Patent-to-Patent Similarity:

A Vector Space Model

Kenneth A. Younge*
College of Management of Technology
École Polytechnique Fédérale de Lausanne
kenneth.younge@epfl.ch

Jeffrey M. Kuhn†
Haas School of Business
University of California Berkeley
jeffrey_kuhn@haas.berkeley.edu

July 30, 2016

Abstract

Current measures of patent similarity rely on the manual classification of patents into taxonomies. In this project, we leverage information retrieval theory and Big Data methods to develop a machine-automated measure of patent-to-patent similarity. We validate the measure and demonstrate that it significantly improves upon existing patent classification systems. Moreover, we illustrate how a pairwise similarity comparison of any and every two patents in the USPTO patent space can open new avenues of research in economics, management, and public policy. We make the data available for future scholarship through the Patent Research Foundation.

We thank Alan Marco, Andrew Toole, David Abrams Ludovic DiBiaggio, Neil Thompson, Noam Yuchtman, Rui de Figueiredo, and Tony Tong for helpful comments, and seminar participants at the United States Patent and Trademark Office, Skema Business School, and the Academy of Management.

* The authors thank Google for a generous research grant of computing time on the Google Cloud.

† The author thanks the Hoover Institution Working Group on Intellectual Property, Innovation, and Prosperity at Stanford University for financial support during the preparation of this paper.

1 Introduction

Concepts of technological space, distance, and relatedness are central to the study of invention and innovation. As Arrow points out, "the determination of optimal resource allocation for invention will depend on the technological characteristics of the invention process and the nature of the market for knowledge" (Arrow 1962). Researchers rely on technological space, explicitly or implicitly, to model a wide-range of phenomena, including technological diffusion (Nelson and Phelps 1966), technological opportunity (Rosenberg 1974), endogenous growth (Romer 1990), geographic spillovers (Jaffe, Trajtenberg et al. 1993), research productivity (Henderson and Cockburn 1996), knowledge transfer (Mowery, Oxley et al. 1996), corporate diversification (Silverman 1999), employee mobility (Almeida and Kogut 1999), boundary spanning (Rosenkopf and Nerkar 2001), knowledge recombination (Fleming 2001), knowledge complementarity (Nesta and Saviotti 2005), alliance networks (Gilsing, Nooteboom et al. 2008), competitive rivalry (Bloom, Schankerman et al. 2013), and many others. As technology is difficult to observe on a broad scale, much scholarship relies on the patenting system as a proxy (Griliches 1979).[1]

Empirical studies generally adopt one of two methods for positioning patents within a technological space: 1) manual classification of patents by the patent office (e.g., Jaffe 1986, McNamee 2013), and/or 2) linking patents through prior art citations (e.g., Podolny and Stuart 1995, Singh and Agrawal 2011). For example, research on knowledge spillovers combines both methods by examining patent citations, while controlling for technology classification (Jaffe, Trajtenberg et al. 1993). Researchers have attempted to improve these methods by moving to

[1] Scholars, of course, also use patent data to directly understand the patenting system itself (e.g., Graham and Harhoff, 2014, Hegde 2012). The measures from this paper also contribute to fields such as law, sociology, and public policy.

lower levels of granularity (Thompson and Fox-Kean 2005), incorporating multiple classifications (Benner and Waldfogel 2008), adjusting for relative distances within the classification taxonomy (McNamee 2013), and/or switching to alternative schemes such as the International Patent Classification system. Although such efforts have merit, concerns persist "about the accuracy of the more detailed U.S. patent subclasses" (Belenzon and Schankerman 2013) and one wonders if the field is beating a dead horse (or at least one that does not run in the way hoped for by researchers). For their part, patent attorneys indicate that they now largely ignore the classification system in favor of text-based searching.[2]

Scholars working with patent data often need to identify, compare, and/or match patents from similar domains of technology (Singh and Agrawal 2011, for example, use matching methods to compare firms and inventors with similar patent portfolios). Several problems arise from using the patent classification system for such purposes. First, patent classes may not be independent from the phenomena of interest. Patent classes are created, grouped, merged, split, and/or reassigned in ways that can correlate to changes in the technological space, and such effects can introduce a sample selection bias into the analysis. Second, patent classes may be too broad for the types of counterfactual comparisons attempted by researchers. For example, Schmookler observed patents for both a toothpaste tube and a manure spreader in the same "dispensing of solids" patent class/subclass (Schmookler 1966: P. 20, Jaffe 1986). Similarly, we observe patents for a compact catheter assembly (U.S. Patent No. 8,556,884) and a device for practicing golf (U.S. Patent No. 5,853,334) in class/subclass 242/370 because they both involve "winding, tensioning, or

[2] We interviewed ten patent attorneys; no one indicated that they looked at, or relied on, the classification system. Discussions with the USPTO, however, indicated that the patent classification system is helpful to the USPTO for assigning patents to examining divisions.

guiding." Third, patent classes may at times be too ***precise***, leading to a curse of dimensionality that lowers recall and the ability to find relevant comparisons (Benner and Waldfogel 2008). Narrow subclasses can also lead to measurement error as the USPTO attempts to pick the "best" subclass out of 150,000 options (Henderson, Jaffe et al. 2005). Overall, the econometric problem is that categorical classification can select systematic differences into a sample, which then are difficult to identify through sensitivity and robustness tests. As such, we agree that "the best way forward is to devise identification systems that avoid entirely technology classification systems" (Thompson and Fox-Kean 2005b: Page 466).

In this project, we demonstrate that the United States Patent and Trademark Office (USPTO) and World Intellectual Property Organization (WIPO) patent classification systems are too coarse for many of the comparisons sought after in academic research. Although this critique of the classification system is not new, we differ from previous work by leveraging information retrieval theory to develop a machine-automated measure of patent-to-patent similarity (one that is independent of institutionalized systems of classification) as a potential solution to the problem. Moreover, we find that a text analysis of the technical description (i.e., "specification") can discriminate between patents in the immediate space around a focal patent. In contrast, patent classes and subclasses include a considerable range of diversity, such that patent classes may be more appropriate for identifying broad, functional *categories,* as opposed to technological similarities. The measure we propose is also an initial step toward exploiting Big Data methods in patent research – methods that hold great promise for analyzing phenomena of high dimensionality (Fan, Han et al. 2014).

We make several contributions. First, we introduce and validate a *continuous* measure of technological similarity; prior research has largely relied on discrete divisions between categories for such measures. Second, our continuous measure is calculated at the dyadic, *patent-to-patent* level across the entire population of patents, whereas prior research has relied on small subsets of data, or "profile" measures aggregated up to the level of the firm. Third, we demonstrate that a text-based similarity measure provides greater accuracy, specificity, and generality than the patent classification system for many research questions. Fourth, we demonstrate how a continuous measure of patent-to-patent similarity can shed light on a wide range of problems in economics and management. Finally, we provide access to the data for follow-on research. Because the data resulting from this project is truly ("big" (bordering on massive), we formed a new organization to curate the data and provide open access for academic use.[3]

The plan for the paper is as follows. First, we define the vector space model ("VSM") and explain how it is calculated. Next, we validate the measure and compare it to the patent classification system. Finally, discuss opportunities for future research.

2 The Vector Space Model

Vector space models are one of the most robust methods in the field of Information Retrieval (Manning, Raghavan et al. 2008). Originally developed for the SMART retrieval system (Salton, Wong et al. 1975), vector space models now motivate nearly every modern search engine (Turney and Pantel 2010). This project, however, is the first we know of to map the entire USPTO patent space into a single vector space model at the patent-to-patent level.

[3] We thank Google for a generous research grant. See the Patent Research Foundation (www.patrf.org) for information on open access.

Jaffe (1986, Jaffe 1989) pioneered the use of vector space models at the level of the firm to position firms based on their patent portfolios. The 'Jaffe Method' constructs a profile for each firm, as vector $F = (F_1 \ldots F_K)$ across K patent class dimensions, where F_k equals the fraction of patents held by a firm in the k^{th} patent class (e.g., Ahuja 2000, Song, Almeida et al. 2003). Given a vector for each firm, the Euclidean distance (e.g., Rosenkopf and Almeida 2003), angular separation (e.g., Agarwal, Ganco et al. 2009), correlation (e.g., Benner and Waldfogel 2008), min-complement (e.g., Bar and Leiponen 2012), or Mahalanobis distance (e.g., Bloom, Schankerman et al. 2013) can then be calculated between any two firms.

While the above methods build on a "vector space model," they have important limitations. First, prior research relies on a discrete categorization within the patent classification system to define the dimensionality of the model (e.g., Chen and Chiu 2011). Later in the paper we will examine problems that arise from using the classification system and discrete comparisons to measure technological distance. Second, prior research aggregates data to the firm. Aggregation to the firm discards useful information at the patent level and conflates firms of different sizes, diversification, focus, and other characteristics. Moreover, the combination of categorization and firm-level aggregation leads to difficult tradeoffs in determining the optimal distance metric to use for a given research question (see Table X in Bloom, Schankerman, Van Reenen, 2013: page 1388).

In this project, we calculate the similarity of every possible *patent-to-patent* similarity comparison, for over 14 trillion calculations (an admittedly brute-force approach). We also construct the dimensionality of our model from the vocabulary used in patent documents, rather than the assignment of patent classifications. Doing so allows us to capture the similarity between

technological fields where patent classifications do not overlap, a key concern raised by Bloom et al. (2013). Moreover, lower-level data allow researchers to explicitly construct, and then aggregate, measures up to the firm level to address the particular research questions of their study. While there have been earlier attempts in this direction (e.g., Nanda, Younge et al. 2015), studies to date have focused on small portions of the overall space and have lacked comprehensive validation.

