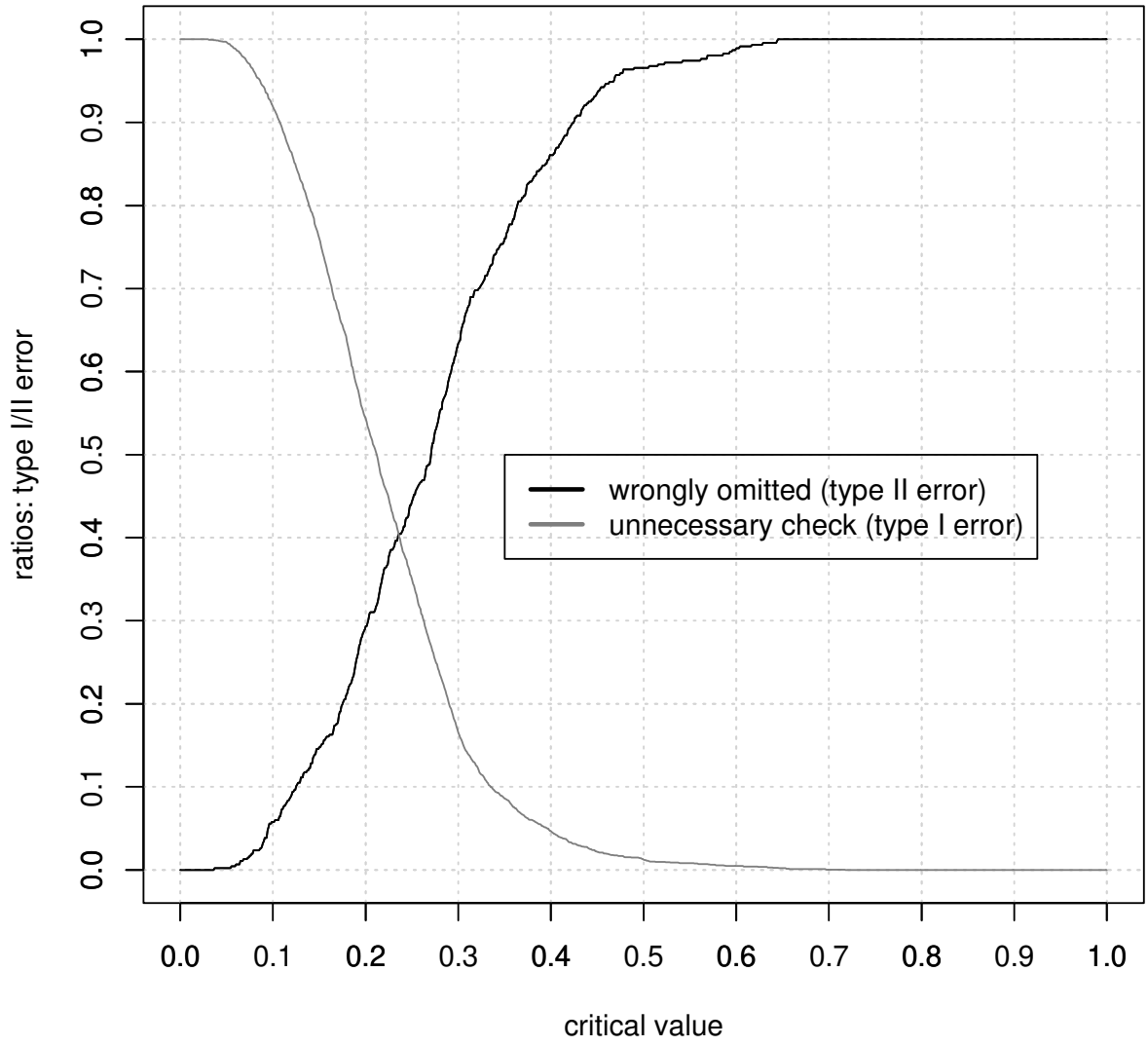
Figure E.5. False positive/negative ratios as functions of the critical value for the model with group dummies (6').



