

Close to You?
Bias and Precision in Patent-Based Measures of Technological Proximity

Mary Benner
2010 Steinberg Hall-Dietrich Hall
The Wharton School
Philadelphia, PA 19104

benner@wharton.upenn.edu
215-746-5719

Joel Waldfogel
Business and Public Policy Department
The Wharton School
Philadelphia, PA 19104

waldfogj@wharton.upenn.edu
215-898-7148

July 25, 2007

Close to You? Bias and Precision in Patent-Based Measures of Technological Proximity

Abstract

Patent data have been widely used in research on technological innovation to characterize firms' locations as well as the proximities among firms in knowledge space. Researchers could measure proximity among firms with a variety of measures based on patent class data, including Euclidean distance, correlation, and angle between firms' patent class distributions. Alternatively, one could measure proximity using overlap in cited patents. We point out that measures of proximity based on small numbers of patents are imprecisely measured random variables. Measures computed on samples with few patents generate both biased and imprecise measures of proximity. We explore the effects of larger sample sizes and coarser patent class breakdowns in mitigating these problems. Where possible, we suggest that researchers increase their sample sizes by aggregating years or using all of the listed patent classes on a patent, rather than just the first.

Patent data have been widely used in organizational research in recent years to study firms' technological innovation (e.g. Rosenkopf & Nerkar, 2001; Benner & Tushman, 2002; Sorenson & Stuart, 2000; Helfat, 1994; Ahuja, 2000; Ahuja & Katila, 2001), knowledge flows between firms (e.g. Rosenkopf & Almeida, 2003; Song, Wu, & Almeida, 2003; Mowery, Oxley, & Silverman, 1996), diversification (e.g. Silverman, 1999; Miller, 2004; Argyres, 1996; Piscitello, 2004), and technological spillovers (e.g. Jaffe, 1989; 1986; Jaffe, Trajtenberg, and Henderson, 1993). A common use of patent data has been to characterize firms' technological positions and calculate the proximities of firms in technological space (e.g. Jaffe, 1986; 1989; Stuart & Podolny, 1996; Rosenkopf & Almeida, 2003).

This research frequently makes use of one of two approaches for measuring proximity. First, researchers use the 3-digit patent classifications listed on each firm's patents, provided by the U.S. Patent and Trademark Office (USPTO). The distribution of patents across classes is taken to characterize a firms' location in knowledge space, and the distance between firms' technological resources is frequently assessed by calculating the distance between vectors of patent class listings (e.g. Jaffe, 1986; Ahuja, 2000; Song, Almeida & Wu, 2004). Two firms are close if they tend to patent in the same technological domains, that is, if their shares of patents in various patent classes are similar. Second, researchers use measures of overlap in patent citations between firms (Stuart & Podolny, 1996; Mowery, Oxley, & Silverman, 1996).

The increasing use of patent data in strategy and technology research has given rise to a need to better understand the accuracy and appropriate use of such data. In this paper we explore the bias and imprecision of estimates of various measures of proximity between firms when the measures are calculated on small samples. We start with firms holding large patent portfolios, allowing us to estimate their distances relatively accurately from the large number of patents in

their full samples. We then compute inter-firm distance measures based on random samples of the patent classes listed on each firm's patents. The resulting proximity measures are random variables, and the apparent distance between two firms depends on the samples drawn. To evaluate the performance of estimates of proximity measures based on, say, 25 patent class listings per firm, we calculate measures repeatedly using different random samples of the firms' listings, a process known as "bootstrap sampling." This method, which has seen increasing use in social science recently, allows us to examine properties of estimators empirically even when – as is the case here – analytical derivations would be tedious or difficult. The resulting empirical distribution of proximity measures shows how estimators of proximity perform, in bias and precision, relative to the full-sample values. We perform this exercise for increasingly large sample sizes, from 25 to 5000, to allow us to compare how bias and precision improve with larger samples.

The main result in this paper is that for sample sizes commonly employed in the literature, patent-based proximity measures have substantial bias and imprecision. Using larger samples of patents or patent class listings, aggregating data across years, or aggregating to coarser groupings of patent classifications can mitigate some of these problems.

Our paper joins recent studies, including Alcacer and Gittleman (2006) and Fogarty, Jaffe and Trajtenberg (2000) in providing insight into current practice and caveats associated with the use of patent data. We are interested in better understanding the consequences for typical uses of proximity measures based on patent data. Our goal is to guide future research using patent data by outlining the conditions under which bias and precision issues might emerge. To the extent the potential for these issues can be recognized and mitigated, patent information can become even more reliable and valuable in research.

We proceed in four sections. Section 1 provides background on the USPTO's patent classification system along with a discussion of the methods for measuring proximity among firms. Section 2 describes the data used in this study, a sample of 118,350 photography-related patents granted to firms in the photography industry between 1980 and 2000 that includes all the patent classifications listed for each patent. Section 3 presents results, and we present our conclusions in Section 4.

I. Patent Classifications

The USPTO's patent classification system, comprising over 400 3-digit (main) patent classes in addition to over 120,000 nested subclasses, is designed to facilitate the storage and retrieval of patent documents to "assist patent examiners performing patentability searches" for related claims in prior patents (USPTO Handbook of Classification, 2006). Unlike taxonomies such as SIC or NAICS codes that organize industries hierarchically so that digits convey information about industries' relatedness, patent classes were devised as a filing system. While patents within a particular patent class are related to one another, patents in numerically adjacent patent classes are not necessarily more related to one another than are patents in distant classes.

Each patent document must be classified by a patent examiner into one "Original Classification" or "OR," which is the first patent class listed on the patent document. There can be only one OR, even if the "claims"¹ on a patent or information about the invention could easily be assigned to multiple different 3-digit patent classes. The USPTO's Patent Classification

¹ The specific ways the current invention represents an advance over prior art.

Office has extensive rules for determining the OR, selected to correspond to the “controlling claim” in the patent information.²

Patents can list additional patent classes beyond the OR, called “cross-reference classifications” or XRs. These are “mandatory” when there is information about the invention (claimed or unclaimed) in the patent that pertains to a different 3-digit patent class. Patents can also include “discretionary” XRs if the examiner deems them useful for searching for the information on the patent – the information in this case is not part of the specific information about the invention (USPTO, Patent Examiner’s Handbook).³ Thus a patent document may list several different 3-digit patent classes. In our sample, discussed in greater detail below, patents contained as many as – in one case – 130 patent class listings, with an average of four patent classes. This ratio is not unusual.

Although most patents are classified into more than one 3-digit patent class, with a few exceptions (Fleming & Sorenson, 2001; 2004; Benner & Tushman, 2002), the research that has utilized patent class information includes only the first patent class listed for each patent and ignores the subsequent listed patent classes. The NBER patent data file, which is easily available and widely used by researchers, also includes only the first listed patent class, termed the “original 3-digit patent class” (Hall, Jaffe, & Trajtenberg, 2001).⁴

Utilizing only the first listed patent class in research on technology and innovation relies on an (often unstated) assumption that the first classes are representative of the underlying technological domain of a firm. Yet a review of the USPTO’s stated process for selecting the

² In Appendix 1, we have included the detailed hierarchy of rules from the Patent Examiners’ Handbook for determining the controlling claim and corresponding OR.

³ At this time there is no way for users of patent data to determine which classes are mandatory or discretionary.