The Corpus. In the field of information retrieval, vector space models represent a high-dimensional space based on the full text of a given sample (or "corpus") of documents. Our corpus is the complete set of utility patents granted by the USPTO from 1976 through 2014. While most research samples from a population, we were able to calculate the pairwise similarity between *every* member of the corpus, for a full population analysis. We scraped HTML for each granted patent from the USPTO website, converted it to plain text, and extracted the section for the technical description.[4] Scraping from patent number 3,552,244 to 8,952,843 (a set of 5,400,599 potential patents), we successfully acquired 5,298,356 patents (98%). We investigated missing patents and all were withdrawn patents, missing from the USPTO system, or malformed records.

----- Insert Table 1 about here -----

The Vocabulary. Using a 'bag of words' methodology (Turney and Pantel 2010), we defined VSM dimensions based on the vocabulary of terms used by patents within the sample. We tokenized the technical description for each patent, removed punctuation, stemmed variations with an English Porter2 ("snowball") algorithm, and truncated terms at 25 characters. We dropped

[4] We focused on the full technical description (not the abstract), because the editorial process of condensing text down can introduce selection effects that reduce precision or otherwise bias the analysis. We excluded text from metadata and claims for such text may not reflect similarity in a technical sense, but instead similarity in terms of sharing the same legal firm, style of drafting claims, etc.

terms with numbers,[5] genetic sequences, a high proportion of repeating characters, natural-language stop words, legal stop words identified through interviews with patent attorneys, and terms with less than 3 characters. Finally, we dropped terms with extremely high document-frequencies and terms appearing in fewer than 10 patents (misspellings or unnatural language). Overall, iterative filtering helped to optimize the balance between Type-II errors (i.e., missing overlap), Type-I errors (i.e., erroneous overlap), and efficiency (Ming, Wang et al. 2010).

After building the vocabulary, we generated a weighted vector for each patent into the "space" based on the term-frequency of each term for a patent, scaled by the inverse document-frequency of each term across the corpus. Called term-frequency-inverse-document-frequency (Aizawa 2003), the "*tf-idf*" method allocates more importance to terms used frequently within a patent, while attenuating the importance of terms appearing widely across the corpus. In other words, the weight of a term for a given patent represents the product of the importance of the term for the patent, and the amount of information embedded in the term as determined by the corpus. *Tf-idf* weighting is one of the most widely used methods in text analysis; theoretical work suggests it should be efficient and consistent for most retrieval tasks (Aizawa 2003, Wu, Luk et al. 2008), and empirical work suggests it works well in practice (Elkan 2005).

The Similarity Measure. To measure similarity, we calculated the cosine of the angular separation between every two patents in the population. Figure 1 illustrates a simple example in three dimensions: vector A and vector B point out from the origin, θ is the angular separation between A and B, and cosine(θ) measures the "similarity" of the two vectors. The cosine of θ

[5] A case can be made for retaining alpha-numeric terms. "CAS9," for example, is an important protein. Cross-validation, however, found that alpha-numeric terms introduced an unacceptable rate of false-matching on arbitrary dimensions.

varies from zero (completely orthogonal) to one (completely overlapping). As reported in Table 1, most calculations for the full population were in fact close to zero (*Mean*=0.0177), and 97.27% of calculations had a raw cosine similarity score of less than 0.10. Manual review of similarity scores less than 0.10 indicated that variation below that threshold was essentially noise and that such measurements could be discarded from the dataset (i.e., not stored) and imputed later. Dropping data below a score of 0.10 reduced the storage and query requirement from 296.3 TB to 8.1 TB.[6]

----- Insert Figures 1 and 2 about here -----

Finally, although it is tempting to think of similarity as a percentage, manual inspection suggests that cosine similarity becomes informative around 0.20, and that two patents are quite similar by 0.50. We therefore transform results by stretching them from the captured 0.10 - 1.0 interval back to the 0.0 - 1.0 interval, and then radially stretching results away from the ends of the interval while maintaining rank order.[7] Summary statistics for the original Full Population, and transformed Final Sample are presented in Table 1 and graphed in Figure 2.

3 Validation

In this section, we validate four important aspects of the patent-to-patent measure of similarity: internal validity, content validity, external validity, and predictive validity.

Internal Validity. Internal validity refers to the degree to which a measure minimizes systematic error (Nunnally and Berstein 1994, Short, Broberg et al. 2010). Systematic error can

[6] The dropped data, however, must be imputed back into the distribution for many calculations. When it is necessary to impute values back into the dataset for certain calculations, we do so by adding missing values back at the 0.10 level — doing so should be conservative for most analysis as it winsorizes observations to a shorter interval.

[7] The final *Similarity* score was calculated as: *Similarity* = arcsine(square-root(((*rawresult* - 0.10)/0.90)))/arcsin(1)

result from malformed theory (where factors within a construct do not aggregate or interact as expected) or from errors made during computation. Below, we review the steps taken at both the theoretical level and computational level to avoid such problems.

When defining the vocabulary space, there is a tradeoff between identifying attributes that are rare and therefore informative, and giving too much power to erroneous terms. If the vocabulary space is too small, it will overlook important information (akin to a Type-II error); if the vocabulary space is too large, it will draw false connections (akin to a Type-I error). For example, a rare form of spelling mistake can, once weighted by *tf-idf,* increase the similarity between patents that share the spelling mistake; this is especially true when the spelling mistake is made on a common term that otherwise has little weight. Alternatively, failure to identify rare but highly distinguishing terms (e.g., "CRISPR"), will fail to identify a connection that could be clear to a human reader. As the optimal solution to this dilemma is difficult to determine *a priori*, we parsed, cleaned, and constructed the vocabulary through many iterative passes. After each pass (but before calculating the VSM), we inspected, tested, and recalibrated the vocabulary. Future work may benefit from *a posteriori* analysis of competing vocabularies, but budget limitations prevented us from making multiple passes at the entire similarity calculation itself.

A truly Big Data calculation (273 TB; n=14,036,290,800,546), our procedure was distributed over thousands of computers running on the Google Cloud platform for many weeks. Errors can arise for many reasons, including machine failures, network failures, and programming errors. When that happens, small ranges of data go missing. If processing is ordered by a variable of interest (such as time, as it often is for computational efficiency), then missing data will no longer

be 'missing at random,' and missing data can bias subsequent results. One cannot visually scan Big Data for errors, so we took additional steps to verify the results during and after computation. We also checked computational validity throughout the pipeline by verifying that counts and distributions of output data were consistent with the values expected from the input data.

Content Validity. The text analysis approach used for the vector space model derives directly from the content of patents themselves, and as such one would expect the content validity of the VSM to be intrinsically high. Nevertheless, we performed several face validity checks to ensure that the computed similarity values reflect real world phenomena. First, we checked if the average VSM similarity within a sample rises as one moves from a random sample to samples from companies with a narrower range of a given technology; we found that it did. Second, we checked the average VSM similarity of patent "families." Given that patent families share some or all of the text of the technical description amongst family members (through "patent priority claims"), one would expect the VSM similarity of patent families to be particularly high; we found that it was.

Our first test of face validity was whether VSM similarity rises as one examines the patent portfolios of increasingly focused firms. We plotted the pairwise VSM similarity between every two patents held within several firms in the medical instruments industry (Figure 3). We start with a random sample, and then move through firms known to have an increasing focus on a narrower and narrower slice of the medical instruments space. Starting with a random sample, the great majority of pairwise comparisons ($> 97\%$) share essentially no similarity (i.e., they have a similarity score < 0.20). Next, moving to Medtronic - a large medical instruments company with R&D in many different areas of the medical instruments space - the proportion of patents held by

the firm with a pairwise similarity score of 0.20 to 0.40 rises, and a small share appears at even higher levels of similarity. Next, moving to Intuitive Surgical - a midsized company focused on robotically assisted surgery - the share of unrelated patents continues to fall, the share of somewhat related patents continues to expand, and the share of highly related patents > 0.40 continues to expand. Finally, moving to Hansen Medical - a small and tightly focused company that patents exclusively in intravenous, catheter-based surgical systems - the share of unrelated patents continues to fall, the share of related patents continues to grow, and the share of highly-related patents (i.e., pairwise patents with a similarity score > 0.40) continues to grow to be the largest of the samples. We interpret the rising distribution of similarity as one moves from less-to-more focused R&D activities as face validation for the content validity of the VSM measure.

----- Insert Figure 3 about here -----

A second test of content validity is whether VSM similarity scores higher when one knows that the actual content of two patents should be more similar. Here we examine "patent families," a phenomenon driven by patent priority.[8] An applicant may file an application with a priority claim to expand upon claims filed in an earlier-filed application, while removing some prior art from consideration that would otherwise block the patent application. Importantly, a priority claim is valid only insofar as the later-filed application claims an invention that was already disclosed by the earlier filed application. Thus, two applications linked by a priority claim should be highly similar.

[8] A patent priority claim is an instance in which a later-filed patent application claims priority to an earlier-filed application with at least one inventor in common, effectively extending the priority date of the later-filed application earlier than its filing date, back to the priority date of the earlier-filed application.

We include in a patent family all patents that are linked by a U.S. patent priority claim. Patents with the most restrictive priority claims, referred to as continuation or divisional patents, should be virtually identical to their earlier-filed parent applications. We term these "clones." A somewhat less restrictive grouping includes clones, but also pairs of patents that share a priority claim to a common provisional patent application. We term these "immediate families." The loosest type of priority relationships are those established by a "continuation-in-part," in which the application declares that the later-filed application substantively expands upon the specification of the earlier-filed application. We term groups of patents joined by such linkages "extended families."

