Prospective Patent Thicket Identification

Mateusz Gątkowski[*a], Marek Dietl[*b], Lukasz Skrok[b], Ryan Whalen[c], Katharine Rockett[ad+]

[a] *University of Essex, Department of Economics, Wivenhoe Park, Colchester CO4 3SQ, UK*
[b] *Faculty of Law, University of Hong Kong, Hong Kong*
[c] *Warsaw School of Economics, Department of Business Economics, Warsaw, Poland*
[d] *CEPR*

Abstract

Patent thickets have been identified by various citations-based techniques, such as Graevenitz et al (2011) and Clarkson (2005). An alternative direct measurement is based on expert opinion. We use natural language processing techniques to measure pairwise semantic similarity of patents identified as thicket members by experts to create a semantic network. We compare the semantic similarity scores for patents in different expert-identified thickets: those within the same thicket, those in different thickets, and those not in thickets. We show that patents within the same thicket are significantly more semantically similar than other pairs of patents. We then present a statistical model to assess the probability of a newly added patent belonging to a thicket based on semantic networks as well as other measures from the existing thicket literature (the triples of Graevenitz and Clarkson's density ratio). We conclude that combining information from semantic distance with other sources can be helpful to isolate the patents prospectively that are likely to be members of thickets.

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

[*] Part of the research was supported by the National Science Center (NCN), grant No. DEC-2013/11/B/HS4/00682 "A new method for identification of patent thickets"
University of Essex, Department of Economics affiliation is an honorary agreement
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 provided invaluable ideas and advice.
[+] Corresponding Author: kerock@essex.ac.uk. Address: Dept of Economics, University of Essex, Wivenhoe Park, UK CO43SQ

1. Introduction

Measurement of patent thickets is an important issue within the intellectual property field. Fragmentation of claims ownership dispersed in patent space has been used to measure the existence of patent thickets. Measurement has been based on qualitative methods such as interviews with executives on patenting strategies (Hall and Ziedonis (2001)) or on examination of citations to prior art and its fragmentation as measured by a Herfindahl concentration index (Ziedonis (2004), Galasso and Schankerman (2010)). Clarkson (2005) and Clarkson and De Korte (2006) suggest calculating measures based on network density of citations. Graevenitz et al (2011) suggests identification of critical references and calculating the density of "triples", which are specific networks of these references.

The citation based methods propose measures that return a proxy for overlap in a given patenting space. Graevenitz et al (2011) recognise the need to show external validity of these proxies, and do so by noting the match between occurrence of triples and the complexity of the technology, identified by Cohen et al (2000) in surveys of managers.

An alternative means of external validation is to return to the qualitative methods that were present early in the literature, however, by surveying field experts. This can be used to create a thicket measure to compare to citations-based measures. Advances in natural language processing techniques facilitate this work, as they can be used to capture formally the networks created by expert classification of patents. This is the exercise we undertake here. We also combine the different thicket measures in a statistical model to predict thicket membership based on the pool of past patent characteristics.

A difficulty of such an exercise is to give the experts sufficient guidance to detect what the researcher means by a "patent thicket". Here, the literature gives mixed signals. Patent thickets have been referred to variously as "blocking patents", "patent floods", or "patent clusters" in (IPO, 2011). In a recent review of the literature, Egan and Teece (2015) enumerate the many

different definitions that have been used, each definition associated 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 the 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). A popular citations based measure of 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 we provide to our experts is closely related to both.

Indeed, patenting at high volumes appears to be occurring: according to the World Intellectual Property Organization (WIPO) there was respectively a 9.2%, 9% and 4.5% year-on-year increase in the number of patent applications in 2012, 2013, and 2014. The total number of new patent applications in 2015 was 2.6 million (see WIPO, 2013, 2014, 2015). To the extent that these patents do not represent completely new areas of work, overlapping or even conflicting claims may rise with this increase in total applications (Barnett, 2014). If in addition there is a "lack of resources and misaligned incentives at patent offices dealing with a flood of patents" (Hall et al., 2013), then a contributing factor in generating thickets may be low quality drafting. Indeed, Hall and co-authors note that the critical references that form the basis of the Graevenitz et al (2011) exist precisely to allow patent examiners to redraw claims so that the underlying wording in the claims points to less patent overlap. An alternative approach departing from the same starting point is that, as the patent overlap derives from the wording in the claims, it should be traceable to those words via semantic analysis.

The view that claim writing may be behind thickets has been noted by others, who question the quality of patent documents. Holman (2006) proposes that junk patents, which arise when patents are granted too broad claims, may contribute to thickets by creating patent interferences via extending their scope beyond the underlying innovation. The problem could be alleviated if

patent offices examined the patent claims more thoroughly but this requires more effort and scarce resources. Indeed, patent examiners spend very little time reviewing each application. According to Lemley (2001), overall time spent per application is about eighteen hours spread over the months of a patent granting process. He concludes that an automated method of comparing claims of a patent or an application against the pre-existing set of patents would increase significantly the efficiency of the patent thickets screening process. We will return to this point below[1]. Indeed, while some might be tempted to rely on litigation or the patent fee structure to weed out low quality patents, a recent paper by Schankerman and Schuett (2018) casts doubt on this view. This makes efficient screening more pressing.

Semantic analysis of patents has already used in a few papers in the patent literature. A study by Preschitschek et al. (2013) studied semantic similarities in text of chosen USPTO patents and showed predictive power of semantic analysis for technology convergence. Yoon and Park (2004) was one of the first pieces of work where a network of keywords found in patents belonging to a technology field was used. Gerken and Moehrle (2012) used semantic analysis for detecting the novelty of innovations. Recently, Bergeaud et al. (2017) presented a method for classifying patent technologies with an automated model for analysing semantic contents of patent abstracts. Khun and Thompson (2017) analysed the word-count of the patents' first claim in order to assess patent scope. Our study adds to these but focuses on patent thicket measurement.

Document similarity can be measured in a wide variety of ways (Harispe, Ranwez, Janaqi, & Montmain, 2015). Many of the most common rely on document metadata to infer content and

[1] Consistent with the interpretation that thickets results 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).

perform similarity comparisons. In the context of patents this is often done by categorizing inventions according to their technical classification (e.g. USPC or IPC) and treating inventions of the same category as similar (see e.g., Fleming, 2001). However, these metadata-oriented approaches essentially gloss over intra-category variation between documents and as such provide only coarse measures of document similarity.

To achieve more precise similarity measures, one must look to the contents of documents and compare them to one another. 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 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). These relatively simple text comparison methods are, however, hampered by their inability to detect latent similarities between documents that might contain content on a similar area, but use a somewhat different vocabulary to discuss it.

More sophisticated natural language processing techniques address many of the weaknesses of these comparatively simple document comparison methods. Latent Semantic Analysis (LSA)[2] is a well-established method to detect latent similarities between texts and compress them to a common set of dimensions that can then be compared to one another (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990; Landauer, Foltz, & Laham, 1998). LSA takes account of word co-occurrence to detect latent similarities in the way vocabularies are used (e.g. the fact that "car" and "automobile" appear in similar contexts) and as such, the resulting reduced-dimensional vector representation of the documents can be used to more accurately identify similar or dissimilar documents.

To the best of our knowledge ours is a first attempt to use the semantic distance between patent documents to identify patent thickets. We calculate pair-wise patent semantic distance, creating a semantic patent network, where links between patents are weighted by the semantic

[2] More information on LSA can be found in appendix – Latent Semantic Analysis

distance. In order to form a more accurate picture, we use the full body text of patents including claims, rather than focusing solely on keywords or abstracts. We hypothesize that the overlapping rights indicative of patent thickets will result in semantic similarity between patents that occupy the same thicket. To test this hypothesis we benchmark semantic similarity measures against a set of expert identified patent thickets, which we implicitly take as accurate measures that we are attempting to replicate by our semantic analysis.

Apart from an identification of a group of patents as a patent thicket, the relationship of patents within the group is important as well. A standard list of possible relationship types would include blocking, complementary, independent, or substitute patents (Clarkson, 2005). While the Hall and Ziedonis (2001) and Ziedonis (2004) methods focus on general blocking relationships, Clarkson (2005) looks instead at substitutes, which he argues can also generate hold-up. The method presented in this paper, based on semantic networks could potentially cover either substitutes or complements, although our treatment is linked more to the former. The method is also directly linked to the Shapiro (2001) definition of patent thickets via our instructions to our experts on thicket identification. Being a contents-based method we derive linkages directly from overlapping content of claims rather than via citations. Our work includes a measure of fragmentation, which is also used to proxy the portion of patent thickets reflecting hold-up rather than defensive patenting concerns in some work (Noel and Schankerman, 2006, Galasso and Schankerman, 2010). We have a separate measure for the thicket itself, however, with fragmentation added alongside our semantic indicator.

We find that, indeed, patents belonging to the same expert-identified thicket are closer semantically than other combinations of patents: in other words, semantic distance varies with the membership of any pair of patents in different thickets, or membership in a thicket while the other patent is not a member of a thicket, or whether neither belongs to a thicket. Patents within 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. Finally, we find that the semantic distance between patents in discrete technological

areas is shorter than it is for complex technologies, which accords with the intuition that in complex technologies patent thickets may cover a wider range of patent claims.

Our findings suggest the possibility of creating quantitative semi-automated methods for preliminary patent thicket screening in the same vein as the critical references discussed by Hall et al (2013), based on semantic distance to the extent that it captures the evaluation an expert would make. An advantage of such a method is that it is quick and can be done prospectively: there is no need to wait for patents to accumulate to determine if a patent is likely to be a thicket member and the identity of that thicket. We propose such a statistical model, which provides a certain probability that a given patent belongs to a thicket and test it, showing that semantic distance can be informative of expert opinion, whereas triples and weighted average patent citation density are broadly indicative but not as closely tied. This may be due to the definition of patents used in the expert instructions: our point is that expert opinion can be represented by semantic analysis, and this can be used to represent the thicket definition that is selected. If the resulting thickets differ from the patents flagged by other methods, it may indicate that those methods correspond to a different definition. Whether that alternative definition is better or worse for the purposes of the analysis is an issue we do not address. As pointed out by Egan and Teece (2015), the appropriate definition will depend on the policy question that is posed.

In the remaining part of the paper, section 2 introduces the methodology used, section 3 presents the results, section 4 contains results on an overlap between patent thickets, triples (Graevenitz et al., 2011), and weighted average patent network density (Clarkson, 2005), section 5 introduces an analytical model for thicket recognition based on semantic networks and section 6 concludes.

2. Methodology

We use data from the USPTO on 12,312 patents from 58 patent groups (subclasses within the USPC classification scheme), sampled as for 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 8 subject matter experts and asked them to review patents in the 58 patent groups under study. The experts were assigned to patent groups relevant to their fields of expertise. They were each asked to review patents within their field and identify those belonging to patent thickets. More precisely, these thicket identifiers were experts in the fields of electrical systems, chemical engineering, material engineering, electricity: measuring and testing, electrolytic coating, nanostructures, dentistry, drugs, medical chemistry, surgery, image processing.

To ensure they were working with similar definitions of what constitutes a patent thicket, we defined patent thickets for them using 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". We incorporated the additional words based on feedback on confusions that could arise from using the shorter original definition. We do allow for the patents to be associated with a single firm, so that defensive as well as hold up reasons for thicket generation can potentially be taken into account. Hence, we allow for the full scope of issues that Egan and Teece flag as being evoked by this definition.

Upon completing their review of the patents within each of their technical areas, the subject matter experts identified 307 patent thickets containing 2732 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.

With this expert-identified set of patent thickets in hand, we then set out to attempt to measure the semantic distance between patents identified as thicketed and non-thicketed patents. In order to determine the distance between patents, we first performed a latent semantic analysis (LSA) of the entire corpus[3] of granted utility patents issued from 1976 until late 2014. This LSA was

[3] First, extremely common and uncommon terms were removed from each documents term vector. Any term

performed by extracting the text from each patent's abstract, description, and claims fields. These fields were combined, and then standard natural language processing transformations were made across all of the term vectors within the corpus. The resulting re-weighted term vectors were used to compute a 500-dimension LSA model.