⁴ Hall, Jaffe, and Trajtenberg (2001) argue: “For the vast majority of uses one is likely to resort only to the original, 3-digit patent class, and hence we include only it in the **PAT63_99** file.” The NBER patent data are available at <http://www.nber.org/patents/>.

Original Classification suggests that a patent's (or correspondingly, a firm's) full technological breadth may not be represented accurately when only the first patent class is included.

For patents that defy easy classification, the patent examiners' instructions (partly summarized in the appendix) make clear why the OR does not fully represent the patent's location in technological space. When a patent spans multiple technologies, examiners are instructed to choose the controlling claim and associated patents class by employing the following hierarchy of subject matter:

1. Relating to maintenance or preservation of life
2. Chemical subject matter
3. Electrical subject matter
4. Mechanical subject matter
 - a. Dynamic
 - b. Static

When a firm's patent portfolio is summarized by its distribution of ORs, this algorithm makes it appear that the firm is more involved in "preservation of life," relative to chemistry, electricity, or mechanical subject matter, than is actually so.

Second, the patent office makes a concerted effort to limit cross-referencing, in part, by encouraging the use of subclasses within each patent class which can be used to further delineate the claims related to a particular technology (USPTO, Handbook of Classification, 2006).⁵ This suggests further that if additional 3-digit patent classes do appear in a patent document, despite the USPTO's efforts to limit them, it is because the patent pertains to a technology that is outside the first 3-digit patent class and cannot be characterized by classifying it further in subclasses within the first patent class.

⁵ The text referring to the assignment of a patent to additional cross-reference patent classes is also included in Appendix 1, in section V.

Moreover, some patent classes, such as Nanotechnology (Patent Class 977), are cross-reference classes only. Patents pertaining to nanotechnology must be assigned to another Original Classification, as patent class 977 never appears as the first patent class on a patent. As a result, research on innovation pertaining to nanotechnology must utilize the full set of cross reference patent classes to identify such patents.

In light of the patent office's classification system, using only the Original Classification, i.e. the first patent class listed on a patent has two problems. First, it will not accurately characterize the technological footprint of that patent. Second, the failure to use the remaining cross-reference patent classes listed on the patent discards information that would be useful for producing the attendant benefits of larger sample sizes.

1. Locating Firms in Technological Space

The idea underlying distance measurement is that each firm has a location in N -dimensional knowledge space, where the N dimensions should in principle represent kinds of technology (e.g. electrical, chemical). It is impossible to observe a firm's knowledge directly, but patents provide a convenient window into its knowledge. Following the large literature using patent data to measure distance among firms, we can view the patent class listings on a firm's patents as draws from the firm's knowledge base. For firms that patent a lot – and thus have many patent class listings – their many patents provide many draws from their knowledge urn. In that case, researchers can form relatively accurate estimates of those firms' technological locations. Firms that patent less frequently – and therefore have a smaller number of patent class listings – offer fewer draws from their urns. It is harder for researchers to accurately determine the technological locations of those firms.

Like each firm, each patent also has a location in technological space. Patents contain multiple pieces of information usable for their classification: these include the assignment to any of roughly 120,000 patent class/subclass combinations, or to coarser breakdowns such as the 400 3-digit patent classes, or to even coarser divisions. Hall, Jaffe, and Trachtenberg (2001) have provided two such higher level classifications or partitions within the NBER patent database. The first aggregates all of the roughly 400 patent classes into 36 two-digit technological subcategories, and the second further groups the subcategories into an even coarser set of six main categories, including: Chemical (excluding Drugs); Computers and Communications (C&C), Drugs and Medical (D&M); Electrical and Electronics (E&E); Mechanical; and Others. In what follows we work with all three of these possible descriptions of knowledge space.

The first step in class-based distance measurement is choosing the dimensions which, in this context means choosing whether to classify patents by class and subclass (approximately 120,000 dimensions), 3-digit class (approximately 400 dimensions), or by the coarser 36 or 6 class groupings. While low-cost computing eliminates technical limits on the number of dimensions, adding dimensions does not necessarily improve the measures. An heuristic geographic analogy helps illustrate this. Suppose we are trying to characterize whether two groups of people in the US live near one another using data on the populations of the groups in each of the roughly 30,000 US postal zip codes. We also know the groups' populations by county, state, and region. How should we calculate the proximity of their residential patterns?

Using the most finely detailed (zip code) information, we could calculate proximity measures based on vectors showing the share of each group in each zip code. But this can give misleading results. Suppose that all members of both groups live in the state of Wyoming, but they never live in the same zip code. The use of zip codes as the dimensions of "location space"

shows the groups' residential patterns to be as different as if all members of one group lived in New York City while all members of the other group lived in Los Angeles, even though they both reside only in Wyoming. The example shows that spaces with more dimensions do not necessarily provide more accurate measures of distance. Depending on the question one seeks to answer, using counties, states, or even regions consisting of groups of states could provide a more sensible measure. In this example, using zip codes as dimensions makes sense only if different zip codes are as conceptually distinct as, say, vastly distant regions of the country.

The lesson of this example for our technological proximity problem is that one does not necessarily want to use the finest partition possible such as the thousands of patent classes in conjunction with subclasses. Patent classes, like zip codes, are most useful in assessing technological distances between firms to the extent they are conceptually distinct. Patent class partitions should instead group like technologies together into coarser partitions.

Beyond the choice of what to use as dimensions researchers can choose among a variety of ways of locating a particular patent in knowledge space. One approach is to treat each patent as a point along the dimension identified by the OR. This is a common approach in the literature (e.g. Rosenkopf & Almeida, 2003; Song, Almeida, & Wu, 2003; Ahuja, 2000). Patent k 's location in technological space is represented with a 1 on the dimension associated with the class of its OR, along with zeroes elsewhere in the vector. The firm's location in knowledge space is estimated by summing the vectors representing each of the M patents, element by element, then dividing the elements by M . The elements of each firm's vector p sum to 1, and the elements of p estimate the shares of the firm's knowledge that reside in each of the technological domains.

For reasons detailed above, it is conceptually preferable to use all listed patent classes rather than just the first. If all listed patent classes are used, rather than simply the Original

Classification, then the vector representing a particular patent's location in knowledge space has multiple positive elements. That is, it is a vector with a 1 for each listed class and zeroes elsewhere. The elements of p are then estimated by summing the vectors and dividing them by the number of classes listed across all M patents used in the calculation.

Once a researcher selects both the dimensionality and the approach for calculating the p vector describing each firm's location, distances among firms may be calculated. There are three common methods. First, researchers calculate Euclidean distance (E) to compare two vectors representing the set of first patent classes listed on each firm's patents (Ahuja, 2000; Rosenkopf & Almeida, 2003). Technological proximity between firms i and j is defined as:

$\sqrt{\sum_{c=1}^N (p_i^c - p_j^c)^2}$, where N is the number of dimensions (patent classes) in the partition. In this work, N is either 400, 36, or 6.

A second possible measure of distance is the angle (A) between firm i and j :

$$\frac{\sum_k p_{ik} p_{jk}}{\sqrt{\sum_k p_{ik}^2 \sum_k p_{jk}^2}}, \text{ where indexes are summed over the } N \text{ dimensions (see Jaffe, 1989).}$$

When the firms have the same location, this measure equals 1, when their knowledge vectors are orthogonal, it is zero. A third possible measure is correlation between vectors, which we term C . The correlation runs from -1 to +1.