Next we show that the VSM correctly identifies family linkages. Figure 4 includes the breakdown of similarity values for random samples of patent dyads that share clone, immediate, and extended family relationships. Virtually all patent pairs linked by continuation or divisional are highly similar, with over 96% having a similarity value of at least 0.99. As expected, the correlation between family relationship and similarity decreases slightly but remains quite high as we relax the definition of families to include immediate and extended family members.

Because the similarity measure matches independent phenomena both inside and outside the patent system, we conclude that the similarity measure exhibits substantial content validity.

----- Insert Figure 4 about here -----

External Validity. To test the external validity of the VSM against human perceptions, we recruited a set of crowd-sourced raters, a U.S. patent attorney, and an expert inventor and asked each of them to rate a random sample of pairwise-comparisons from our VSM results. Because a simple random sample of pairs from our dataset would not select enough cases of high-similarity to

assess the high end of the scale, we selected a random sample stratified at similarity levels of 0.10, 0.30, 0.50, 0.70, and 0.90, with an approximately equal number of comparisons at each level.[9]

First, we tested the correlation of the vector space model to individuals who are ***not*** experts in the patenting system. We recruited seven workers from the Amazon Mechanical Turk crowdsourcing system [10] and asked them to rate each of the 150 pairwise-comparisons on a scale of: (9) extremely similar, (7) very similar, (5) somewhat similar, (3) not very similar, and (1) extremely dissimilar. Prior research demonstrates that a measure of central tendency by multiple, independent raters, can generate a surprisingly accurate assessment (Fuxman, Tsaparas et al. 2008). We therefore advertised that we would double the payment for ratings that agreed with the consensus rating of other raters assigned to the same comparison. We took the mode rating from seven raters as the consensus rating. Panel A of Figure 5 plots the results, with the number of consensus ratings at a given level of VSM similarity indicated on the graph, the size of each circle plotted proportional to the number of ratings, and the OLS best-fit line is overlaid as a dashed line. We find the Pearson correlation between the consensus rating and the vector space model to be moderate [r=0.63, $\chi^2(16)$=90.24], the inter-rater reliability to be high [*Cronbach* α=0.77], and the slope of the OLS best-fit line to be highly significant [p<0.0001].

----- Insert Figure 5 about here -----

Second, to test the external validity of the VSM measure against one skilled in the art of prosecuting patents, we hired an independent US patent attorney to manually rate the same

[9] Variation within stratified random sampling, however, led to a few more/less observations at each level.

[10] All M-Turk raters resided in the United States and had a "master-level" status. Amazon tracks workers and assigns/revokes a "Master-level" certification based on their reliability on previous categorization tasks. Because Master-level workers command a premium in the M-Turk marketplace, they presumably are motivated to work harder, and be accurate, to retain their status.

sample of patent comparisons presented to the crowd-sourced workers, using the same scale. Panel B of Figure 5 plots the results. We find the Pearson correlation between the attorney rating and the vector space model to again be moderate (r=0.60, χ^2=87.96), the inter-rater reliability to be high (*Cronbach* α=0.75), and the slope of the OLS best-fit line to be highly significant (p<0.0001). We note a remarkable consistency between results from the crowd-sourced and patent attorney validation test. In a three-way test of inter-rater reliability between the expert, crowd-sourced, and VSM rating, we find an overall inter-rater reliability of *Cronbach* α=0.84, a very high level even for psychometric studies (Cronbach 1971, Cortina 1993, Nunnally and Berstein 1994).

Finally, to test external validity in a particular domain, we took a random sample of patents from Intuitive Surgical Inc. (as described earlier under content validity), and recruited an inventor familiar with those technologies to rate a random sample of pair-wise comparisons. The sample was again stratified at similarity levels of 0.10, 0.30, 0.50, 0.70, and 0.90. Time constraints with the (volunteer) inventor prevented us from testing an equivalently sized sample as with the (paid) patent attorney and crowd workers, so we reduced the sample size from 150 to 30. Panel C of Figure 5 plots the results. We find the Pearson correlation between the inventor and the vector space model to increase substantially (r=0.73, χ^2=29.17), the inter-rater reliability to also increase (*Cronbach* α=0.85), and the slope of the OLS best-fit line to steepen somewhat (p<0.0001).

The consistency of external validity tests across different types of external raters lead us to conclude that the vector space model generally conforms to what a human rater would expect from the measure. In our post-rating discussions with the attorney and inventor, however, we should note that the VSM measure of technological similarity does *not* necessarily demark the

areas of intellectual property that are protected by a given patent, or how one patent might block another. Instead, the measure identifies technological elements in common between two patents.

Predictive Validity. Predictive validity is the extent to which a test measure can predict a criterion measure. In patent data, the clearest criterion measure is whether a pair of patents is also classified as similar by patent applicants and/or the USPTO. Patent applicants and/or the USPTO may declare two patents to be similar in one of two ways. First, a previously-issued patent may be cited during the prosecution of a later-issued patent. Second, two patents may be similarly classified by the patent office, either explicitly through the USPTO's patent classification system or implicitly through assignment into the same administrative patent examining division.

We begin by using patent similarity to predict patent citations. McNamee (2013) provides the most powerful test to date for predicting patent citations: he constructs a measure of the degree to which a pair of patents shares the same subclass(es) under the USPTO's patent classification system. To test the measure, McNamee constructed a sample of all USPTO patents from class 704 (either primary class or additional class). He split the dataset into two subsets by grant year: the first subset (1997-2001), was taken as the potential issuer of a citation; the second subset (2002-2006), was taken as the potential receiver of a citation. We duplicate the sample in McNamee and compare our variable VSM *Similarity* to the subclass overlap variable reported in McNamee.

Table 2 reports predictions of patent citation for the VSM measure of patent similarity. All models were estimated with `relogit`, a rare events logistic regression package in Stata (King and Zeng 2001). Although citation is a very rare event (less than 0.001 of patent-to-patent pairs), the sample size for the rare event was greater than 16,000, which should be sufficient to avoid small-

sample bias in logistic regression. Columns 1 and 2 predict whether a given dyad also exists as a citation (no: 0; yes: 1). Column 1 copies results from (McNamee, 2013) based on his measure of extended patent subclass overlap. Column 2 replicates the sample reported by McNamee, and performs an equivalent analysis using the VSM measure of textual similarity. The aforementioned sample construction resulted in approximately 15 million potential citation dyads for analysis. In Column 1, McNamee finds extended subclass overlap to be a highly significant predictor of citation, with subclass similarity predicting 12% of variation. In column 2, we find that the VSM similarity measure is also a highly significant predictor of citation, explaining 14% of variation.

----- Insert Table 2 about here -----

Table 2 indicates that the vector space model outperforms subclass similarity, even when looking within a sample of dyad pairs pre-selected to share the same main class. Selecting a sample where patents share the same class, and then using a class-based measure to predict citations, is also a form of selecting on the dependent variable and likely overstates the power of the taxonomy. We therefore take issue with McNamee's conclusion that "the type of analyses done in dataset #1 are not possible without taxonomically appropriate methods like those presented in this paper" (McNamee 2013: p. 871). Additionally, an important limitation of relying on the patent classification system is that one cannot analyse the strengths or weaknesses of the patent classification system itself. An advantage of the VSM similarity measure is that it is not dependent upon any explicit classification scheme; it works with a randomly selected sample just as well as with a sample selected based on patent class. We therefore move in the next section to a regression framework where we use patent similarity to predict patent classification.

In Table 3 we analyse whether similarity predicts an overlap in patent classification. The sample is drawn from a random sample of 16,000 patents. After constructing a unique list of all pairwise combinations between the 16,000 patents (i.e., the bottom triangle of the cross-joined matrix), the sample resulted in 1,675,563,971 observations. The transformed *Vector Space Similarity* variable was merged in from the vector space model, and indicator variables were merged in from USPTO metadata to indicate whether each pairwise comparison was assigned to the same Art Unit for examination (DV in Column 1), shared a primary USPTO main class (DV in Column 2), shared a primary USPTO main class and subclass (DV in Column 3), or shared an Art Unit, primary USPTO main class and subclass (DV in Column 4).[11] The pseudo R^2 for each model was calculated using the `relogit` function in Stata. As is shown in Table 3, the VSM measure is a strong predictor of shared main class (Column 1) and of main class and subclass (Column 2). The transformed VSM variable is highly significant at the p=.001 level, and explains approximately 19% of the variation in whether or not a patent dyad shares the same primary USPTO main class and subclass. The predictive validity improves even more when predicting dyads that not only share the same primary main class and subclass, but are also grouped by the USPTO into the same Art Unit. In this case, the VSM measure explains 22% of the variation.

----- Insert Table 3 about here -----

Based on the results reported above, a vector space model based on the technical description of patent applications appears to be a reliable and consistent measure of similarity. First, the VSM measure is a highly significant predictor of pre-existing measures of patent similarity such as

[11] McNamee (2013) relies on the patent classification system and therefore is undefined for this problem and not tested.

shared main class or subclass. Second, the VSM measure explains a greater proportion of the variation than pre-existing measures of patent similarity. Third, the VSM measure is not reliant on human-based classification schemes such as patent class and thus is not subject to selection effects. Fourth, by being independent of human-based classification schemes, the VSM measure can be compared with existing classification schemes without selecting on the dependent variable. Further research is needed to assess the full merits of the measure for applied work, but initial tests suggest that it may be a powerful and broadly applicable tool for understanding the similarity and relatedness of patents across a range of contexts.