With the model computed, each document is then assigned a 500-dimension vector representing its composition of topics within the entire patent semantic space. We then use these vectors to compute the pairwise cosine distance between patents of interest, the most commonly-used measure for distance within vector space models (Turney & Pantel, 2010). Patents with a low score are 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 calculate scores for each patent pair within the 58 patent groups[4] and then compare the average distances for four different sets of the pairs using Welch's unequal variances t-test test for mean equality[5] (Welch, 1951). Pairs were divided in four sets[6]: 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 none of the patents belong to a thicket.

3. Results

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

Our primary finding is that the average 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

---

appearing fewer than 5 times across the entire corpus was dropped, as was any term appearing in over 50% of all documents. We then performed a TF-IDF transformation on the resulting term vectors, so that particularly important terms would be more heavily-weighted while less important terms would be more lightly-weighted.

[4] There were overall more than 3.7 million patent pairs

[5] It is a version of a Student-t test. It is more robust when samples have unequal variances and sizes.

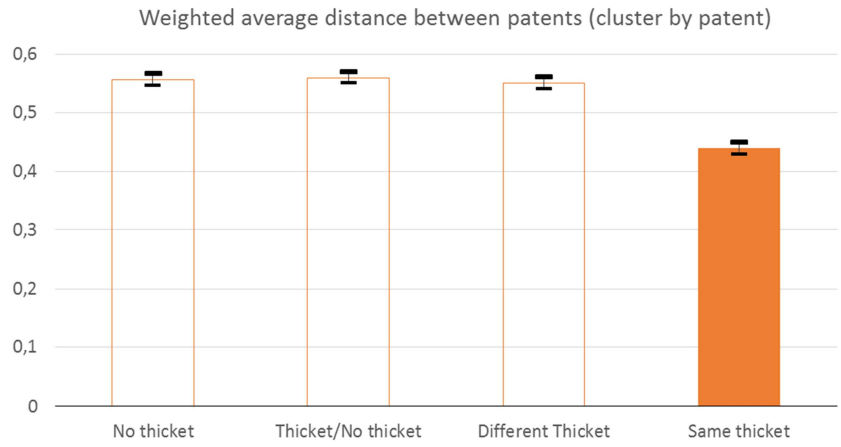
[6] We use "set" to describe groups of pairs of patents – depending whether patents belongs to a thicket or not; we use "patent group", when we mention the USPTO patent classification.

than pairs of patents that do not inhabit the same thicket, and that this difference is detectable using natural language processing techniques.

We demonstrate this in Figure 2, 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 (as defined in section 2), calculated as the average of the distances in each of the 58 patent groups, weighted by the number of patents. We have calculated the significance of differences using the linear OLS regression model with errors clustering. The base scenario is set I (same thickets). Dummies are used for the remaining sets. We have clustered errors by patents (the same patent can belong to more than one pair). All the coefficients are significant at a high level – more than 99.999%. We have also tested the regression with clustering of errors by patent groups (58 clusters) and the results hold with a similar significance level (the lowest being 99.997%).

Figure 1 shows confidence intervals with significance level of 95%.

Figure 1. Weighted average distance between patents in orange bars (errors clustered by patent), with confidence intervals illustrated by black bars (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	0.0054215	0.0050359	0.0051052	0.0052137
No of pairs in the sample	2 425 272	1 100 243	162 910	30 420

Source: Own calculations

Table 1. Results of the test for mean equality of semantic distance: the percentage of the patent groups (out of 58), 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
63.8%	**60.3%**	**55.2%**	36.2%	48.3%	48.3%	<=0.0001
70.7%	**62.1%**	**56.9%**	43.1%	58.6%	53.4%	<=0.001
72.4%	**70.7%**	**63.8%**	53.4%	60.3%	56.9%	<=0.01
82.8%	**77.6%**	**69.0%**	63.8%	70.7%	70.7%	<=0.05
13.8%	**19.0%**	**19.0%**	24.1%	17.2%	29.3%	>0.05
3.4%	**3.4%**	**12.1%**	12.1%	12.1%	0.0%	Test not possible

Note: The grey cells contain cases, where test was not passed with more than 95% significance or it was not possible to perform because there were not enough thickets in the group. Bold columns show results for the differences of mean semantic distance between "same thicket" and other sets.
Source: Own calculations

While figure 2 strongly suggests that patents identified to be in the same thicket are semantically more similar, we additionally confirm that these differences are statistically significant by testing for the mean equality between sets of pairs of patents using the Welch test. We perform this test on six different combinations starting with 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 each of the 58 patent groups.

Table 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 95% significance level as a cut-off value for the test. In some cases, where there was only one or no thicket in a patent group, it was not possible to conduct the test.

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

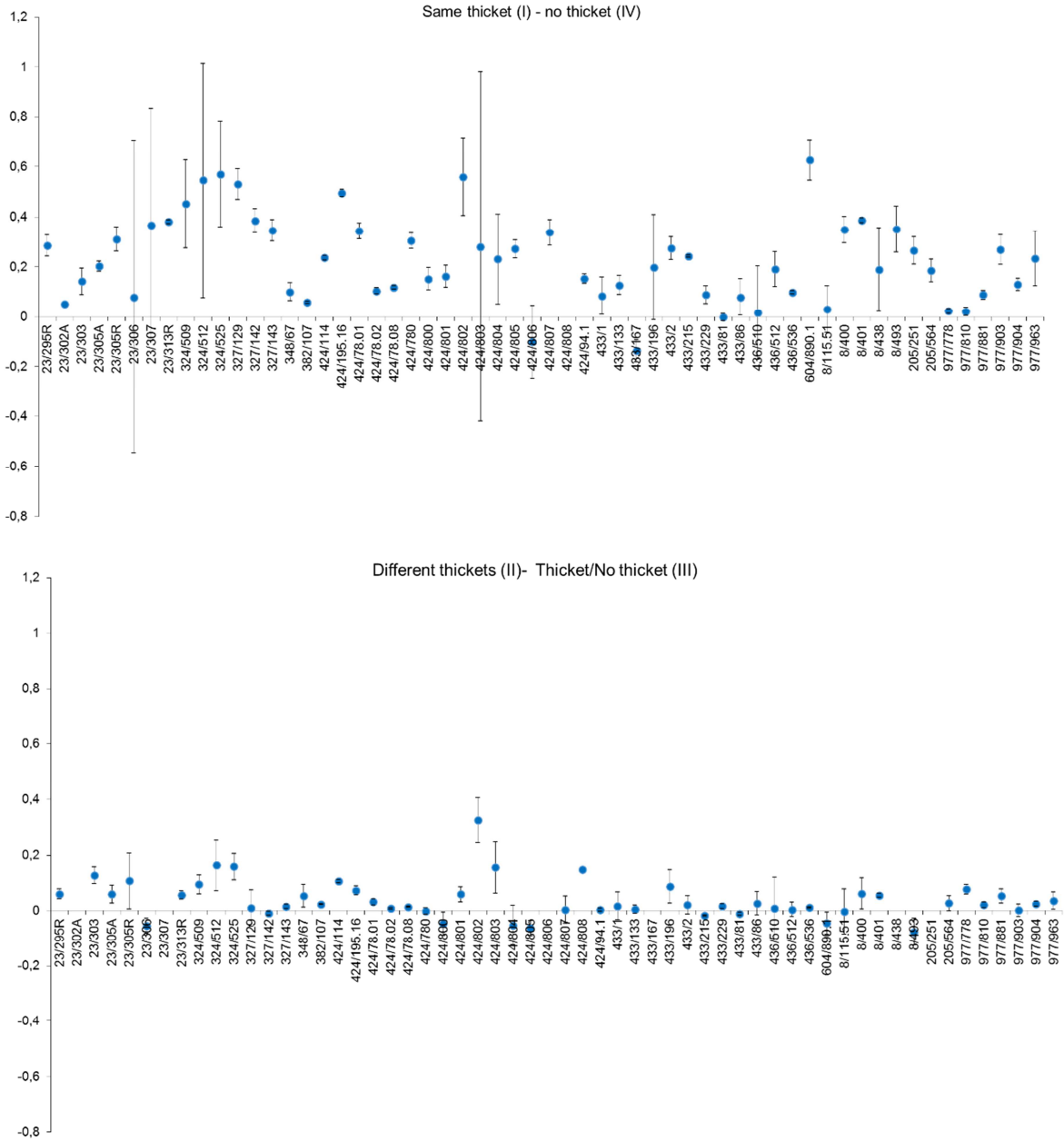
setup, from the 58 patent groups, 82.8% (when testing for difference between averages in sets I and IV, i.e. patents belonging to the same thicket and patent outside any thicket) to 69% (I-II, i.e. patents from the same thicket compared with patents from different thicket) of groups have passed the test for difference in average semantic distance with 95% significance. What is more, the results are strong, as 63.8% to 55.2% of the results were significant with more than 99.9999% significance.

The differences in average semantic distances between other sets are also evident. However, the difference is least significant for the sets II – III, that is between different thickets and thicket/no thicket sets. There were 36.2% of groups with non-significant results of Welch test or where test was not possible. Also there were only 36.2% of groups significant at the level of 99.9999% or more, compared to 48.3% for differences between sets II-IV and III-IV. Nevertheless, tests show that for the majority of groups, all four sets are distinguishable.

Figure 2 presents two charts showing differences in average semantic distances for different patent groups (blue dots) with 95% confidence intervals. The sets with the strongest and the weakest differences were selected for the graph[7]. The first chart shows the difference between "same thicket" and "no thicket" sets, the second "different thickets" and "thicket/no thicket". As shown in the table above, the first case exhibits larger differences in average semantic distance than the second case. It also shows that there are also groups where the semantic distance in set I ("same thicket") is not the shortest (in total there are six groups where this is a case), but the general tendency for set I to have the shortest semantic distance is clearly visible.

Figure 2. Average semantic distances (blue dots) between chosen sets with confidence intervals (black bars). Where the confidence interval overlaps with 0, the result is statistically insignificant (at 95%).

[7] The remaining charts for the four tests are included in the appendix.



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 with the

results for three other sets. Therefore, to confirm the above results, we perform the tests using a Bonferroni correction, which is one of the methods to compensate for multiple comparisons. In this context, the Bonferroni correction requires that in order to reach 95% statistical significance of the 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 = 0.01667, i.e. be significant at 98.333% level.

When the Bonferroni correction is taken into account, the semantic distance in set I, "Same thicket", remains significant at 95% level in 37 patent groups; set II "Different thickets" in 26 groups; set III "Thicket/No thicket" in 21 groups and set IV "No thicket" in 26 groups. For each set the total number of groups for which all three tests could be performed was 51, so, expressed in percent: set I – difference is significant in 72,5% of groups; set II – 51%; set III – 41,2%; set IV – 51%. 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.

The main finding from this analysis is that patents belonging to the same thicket as identified by our experts are semantically more similar to one another than patents that belong to two different thickets. Furthermore, when two patents belong to different thickets the distance between them is greater, but also much closer to that observed to the case when there is no thicket or just one of the patents belong 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 so it is perhaps unsurprising that overlapping IP rights corresponds with detectable semantic similarity. As the results show that semantic similarity is a good proxy for expert identification, however, it suggests that we can use semantic similarity to identify potential patent thickets, taking expert opinion as reflecting a valid definition of a thicket. We explore this further in section 4, but for the remainder of section 3 we further explore the characteristics of our semantic groupings.