Citation overlap measures provide an alternative way of measuring proximity among firms (Podolny & Stuart, 1996; Mowery, Oxley, & Silverman, 1996). These measures ask how many of the patents that one firm cites are also cited by another firm. Suppose the universe of

cited patents runs from 1 to T , and δ_i^j is an index that is 1 if firm j cites patent i and 0 otherwise.

Then the overlap between firm j and k , which we term O , is
$$\frac{\sum_{i=1}^T \delta_i^j \delta_i^k}{\sum_{i=1}^T \delta_i^j}.$$

This is the share of the patents cited by firm j that are also cited by firm k . Note that this measure is not symmetric (the denominator is the number of patents cited by either firm k or firm j).

In what follows we calculate both class-based proximity measures (A , C , E) with p vectors of various dimension as well as the overlap measure (O).

II. Data

To address the questions of bias and precision in calculating technological proximity using patent class information, we rely on patent data collected for prior studies of innovation among firms in the photography industry (e.g. Benner & Tushman, 2002). The sample includes patents that fall within any of 12 technological classes broadly related to photography and imaging granted to 64 firms in the photography industry spanning the period 1980 to 2002.⁶ The included patent classes are listed in Table 1. A patent was selected for the sample if it was held by one of the 64 firms with operations in the photography industry and if any of its listed patent classes fell into the 12 classes in the study.

Because these data were obtained for other purposes, it is a convenience sample for this study. However, the approach used to collect the data follows other research using patent data and patent class information. In prior work, researchers often select an industry or technological

⁶ The data were obtained from CHI Research, but similar patent data including all listed patent classes are available in the Micropatent data used by Stuart & Podolny (1996) and others.

area to study and choose the relevant patent class(es) (e.g. Song, Almeida, & Wu, 2003; Rosenkopf and Nerkar, 2001). Patent data is then collected from within those patent classes for a set of firms participating in the industry. For example, Rosenkopf and Nerkar (2001) select one patent class and a set of nested subclasses pertaining specifically to the optical disk industry for collecting the relevant patent data for their study of firms' exploration. Similarly, Ahuja's (2000) study of the effect of alliances on innovation explored firms' patents related only to chemicals patent classes. Beginning by narrowing the technological area under study allows for comparable measures across firms of activities in related business areas, especially important when comparing the activities of large, multibusiness firms. Our data were collected in a similar way, and thus can shed light on the typical current use of the patent class information. Our dataset draws the patent data from an unusually broad range of 12 patent classes, however, yielding 118,350 patents, more than the usual number in studies such as this.

In addition, each of our patent records includes all of the patent classes listed on the patent, that is, the Original Classification as well as all Cross-Reference Classifications, allowing us to compare the implications of using only the first patent class versus all the listed patent classes. The data include 463,611 patent class listings. The mean number of patent classes listed on each patent is 3.9, and the median is 3.

To determine whether the number of listed classes per patent in our sample is typical, we tabulated the mean and median number of patent classes listed per patent for a random sample of IBM patents. IBM was chosen because of their extensive patenting behavior. A search of IBM patents on the USPTO website yielded 46,546 patents granted to IBM between 1976 and the present (as of May 31, 2007). We randomly sampled by choosing every 500th patent in the list of IBM patents. We then accessed the patent document, and manually read and coded the patent

classes listed on the front page. This process resulted in a set of 93 patents. In this sample, the average number of patent classes per patent was 3.9 and the median was 3, identical to our sample. The number of patent classes in the IBM sample ranged from one per patent to 14. In our sample the maximum number of patent class listings is 130.⁷ The 99th percentile is 16.

III. Analysis

The main analysis in this paper assesses how estimates of distance between firms perform, in terms of bias and precision, as the samples underlying their calculation increase in size. We begin with the class-based proximity measures (A , C , E) and then turn to the overlap measures (O).

It is intuitively clear that a distance measure between two firms using many patent class observations will be more accurate than measures based on only a few patent class observations. Accurate measurement of distances (discussed in the previous section) assumes that the p 's are measured with precision.

Of course the precision with which the p 's are measured is a direct function of the number of listings sampled (and relatedly, of the number of patents sampled). The proximity measures are nonlinear functions of various proportions, so it is not immediately obvious how sample sizes translate to their precision.

Each element of the p vector is a proportion, such as the proportion of a firm's patent class listings in class 396. If π is the share of the firm's patent class listings that would fall into class 396 with an infinitely large patent sample, then the standard deviation of an estimate of this

⁷ While 130 is clearly a large number of listed patent classes, one should bear in mind that this number is the maximum across about 300,000 sample patents, while the IBM number is the maximum across fewer than 100 IBM patents.

true share based on n patent class listings is $\sqrt{\pi(1-\pi)/n}$. As n increases, the precision of the estimated proportion increases. The proximity measures A , C , and E are complicated nonlinear functions of the p vectors' elements, and it is difficult to derive properties of their estimators analytically. Fortunately, fast and inexpensive computing makes it possible to examine them empirically using the bootstrap to produce empirical distributions of proximity estimates for different-sized samples.⁸ To see how this works in our context, consider two firms in our sample, Kodak and Canon, both of which have thousands of patents over our 22-year period. Because they have so many patents (and therefore, patent class listings), their p vectors estimated on their full samples are precise measures of their distributions of patent portfolios across the partition chosen; and, by extension, the proximity calculations are precise measures of the distance between them.

1. Kodak-Canon Proximity

To see how sample size affects precision, we perform the following exercise: draw samples of, say, 25 patent class listings on Kodak patents and 25 class listings on Canon patents. Using these samples, we compute the associated proximity measures between Kodak and Canon. We then repeat with new random samples of 25 listings per firm. We perform this exercise 500 times with replacement, and we plot the empirical distribution of estimated proximity measures based on samples of 25. The distribution of these 500 estimates shows how the estimators perform with 25 observations per firm. We then repeat the entire exercise with different-sized samples, between 50, 100, 500, 1000, and 5000 patent class listings per firm. We explore how

⁸See Efron, B, and Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Monographs on Statistics and Applied Probability, No. 57. Chapman and Hall, London.

the resulting distributions of estimated proximity measures change, in terms of bias and precision, as sample sizes increase.

Tables 2-4 illustrate the results of this exercise for our three distance measures computed on p vectors of three numbers of dimension (based on 3-digit patent classes as well as the coarser aggregations included in the NBER database (Hall et al, 2001)). The top panel of Table 2 shows how the distribution of the Euclidean distance measure (E) between Kodak and Canon based on 3-digit patent classes changes as the size of the patent class sample from each firm increases from 25 to 5000. The first row, for example, summarizes the 500 estimates of Euclidean distance measures based on random draws of 25 patent class listings per firm. While the more accurate distance calculation based on the full sample (E) is 0.329, the mean of the 500 bootstrapped estimates of the 25-draw E measures based on patent classes, which we term E_{25}^{class} , is 0.403. The average across 500 estimates does not converge to the full-sample value. Instead, the mean of the estimates is substantially biased and is 22 percent above the full-sample value. The estimates based on 25 observations are also quite imprecise. The fifth percentile E_{25}^{class} is 0.265, while the 95th percentile E_{25}^{class} is 0.560, meaning that the 90 percent confidence interval on a measure whose full-sample value is 0.329 is 0.295.