4 Classification

We now turn to consider how automated measurements of similarity by a vector space model compare to manual classification of patents by the patent office. Given that the dominant method for scholars to position patents in a technological space relies on the manual classification of patents by the patent office (Benner and Waldfogel 2008), it is important for research to assess and understand the relative strengths and weaknesses of the classification system. As reported below, we find that the VSM is better suited for identifying near-neighbors, while the patent classification system is better suited for identifying broad, categorical groupings.

Recently, scholars have highlighted the fact that most patents are actually cross-classified into many different classes and subclasses (Benner and Waldfogel 2008, Bar and Leiponen 2012, McNamee 2013), not just the primary one referenced in most studies. The NBER patent data file (Hall, Jaffe et al. 2001), for example, drops multiple classifications altogether and retains only the

primary classification for analysis. Given that the NBER patent file is the most frequently used patent dataset, it is perhaps unsurprising that most studies do not address multiple classification assignments. To explore the issue, we processed the XML metadata files from the USPTO from 2005 through 2014. We identified 491 distinct USPTO main classes and 82,520 distinct subclasses at the six-digit level.[12] (Henderson, Jaffe et al. 2005). For patents granted after 2005, 20.1% of patents were assigned to just one classification (the "primary" main class / sub class combination), 23.4% were assigned two classifications, and 56.5% were assigned to three or more classifications.

While early research focused on the primary ***main*** class (e.g., Jaffe, Trajtenberg et al. 1993), later research argued that main class alone may be too coarse (e.g., Thompson and Fox-Kean 2005). While some research has moved to use subclasses (e.g., Belenzon and Schankerman 2013), others have argued that subclasses may increase measurement error (Henderson, Jaffe et al. 2005). Yet other work suggests that measures of overlap for one subclass suffers small-sample bias, such that scholars should use the secondary (multiple) subclass classifications to cast a wider net and reduce sampling variation (Benner and Waldfogel 2008). And Bar and Leiponen (2012) point out that multiple classifications can be irrelevant when making portfolio-to-portfolio comparisons.

To better understand the relative merits of using subclasses, we first examine the distribution of similarity for patents that share such categorization. Figure 6 plots the cumulative percent of pairwise comparisons (where both patents fall *within* the same classification) for given ranges of similarity. We constructed a random subsample for each column where each patent pair in the sample shares the grouping indicated in the figure. A random sample from across the entire

[12] Henderson et al. (2005) report there being approximately 150,000 subclasses in the USPTO classification system. Subclasses however morph over time, and our data for this cross-check derive from a short interval from 2005 to 2014.

dataset is included for comparison. For the sample of dyads that share the same USPTO main class and subclass, we find that 30% of dyads fall below a similarity value threshold of 0.20, and only 10% of dyads have similarity values of greater than 0.40. Viewed another way, a plot of the cumulative distribution of patent-to-patent pairs within a shared primary USPTO subclass indicates that over one-third of patents within a primary USPTO subclass are too dissimilar to be accurately rated by the vector space model, resulting in similarity values of approximately zero.

----- Insert Figure 6 about here -----

Earlier we reported results from a logistic model using the VSM measure of patent text similarity to predict patent classification at the main class plus subclass level; that model explained 19% of the variation in patent class assignment. The unexplained variation, however, is not necessarily due to measurement error or inconsistency in the VSM predictor variable. Instead, most dyads within a given patent classification simply have a low level of similarity. Due to the *tf-idf* weighting of vectors, the VSM measure is better suited for identifying differences between near-neighbours than it is suited for discriminating between distant relations, in part because relatively few key terms determine the overall variation in the scale.

In all of our empirical results, we find considerable variation within subclasses for our measure of similarity. The question therefore arises as to whether the patent classification system can be exploited in some way to accurately center around a patent of interest. To illustrate the point, Figure 7 presents four Venn diagrams to illustrate various problems that arise from overlapping patent classifications. A focal patent of interest is represented by a red star, and other patents are distributed as black dots in spatial relation to their VSM similarity. Panel A illustrates the full set

of patents associated with a hypothetical primary patent class A, and panel B illustrates the set of patents associated with sub-class A.1. The problem here is that a patent near the boundary of the class can actually have a lower similarity to other members of its subclass, as compared to some members of its main class. Furthermore, Panel C illustrates that taking the set of patents formed through the union of sub-classes A.1, B.1, and C.1, as suggested by Benner and Waldfogel (2008), can increase the problem. By comparison, Panel D illustrates that a multidimensional set of patents, as identified by patent similarity, can potentially improve the selection by identifying both ***new*** patents (plotted as green triangles), while including some previously identified patents (plotted as blue squares), and while excluding distant/irrelevant patents. Although the exact mix and number of new patents, retained patents, and excluded patents will depend on the coarseness and accuracy of the patent classification system in a particular domain, the illustrations of Figure 7 demonstrate the potential for systematic error in the classification scheme.

To the extent that a vector space model cuts across arbitrary selection boundaries, we expect the VSM to generate a more precise – and more reliable – measure of technological space. Nevertheless, we note that the validity of patent classification depends on the use case for such data. The patent classification system may work well for identifying functional characteristics of a focal patent, but our results suggest that it does not work well for identifying near-neighbors to a focal patent. In other words, a categorical assignment may be helpful for the USPTO when assigning a patent application to an art unit for examination, but be problematic for scholars when examining problems that are relational.

----- Insert Figure 7 about here -----

5 Future Research and Conclusion

A precise measure of technological space is important for many core questions in economics, strategy, and innovation. In this paper, we develop and validate a new measure of technological space, and the underlying data is already supporting a range of additional research endeavors: Kuhn and Younge (2016) use similarity data to show how patent citation patterns have become dramatically noisier over time; Younge and Römer (2016) use similarity data to control for the pre-existing distribution of inventive activity and contribute new evidence to the debate on the geographic localization of knowledge spillovers (Jaffe, Trajtenberg et al. 1993, Thompson and Fox-Kean 2005); and Thompson and Kuhn (2016) use similarity data to evaluate the effect of winning a patent race on cumulative innovation. As we continue to share the patent-to-patent similarity data with other researchers, we expect the list of derivative research to continue to grow. Finally, we comment on a few of the many possibilities for using the measure in future research (Table 5).

Patent similarity can be used to position firms (see, for example, Stuart and Podolny 1996): scholars can compare the countervailing effects of spillovers in technology space and business stealing in the product market space (Bloom, Schankerman et al. 2013); market entry can be detected by identifying cases when a firm first files a patent similar to portfolios by incumbents (Hall and Ziedonis 2001); and patent portfolios of allied (or competing) firms can be compared by patent similarity to measure the extent of overlap or divergence between firms. (Mowery, Oxley et al. 1996). Patent similarity can also be used to measure innovation: researchers can compare the similarity of a focal patent to patents issued before/after it to measure breakthroughs (Ahuja and Lampert 2001); and patent similarity can be used to measure the effects of venture capital funding

by observing firms' technological contributions (Kortum and Lerner 2000). Patent similarity can also be used to investigate the patent system itself: patent similarity can be used to identify patent thickets and/or holdup by investigating whether dense webs of patents deter product development (Lemley and Shapiro 2006); patent similarity can be used to analyze patent scope by comparing the effects of boundary-spanning patents on follow-on innovation (Merges and Nelson 1990); and the effect of winning or losing a patent race can be investigated by comparing outcomes by highly similar patents by different firms (Lerner 1997). The VSM similarity measure may shed new light on selection effects that are inherent to many applications of patent data: patent similarity can be used to construct robust counterfactuals for investigating phenomena such as geographic spillovers (Jaffe, Trajtenberg et al. 1993, Thompson and Fox-Kean 2005), and patent-to-patent similarity can also be used to better understand patent citations (Trajtenberg 1990).

----- Insert Table 5 about here -----

5.1 Conclusion

Scholars have argued that existing measures of technological space are too coarse for many research questions (Thompson and Fox-Kean 2005), and we provide descriptive evidence in support of that argument. In addition, we develop a new measure of technological space and validate it for future research. Although classification-based measures remain useful for some uses, we argue that textual similarity is a better measure for researchers to use for the following reasons.

First, the VSM measure generates a single, standardized value that represents the same concept, even in different contexts and across different samples. It is conceptually simpler than patent classification, as there is only one number, it works the same for any comparison, and there

are fewer (arbitrary) decisions to be made about the correct primary/secondary, class/subclass, and classification scheme to follow. Nevertheless, researchers also can use patent-to-patent similarity at a low-level to construct derivative measures at a higher level to meet specific research needs. Second, the VSM measure is continuous, bounded zero to one, free of breakpoints and divisions, and calculated in a way that is consistent across domains. Although measures based on a patent classification scheme may sometimes appear to be continuous (due to aggregation), classification itself is always discrete, for a patent either belongs to a class or it does not. Third, the VSM measure is free of selection effects and errors made during classification, as well as the lack of observations associated with many patent subclasses. Fourth, we demonstrate the validity of patent-to-patent similarity using a range of techniques. The measure is based on recent theoretical and empirical work in information retrieval theory. It correlates well with inventor, patent attorney, and lay person notions of relatedness; it predicts patent citations better than existing class-based measures; and it matches with patent phenomena both internal (patent families) and external (firm technological focus) to the patent system.

Finally, although raw patent data is technically available to all, the field has advanced in large part due to efforts by others to generate datasets that are easy to use and widely available (Hall, Jaffe et al. 2001, Li, Lai et al. 2014). In a similar vein, we provide access to our data via the Patent Research Foundation (www.patrf.org). We believe the field benefits the most from open access and the freedom to explore research questions without the selection effects and constraints that come from co-authorship. We therefore encourage other scholars to use this article as a guide to the data, and to use the data in their own research.