Source: Own calculations

Furthermore, using only patent-specific variables (discarding group characteristic/dummies) mean that the model is not very useful:

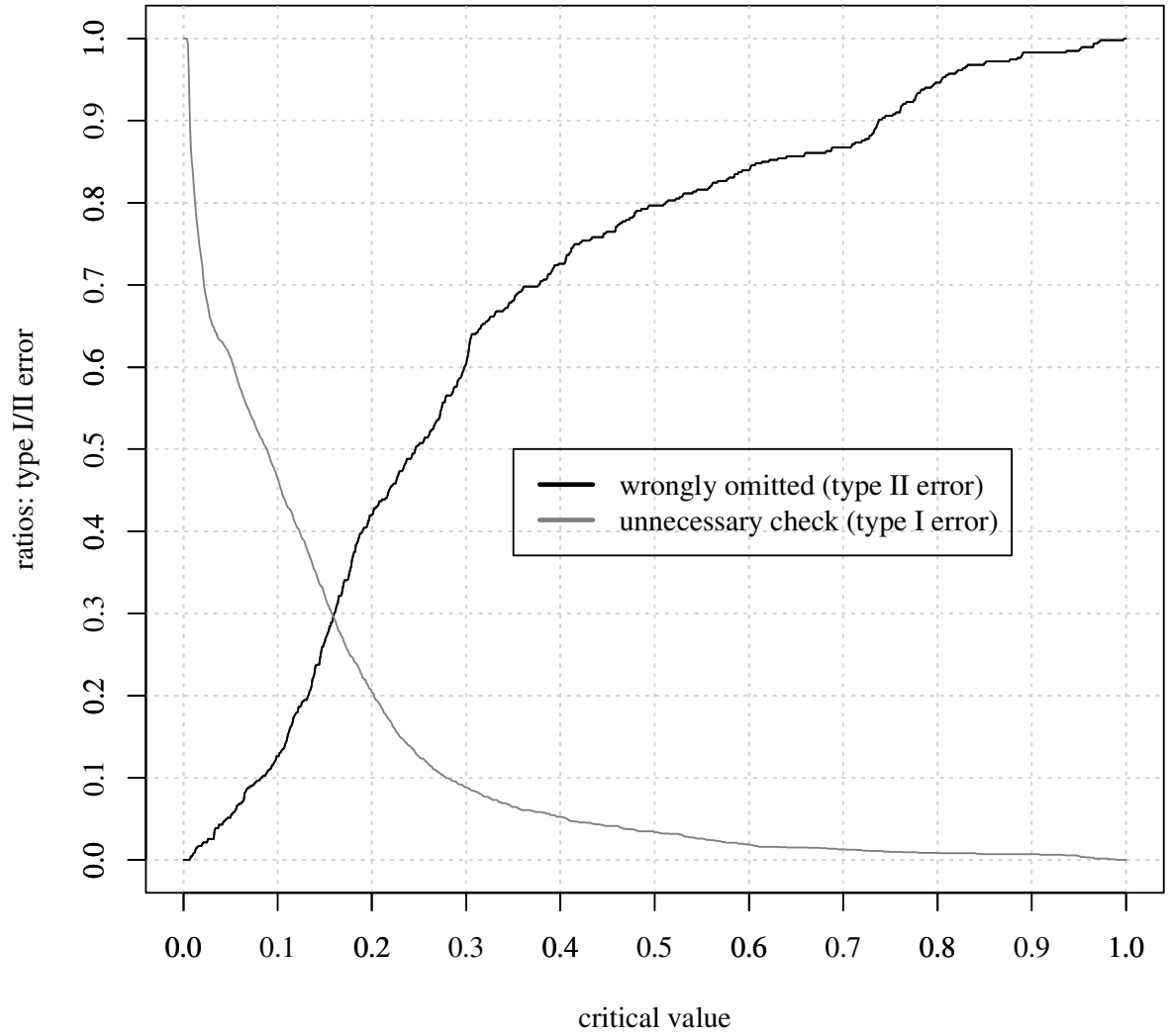
Figure E.6. False positive/negative ratios as functions of the critical value for the model without group-specific variables (7').