The next row summarizes the distribution of the 500 estimates based on random draws of 50 (E_{50}^{class}). The mean falls to 0.374, which is closer to the full-sample value of 0.329, but the estimator remains biased, with a mean that is 12 percent above the full-sample value. Now 90 percent of the estimates fall between 0.262 and 0.503. As the number of patent class listings used in the distance calculation increases to 5000, the mean declines all the way to the full-sample value, while the dispersion of estimates shrinks.

It is not surprising that small-sample proximity measures are imprecise. The source of their bias is less obvious. To understand why the bias varies systematically with sample size, note that the proximity measure E contains squared differences of patent class shares across firms. While errors cancel in linear procedures such as calculating means, errors propagate in nonlinear procedures such as squared differences. Imprecise estimates of the underlying p 's lead to larger (upwardly biased) estimated differences between firms when firms are truly close, and lead to smaller (downwardly biased) estimated differences when firms are actually distant from each other. To see this, suppose that there are two patent classes and one firm's p is $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ while the other firm's p vector is $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$. The true distance E between them is $\sqrt{2}$. Because each element of p must be between 0 and 1 (inclusive), estimates of the p vectors will tend to be closer than $\sqrt{2}$. On the other hand, if both firms' true p vectors are $\begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$, then the true distance is 0; and estimated distances will in general be higher. This indicates that bias will vary for different pairs of firms. When a pair is truly close, imprecise p 's give rise to over-estimates of E , and *vice versa*.

The second and third panels of Table 2 repeat the exercise of the top panel using Euclidean distance measures estimated from the NBER database's (Hall et al, 2001) coarser 36 subcategory and six-category groupings. Similar patterns emerge: bias and imprecision decrease as sample size grows. The degree of bias is smaller when the coarser partition is used. The mean of the Euclidean distance measure based on 25 random draws with the 6 patent categories, which we term E_{25}^6 , is only 11 percent off of the full sample value, while the mean of E_{25}^{class} is 22 percent off its full sample value.

Bias declines with coarser aggregations because grouping related patent classes together in the coarser partitions allows for some averaging of the random variation in sample measures of the shares of patent class listings in particular 3-digit classes. Using our heuristic geographic example, a random sample of 100 Americans can give rise to an accurate measure of the fraction of Americans living east of the Mississippi but only a poor estimate of the fraction of Americans living in Kentucky.

Analogous patterns are repeated in Tables 3 and 4 for proximities measured with vector angles and correlation measures. Here the estimates rise with larger samples toward the full-sample values with larger samples because, for example, higher correlation, like lower distance, denotes greater proximity. In short, the smaller the samples, the more the estimates are both biased and imprecise.

Table 5 reports prior citation overlap measures between Kodak and Canon, based on repeated draws from samples of different sizes. For example, for the 25 observation samples, we draw 25 random patents from Kodak's portfolio and 25 from Canon's. We then calculate the number of patents cited in each of these 25-patent portfolios, as well as the number of patents cited by both Kodak and Canon. We repeat this exercise 500 times, with replacement. Using Kodak (Canon) as the denominator, the mean overlap is 0.0011 (0.0016). The 90 percent confidence intervals are about five times the mean and runs from 0 (0) to 0.0054 (0.0070).

As sample size increases, overlap estimates increase, reaching means of 0.0884 (0.0994) for samples of 5000 patents per firm. That the mean estimates change substantially with sample size reflects an important source of bias in overlap measures. To see how this arises, suppose we are interested in the overlap between Kodak and Canon, in particular in the share of patents cited by Kodak that are also cited by Canon.

If we have all of Canon's patents but only a small sample of Kodak patents, we will produce a noisy but unbiased estimate of the share of patents cited by Kodak that are also cited by Canon. But suppose we have a *sample* of Canon's patents that collectively cite only half of the patents that all of Canon's patents cite. Then even if we have all of Kodak's patents, we will not produce an accurate estimate of share of patents cited by Kodak that are also cited by Canon when examining the overlap between all of Kodak's and some of Canon's patents. Rather, we will produce an estimate of the share of the patents cited by Kodak that were cited by *this subset* of Canon patents. This will produce a smaller share. And indeed, when we compute measures of overlap based on patent samples of varying sizes, the estimates shrink as the Canon samples decline.

The difference between class-based and citation measures arises because "the patents cited by Canon this year," unlike "patent class 396," is a moving target. What constitutes overlap depends on the cited patents in the samples for each firm, and the target grows as sample sizes increase. Patent class 396, by contrast, is a fixed target. That is, the expected value of the share of Kodak patents in class 396 is the same regardless of the Kodak patent sample size, while the expected value of the share of Kodak patents also cited by Canon depends on how many of Canon's patents are used to constitute the set of Canon-cited patents.

When we try to calculate overlap measures between firms with small patent portfolios, we will find small measures of overlap for the mechanical reason that the "target" is small. The question, what share of the patents cited by firm A are also cited by a small firm B, is by nature going to be small. If this is because all of firm B's knowledge is summarized by its citation behavior – but they have little knowledge – then this will be a meaningfully small number. On the other hand, if (small) firm B has lots of knowledge but simply hasn't patented much of it yet

(and therefore hasn't cited many patents yet), then the small overlap measure will understate the extent to which the two firms share sources of knowledge.

2. Proximity among Many Firm Pairs

It is clear that proximity sample statistics provide biased and noisy measures. But how much of a problem are these sampling issues? Tables 2-4 provide one set of answers concerning the Kodak-Canon distance using class-based measures. For the 3-digit patent class based measure of Euclidean distance, when the full sample distance between Canon and Kodak is 0.329, 90 percent of distance estimates based on 25 randomly drawn patent class listings are between 0.265 and 0.560.

Tables 2-4 report results on only the Kodak-Canon distance. But we can perform similar exercises for any pairs of firms with sufficient data that the full-sample distance measures provide reasonably precise estimates. To get a better idea of the extent and magnitude of the issues associated with small sample sizes, we chose the 16 firms in our sample that had over 10,000 patent class listings over the sample period.⁹ This resulted in 120 ($16 \cdot 15/2$) pairs. The full sample distances between these firms varies from 0.069 for Kodak and Fuji Photo to 0.680 for Nikon and Konica using the patent class measure (E^{class}). For each pair we calculate 500 bootstrapped estimates of Euclidean distance using the finest (400 patent class) and coarsest (6 category) partitions, based on 25 and 500 observation samples. That is, we calculate E_{25}^{class} , E_{25}^6 , E_{500}^{class} , and E_{500}^6 .

⁹ The 16 firms are 3M, Canon, Dupont, Fuji Photo, Hitachi, Kodak, Konica, Matsushita, Mitsubishi, NEC, Nikon, Philips, Samsung, Sony, Toshiba, and Xerox.

We illustrate the results of this broader analysis in Figure 1. Here we present the distributions of E_{25}^{class} for each firm pair. That is, for each firm pair we estimate E_{25}^{class} 500 times to characterize the distribution of resulting estimates. In the figure the full-sample distance between a firm pair appears on the horizontal axis (and the 45 degree line shows where this distance appears on the vertical axis). Each vertical line in the diagram, located horizontally at a pair's full-sample distance, describes the range from the 5th to the 95th percentile of the 500 bootstrapped E_{25}^{class} estimates for the firm pair. The circles in the figure represent means.