References

Agarwal, R., M. Ganco and R. H. Ziedonis (2009). "Reputations for toughness in patent enforcement: Implications for knowledge spillovers via inventor mobility." Strategic Management Journal **30**(13): 1349-1374.

Ahuja, G. (2000). "Collaboration Networks, Structural Holes, and Innovation: A Longitudinal Study." Administrative Science Quarterly **45**(3): 425-455.

Ahuja, G. and C. M. Lampert (2001). "Entrepreneurship in the large corporation: a longitudinal study of how established firms create breakthrough inventions." Strategic Management Journal **22**(6-7): 521-543.

Aizawa, A. (2003). "An information-theoretic perspective of tf–idf measures." Information Processing & Management **39**(1): 45-65.

Allison, J. R. and M. A. Lemley (2002). "Growing Complexity of the United States Patent System, The." BUL Rev. **82**: 77.

Almeida, P. and B. Kogut (1999). "Localization of Knowledge and the Mobility of Engineers in Regional Networks." Management Science **45**(7): 905-917.

Arrow, K. (1962). Economic Welfare and the Allocation of Resources for Invention. The Rate and Direction of Inventive Activity: Economic and Social Factors. R. Nelson. Princeton, NJ, Princeton University Press**:** 609-626.

Bar, T. and A. Leiponen (2012). "A measure of technological distance." Economics Letters **116**(3): 457-459.

Belenzon, S. and M. Schankerman (2013). "Spreading the word: Geography, policy, and knowledge spillovers." Review of Economics and Statistics **95**(3): 884-903.

Benner, M. and J. Waldfogel (2008). "Close to you? Bias and precision in patent-based measures of technological proximity." Research Policy **37**(9): 1556-1567.

Bloom, N., M. Schankerman and J. Van Reenen (2013). "Identifying technological spillovers and product market rivalry." Econometrica **81**(4): 1347-1393.

Chen, Y.-L. and Y.-T. Chiu (2011). "An IPC-based vector space model for patent retrieval." Information Processing & Management **47**(3): 309-322.

Cortina, J. (1993). "What Is Coefficient Alpha? An Examination of Theory and Applications." Journal of Applied Psychology **78**(1): 98-104.

Cronbach, L. (1971). Test validation. Educational measurement. R. L. Thorndike. Washington, DC, American Council of Education 443-507.

Elkan, C. (2005). Deriving TF-IDF as a Fisher Kernel. String Processing and Information Retrieval. M. Consens and G. Navarro, Springer Berlin Heidelberg. **3772:** 295-300.

Fan, J., F. Han and H. Liu (2014). "Challenges of Big Data analysis." National Science Review **1**(2): 293-314.

Fleming, L. (2001). "Recombinant Uncertainty in Technological Search." Management Science **47**(1): 117-132.

Frakes, M. D. and M. F. Wasserman (2014). "The Failed Promise of User Fees: Empirical Evidence from the U.S. Patent and Trademark Office." Journal of Empirical Legal Studies **11**(4): 602-636.

Fuxman, A., P. Tsaparas, K. Achan and R. Agrawal (2008). Using the wisdom of the crowds for keyword generation. Proceedings of the 17th international conference on World Wide Web, ACM.

Gilsing, V., B. Nooteboom, W. Vanhaverbeke, G. Duysters and A. van den Oord (2008). "Network embeddedness and the exploration of novel technologies: Technological distance, betweenness centrality and density." Research Policy **37**(10): 1717-1731.

Graham, S. J. H. and D. Harhoff (2014). "Separating patent wheat from chaff: Would the US benefit from adopting patent post-grant review?" Research Policy **43**(9): 1649-1659.

Griliches, Z. (1979). "Issues in Assessing the Contribution of Research and Development to Productivity Growth." The Bell Journal of Economics **10**(1): 92-116.

Hall, B. H., A. B. Jaffe and M. Trajtenberg (2001). The NBER patent Citations Data File: Lessons Insights and Methodological Tools, NBER.

Hall, B. H. and R. H. Ziedonis (2001). "The patent paradox revisited: an empirical study of patenting in the US semiconductor industry, 1979-1995." RAND Journal of Economics **32**(1): 101-128.

Hegde, D. (2012). "Funding and performance at the US Patent and Trademark Office." Nature Biotechnology **30**(2): 148-150.

Henderson, R. and I. Cockburn (1996). "Scale, Scope, and Spillovers: The Determinants of Research Productivity in Drug Discovery." The RAND Journal of Economics **27**(1): 32-59.

Henderson, R., A. Jaffe and M. Trajtenberg (2005). "Patent Citations and the Geography of Knowledge Spillovers: A Reassessment: Comment." American Economic Review **95**(1): 461-464.

Jaffe, A. B. (1986). "Technological Opportunity and Spillovers of R & D: Evidence from Firms' Patents, Profits, and Market Value." The American Economic Review **76**(5): 984-1001.

Jaffe, A. B. (1989). "Characterizing the "technological position" of firms, with application to quantifying technological opportunity and research spillovers." Research Policy **18**(2): 87-97.

Jaffe, A. B., M. Trajtenberg and R. Henderson (1993). "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations." The Quarterly Journal of Economics **108**(3): 577-598.

King, G. and L. Zeng (2001). "Logistic Regression in Rare Events Data." Political Analysis **9**(2): 137-163.

Kortum, S. and J. Lerner (2000). "Assessing the contribution of venture capital to innovation." RAND journal of Economics: 674-692.

Kuhn, J. and K. A. Younge (2016). Patent Citations: An Examination of the Data Generating Process. Working Paper: University of California Berkeley/École Polytechnique Fédérale de Lausanne, Available on SSRN: http://papers.ssrn.com/abstract_id=2709238.

Lemley, M. A. and C. Shapiro (2006). "Patent holdup and royalty stacking." Tex. L. Rev. **85**: 2163.

Lerner, J. (1997). "An Empirical Exploration of a Technology Race." The RAND Journal of Economics **28**(2): 228-247.

Li, G.-C., R. Lai, A. D'Amour, D. M. Doolin, Y. Sun, V. I. Torvik, A. Z. Yu and L. Fleming (2014). "Disambiguation and co-authorship networks of the U.S. patent inventor database (1975–2010)." Research Policy **43**(6): 941-955.

Manning, C. D., P. Raghavan and H. Schütze (2008). Introduction to information retrieval. Cambridge, U.K., Cambridge University Press.

McNamee, R. C. (2013). "Can't see the forest for the leaves: Similarity and distance measures for hierarchical taxonomies with a patent classification example." Research Policy **42**(4): 855-873.

Merges, R. P. and R. R. Nelson (1990). "On the complex economics of patent scope." Columbia Law Review: 839-916.

Ming, Z. Y., K. Wang and T. S. Chua (2010). Vocabulary filtering for termweighting in archived question search. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). **6118 LNAI:** 383-390.

Mowery, D. C., J. E. Oxley and B. S. Silverman (1996). "Strategic alliances and interfirm knowledge transfer." Strategic Management Journal **17**: 77-91.

Nanda, R., K. A. Younge and L. Fleming (2015). Innovation and Entrepreneurship in Renewable Energy. The Changing Frontier: Rethinking Science and Innovation Policy. A. B. Jaffe and B. Jones. Cambridge, MA, NBER/University of Chicago Press.

Nelson, R. R. and E. S. Phelps (1966). "Investment in humans, technological diffusion, and economic growth." The American Economic Review: 69-75.

Nesta, L. and P. P. Saviotti (2005). "Coherence of the Knowledge Base and the Firm's Innovative Performance: Evidence from the U.S. Pharmaceutical Industry." The Journal of Industrial Economics **53**(1): 123-142.

Nunnally, J. C. and I. H. Berstein (1994). Psychometric Theory. New York, NY, McGraw-Hill.

Podolny, J. and T. Stuart (1995). "A role-based ecology of technological change." American Journal of Sociology: 1224-1260.

Romer, P. M. (1990). "Endogenous Technological Change." The Journal of Political Economy **98**(5): S71-S102.

Rosenberg, N. (1974). "Science, invention and economic growth." The Economic Journal: 90-108.

Rosenkopf, L. and P. Almeida (2003). "Overcoming Local Search Through Alliances and Mobility." Management Science **49**(6): 751-766.

Rosenkopf, L. and A. Nerkar (2001). "Beyond local search: boundary-spanning, exploration, and impact in the optical disk industry." Strategic Management Journal **22**(4): 287-306.

Salton, G., A. Wong and C.-S. Yang (1975). "A vector space model for automatic indexing." Communications of the ACM **18**(11): 613-620.

Schmookler, J. (1966). Invention and Economic Growth. Cambridge, Harvard University Press.

Short, J. C., J. C. Broberg, C. C. Cogliser and K. H. Brigham (2010). "Construct Validation Using Computer-Aided Text Analysis (CATA)." Organizational Research Methods **13**(2): 320-347.

Silverman, B. S. (1999). "Technological Resources and the Direction of Corporate Diversification: Toward an Integration of the Resource-Based View and Transaction Cost Economics." Management Science **45**(8): 1109-1124.

Singh, J. and A. Agrawal (2011). "Recruiting for ideas: how firms exploit the prior inventions of new hires." Management Science **57**(1): 129-150.

Song, J., P. Almeida and G. Wu (2003). "Learning-by-Hiring: When Is Mobility More Likely to Facilitate Interfirm Knowledge Transfer?" Management Science **49**(4): 351-365.