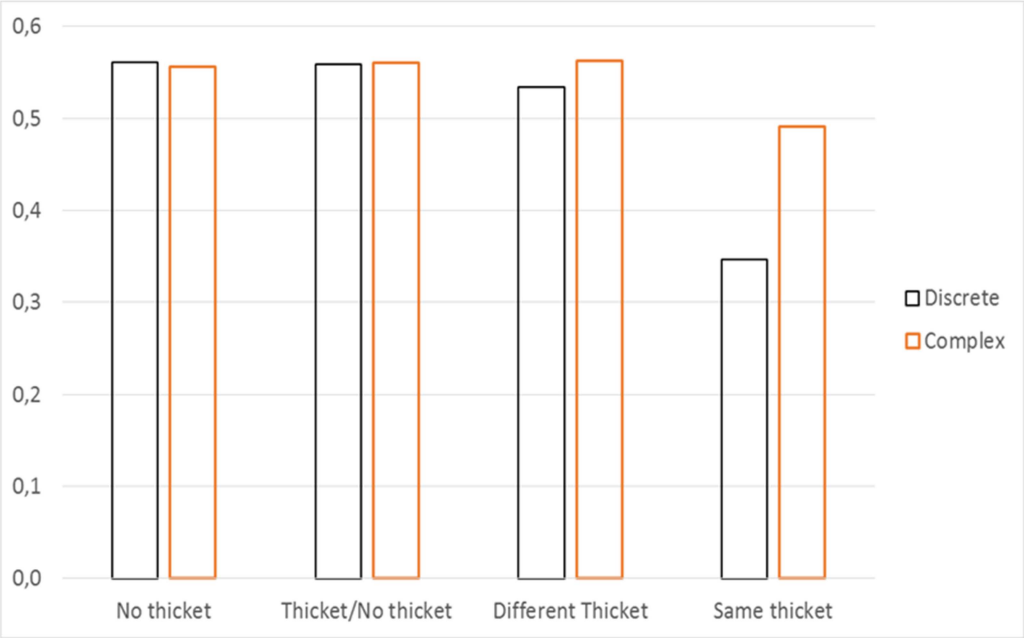
3.2. The semantic distance effect is greater in discrete than in complex technology areas

In addition to comparing the semantic distance between patents in and outside thickets, we can also explore how the 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 implicated 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. The list of patent groups and their membership in complex or discrete technology type can be found in the appendix.

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, as we mentioned above.

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	0.000062	0.000071	0.000214	0.000855
Complex	0.000048	0.000058	0.000167	0.000839



Source: Own calculations; Note: Confidence intervals are so small that we don't show them on the graph

Table 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 the overall conclusions from the full sample presented in Table 1. The overall tendency, however, is that there is a higher percentage of groups with statistically significant differences at 95% for complex than for the 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 2. Results of the test for mean equality for discrete and complex areas: the percentage of the patent groups (out of 58), 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						
68,8%	**65,6%**	**56,3%**	37,5%	46,9%	40,6%	<=0.0001
71,9%	**65,6%**	**59,4%**	43,8%	56,3%	46,9%	<=0.001
71,9%	**68,8%**	**62,5%**	50,0%	56,3%	50,0%	<=0.01
78,1%	**75,0%**	**71,9%**	62,5%	62,5%	62,5%	<=0.05
15,6%	**18,8%**	**9,4%**	18,8%	18,8%	37,5%	>0.05
6,3%	**6,3%**	**18,8%**	18,8%	18,8%	0,0%	Test not possible
Complex						
57,7%	**53,8%**	**53,8%**	34,6%	50,0%	57,7%	<=0.0001
69,2%	**57,7%**	**53,8%**	42,3%	61,5%	61,5%	<=0.001
73,1%	**73,1%**	**65,4%**	57,7%	65,4%	65,4%	<=0.01
88,5%	**80,8%**	**65,4%**	65,4%	80,8%	80,8%	<=0.05
11,5%	**19,2%**	**30,8%**	30,8%	15,4%	19,2%	>0.05
0,0%	**0,0%**	**3,8%**	3,8%	3,8%	0,0%	Test not possible

Note: The grey cells contain cases, where test was not passed with more than 95% significance or it was not possible to perform, because there were enough thickets in the group. Bold columns show results for the differences of mean semantic distance between "same thicket" and other sets.

Source: Own calculations;

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, 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 methodology. 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 which belong to thickets in these areas are more diverse, i.e. these thickets are also more complex, covering a larger variety of rights.

3.3 The above results hold if we control for experts

As we have used expert opinion as our base, we repeat our analysis controlling for the identity of each expert so as to see if there are any outliers in behaviour. The control for expert identity is both interesting in itself and a way for us to be sure that errors in individual judgement were not driving our overall results. We find that our results still hold: accordingly, Table 3 below breaks down the tests by expert.

There is no difference in the main conclusions presented in 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.

Table 3. 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%, by expert. 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.

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).
Source: Own calculations;

4. Comparison of an expert-based method of patent thicket recognition, triples and network density

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 has recently attracted significant attention. Triples are 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 there are overlapping portfolios of three firms, the negotiation process between them or with another entity is costly in terms of resources. 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 the expert-based method of 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. We reproduce the triples thicket identification method on patents granted by the EPO[8] and map these, where possible, with patents from our USPTO sample, using the PATSTAT database. Triples are calculated in the OST-INPI/FhG-ISI technology areas in order to keep the comparability with the original Graevenitz et al. (2011) paper and because the measure requires broad samples. Subsequently, we compare patents that belong to triples with patents that were identified as belonging to thickets by field experts. When interpreting the results one must remember that 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.

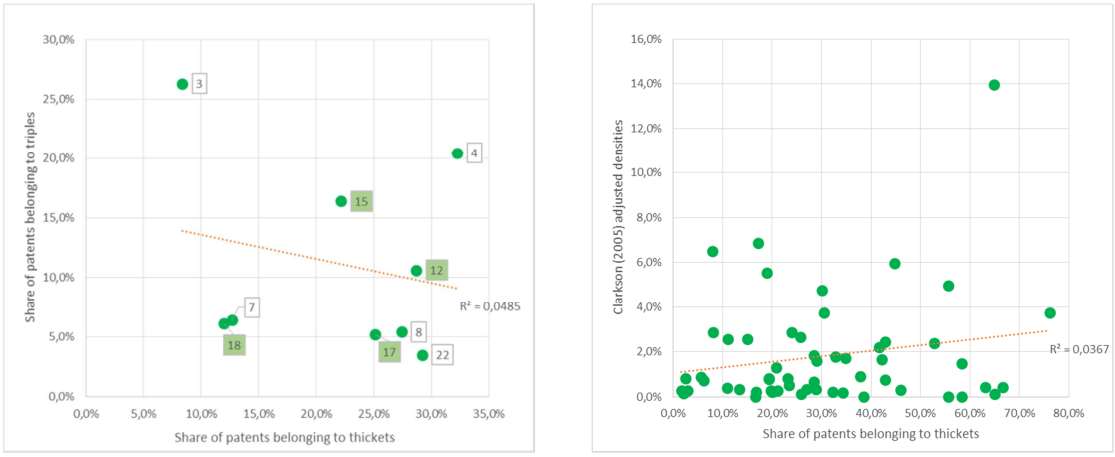
The comparison shows that only 3.7% percent of patents in expert-identified thickets belong to the triples. This is barely higher than the baseline 3.2% thicket membership we observe when we look at all of the patents, from our USPTO sample that were mapped to EPO patents. This small increase in the 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. The left pane of Figure 4 bolsters this point by examining individual technology areas and plotting the share of patents in expert-identified thickets against the share of patents belonging to triples identified using the Graevenitz et al. (2011) method. A simple regression run on the data shows little overlap between the two with R^2=0.0485.

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 and given the definition we provided to the experts. The two have identified quite different sets. At the

[8] Triples may be calculated only on the EPO database as they require cited patents to be assessed whether they constitute a critical innovation.

same time, the USPTO sample we used was small, so our findings should be expanded with other expert-based analyses, ideally using a larger sample of patent groups and drawing on the EPO dataset, which would not require mapping to USPTO data.

Figure 4. Green dots indicate share of patents belonging to thickets vs share of patents belonging to triples in different technologies (left panel) or vs Clarkson's (2005) adjusted densities in various patent groups (right panel). Discrete technology areas are coloured 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

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 an idea that patents in a potential patent thicket should cite one another more densely than patents not belonging to the thicket. This set-up results in higher density values for substitutes, than for complementary patents even though both types could potentially result in the sort of hold-up that has been associated with thicket "problems". For example, the author 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 rather substitute patents, describing similar approach to the same technology[9].

In order to compare Clarkson's density with our expert-based method we have calculated Clarkson's measure[10] on the groups from USPTO classifications. The right pane on Figure 4 plots the 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=0.0367. In order to account for 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 the Clarkson's measure on respective patent classes and on the OST-INPI/FhG-ISI technology areas. In none of the cases can the share of patents belonging to the expert-identified thickets be related to the Clarkson density in a statistically significant manner.

Our findings suggest 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. This suggests that semantic similarity may be a useful tool in identifying patent thickets, albeit not necessarily those identified by citation measures. This may simply make Egan and Teece's point more formally: there may be a number of concepts corresponding to thickets and, as not all reflect the same concerns, not all have the same features.

5. A semantic network model for thickets recognition

A way to incorporate the divergence of these methods as well as their value is to propose a logit model based on the network of pairwise semantic distances and drawing from information

[9] Régibeau et al. (2012) indicate that Clarkson density is a noisy measure, its value depending strongly on how broadly the patent network, i.e. technology, is defined.

[10] We use the weighted average patent network density described by formula (6) in Clarkson (2005).

contained in other methods, specifically triples and Clarkson's weighted average patent network density. The model is aimed at explaining the newly added patent's probability of membership in an existing thicket within a given patent group. The result returned by the model could, for example, serve as a prompt indicating whether a patent or patent application overlaps existing patents heavily.

The model has been estimated on an "in-sample" dataset and tested on "out-of-sample" data. The quality of the model has been assessed on the out-of-sample portion of the data, for which we have forecasted the theoretical probability of being in a thicket given specific characteristics of a patent application in question. Year 2001 was chosen a break point between in-sample and out-of-sample portions of the data[11]. More precisely, the estimation sample consists of patent applications from years 1976-2000 (5,482 patents of which 1,088 are in thickets), while the testing sample contains patents from the period 2001-2010 (3.089 patents of which 467 belong to thickets).

The logit regression model was estimated using the generalized linear model and its implementation in R. The results of the estimation are presented in Table 4, where the dependent variable is the membership of a patent in a thicket[12]. The independent variables of interest were: minimal semantic distance – distance to the most similar earlier patent; Clarkson's ratio for a group (at the moment of filing); 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 Graevenitz et al's (2011) method translated to our dataset. We also include 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 the 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

[11] The rule for selecting the break year was that it must be a first year which for which there was at least 60% of patents in sample and 70% of patents in the thickets in the estimation dataset.

[12] 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, so the first patent does not belong to any thicket at the moment of its filing.

patents for a given group (at the moment of filling, based on filing dates of eventually successful patents); number of prior (eventually successful)[13] prior fillings by assignee[14]; 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.

Compared to model (1), the first four models listed in Table 4, models (2)-(5), differ from the first model by one variable (or one group of dummies) only. Respectively, these additions are: dummies for class of patents (2), thicket ratio (3), Clarkson's ratio (4) or triples ratio (5). Model (6) is a model with dummies for patent groups instead of various 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. Model (10) is a simplified version of (9) without information on number of prior applications and grants in a given patent group.

Table 4. 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	0.646***	0.706***	0.482***	0.652***	0.647***	0.562*	0.575***	0.709***	0.617***	0.691***

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

[14] 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. R package by van der Loo, M (2014) was utilised.

groups	(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 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 **p ***p<0.01

Source: Own calculations

Analysis of the results suggests several conclusions.

First, the models suggest that applications belonging to groups that, at a given moment in time, tend to include patents belonging to thickets, are substantially more likely to be in a thicket as well. This suggests both that thickets are, to some extent, a characteristic of a patent group and that a larger pool of thicket patents breeds a higher likelihood that further work will impinge on those existing thickets.