Source: Own calculations

Discarding semantic distance seems to worsen the model marginally:

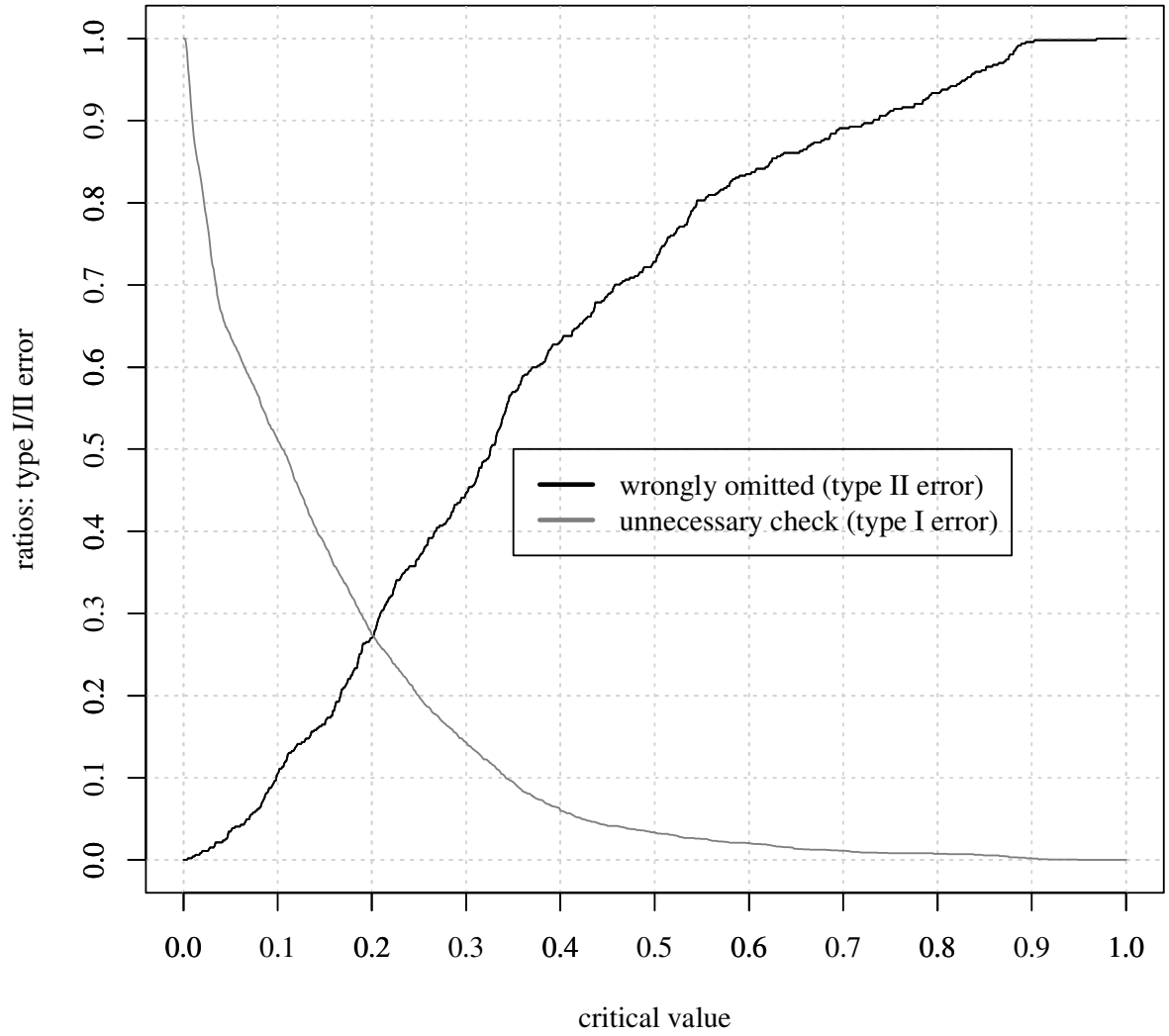
Figure E.7. False positive/negative ratios as functions of the critical value for the model without semantic distance (8').