These ranges are quite large: they are on average larger than the associated full sample values. The ratio of the 95th – 5th percentile range to the full-sample value provides one measure of precision. For the E_{25}^{class} measure, the average ratio across the 120 firm pairs is 1.18. Bias is also systematically related to full-sample distance. Figure 1 also shows us something we could not see when looking at just Kodak-Canon proximity. With a range of full-sample distances among firm pairs, we can see that the distribution of E_{25}^{class} is upwardly biased for pairs whose full-sample values are close, and the distributions are downwardly biased for pairs with more distant full sample values. The average of the 120 full-sample distances is 0.330. The average of the mean E_{25}^{class} estimates is 0.265.

Figure 2 repeats the exercise with the larger-sample E_{500}^{class} measure. As with the Kodak-Canon results, the precision is now much greater, which we see here in the smaller ranges of estimates for each full sample pair distance. The ratio of the 95th-5th percentile gap to the full sample distance is 0.339, indicating much greater precision than the E_{25}^{class} distance estimate allows. Bias is also much smaller with a larger sample. This is evident from the fact that the

means are near the 45 degree line in Figure 2. The averages of these firm-pair means is 0.323, much closer to the average of the full-sample distance measures (0.330).

Figures 3 and 4 repeat the entire exercise of Figures 1 and 2 with E_{25}^6 and E_{500}^6 measures. As with the distance measures based on patent classes, the distance measures based on 6 patent class groupings are imprecise. The average ratio of the gap between the 95th and 5th percentile to the full-sample distance is 1.31 for E_{25}^6 and 0.339 for E_{500}^6 . Although the estimates based on coarser groupings do not improve on the finer class based measures in precision, the E^6 measures are less biased, as we see in the fact that the estimate ranges in Figure 3 are more centered on full sample values than are the ranges in Figure 1. Using the 6-categories, the mean of the full-sample distances among these firms is 0.425. The mean of the E_{25}^6 measures is 0.428, which is not only very close to the full-sample value but also nearly identical to the mean of the E_{500}^6 estimates.

IV. Conclusion

If all firms had, say, over a thousand patents per year, then measures computed from annual patent data would be fairly precise measures of the differences in vectors of patent classes or prior citations. In reality, patent portfolios are much smaller than this. IBM tops most lists of patenting firms. The firm received 3,411 patents in 2001 and 2886 in the previous year.¹⁰ IBM's p vector for 2001 is presumably a precise estimate of its true p . But most firms studied in patent research receive far fewer patents. Indeed, only 437 organizations (including both firms and academic and other research institutions) received 40 or more US patents during 2004.¹¹ Many

¹⁰ See <http://www.uspto.gov/web/offices/com/speeches/02-01.htm>, accessed April 25, 2006.

¹¹ See http://www.uspto.gov/web/offices/ac/ido/oeip/taf/topo_04.htm#PartB, accessed April 25, 2006.

contexts receiving academic scrutiny have sufficiently small patent samples so that precision is an issue. As a result, the concerns raised in this study may be important for research employing patent-based measures.

While our results are in some ways negative, we can also offer some helpful prescriptions for more accurately assessing distances based on patent characteristics. For the imprecision that arises from employing small samples, the solution is to increase sample size, and researchers have at least two easy means for doing this. First, researchers can roughly quadruple their samples by using all, as opposed to just first patent class listings. Quite apart from the larger sample size, this will better reflect the technological footprint of firms' patent portfolios. Second, and possibly in addition, researchers can aggregate their data. Rather than computing measures such as Euclidean distance based on firm-year data, researchers can aggregate data across years into 5 or 10-year periods, as in Stuart & Podolny (1996). This is a useful approach if a researcher has reason to believe the technological footprints of the firms they study have remained stable. The greater precision from larger samples of patents, patent classes, or years can mitigate both bias and precision problems. When it's appropriate to the research question, using coarse partitions can improve the bias of proximity measures.

But researchers should at a minimum be aware of the issues we raise so that they can better interpret their results. When comparing proximities of firms, researchers should be wary of calculations involving firms with small patent portfolios. Such a firm might appear very close or very far from another firm simply because its small patent portfolio provides few draws from its knowledge urn.

References

- Ahuja, G. 2000. Collaboration networks, structural holes and innovation: A longitudinal study. *Admin. Sci. Quart.* **45** 425-455.
- _____, R. Katila. Technological acquisitions and the innovative performance of acquiring firms: A longitudinal study. *Strategic Management J.* **21** 197-220.
- Alcacer, J., M. Gittelman, 2006. Patent citations as a measure of knowledge flows: The influence of examiner citations. *Review of Economics and Statistics*, **88** 774-
- Argyres, N. 1996. Capabilities, technological diversification and divisionalization. *Strategic Management J.* **17** 395-410.
- _____, B.S. Silverman. 2004. R&D, Organization structure, and the development of corporate technological knowledge. *Strategic Management J.* **25**: 929-958.
- Benner, M., M. Tushman, 2002. Process management and technological innovation: Evidence from the photography and paint industries. *Admin. Sci. Quart.* **47**: 676-706.
- Fleming, L., O Sorenson. 2001. Technology as a complex adaptive system: evidence from patent data. *Res. Policy.* **30** 1019-1039.
- _____, _____. 2004. Science as a map in technological search. *Strategic Management J.* **25** 909-928.
- Hall, B.H., A.B. Jaffe, M. Trajtenberg, 2001. "The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools." NBER Working Paper 8498.
- Helfat, C.E. 1994. Evolutionary trajectories in petroleum firm R&D. *Management Sci.*, **40**: 1720-1747.
- Jaffe, A. 1986. Technological opportunity and spillovers of R&D: Evidence from firms' patents, profits, and market value. *Amer. Econom. Rev.* **76** 985-1001.
- _____. 1989 Characterizing the "technological position" of firms, with application to quantifying technological opportunity and research spillovers. *Res. Policy* **18** 87-97.
- Jaffe, A., M. Trajtenberg, M. Fogarty, 2000. Knowledge spillovers and patent citations: Evidence from a survey of inventors. *The American Economic Review*, **90** 215-218
- Jaffe, A., M. Trajtenberg, R. Henderson, 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. *The Quarterly Journal of Economics*, **108**(3) 577-598.

- Miller, D.J. 2004. Firms' technological resources and the performance effects of diversification: A longitudinal study. *Strategic Management J.* **25** 1097-1119.
- Mowery, D.C, J.E. Oxley, B.S. Silverman, 1996. Strategic alliances and inter-firm knowledge transfer. *Strategic Management J.* 17 77-91.
- Piscitello, L. 2004. Corporate diversification, coherence and economic performance. *Industrial and Corporate Change*, **13**(5) 757-787.
- Rosenkopf, L., P. Almeida, 2003. Overcoming local search through alliances and mobility. *Management Sci.*, **49**(6) 751-766.
- _____, A. Nerkar. 2001. Beyond local search: Boundary-spanning, exploration and impact in the optical disc industry. *Strategic Management J.* **22** 287-306.
- Silverman, B. 1999. Technological resources and the direction of corporate diversification: Toward an integration of the Resource-Based View and Transaction Cost Economics. *Management Sci.* **45**(8) 1109-1124.
- Song, J., P. Almeida, G. Wu. 2003. Learning-by-hiring: When is mobility more likely to facilitate interfirm knowledge transfer? *Management Sci.* **49**(4) 351-365.
- Sorenson, J., T.E. Stuart, 2000. Aging, obsolescence, and organizational innovation." *Admin. Sci. Quart.* **45**: 81-112.
- Stuart, T.E., J.M. Podolny. 1996. Local search and the evolution of technological capabilities. *Strategic Management J.* **17** 21-38.