Second, the closer (semantically) is the patent application to an earlier most similar patent, the greater the probability that it will belong 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 part of a thicket with a higher probability. Complexity is unlikely to underlie this, as the relation holds when we control for complexity[15]. 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, even if one does not control for semantic distance. Indeed, without group descriptors, the claims variable picks up the group effects that are captured by other group descriptors in the other equations.

Fifth, Clarkson's density ratio, and the triples ratio do not carry significance in most cases. Furthermore, the negative coefficient on triples indicates that the presence of existing triples in the group actually are 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

[15] 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 points of conflict in a patent. While this may be related to fragmentation, as in their work, it could also possibly be related to large applicability, which could be indicated by membership in a large number of groups.

the patent examiner's as well as the assignee's task and so translate into higher quality patents overall.

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) when 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, mentioned above. Furthermore, the magnitude of the coefficient suggests a relatively strong effect.

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. Hence, for a given quality of review more patents that overlap would tend to occur in such a trajectory.

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 (sufficiently for the number of patents granted to be greater than the patents filed for before the examined patent), 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.

To facilitate interpretation, the odds ratio for thicket membership has been estimated for model (9), and is presented in Table 5. For continuous variables the incremental step of one standard deviation has been chosen to evaluate the change in odds due to a change in a given variable, for example we can see from the table below 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%.

Table 5. Estimates of odds ratios for model (9).

Variable	Odds ratio	CI low (2.5 %)	CI high (97.5 %)	Incremental step
Semantic distance	0.579	0.528	0.634	0.159
Number of backward citations	1.198	1.114	1.291	2.683
Number of claims	1.016	0.942	1.094	13.979
Number of groups	1.156	1.082	1.233	0.234
Thicket ratio for a group (in %)	2.183	1.968	2.429	0.182
Clarkson ratio for a group	1.020	0.935	1.103	0.033
Complex group (average)	9.463	0.042	2,269,290.0	1
Triples incidence in group	0.963	0.862	1.074	0.025
HHI	0.906	0.813	1.003	0.110
Prior applications of assignee	1.142	1.060	1.228	9.477
Prior applications in the group	1.494	0.981	2.328	190.862
Prior patents in the group	0.470	0.299	0.717	192.733
Class 23	1.152	0.762	1.751	1
Class 324	0.013	0	3.245	1
Class 327	0.027	0	6.148	1
Class 348	0.201	0	49.457	1
Class 424	0.890	0.630	1.265	1
Class 433	0.175	0	38.821	1
Class 436	0.120	0	26.743	1
Class 604	2.845	0.690	12.489	1
Class 977	0.119	0	26.057	1

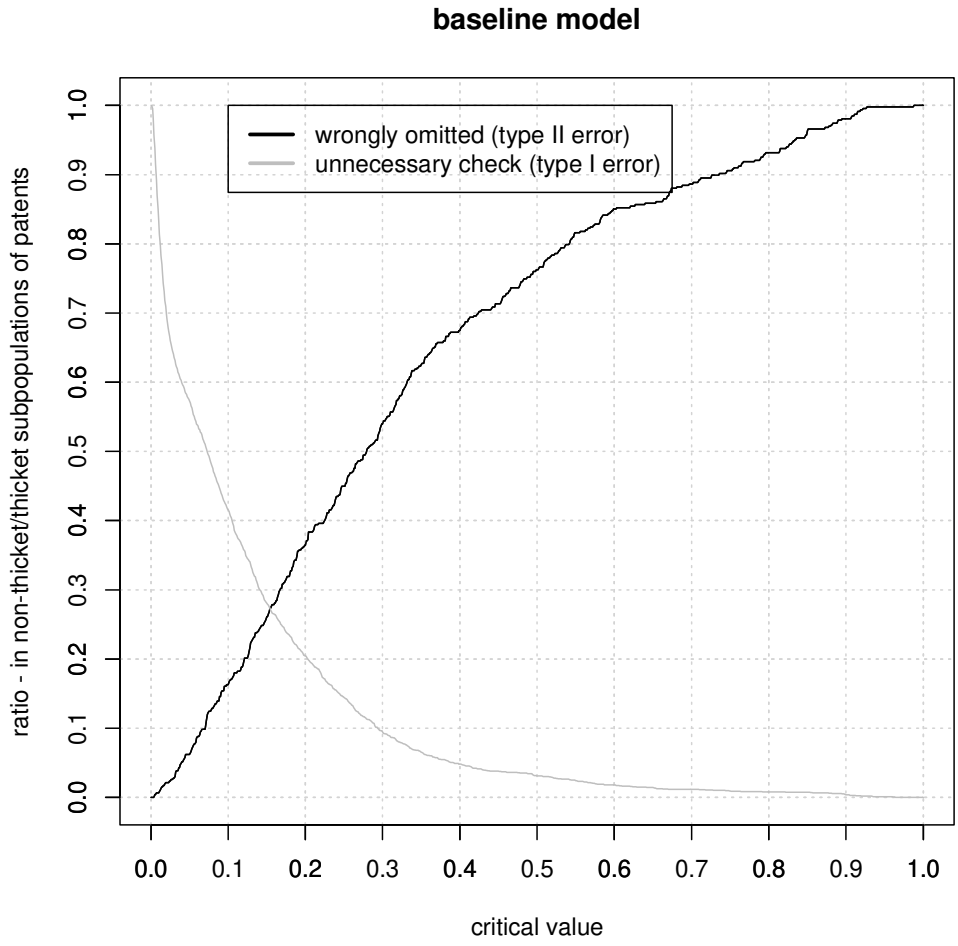
Note: Incremental step is equal to standard deviation of variable or 1 for dummies and 'complex group' (which is a dummy variable averaged over all the groups for a given patent – in almost all cases it is either 0 or 1).
Source: Own calculations

To assess the performance of the model 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 thickets are actually not in a thicket; (2) a "false negative" ratio – which shows how many patents in thickets would be wrongly omitted, i.e. how many patents flagged by

the model as not in a thicket, 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 6, below. Another way to see the exercise depicted in this figure is that we ask: for a given tolerance level, which model produces the fewest false positives and negatives? For example, if one were to apply this model to checking whether a patent should be reviewed for quality (ie, for perhaps contributing to a thicket) Figure 6 suggests that when the critical value is 0.1, around 15% of patents that are indeed members of thickets are wrongly classified as those that are not and 40% of patents that indeed do not belong to a thicket would be classified as belonging to one (which means that 60% of these could be subject to a quick-check only). Setting a critical value at 0.2 would change the values to be over 35% and around 20% respectively. The chart for the baseline model is presented in Figure 5, while the remaining charts can be found in the appendix.

In this part, model (9) has been chosen as a baseline model due to the fact that year dummies are not useful for forecasting. For the sake of comparison, results for models (2)-(8) without yearly dummies and model (10) are presented in table 6 and in appendix D as well.

Figure 5. False positive/negative ratios as functions of the critical value for the baseline model (9).



Source: Own calculations
Note: The lines are not smooth as they are derived from the tests on out-of-sample datasets.

Table 6. 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								
	baseline model (9)	model w/o class dummies (2')	model w/o thicket ratio (3')	model w/o Clarkson ratio (4')	model w/o triplets 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 4 without the year dummies.

Table 6 summarises the performance of all the models. 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). 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. Omission of the thicket ratio worsens results substantially (3' and 8' vs 9), while omission of historic group thicket information worsens performance less (or, one could argue, enhances it; 9 vs 10). Similarly, impact of omission of the class dummies is substantial but inconsistent (2' vs 9). Omission of the Clarkson ration and Triplets reduces performance inconsistently (4' and 5' vs 9). What is also interesting using (time-static) group dummies usually does not work better than using (dynamic) group-specific variables (6' vs 9).

6. Conclusions

If we assume that the legal definition of thickets as "an overlapping set of patent rights" is necessarily reflected in the body of the patent's text, then semantic distance should be able to identify this feature. Accordingly, we calculate a semantic distance for pairs of patents based on the content similarity between the patent document text, creating thus a semantic network. We then use these distance scores to determine whether patents identified by a group of experts as belonging to thickets using a standard definition are semantically more similar to one another than other patents by comparing the mean semantic distance for patents in and outside thickets. Our key conclusion is that patents belonging to the same thicket as identified by our experts are closer semantically than other pairs of patents. This result is clearly dependent upon the definition provided to the experts and is a costly method of detection; however, semantic distance appears to be a good way of proxying the view that would be obtained by a careful reading of the patent document. Given the availability of computing power and natural language

processing tools, the performance of the proxy compared to expert view creates an interesting possible alternative to expert opinion.

Our semantic measure also allows us to check on whether existing citation-based methods of identification also perform well to proxy expert view, as represented by semantic distance. Accordingly, we compare this content based method to two citations-based methods for identifying patent thicket membership as proxies for expert opinion, finding that the semantic method correlates more closely than the citations methods with the expert views. This does not indicate that either method is "wrong": they may measure different things. The expert views are based on interpretation of a standard thicket definition and so may measure effects that are not the focus of the citation-based methods. In particular the definition would tend to include patents obtained for defensive reasons much more than the citation based methods.

We then combine the various measures into a single model of thicket identification, and evaluate its performance 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). We identify a model that performs relatively well and that also combines several measures of thickets, including citation based measures and controls for fragmentation, all of which have figured prominently in the thicket literature and all of which find some support within the model. Crowding and technology group complexity also play a role in the likelihood that a patent will belong to a thicket and enter into the specification.

We view our work as contributing to patent thicket measurement on several levels. First, external validity of citations based measures has not been investigated thoroughly, although it has been mentioned as an issue. We have taken such a further step in this investigation.

We have also illustrated the value of semantic analysis in capturing expert views on this topic. Shorter semantic distance does appear to identify thickets well in the sense of capturing this view. We also 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 in 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 are thus not influenced by a single expert whose expertise was misjudged.

We find also that there is no significant overlap between individual patents indicated by experts as belonging to a thicket and patents belonging to triples (methodology from Graevenitz et al., 2011). Similarly, patent network density measure introduced by Clarkson (2005), shows no significant relation to the share of individual patents in these thickets. At the same time, we show that these alternative measures can be combined to create a quantitative model that identifies patent thickets. Such a model can provide a 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 examined by Gallini (2017) in the light of the theory of innovation incentives in the presence of cumulative innovation. We examine a logit model, which assesses the probability that a newly added patent would form a thicket. The model shows that the semantic distance combined with other information can be helpful in assessing a newly filed application. Important conclusions also include the fact that a key indicator of a patent belonging to a thicket is the previous density of thickets within a patent group. Fragmentation also is an indicator of thicket formation, which has been emphasised by the literature.

Our method exploits expert opinion to identify thickets. This is not as precise as methods that have use data sets based on court cases. We have, then, sacrificed some *ex post* information confirming thicket membership in exchange for dataset size in some sense. While we investigate individual expert error as a source of our results and find that it is not, we should emphasise that what we aim for is not a tool enabling infallible recognition of patent thickets, but a method for

delineating a set of patents which with high probability are 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.

7. References

1. Barnett J. M. (2014). From Patent Thickets to Patent Networks: The Legal Infrastructure of the Digital Economy, 55 Jurimetrics J. 1–53.
2. Bergeaud A., Poitiron Y., Raimbault J., (2017). Classifying patents based on their semantic content, PLOS ONE, 26th April 2017.
3. Bessen J. (2003). Patent Thickets: Strategic Patenting of Complex Technologies, Available SSRN 327760, pp. 1–30.
4. Bradford, R. B. (2008). An Empirical Study of Required Dimensionality for Large-scale Latent Semantic Indexing Applications. In Proceedings of the 17th ACM Conference on Information and Knowledge Management (pp. 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). Working Paper 7552, NBER.
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. Egan, E., Teece, D. (2015). Untangling the Patent Thicket Literature. Tusher Center for the Management of Intellectual Capital Working Paper 7. March.
9. Fleming, L. (2001). Recombinant Uncertainty in Technological Search. Management Science, 47(1), 117–132. doi:10.1287/mnsc.47.1.117.10671
10. Galasso, A., Schankerman, M. (2010) Patent Thickets, Courts, and the Market for Innovation. Rand Journal of Economics. 41(3): 472-503.
11. Gallini, N. (2017). Do Patents Work? Thickets, Trolls, and Antibiotic Resistance. Canadian Journal of Economics, 50(4), 893-926.
12. Gerken, JM., Moehrle MG. (2012). A new instrument for technology monitoring: novelty in patents measured by semantic patent analysis. Scientometrics. 91(3):645–670.