Source: Own calculations

Omission of historic data on number of fillings and awarded patents has a negligible effect:

Figure E.8. False positive/negative ratios as functions of the critical value for the model without historic data on number of fillings or granted patents (10).



Source: Own calculations

Table E.1. Relationship between a given value of the "false negative" ratio (type II error) and "false positive" (type I error) for different logit models

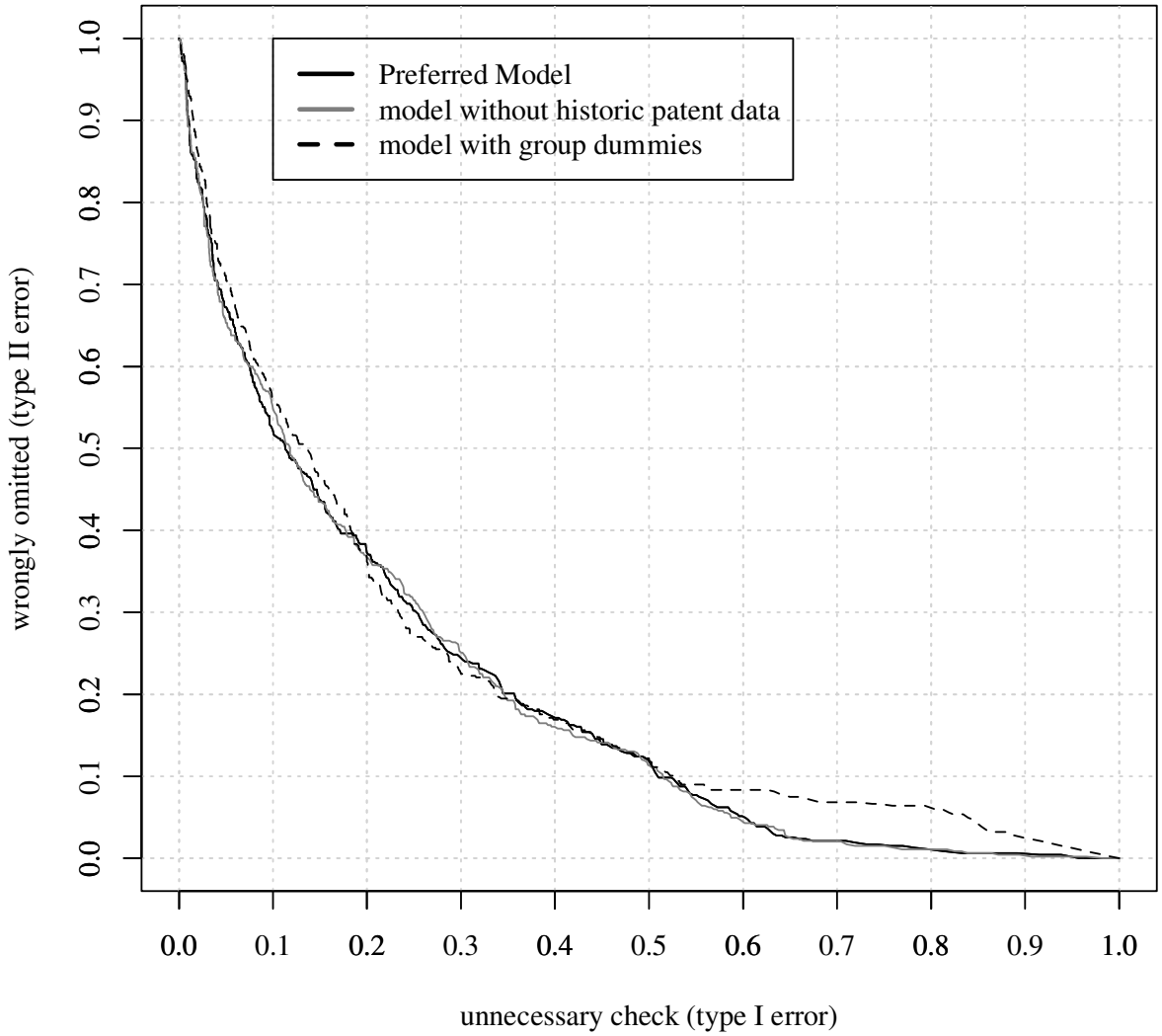
Type II error	***Type I error***								
	Preferred ModeI (9)	model w/o class dummies (2')	model w/o thicket ratio (3')	model w/o Clarkson ratio (4')	model w/o triples ratio (5')	model with group dummies (6')	model w/o group specific variables (7')	model w/o semantic distance (8')	model w/o total patent numbers (10)
0.05	0.603	0.666	0.605	0.603	0.600	0.839	0.934	0.622	0.587
0.10	0.510	0.483	0.535	0.511	0.519	0.527	0.850	0.514	0.516
0.15	0.443	0.365	0.469	0.440	0.452	0.442	0.751	0.431	0.419
0.20	0.358	0.321	0.424	0.355	0.362	0.337	0.656	0.375	0.347
0.38	0.200	0.182	0.232	0.200	0.200	0.191	0.447	0.241	0.191
0.44	0.149	0.142	0.190	0.151	0.149	0.166	0.354	0.178	0.148
0.52	0.101	0.107	0.155	0.100	0.099	0.119	0.261	0.116	0.112
0.67	0.052	0.047	0.094	0.053	0.052	0.061	0.146	0.069	0.047