Stuart, T. E. and J. M. Podolny (1996). "Local Search and the Evolution of Technological Capabilities." Strategic Management Journal **17**: 21-38.

Thompson, N. and J. Kuhn (2016). Patent Races in The Real World: They're Different Than Theory Suggests. Working Paper: Massachusetts Instititute of Technology/University of California Berkeley.

Thompson, P. and M. Fox-Kean (2005). "Patent Citations and the Geography of Knowledge Spillovers: A Reassessment." The American Economic Review **95**(1): 450-460.

Thompson, P. and M. Fox-Kean (2005). "Patent Citations and the Geography of Knowledge Spillovers: A Reassessment: Reply." American Economic Review **95**(1): 465-466.

Trajtenberg, M. (1990). "A Penny for Your Quotes: Patent Citations and the Value of Innovations." The RAND Journal of Economics **21**(1): 172-187.

Turney, P. D. and P. Pantel (2010). "From frequency to meaning: Vector space models of semantics." Journal of artificial intelligence research **37**(1): 141-188.

Wu, H. C., R. W. P. Luk, K. F. Wong and K. L. Kwok (2008). "Interpreting TF-IDF term weights as making relevance decisions." ACM Trans. Inf. Syst. **26**(3): 1-37.

Younge, K. A. and B. C. Römer (2016). The Geography of Knowledge Spillovers: A Reassessment Based on Patent-to-Patent Similarity. Working Paper: École Polytechnique Fédérale de Lausanne/Colorado State University.

Table 1: Summary statistics.

a. Source data.

Patents	Total count	5,298,356
Patents	Date range	1976 through 2014
Patents	Numerical range (USPTO patent numbers)	3552244 through 8952843
Vocabulary	Count of terms (unique)	6,333,830
Vocabulary	Count of terms used in VSM (dimensions)	743,253

b. Sample construction.

	(1)	(2)	(3)
	Full Population*	Retained Calculations	Final Sample
Observations	14.0 trillion	370.8 billion	14.0 trillion
Percent of full pop.	100%	2.73%	100%
Data storage	296.3 TB	8.1 TB	8.1 TB
Similarity – Mean	0.0177	0.1748	0.0051
Similarity – S.D.	0.0335	0.0844	0.0331
Similarity – Skew	5.87	2.28	7.80

Notes: The Full Population includes all unique pairwise comparison calculations of *Similarity* in the 5,298,356 x 5,298,356 cross-joined matrix for all USPTO patents available at the time of the study. Because similarity comparisons are bi-directional (the same, regardless of direction), only the bottom left triangle of the matrix was actually calculated. Because values below a threshold of 0.10 were dropped post-calculation to make the data storage tractable, distributional statistics for the mean, standard deviation, and skew of *Similarity* were estimated for the Full Population from a random sample of 10,000 observations where *all* observations were retained. Column 2 for Retained Calculations summarizes the data initially captured and stored to disk after dropping values below a threshold of 0.10. The Final Sample imputes dropped values back into the distribution at the minimum level and transforms the distribution across the entire unit interval of [0,1] by recalculating every raw result as: *Similarity* = arcsine(square-root(((*rawresult* - 0.10)/0.90))).

Table 2: Predictions of patent citation.

This table reports regressions results for predictions of patent citation between pairs of patents. Following McNamee (2013), the sample was constructed from all patents classified by the USPTO under class 704 (for either the primary class or any additional class) and issued between 1997 and 2006 (inclusive). The sample was then split into two sets: the first set runs from 1997 through 2001, inclusive; the second set runs from 2002 through 2006, inclusive. The set from 2002 through 2006 was taken as the potential *issuer* of a citation, and the set from 1997 through 2001 was taken as the potential *receiver* of a citation. The sample replication resulted in approximately 15.7 million dyads for analysis, roughly equivalent to the 15.1 million dyads found by McNamee. The dependent variable is whether a given dyad in the sample is also a patent citation (no=0, yes=1). Column 1 includes results for the predictor *Extended Subclass Overlap* as developed and reported by McNamee. Column 2 changes the predictor to *Vector Space Similarity* as developed in this study. Both columns were estimated with the rare events logistic regression function `relogit` in Stata (King and Zeng 2001). Although citation is a very rare event (less than 0.001 of cases), the sample size for the rare event was greater than 16,000, suggesting sufficient sample size to avoid small-sample bias in logistic regression. Results were cross-checked with results from the `logit` function and found to be nearly identical. Following McNamee, pseudo R^2 was calculated and reported from the `logit` model. (Standard Errors are reported in parentheses.) Statistical significance levels: *** $p<0.001$, ** $p<0.01$, * $p<0.05$.

	(1)	(2)
	McNamee	VSM
Extended Subclass Overlap	0.62***	
	(0.00)	
Similarity		7.7183***
		(0.035)
Constant	-10.08***	-8.2060***
	(0.02)	(0.013)
Total patent-to-patent pairs	15,126,093	15,741,565
Cited patent-to-patent pairs	16,285	16,221
Pseudo R^2	0.12	0.14

Table 3: Predictions of patent classification.

This table reports regression results for predictions of shared patent classification between pairs of patents. The data was drawn from a random sample of approximately 18,000 patents, which was then crossed-joined to generate a unique list of pairwise combinations (i.e., the bottom triangle of the cross-joined matrix) for a sample of 163,004,589 observations. The following indicators were calculated and merged into the sample for dependent variables: an indicator for whether both patents in a pairwise comparison were assigned to the same USPTO Art Unit for examination; an indicator for whether a focal pairwise comparison shared the same USPTO primary main-class; an indicator for whether a focal pairwise comparison shared both a primary USPTO main-class and sub-class; and an indicator for whether a focal pairwise comparison shared all three criteria (art unit, primary main-class, and primary sub-class). All models were estimated with the `relogit` function for rare events logistic regression in Stata (King and Zeng 2001). Although common classification is an extremely rare event (less than 0.00005 of cases), the sample size for the rare event was nevertheless greater than 7,900 in all models, suggesting sufficient sample size to avoid small-sample bias in logistic regression. Results were cross-checked with results from the `logit` function and found to be nearly identical. Pseudo R^2 was calculated and reported from the `logit` model. (Standard Errors are reported in parentheses.) Statistical significance levels: *** $p<0.001$, ** $p<0.01$, * $p<0.05$.

	(1)	(2)	(3)	(4)
Dependent Variable (No=0 / Yes=1)	Shares Art Unit	Shares Main-class	Shares Sub-class	Shares All
Vector Space Similarity	10.3161***	11.6959***	12.3782***	12.9403***
	(0.010)	(0.008)	(0.025)	(0.041)
Constant	-5.9504***	-4.9455***	-9.2609***	-10.6254***
	(0.001)	(0.001)	(0.007)	(0.014)
Total patent-to-patent pairs	163,004,589	163,004,589	163,004,589	163,004,589
Overlap patent-to-patent pairs	542,882	1,527,228	26,967	7,922
Pseudo-R^2	0.08	0.10	0.16	0.20

Table 4: Future research.

Topic / Application	Key Reference
Strategic Position Map the relative positioning of firms in a unified technological space.	(Stuart and Podolny 1996)
Strategic Rivalry Compare spillovers and rivalry in both technology and product market space.	(Bloom, Schankerman et al. 2013)
Strategic Entry and Response Detect technology entry and portfolio competition.	(Hall and Ziedonis 2001)
Strategic Alliances Measure the relatedness and overlap of alliance partners.	(Mowery, Oxley et al. 1996)
Patent Novelty Construct measures of prior activity and breakthroughs in a space.	(Ahuja and Lampert 2001)
Venture Capital Evaluate the impact of venture capital on subsequent innovation.	(Kortum and Lerner 2000)
Patent Law Examine the impact of patent scope.	(Merges and Nelson 1990)
Patent Holdup Identify patent thickets and situations of holdup.	(Lemley and Shapiro 2006)
Patent Races Track the technological trajectory of patent races.	(Lerner 1997)
Geographic Spillovers Control for the pre-existing distribution of inventive activity.	(Jaffe, Trajtenberg et al. 1993)
Patent Citations Assess the quality and relevance of patent citations.	(Trajtenberg 1990)

Figure 1: Angular separation in a vector space model.

This figure illustrates the angular separation of two vectors in a simplified, two-dimensional vector space model. Our measure of similarity is equal to the cosine of the angle between each pair of patents in the vector space. Whereas two dimensions are graphed in this illustration, the actual vector space model encompasses 743,253 orthogonal dimensions.

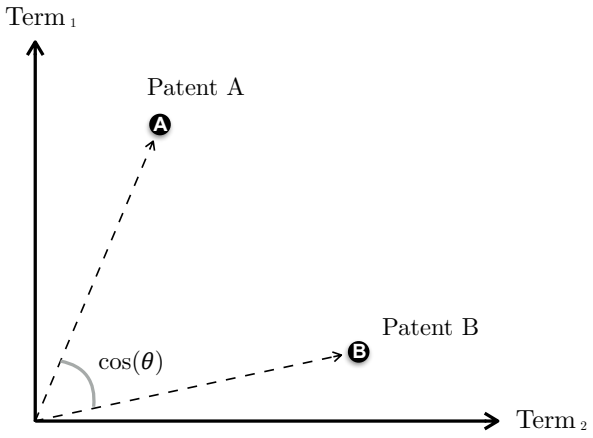
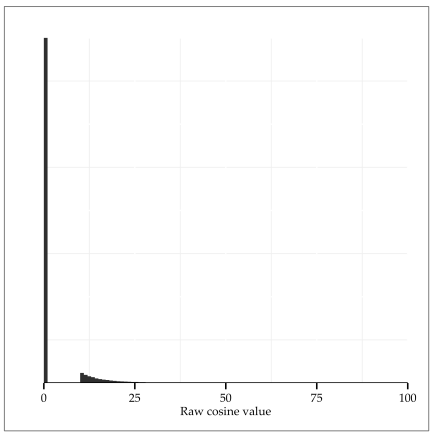
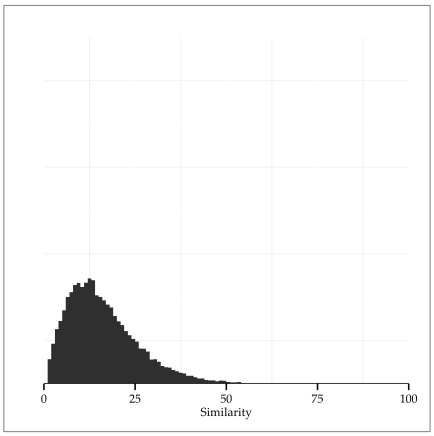


Figure 2: Distribution of similarity by sample.