13. Graevenitz, von G., Wagner S., Harhoff, D. (2011). How to measure patent thickets-A novel approach, Econ. Lett., vol. 111, no. 1, pp. 6–9.
14. 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.
15. Hall B., Helmers C., Graevenitz von G., Rosazza-Bondibene C., (2013): A Study of Patent Thickets, Intellect. Prop. Off., no. 401, pp. 7–76.
16. Hall, B. H., Helmers, C., & Graevenitz, von G. (2015). Technology entry in the presence of patent thickets (No. w21455). National Bureau of Economic Research.
17. 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 101–128.
18. Harhoff D., Graevenitz von G., Wagner S. (2016): Conflict Resolution, Public Goods and Patent Thickets, Management Science, INFORMS, vol. 62(3), pages 704-721, March.
19. 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
20. Holman, C. (2006). Clearing a path through the patent thicket. Cell, 125(4), 629–33
21. Intellectual Property Office, Patent Thickets - An overview, Newport, 2011.
22. 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
23. 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).
24. 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

25. 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.
26. Lang, K. (1995). Newsweeder: Learning to filter netnews. In Proceedings of the 12th international conference on machine learning (pp. 331–339).
27. Noel, M., Schankerman, M. (2006) Strategic Patenting and Software Innovation. Journal of Industrial Economics. 61(3): 481-520.
28. OECD, 1994. Using patent data as science and technological indicators. Patent Manual.
29. Preschitschek N, Niemann H, Leker J, Moehrle MG., (2013). Anticipating industry convergence: Semantic analyses vs IPC co-classification analyses of patents. Foresight. 11;15(6):446–464
30. 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/.
31. 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.
32. 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 (pp. 45–50). Valletta, Malta: ELRA.
33. Rijsbergen, C. J. V. (1979). Information Retrieval (2nd ed.). Newton, MA, USA: Butterworth-Heinemann.
34. Salton, G., & McGill, M. J. (1986). Introduction to Modern Information Retrieval. New York, NY, USA: McGraw-Hill, Inc.
35. Schankerman, M. and Schuett, F. (2018) "Screening for Patent Quality", CEPR Discussion Paper 11688 (updated 2018).
36. Shapiro C. (2001): Navigating the Patent Thicket: Cross Licenses, Patent Pools, and Standard Setting, Natl. Bur. Econ. Res.

37. Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research, 37(1), 141–188.
38. 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>.
39. Welch, B. L. (1951). On the Comparison of Several Mean Values: An Alternative Approach. Biometrika. 38: 330–336.
40. World Intellectual Property Organization (2013), WIPO IP Facts and Figures (WIPO Publication No. 943E/13), p. 44.
41. World Intellectual Property Organization (2014), WIPO IP Facts and Figures (WIPO Publication No. 943E/14), p. 6.
42. World Intellectual Property Organization (2015), WIPO IP Facts and Figures (WIPO Publication No. 943E/15), p. 8.
43. 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
44. 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

8. Appendix – Latent Semantic Analysis

We use latent semantic analysis (LSA) (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990; Landauer, Foltz, & Laham, 1998) to measure the semantic distance between the patents in our dataset. One of the primary advantages of the LSA technique is that it enables the detection of "latent" similarities. That is, it does not require documents to use precisely the same terminology to detect similarities between them. Provided they co-occur with a similar terms, different words related to the same topic will both lead a document to have a higher weight on the same topic. For instance, if one patent uses the term "car" while another opts for "automobile" LSA can detect that these are related terms, and thus treat them similarly.

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, that is then transformed using SVD. We begin the creation of our matrix by generating a term-document matrix with a row for each granted patent (our input documents), a column for each unique term (i.e. word) used across the corpus. The matrix values are 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 remove both 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 the term's frequency in the given document by the logarithmically-scaled inverse document frequency—that is the number of documents in the corpus divided by the number of documents the term appears in. 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. 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 and performs a rank-reduced singular value decomposition on it, creating in the process a term-concept matrix, a singular values matrix, and a k dimensional document-concept matrix. 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 generated by the LSA process.

Once the term-document 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 use the commonly-used cosine distance (Landauer, Laham, & Foltz, 1998) to calculate pairwise distance for the patents in our study. 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 closer 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.

9. Appendix B

Table 7. Names of USPC groups used in the analysis.

Class / group	Name	Classification
327/129	Converting input frequency to output current or voltage. Generating sinusoidal output	Complex
23/295R	Chemistry: physical processes. Crystallization	Discrete
23/302A	Chemistry: physical processes. Crystallization. Alkali method and ammonium compounds. Ammonium compounds	Discrete
23/303	Chemistry: physical processes. Crystallization. Alkali method and ammonium compounds. Common salt	Discrete
23/305A	Chemistry: physical processes. Crystallization. Heavy metal or aluminum compounds. Aluminum compounds	Discrete
23/305R	Chemistry: physical processes. Crystallization. Heavy metal or aluminum compounds	Discrete
23/306	Chemistry: physical processes. Concentration of liquids in liquids	Discrete
23/307	Chemistry: physical processes. Concentration of liquids in liquids. With direct heating	Discrete
23/313R	Chemistry: physical processes. Agglomerating	Discrete
8/115.51	Bleaching and dyeing. Chemical modification of textiles or fibers or products thereof	Discrete
8/400	Bleaching and dyeing. Measuring, testing or inspecting dye process	Discrete
8/401	Bleaching and dyeing. Using enzymes, dye process, composition, or product of dyeing	Discrete
8/438	Bleaching and dyeing. Process of extracting or purifying of natural dye	Discrete
8/493	Bleaching and dyeing. Overal diemnsional modification or stabilization. Modification of molecular structure of substrate by chemical means	Discrete
324/509	Electricity: measuring and testing. Fault detecting in electric circuits and of electric components of ground fault indication	Complex
324/512	Electricity: measuring and testing. Fault detecting in electric circuits and of electric components for fault location	Complex
324/525	Electricity: measuring and testing. Fault detecting in electric circuits and of electric components for fault location by resistance or impedance measuring	Complex
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
205/564	Electrolytic coating (process, composition and method of preparing composition). Preparing single metal. Gallium, germanium, indium, vanadium or molybdenum produced.	Discrete
977/778	Nanostructure. Within specified host or matrix material (e.g., nanocomposite films, etc.)	Complex
977/810	Nanostructure. Of specified metal or metal alloy composition	Complex
977/881	Manufacture, treatment or detection of nanostructure. With arrangement, process, or apparatus for testing. With arrangement, process, or apparatus for testing	Complex
977/903	Specified use of nanostructure. For conversion, containment, or destruction of hazardous material	Complex

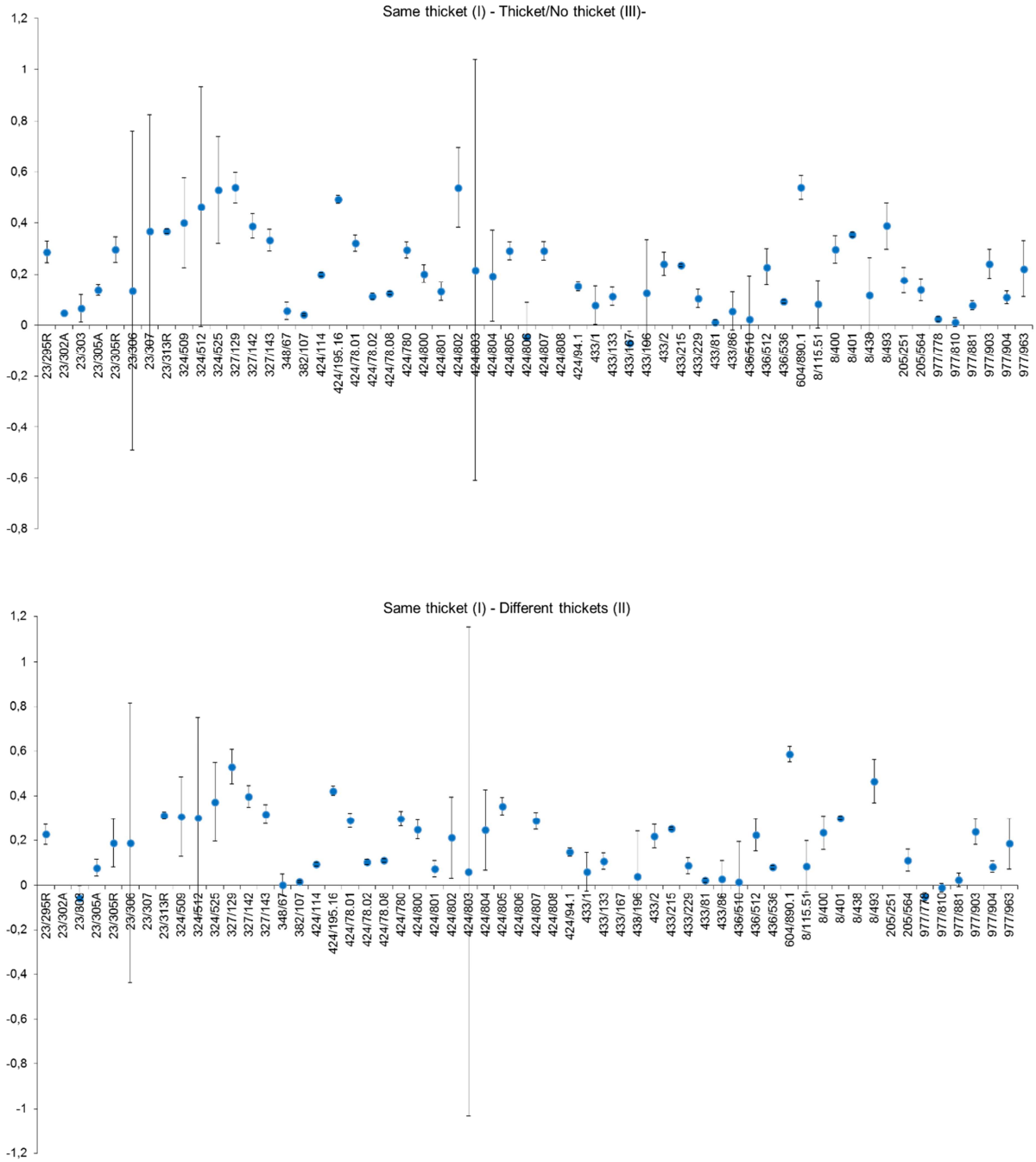
977/904	Specified use of nanostructure. For medical, immunological, body treatment, or diagnosis	Complex
977/963	Specified use of nanostructure. For medical, immunological, body treatment, or diagnosis. Specially adapted for travel through blood circulatory system	Complex
433/1	Dentistry. Veterinary dentistry	Complex
433/133	Dentistry. Apparatus. Having motor or means to transmit motion from motor to tool. Hand-held tool or handpiece. Contra angled handpiece	Complex
433/167	Dentistry. Prosthodontics	Complex
433/196	Dentistry. Prosthodontics. Orienting or positioning teeth	Complex
433/2	Dentistry. Orthodontics	Complex
433/215	Dentistry. Method or material for testing, treating, restoring, or removing natural teeth	Complex
433/229	Dentistry. Miscellaneous	Complex
433/81	Dentistry. Apparatus. Having intra-oral dispensing means. Endodontic	Complex
433/86	Dentistry. Apparatus. Having intra-oral dispensing means. Endodontic. Ultrasonic tool	Complex
424/114	Drug, bio-affecting and body treating compositions. Plural fermentates of different origin	Discrete
424/195.16	Drug, bio-affecting and body treating compositions. Extract or material containing or obtained from a unicellular fungus as active ingredient	Discrete
424/78.01	Drug, bio-affecting and body treating compositions. Digestive system regulator containing solid synthetic organic polymer	Discrete
424/78.02	Drug, bio-affecting and body treating compositions. Topical body preparation containing solid synthetic organic polymer	Discrete
424/78.08	Drug, bio-affecting and body treating compositions. Solid synthetic organic polymer	Discrete
424/780	Drug, bio-affecting and body treating compositions. Extract or material containing or obtained from a micro-organism as active ingredient	Discrete
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
424/801	Drug, bio-affecting and body treating compositions. Involving antibody or fragment thereof produced by recombinant dna technology	Discrete
424/802	Drug, bio-affecting and body treating compositions. Antibody or antigen-binding fragment thereof that binds gram-positive bacteria	Discrete
424/803	Drug, bio-affecting and body treating compositions. Antibody or antigen-binding fragment thereof that binds gram-negative bacteria	Discrete
424/804	Drug, bio-affecting and body treating compositions. Involving IGG3, IGG4, IGA, or IGY	Discrete
424/805	Drug, bio-affecting and body treating compositions. Involving IGE or IGD	Discrete
424/806	Drug, bio-affecting and body treating compositions. Involving IGM	Discrete
424/807	Drug, bio-affecting and body treating compositions. Involving IGM. Monoclonal	Discrete
424/808	Drug, bio-affecting and body treating compositions. Involving IGM. Human	Discrete
424/94.1	Drug, bio-affecting and body treating compositions. Enzyme or coenzyme containing	Discrete
436/510	Chemistry: analytical and immunological testing. Immunochemical pregnancy determination	Complex
436/512	Chemistry: analytical and immunological testing. Involving antibody fragments	Complex
436/536	Chemistry: analytical and immunological testing. Involving immune complex formed in liquid phase	Complex
604/890.1	Surgery. Controlled release therapeutic device or system	Discrete
348/67	Television. Improving the 3D impression of a displayed stereoscopic image	Complex
382/107	Image analysis. Applications. Motion or velocity measuring	Complex