Table 1 – 3 Digit Patent Classes

Patent class	Title
250	Radiant Energy
345	Computer Graphics Processing, Operator Interface Processing, and Selective Visual Display Systems
348	Television (many digital camera patents are included here)
355	Photocopying
356	Optics: Measuring and Testing
359	Optics: Systems (Including Communication) and Elements
360	Dynamic Magnetic Information Storage or Retrieval
382	Image Analysis
386	Television Signal Processing for Dynamic Recording or Reproducing
396	Photography
428	Stock Material or Miscellaneous Articles
430	Radiation Imagery Chemistry: Process, Composition or Product Thereof

Table 2: Kodak/Canon Euclidean Distance Distributions with 500 Draws

N	Type	5 th		
		percentile	mean	95 th percentile
25	3-digit patent classes	0.265	0.403	0.560
50	3-digit patent classes	0.262	0.374	0.503
100	3-digit patent classes	0.267	0.349	0.437
500	3-digit patent classes	0.290	0.333	0.373
1000	3-digit patent classes	0.302	0.331	0.362
5000	3-digit patent classes	0.315	0.329	0.342
Full-sample	3-digit patent classes		0.329	
25	36 sub-categories	0.256	0.410	0.592
50	36 sub-categories	0.238	0.377	0.522
100	36 sub-categories	0.274	0.360	0.458
500	36 sub-categories	0.297	0.342	0.387
1000	36 sub-categories	0.308	0.340	0.369
5000	36 sub-categories	0.326	0.340	0.355
Full-sample	36 sub-categories		0.340	
25	6 categories	0.265	0.454	0.656
50	6 categories	0.295	0.438	0.583
100	6 categories	0.317	0.419	0.530
500	6 categories	0.362	0.412	0.462
1000	6 categories	0.372	0.410	0.442
5000	6 categories	0.394	0.410	0.426
Full-sample	6 categories		0.410	

Table 3: Kodak/Canon Angle Distributions with 500 Draws

N	Type	5 th		
		percentile	mean	95 th percentile
25	3-digit patent classes	0.228	0.536	0.774
50	3-digit patent classes	0.345	0.591	0.785
100	3-digit patent classes	0.458	0.639	0.795
500	3-digit patent classes	0.614	0.688	0.762
1000	3-digit patent classes	0.640	0.694	0.743
5000	3-digit patent classes	0.673	0.700	0.723
Full-sample	3-digit patent classes		0.700	
25	36 sub-categories	0.350	0.625	0.843
50	36 sub-categories	0.457	0.672	0.853
100	36 sub-categories	0.553	0.699	0.828
500	36 sub-categories	0.658	0.725	0.790
1000	36 sub-categories	0.680	0.728	0.772
5000	36 sub-categories	0.707	0.729	0.749
Full-sample	36 sub-categories		0.728	
25	6 categories	0.407	0.652	0.884
50	6 categories	0.493	0.675	0.851
100	6 categories	0.557	0.692	0.817
500	6 categories	0.637	0.701	0.758
1000	6 categories	0.661	0.702	0.749
5000	6 categories	0.683	0.702	0.721
Full-sample	6 categories		0.703	

Table 4: Kodak/Canon Correlation Distributions with 500 Draws

N	Type	5 th		
		percentile	mean	95 th percentile
25	3-digit patent classes	-0.156	0.301	0.698
50	3-digit patent classes	0.115	0.469	0.737
100	3-digit patent classes	0.346	0.572	0.775
500	3-digit patent classes	0.587	0.668	0.748
1000	3-digit patent classes	0.621	0.680	0.732
5000	3-digit patent classes	0.663	0.692	0.715
Full-sample	3-digit patent classes		0.694	
25	36 sub-categories	-0.025	0.399	0.765
50	36 sub-categories	0.226	0.527	0.799
100	36 sub-categories	0.398	0.592	0.783
500	36 sub-categories	0.571	0.656	0.740
1000	36 sub-categories	0.604	0.666	0.723
5000	36 sub-categories	0.655	0.680	0.705
Full-sample	36 sub-categories		0.686	
25	6 categories	-0.673	-0.076	0.652
50	6 categories	-0.568	-0.061	0.518
100	6 categories	-0.459	-0.054	0.426
500	6 categories	-0.234	0.135	0.353
1000	6 categories	-0.148	0.205	0.342
5000	6 categories	0.198	0.238	0.280
Full-sample	6 categories		0.238	

Table 5: Overlap between Kodak and Canon with Different Sized Samples

Number of patents per firm	mean		5 th percentile		95 th percentile	
	Kodak	Canon	Kodak	Canon	Kodak	Canon
25	0.0011	0.0016	0.0000	0.0000	0.0054	0.0070
50	0.0019	0.0025	0.0000	0.0000	0.0061	0.0083
100	0.0039	0.0054	0.0009	0.0012	0.0087	0.0119
500	0.0171	0.0225	0.0126	0.0171	0.0214	0.0284
1000	0.0306	0.0387	0.0263	0.0329	0.0349	0.0444
5000	0.0884	0.0994	0.0851	0.0952	0.0924	0.1037

Note: Each row in this table characterizes the distribution of overlap estimates from 500 draws with different numbers of patents per firm (25, 50, and so on).