This figure plots the distribution of raw cosine scores (Panel A) and transformed *Similarity* scores > 0.01 (Panel B). The sharp spike at zero in Panel A corresponds to the 288.2 TB of calculated results that fell below the threshold of 0.10 and were assumed to be zero. Panel B plots a histogram for the distribution of *Similarity* scores > 0.01 after transforming the raw cosine similarity score above 0.10 with the arcsine-square-root transformation described in the paper and omitting the spike from Panel A for clarity.



(a) Full population (n=14,036,290,800,546)



(b) Final Sample (n=370,842,517,085)

.

Figure 3: Patent similarity within firms.

This figure compares the similarity of patent portfolios for three firms by plotting the cumulative percent of pairwise comparisons *within* each portfolio for a given range of similarity. Firms were selected from the medical instruments industry, where it is generally known that Medtronic has the broadest R&D focus out of the set, Intuitive Surgical has a mid-range R&D focus out of the set, and Hansen Medical has the narrowest R&D focus out of the set. A random sample from across the entire dataset is included for comparison.

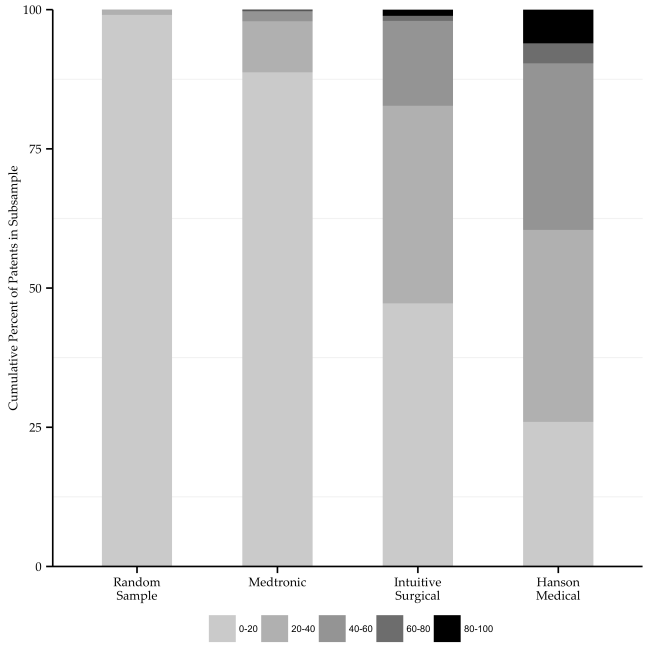


Figure 4: Patent similarity within patent families.

This figure plots the cumulative percent of pairwise comparisons within each subsample for a given range of similarity. "Clone Family" includes pairs of patents that are direct continuation or divisional patents. "Extended Family" includes pairs of patents that share any domestic priority claim, including those based on a continuation-in-part or provisional application. "Immediate Family" includes pairs of patents that share any domestic priority claim except one based on a continuation-in-part.

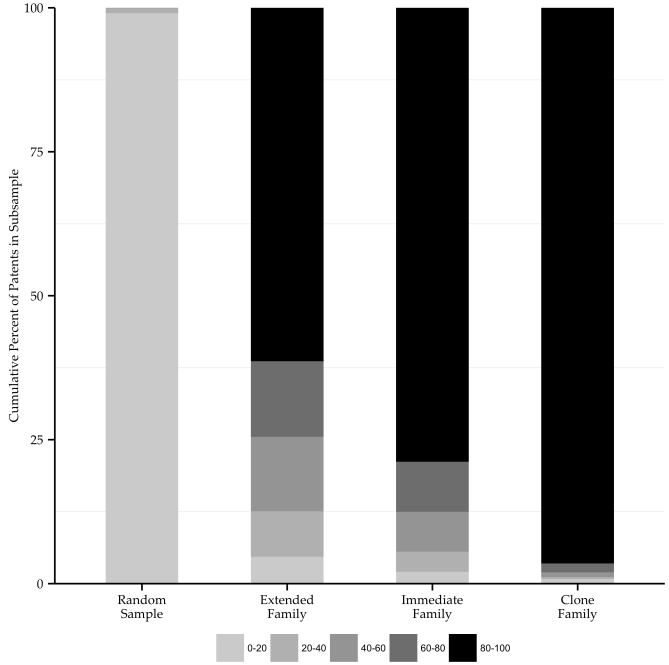
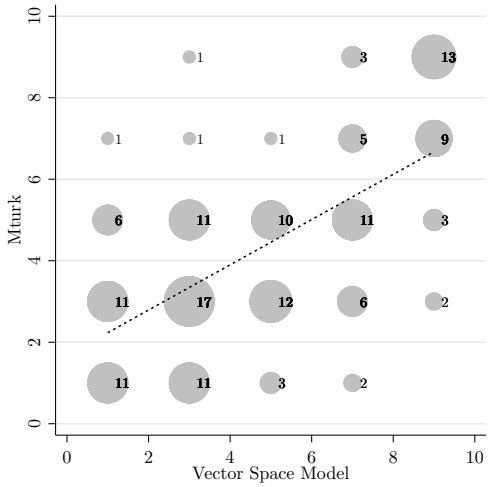
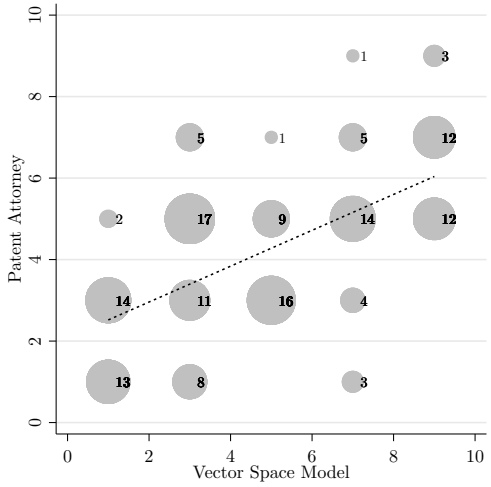


Figure 5: External validation.

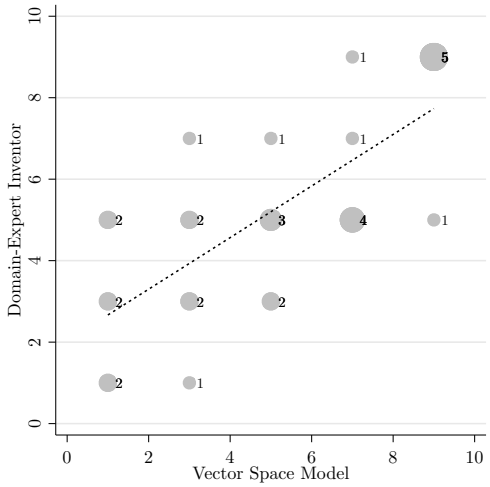
These figures plot a comparison of human-rated (on the vertical axis) and VSM-rated (on the horizontal axis) similarity scores for a random sample of patent-to-patent comparisons drawn at stratified levels (0.10, 0.30, 0.50, 0.70, 0.90) of VSM similarity. Dyads were randomly selected from the same sample used in Figure 3, related to three firms in the medical instruments industry (Medtronic, Intuitive Surgical, and Hansen Medical). Human raters were asked to read the technical description for each patent in the dyad and to then rate the perceived similarity of those two patents on a scale of: (9) extremely similar, (7) very similar, (5) somewhat similar, (3) not very similar, and (1) extremely dissimilar. Results are plotted with circles proportional to the number of ratings obtained at given levels of human-rated and VSM-rated similarity. The OLS best-fit line, between human-rated and VSM-rated scores, is overlaid as a dashed line. For Panel A we hired seven Amazon Mechanical Turk workers to rate each dyad, and we retained the mode rating between the raters as the consensus rating. Here we find Pearson correlation r=0.63; Cronbach Alpha α=0.77; OLS p<0.0001. For Panel B we hired a bar-certified US patent attorney to rate the same 150 dyads as in Panel A. Here we find Pearson correlation r=0.60; Cronbach Alpha α=0.75; OLS p<0.0001. A three-way test of inter-rater reliability between the expert rating, crowd-sourced rating, and vector space model indicates an overall inter-rater reliability of *Cronbach* α=0.84. Finally, for Panel C we recruited an inventor with domain expertise in medical instruments to rate a sub-sample of 30 dyads from the sample sample used in Panels A and B. Here we find Pearson correlation r=0.73; Cronbach Alpha α=0.85; OLS p<0.0001).



(a) Crowd-sourced workers (n=150)



(b) Patent attorney (n=150)



(c) Inventor in domain of expertise (n=30)

Figure 6: Patent similarity within patent classifications.