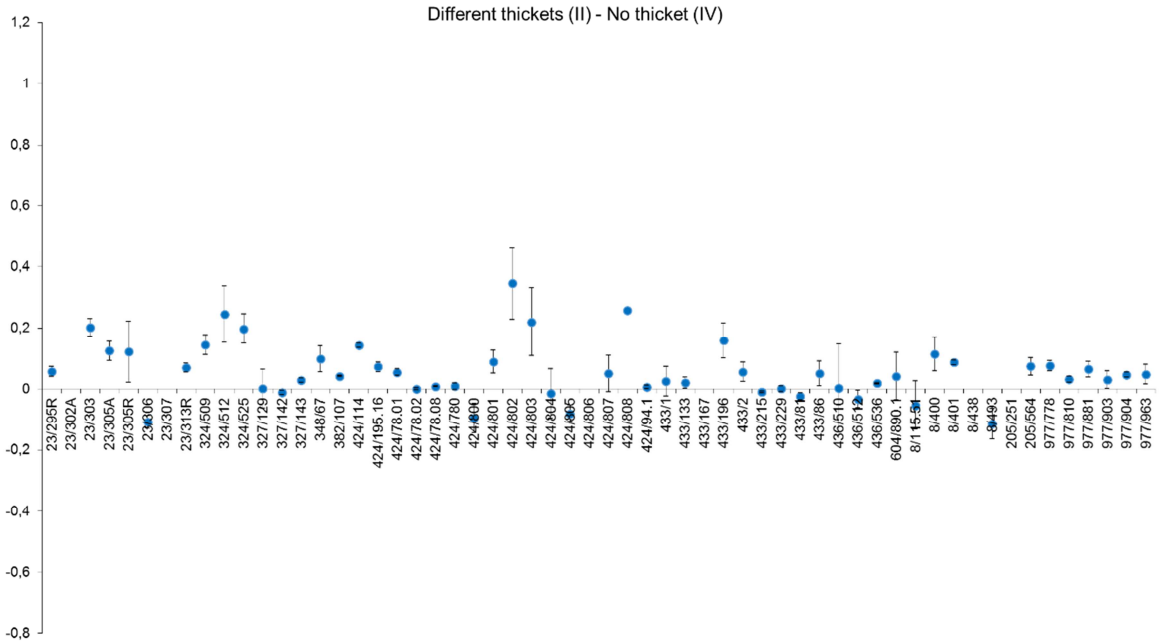
327/142	Converting input frequency to output current or voltage. Synchronizing. Reset (e.g., initializing, starting, stopping, etc.)	Complex
327/143	Converting input frequency to output current or voltage. Synchronizing. Reset (e.g., initializing, starting, stopping, etc.). Responsive to power supply	Complex

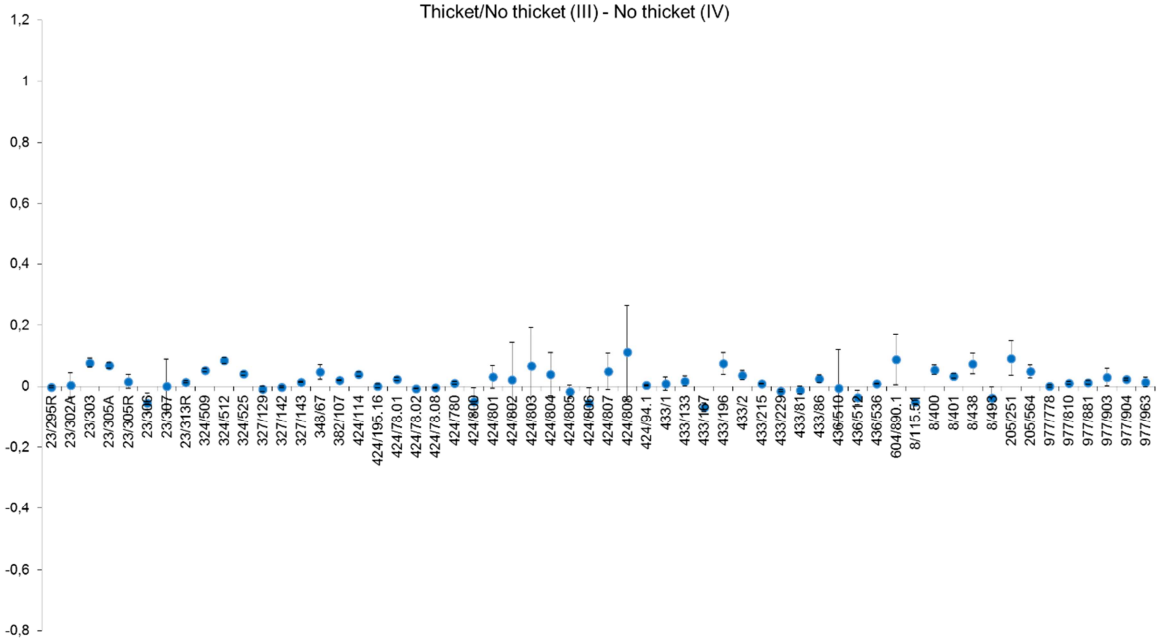
Table 8. Survey questions for the field experts

Question	Range of answers
Does given patent belong to a patent thicket?	Yes/No
To which patent thicket within a patent group it belongs to?	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

Figure 6. 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%).



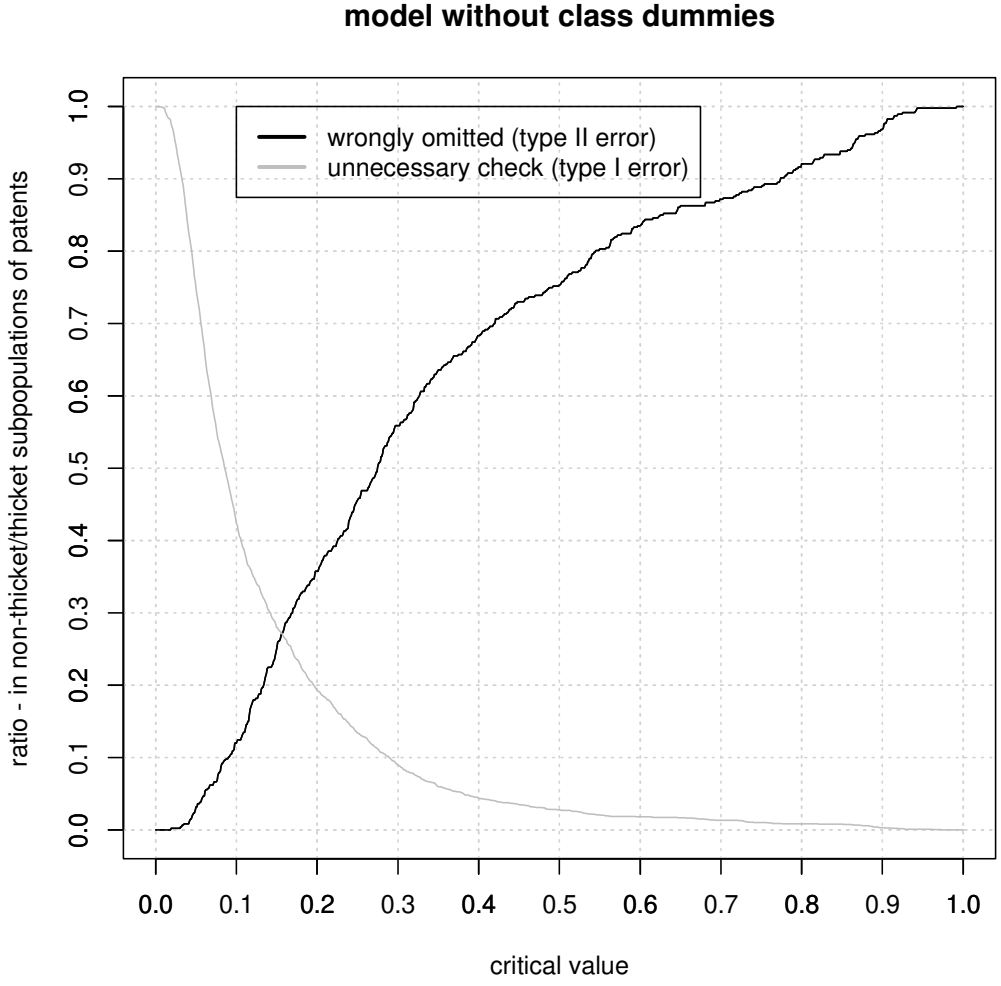

Different thickets (II) - No thicket (IV)
1,2
1
0,8
0,6
0,4
0,2
0
-0,2
-0,4
-0,6
-0,8
23/295R
23/302A
23/303
23/305A
23/305R
23/306
23/307
23/313R
324/509
324/512
324/525
327/129
327/142
327/143
348/67
382/107
424/114
424/195.16
424/78.01
424/78.02
424/78.08
424/780
424/800
424/801
424/802
424/803
424/804
424/805
424/806
424/807
424/808
424/94.1
433/1
433/133
433/167
433/196
433/2
433/215
433/229
433/81
433/86
436/510
436/512
436/536
604/890.1
8/115.54
8/400
8/401
8/438
8/493
205/251
205/564
977/778
977/810
977/881
977/903
977/904
977/963


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
23/295R
23/302A
23/303
23/305A
23/305R
23/306
23/307
23/313R
324/509
324/512
324/525
327/129
327/142
327/143
348/67
382/107
424/114
424/195.16
424/78.01
424/78.02
424/78.08
424/780
424/800
424/801
424/802
424/803
424/804
424/805
424/806
424/807
424/808
424/94.1
433/1
433/133
433/167
433/196
433/2
433/215
433/229
433/81
433/86
436/510
436/512
436/536
604/890.1
8/115.54
8/400
8/401
8/438
8/493
205/251
205/564
977/778
977/810
977/881
977/903
977/904
977/963