Source: Own calculations. Models x' correspond to models x from Table D.1 without the year dummies. The ratios are: number of applications that would not be in thickets flagged as in-thicket patents (i.e. selected for an unnecessary check) and number of applications that would be in thickets flagged as not-in-thicket patents (i.e. wrongly omitted from selection for a check).

Table E.1. summarizes the performance of all the model specifications as an aid to determining which specification performs best as a predictive model. The table lists the share of the patents wrongly omitted from thickets that actually belong to a thicket (type II error) corresponding to the illustrated share of patents that don't belong to a thicket but are wrongly identified as doing so (type I error). By doing so, Table E.1 corresponds to Figures 5, 6 in the text and E.9, providing numerical values. The Preferred Model of the text is listed in the first column to the right of the heavy vertical line, with other comparators listed farther to the right.

Clearly, lack of inclusion of data on the patent groups significantly worsens the predictive power (7'vs 9) in the sense that the type I error increases strongly for a given type II error (see also Figure 5 in the text). Omission of only the thicket ratio or semantic distance almost always worsens results substantially (3' and 8' vs 9; see also Figures 5 and 6 in the text), while omission of only number of past applications and granted patents within the group worsens performance less (or, one could argue, enhances it; 9 vs 10; see also Figure E.9). Similarly, the impact of omission of the class dummies is substantial but inconsistent (2' vs 9). Omission of the Clarkson ratio or triples reduces performance inconsistently and usually negligibly (4' and 5' vs 9; see also Figure 6 in the text). Interestingly, using (time-static) group dummies often does not work better than using (dynamic) group-specific variables (6' vs 9; see also Figure E.9).

Figure E.9. False positive/negative ratios tradeoff for the Preferred Model and alternative specifications (III).



Source: Own calculations