Figure 1

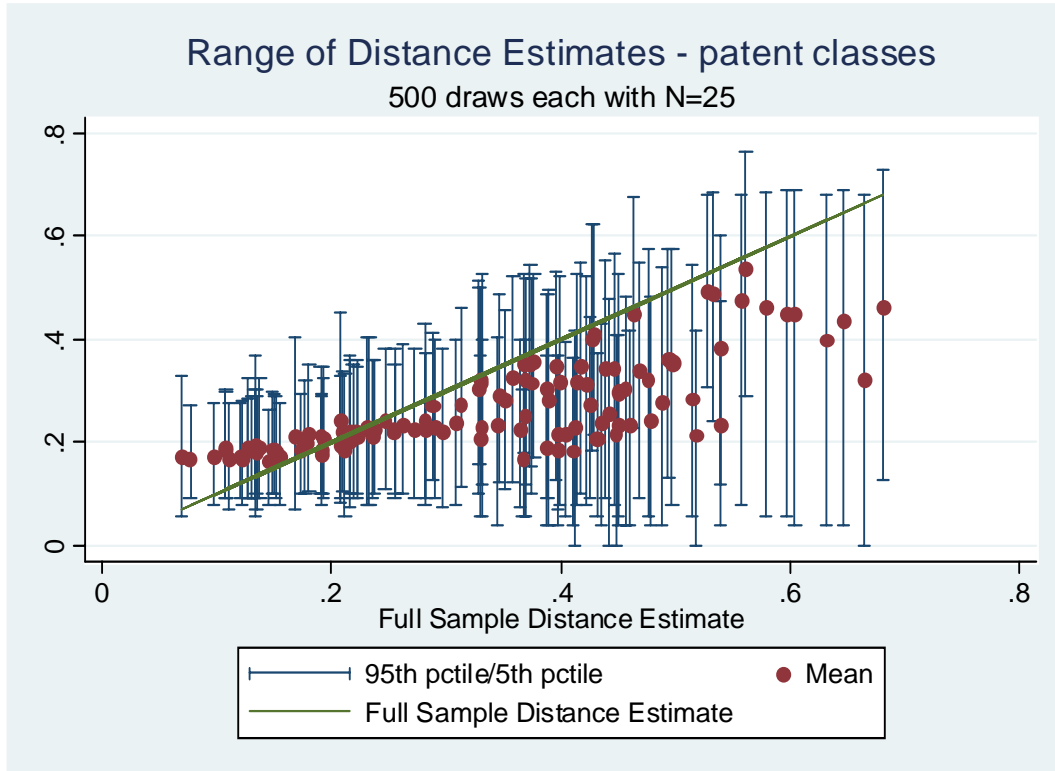


Figure 2

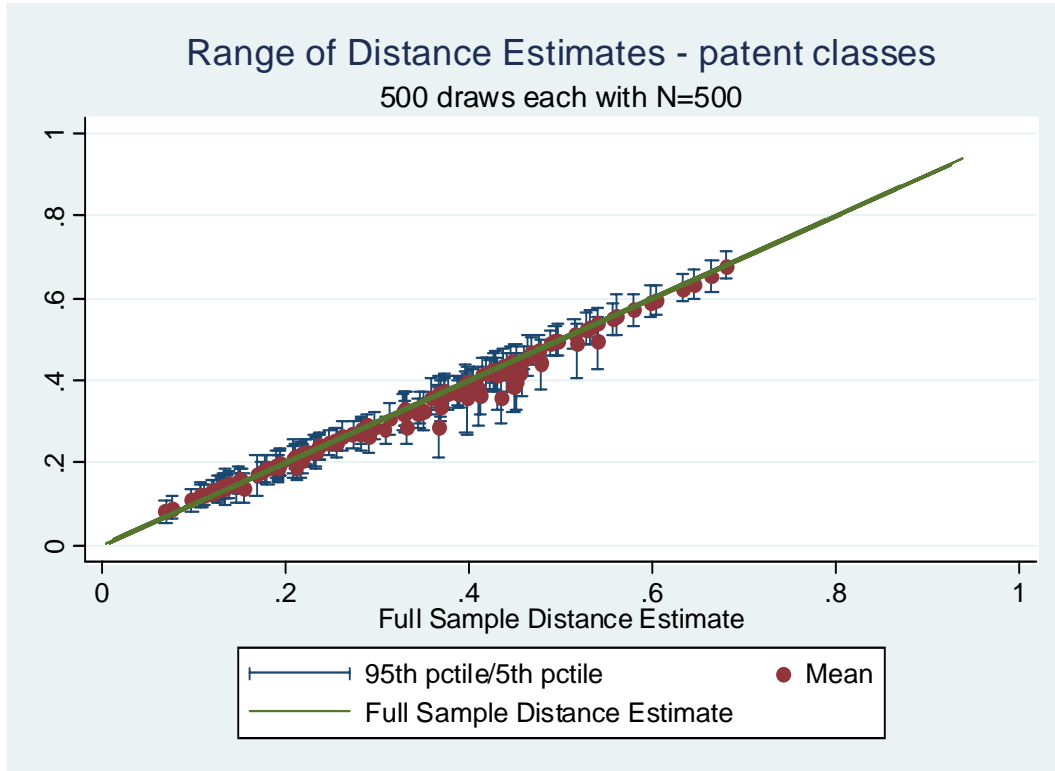


Figure 3

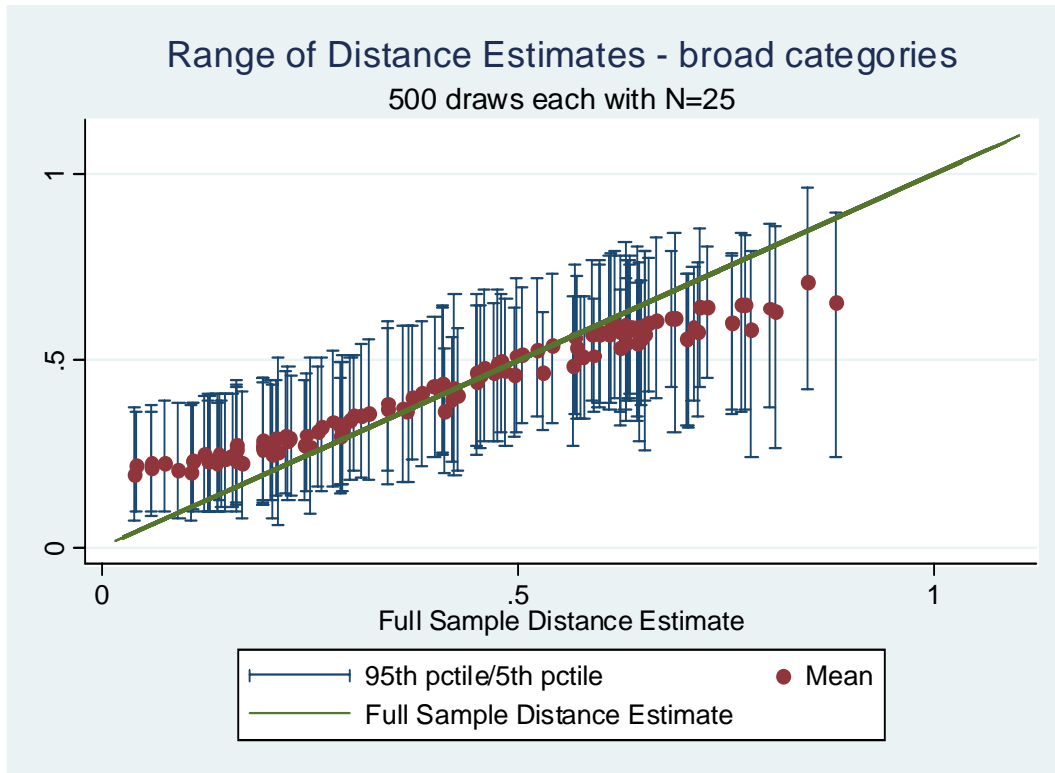
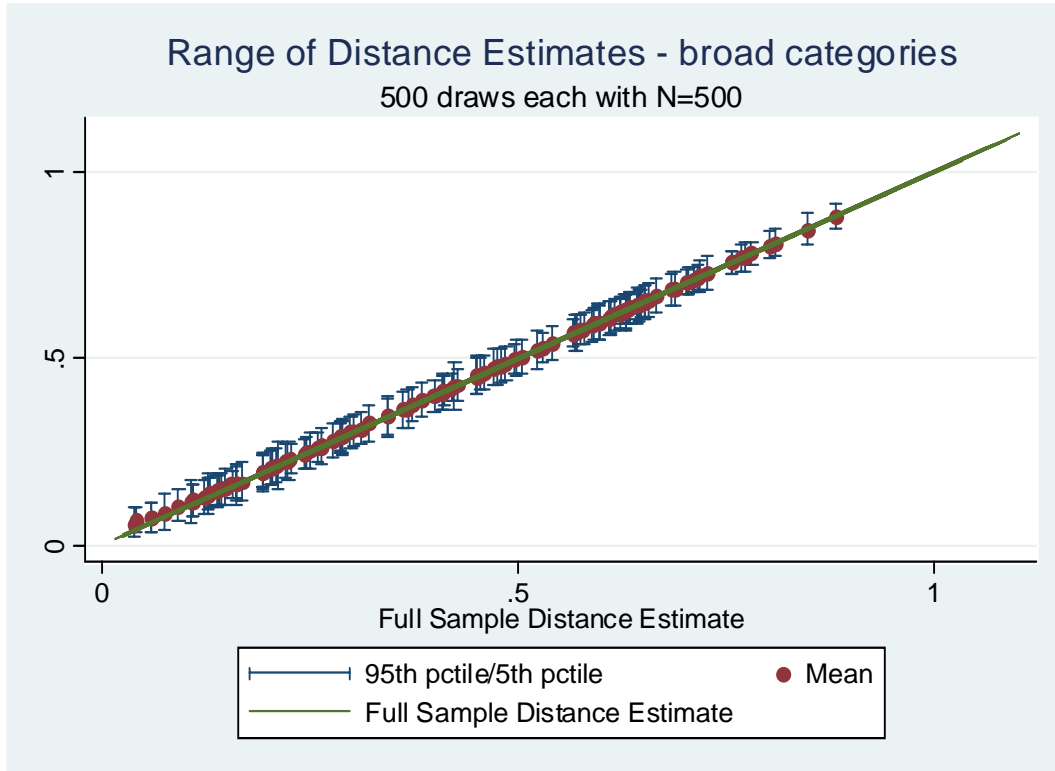