10. Appendix D

Omitting class dummies seems to make the predictive power a bit better when comparing to the baseline model:

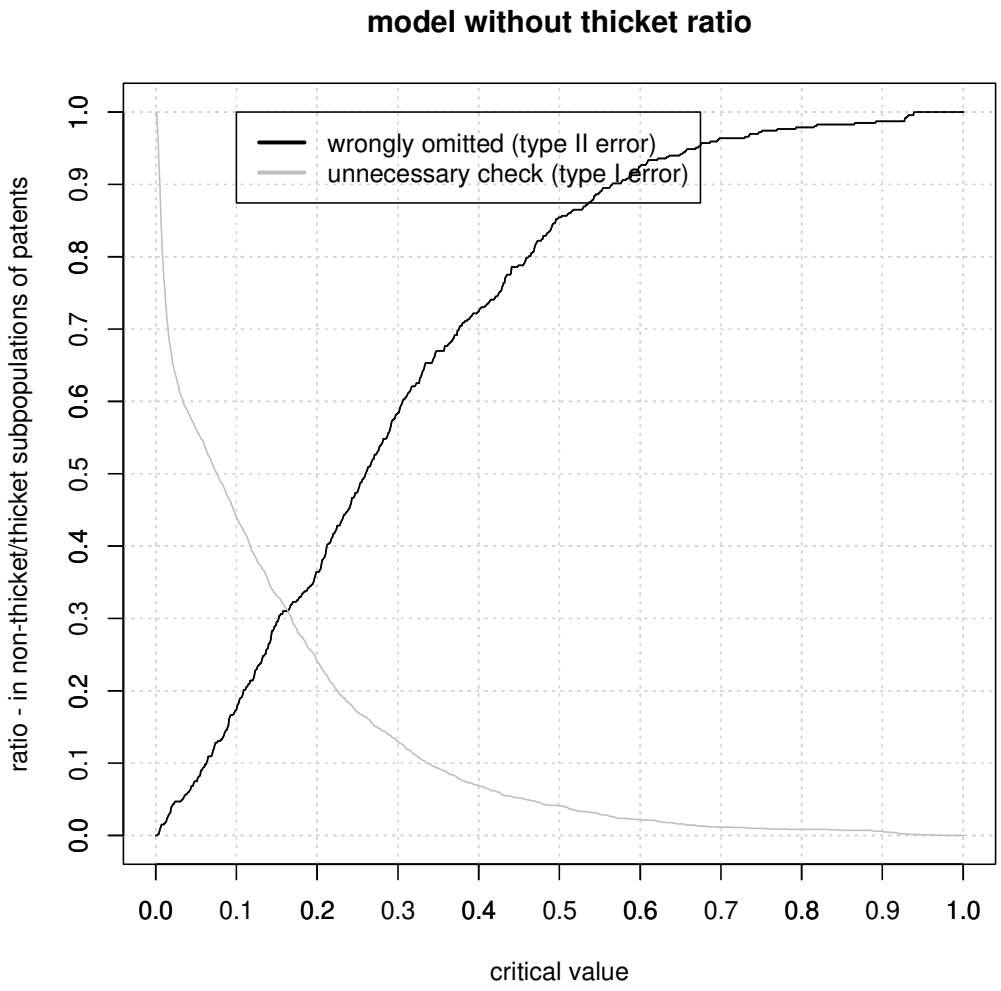
Figure 7. False positive/negative ratios as functions of the critical value for the baseline model without class dummies (2').



Source: Own calculations

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

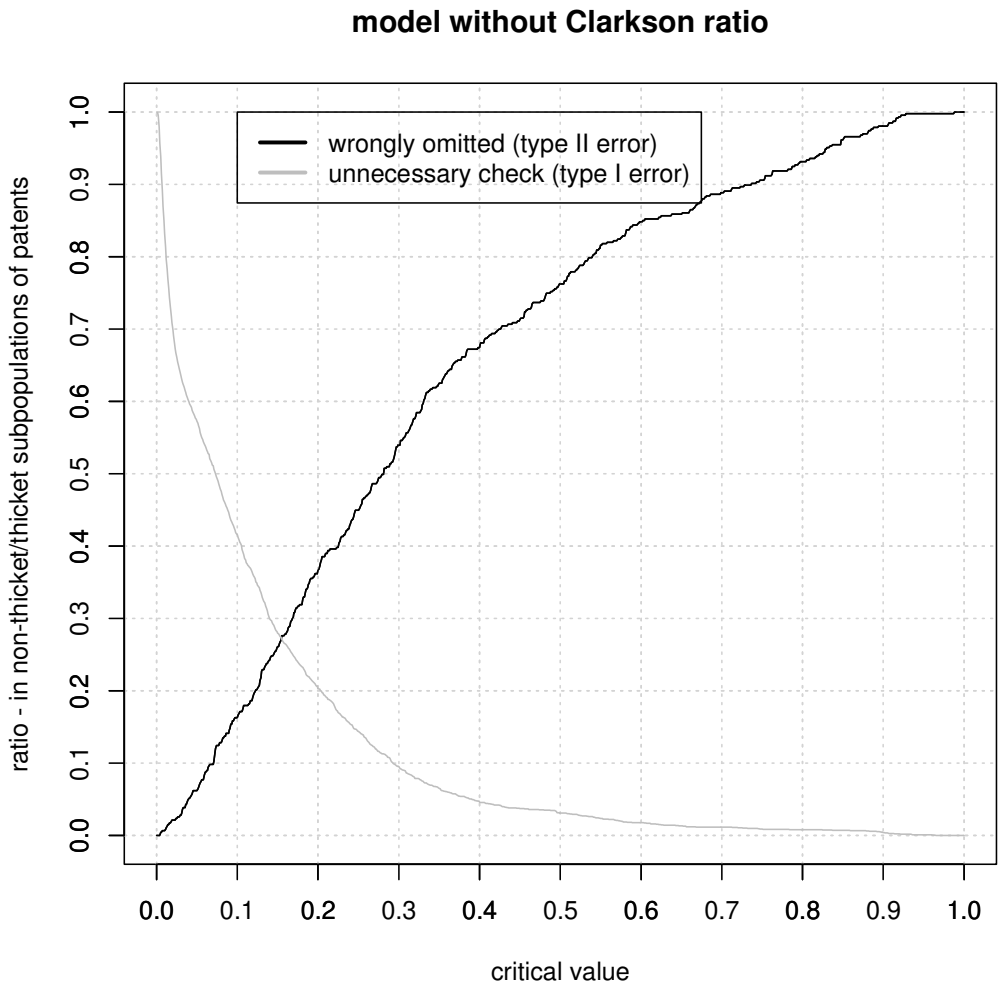
Figure 8. False positive/negative ratios as functions of the critical value for the model without thicket ratio (3').



Source: Own calculations

Omitting Clarkson ratio does not seem to matter much as well:

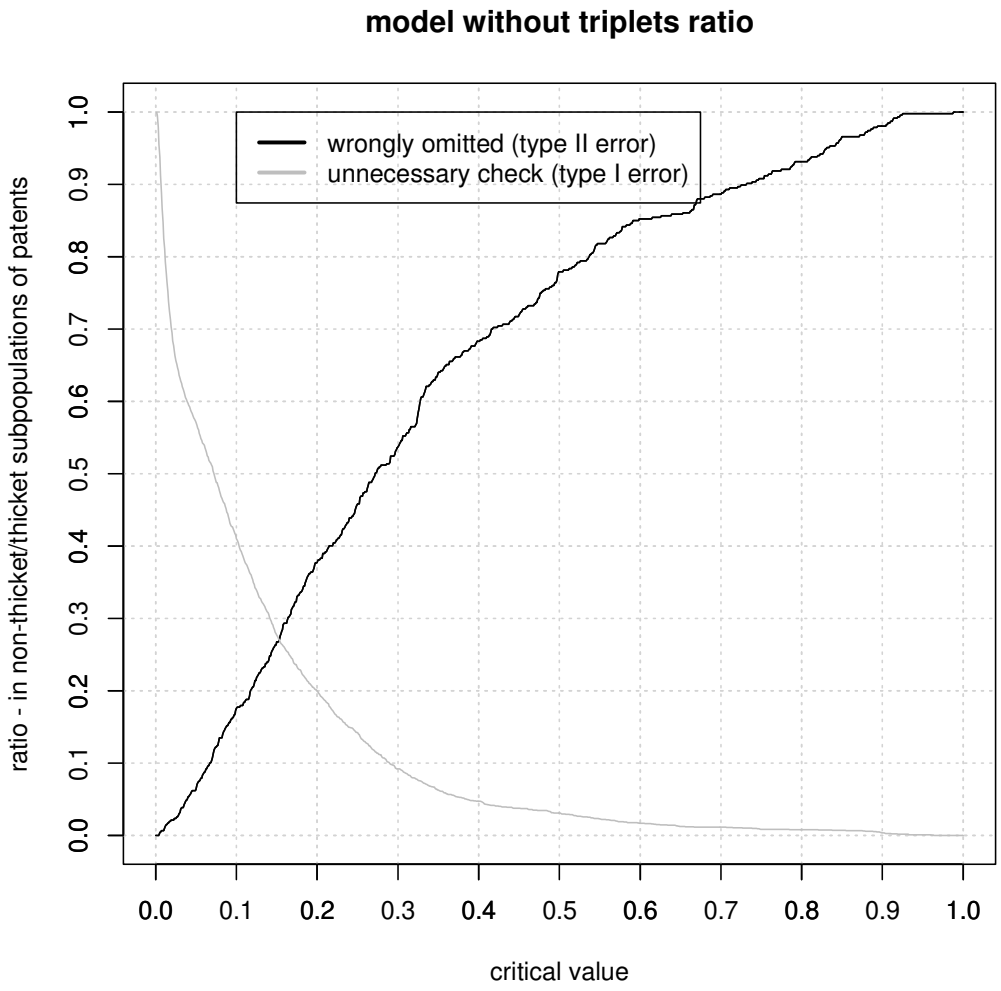
Figure 9. False positive/negative ratios as functions of the critical value for the model without Clarkson ratio (4').



Source: Own calculations

As well as the ratio describing incidence of triplets in the group:

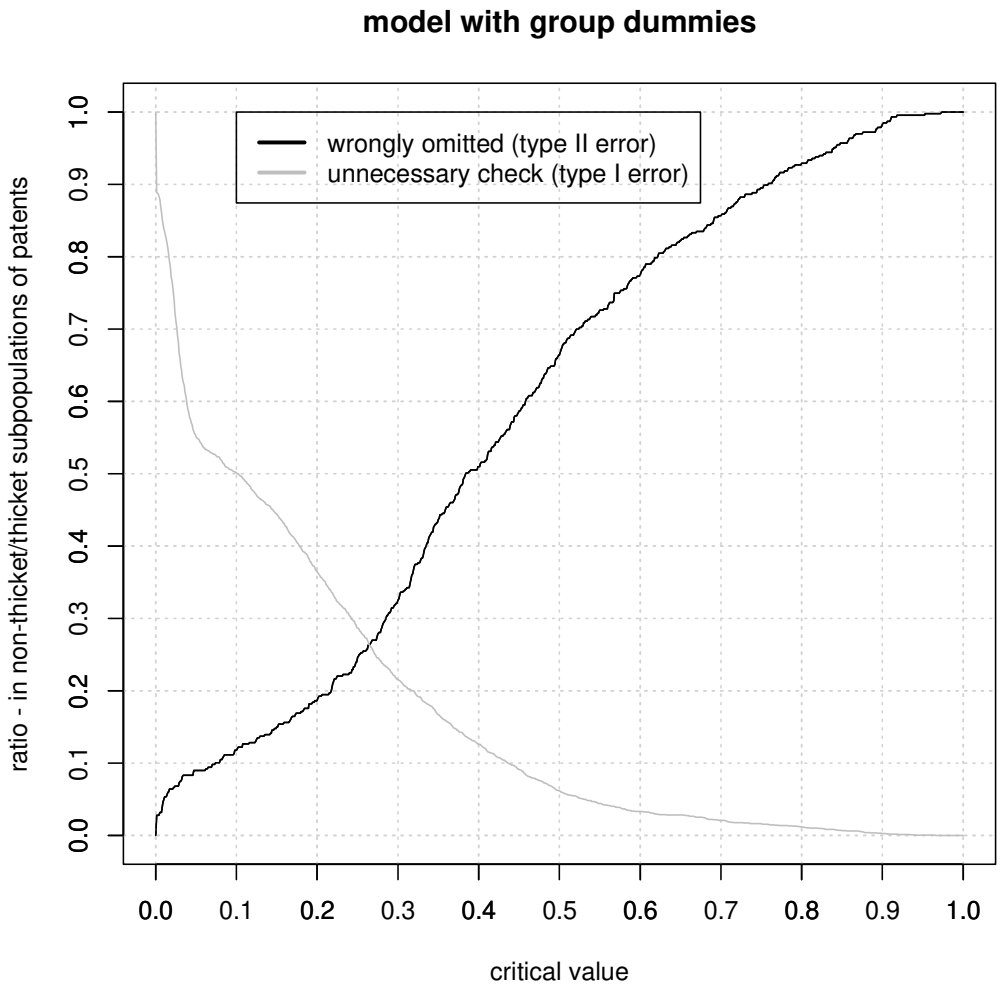
Figure 10. False positive/negative ratios as functions of the critical value for the model without triplets ratio (5').