This figure compares the similarity of different levels of patent classification and/or organizational unit by plotting the cumulative percent of pairwise comparisons *within* the sample that fall into a given range of similarity. A random subsample was constructed for each column where each patent pair in the sample shares the indicated grouping. Each sample is restricted to 2005 and later where data was available for USPTO Art Unit. A random sample from across the entire dataset is included for comparison. The USPTO divides its examining corps into eight "Technology Centers," which are in turn divided into 567 "Art Units." Main classes often span across art units, and a given Art Unit may be assigned to examine patent applications from a number of patent subclasses divided among different patent classes based on the logical similarity of the classified technology. Art Units therefore capture some variation orthogonal to the classification system.

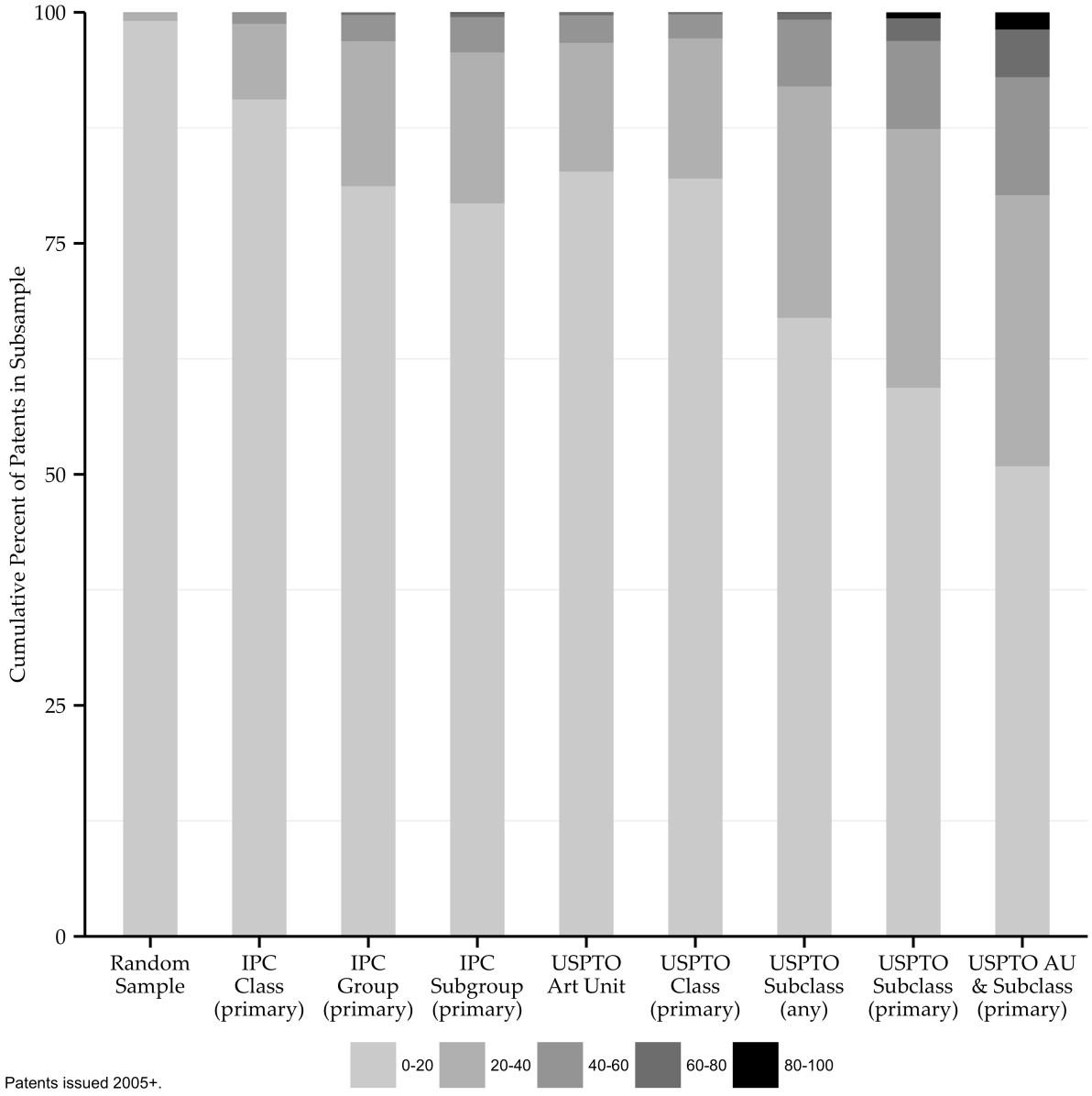
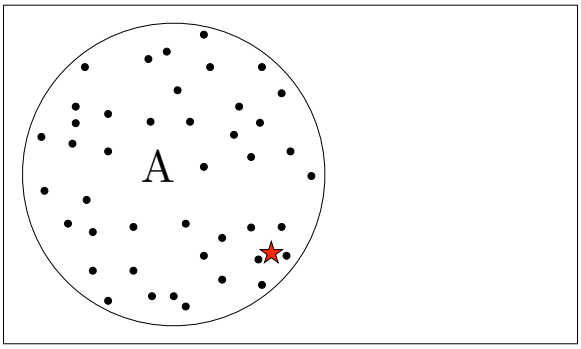
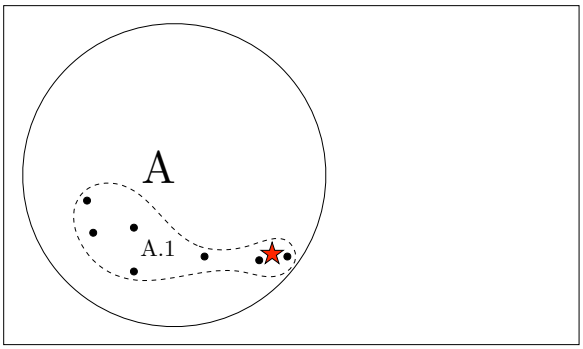


Figure 7: Overlapping patent classifications.

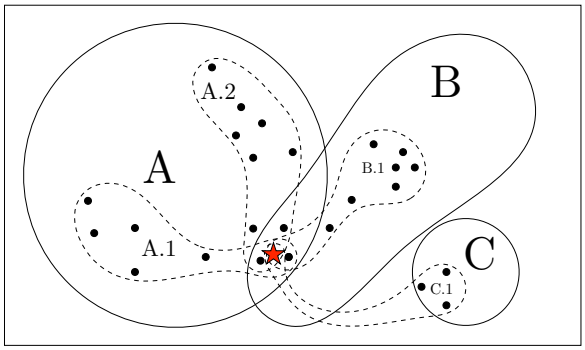
The following Venn diagrams illustrate the problem of overlapping patent classification for a focal patent of interest (red star). Patents (black dots) are distributed in spatial relation to their VSM similarity. Panel (a) illustrates a set of patents associated with a primary patent class A. Panel (b) illustrates the set of patents associated with sub-class A.1. Panel (c) illustrates the set of patents formed through the union of sub-classes A.1, B.1, and C.1. Panel (d) illustrates the set of patents identified by patent similarity, with new patents as green triangles and previously identified patents as blue squares.



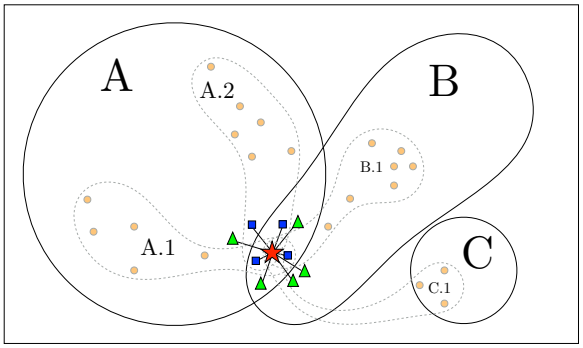
(a) Primary class A



(b) Primary sub-class A.1 in class A



(c) Union of sub-classes A.1, B.1, C.1



(d) VSM similarity around focal patent

APPENDIX A: Calculation of ***tf-idf***.

For clarity, we explain here in greater detail how we calculated the weighted vector for each patent based on the term-frequency, inverse-document-frequency *(tf-idf)*. We calculated *tf-idf* as:

$$\textbf{tf-idf}(t,\ d,\ D) = f_{t,d}\ \text{ x } \log(N/n_t)$$

where $f_{t,d}$ = raw frequency of term t in document d, N = total number of documents, and n_t = the number of documents with term t. When combined into a single vector, the pattern of *tf-idf* weighted values acts as a signature that differentiates (positions) one patent relative to another within the overall vocabulary space (Aizawa 2003). The *tf-idf* weighting identifies particularly representative terms for a given patent, such that one would expect *tf-idf* weighting to far outperform an ad hoc selection of keywords by an individual doing a keyword search. Any given patent will not use most terms in the vocabulary space; moreover, the formula for *tf-idf* pushes the weight for a dimension to zero as a term appears in more documents (i.e., the ratio N/n_t approaches 1 and the logarithm of N/n_t goes to zero). Consequently, vectors are "sparse" (mostly zeros), or weighted toward zero. To illustrate this point, Figure A1 below plots the cumulative percent of total ***tf-idf*** weighting as a Lorenz curve for a random sample of 10,000 vectors. Vectors, on average, have 307 non-zero dimensional weightings, but the distribution of weightings is highly skewed; 20 dimensions contribute a majority of the weight going into the cosine calculation (*Gini coefficient*=0.705).

Figure A1: Relative weight of *tf-idf* terms in a random sample of patent vectors.