Figure 4



APPENDIX 1– USPTO Rules for determining the Original Patent Class From the Examiner’s Handbook and Handbook of Classification

III. DETERMINATION OF A CLASS FOR ORIGINAL CLASSIFICATION OR ASSIGNMENT FOR EXAMINATION

The process for selecting which claimed invention will represent the original classification or assignment designation is effected by considering, in turn, the factors listed below:

- Selection of the most comprehensive claim;
- Selection among categories of subject matter when claims are equally comprehensive, or when the most comprehensive claim cannot be determined;
- Selection among superiority of types of subject matter;
- Selection among generic classes and species classes thereunder;
- Selection among classes in “related subject” listing

1. Selection of the Most Comprehensive Claim

This is illustrated by comparing the following two claims (which not only differ in comprehensiveness, but are also directed to different categories of subject matter) each of which define subject matter properly classified in a different class:

Claim 1

A laminated sheet comprising two panels of aluminum bonded with an intermediate layer of a binder, said binder comprising an elastic, self-vulcanizing rubber-like cement, the adjacent faces of the panels being roughened in a cross-hatched pattern to facilitate adhesion of the cement, two opposite edges of the sheet being notched with corresponding, interfitting dovetail cutout portions to facilitate securing the edges one to the other.

Claim 2

A process comprising scoring the faces of two aluminum panels in a cross-hatched pattern, applying a binder to the scored faces, pressing the coated faces together to secure the panels and forming a sandwich and then bending the sandwich and securing the opposite edges to each other to form a tube.

Explanation

Claim 1 encompasses a product comprising a laminated sheet. Claim 2 encompasses a process of making such a sheet, but includes the additional steps of bending the sheet and securing the edges to form a tube. Process Claim 2 thus contains a greater extent of subject matter than does Claim 1 and is therefore more comprehensive. Assuming that separate classes provided for the subject matter in Claim 1 and 2, respectively, a patent with these two claims should be placed as an original in the class providing for the subject matter of Claim 2.

2. Selection Among Categories of Subject Matter

When a patent document includes separate claims to two or more different categories of subject matter and none of the claims is more comprehensive than the other(s) or if greater comprehensiveness cannot be determined, the Original is classified in the class providing for the claimed category that appears highest in the following list

- 1) Process (of using product 2, e.g. using a fuel or radio transmitter)
- 2) Product (of manufacture, e.g. a fuel or radio transmitter)
- 3) Process (of making product 2)
- 4) Apparatus (to perform 3 or to make 2, e.g. machine, tool, etc.)
- 5) Materials (used in 3 to make 2)

For example, when considering claims to a radio transmitter (category 2) and to a process of manufacturing the same (category 3), the claim to the transmitter would control class assignment. Similarly, a claim to a process of using the transmitter (category 1) would control a claim to the transmitter or process of making it.

3. Selection Among Superiority of Types of Subject Matter

When placement of the Original cannot be determined from considerations of a) comprehensiveness or b) categories of subject matter, placement is next determined by considering the highest category below that provides for claimed subject matter.

- 1) Relating to maintenance or preservation of life
- 2) Chemical subject matter
- 3) Electrical subject matter
- 4) Mechanical subject matter
 - a. Dynamic
 - b. Static

4. Selection Among Classes in “Related Subject” Listing (Last Resort Only)

The number of a class generally has no significance insofar as superiority of one class relative to another. The class number is merely an arbitrary mark of identification. Nor is the class listing the “Classes Arranged in Alphabetical Order” in the Manual of Classification an order of superiority. The title of a class is an accident of language and varies from one language to another.

However, a theoretical organization of the applied sciences into three major areas is published in the front of the Manual of Classification, Section I, titled “Patent Office Classes arranged by Related Subjects.” Within each of the areas the classes have been listed in a hierarchy suggesting an order of superiority.

Where the other bases for selection discussed above cannot be applied, a controlling claim is selected according to this listing. The controlling claim is identified as the one having subject matter provided for by the class that appears highest in such listings.

5. Exceptions:

- a) Where special agreements between groups are in effect, such as for high temperature superconductivity applications and for certain biotechnology areas, these agreements override all other considerations.
- b) Where the historical placement of patents having particular claimed disclosure has been contrary to written definitions and notes, the historical placement overrides all other considerations, except the special agreements mentioned above, and controls placement of like subject matter until corrective reclassification is effected.
- c) Classification definitions (particularly the search notes and lines with other classes) must be read for possible exceptions to the selection procedures discussed in 1-5 above, inasmuch as disclosures in a given areas of technology may have required deviation from these procedures. Any deviation will be mentioned, and explained, in a modern class definition.

V. ORIGINALS VS. CROSS-REFERENCES

The original classification (OR) is a mandatory classification selected from among all the mandatory classifications as being the highest in the schedule (superiority) of the class containing the controlling claim. The remaining mandatory classifications are designated as mandatory cross-references (XR).

A. Discretionary Classifications

Each Technology Center generally has specific information that varies from class to class that they would like to see classified in the USPC, even when the information may not constitute invention information in a patent document. Because this information is useful for other, non-invention search purposes the cross-references assigned to documents based on this invention information must be designated as “discretionary” cross-references.

In creating the USPC system, several techniques have been used to limit the need for discretionary cross referencing. These consist of 1) proper positioning of subclasses in a class schedule and 2) search notes

A search note in the class definition of each of related classes generally precludes the need for discretionary cross-referencing. However, as in all discretionary cross-referencing, there may be a time when it is desirable to cross-reference even if there is cross noting.