Source: Own calculations

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

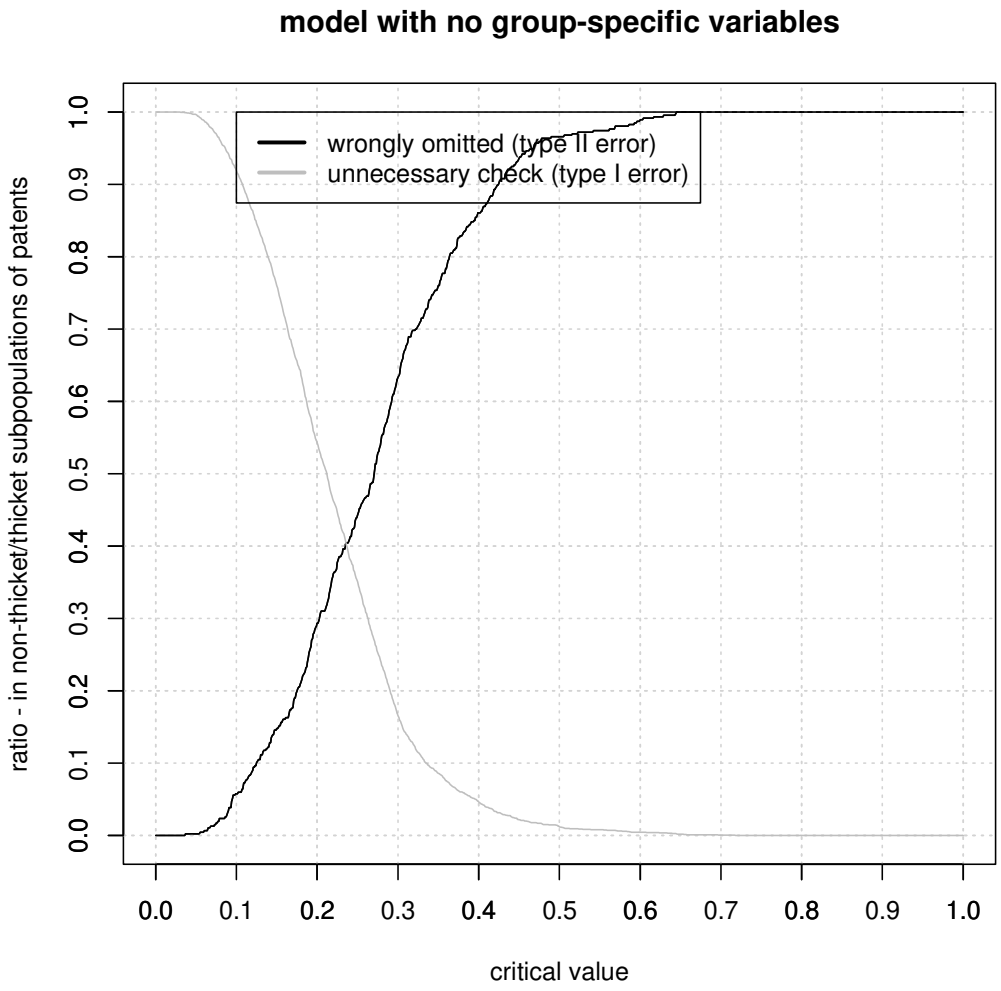
Figure 11. False positive/negative ratios as functions of the critical value for the model with group dummies (6').



Source: Own calculations

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

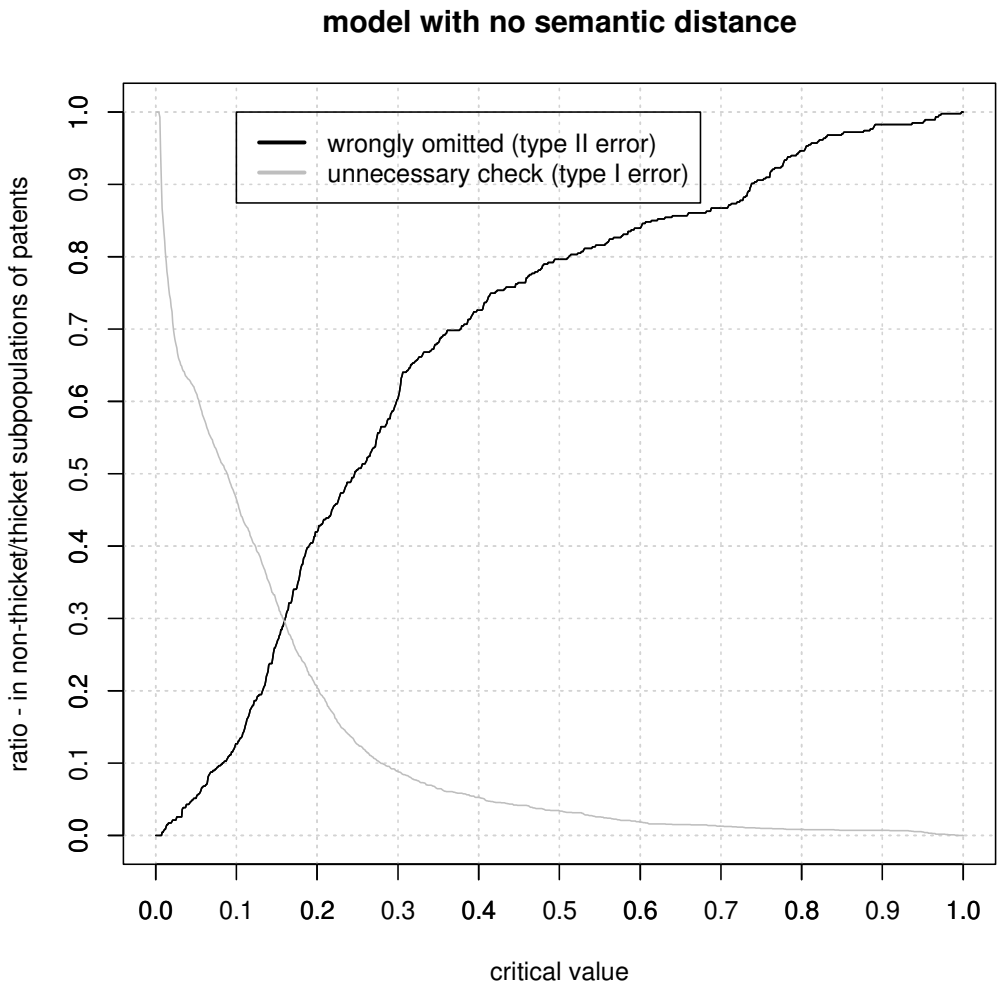
Figure 12. False positive/negative ratios as functions of the critical value for the model without group-specific variables (7').



Source: Own calculations

Discarding semantic distance seems to worsen the model marginally:

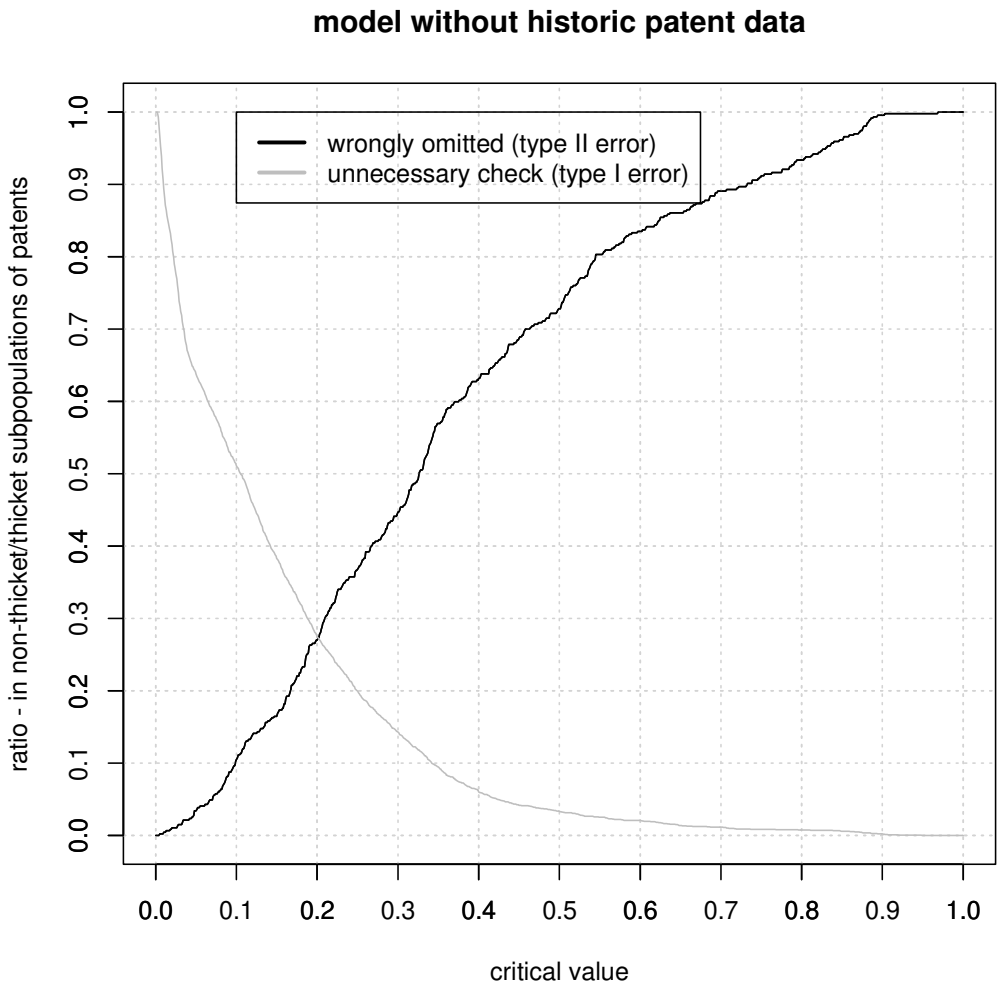
Figure 13. False positive/negative ratios as functions of the critical value for the model without semantic distance (8').



Source: Own calculations

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

Figure 14. 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

11. Appendix C

Table 9. Estimates for different logit models (1-10) of the probability of the membership in an existing thicket for a new patent application - full version.

	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***	-0.004***	-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	-0.249

424/807	(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	-0.104	-0.067	-0.401	-0.082	-0.101	-0.210	-0.289	-0.258

Year applied 1977								
	(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**p***p<0.01