A Machine Learning Analysis of the Geographic Localization of Knowledge Flows

Joel Blit
University of Waterloo

Mikko Packalen
University of Waterloo

June 1, 2019

Abstract

We re-examine whether knowledge flows are localized by applying machine learning to patent texts to map the position of each patent in a vector space representation of the technology space. We first apply this new technology space representation to show that technology classification-based localization analyses are likely to yield biased results as we observe geographical agglomeration within patent classes and subclasses, thus contributing to the debate begun by Jaffe, Trajtenberg, and Henderson (1993) and Thompson and Fox-Kean (2005). We then apply the new technology space representation to re-examine knowledge flow localization. Our results continue to find support for localization. Thus, the bias present in earlier localization analyses was not the sole driver of the results.

Keywords: knowledge spillovers; diffusion; localization; patents; text analysis.

JEL Classification: O33.

1. Introduction

The localization of knowledge spillovers shapes the way that economic activity is organized. Going at least as far back as Marshall (1890), economists have argued that, along with specialized suppliers and labor market pooling, the localization of knowledge spillovers is one of the principal forces driving the agglomeration of technological activity, and therefore the existence of cities. Places like Silicon Valley thrive in spite of their high congestion costs because high-tech firms derive benefits from being near other high-tech firms. The localization of knowledge spillovers also underpins the rationale behind the public support of R&D investments.

Yet showing that knowledge diffusion is indeed at least partially localized has proven a challenge. In their seminal paper, Jaffe, Trajtenberg, and Henderson (1993) (henceforth JTH) use patent citations to measure knowledge flows, which up until that point had been considered to be unobservable. The premise behind this approach is that if an inventor cites a previous innovation, the inventor must have learned of the innovation.

JTH test for the localization of knowledge flows by determining whether citing patents are disproportionately in the same geographic location as the cited patent. Because patents tend to primarily cite other patents in their same narrow technological area, appropriately controlling for the geographic agglomeration of technological activity is crucially important. Even in the absence of localized knowledge diffusion, it would not be surprising to observe Silicon Valley patents (many of which might be computer related) disproportionately citing other Silicon Valley patents, and Detroit patents (many of which might be automobile related) disproportionately citing other Detroit patents.

JTH develop a technology classification-based case-control matching methodology precisely to address this concern. In particular, they control for the geographic agglomeration of technological activity by choosing for each citing patent a control patent with the same application year and technology class. Whether knowledge flows are localized is then determined by comparing the frequency with which the citing patent is in the same location as the cited patent against the frequency with which the control patent is in the same location as the cited patent.

However, as Thompson and Fox-Kean (2005) (henceforth TFK) have pointed out, the technology class may be too coarse a level at which to control because each class encompasses a broad set of different technologies (there are roughly 400 different technology classes). For example, since year 2000 more than half of all U.S. patents pertaining to computer processors (technology class 712) were generated in either San Francisco or Austin. But each region specializes in somewhat different aspects of computer processors; while San Francisco disproportionately specializes in processor architecture (subclasses 1 to 43), Austin disproportionately specializes in instruction dependency checking and monitoring (subclasses 216 to 219). To the extent that patents tend to cite other patents that are most similar to themselves, we would expect San Francisco architecture patents to disproportionately cite other San Francisco architecture patents, even in the absence of knowledge flow localization. Thus, a comparison of the location match rate for the citing and cited patents against the location match rate for the control and cited patents could yield spurious evidence of localization. Moreover, these potential problems apply also to technology subclass-based analyses of localization. This drawback of classification-based delineation of the technology space has formed a key barrier to advances in this area.

In this paper, we pursue an alternative way for identifying which patents are similar to one another and use this approach to conduct a re-analysis of the localization of knowledge flows. We first apply a machine learning algorithm to the text of patents to calculate the position of each patent in technology space. We then show that the proximity of any pair of patents in this technology space is correlated with other measures of similarity, such as whether the pair of patents share a class, subclass, cited patent, assignee, inventor, or location. We next examine heterogeneity among patents in the same technology class and subclass and show that even among patent pairs in the same class/subclass, those that are more proximate in technology space are also more likely to be in the same location. Such heterogeneity confirms that localization analyses that rely on patent classifications are likely to yield evidence of localization even when none exists. Hence, it is important to examine whether the localization hypothesis is still supported when control patents are chosen based on their technology space proximity to the citing patent. Our re-examination of knowledge flow localization continues to find support for the localization hypothesis, indicating that that bias inherent in earlier technology-classification based analyses was not the sole driver of the results.

Following JTH, many subsequent analyses have sought to apply and extend their approach, and our analysis contributes to this important strand of literature. TFK propose that the geographic distribution of technological activity could be better accounted for if control patents matched the citing patent at the much narrower level of the technology subclass (of which there are over 100,000) and find that this approach reduces the estimated degree of localization. Of course, as confirmed in our analysis, even at the subclass level the control patents do not adequately address the geographic distribution of technological activity. The subclass-based approach also suffers from an additional challenge: in many cases the subclasses are so narrowly defined that no admissible control patent exists. Such patents would necessarily be dropped from the analysis, and this could accentuate biases rather than alleviate them, as Henderson et al. (2005) have suggested. Our approach, by contrast, yields a control patent for every patent and is thus not subject to this selection bias. Moreover, our analysis yields an indication of how close the control patent is to the citing patent, allowing us to distinguish between cases where a close control patent is available and cases where one is not.

A second concern raised by TFK is that patents generally have multiple technology classes and focusing on only the primary technology class may result in inadequate controls. Two computer processor patents may be very different if one is applied in the automobile sector and the other in biotech. A further advantage of our machine learning-based approach to delineating the technology space is that we allow the position of each patent to vary continuously along 300 dimensions. Hence the technology space position of each patent can reflect its characteristics on multiple dimensions.

Murata et al. (2014) extend the JTH and TFK analyses in two ways. First, instead of comparing a dichotomous matching variable where the appropriate geographic unit of analysis is unclear, Murata et al. (2014) compare two spatial distance distributions: the geographic distance between citing and cited patent locations and the geographic distance between control and cited patent locations. They conclude that knowledge flows are localized in about 70% (30%) of technology fields when control patents are selected based on technology classes (subclasses). To maintain comparability with JTH, in our analysis we employ matching rate tests instead of the distance density tests employed by Murata et al. (2014).

A second contribution of Murata et al. (2014) is to calculate how different control patents that do not share a technology subclass with the citing patent could be from the citing patent (in

terms of unobserved heterogeneity), in order for the results to still imply localized knowledge flows in the majority of the technology classes. They conclude that the technological exposure of a citing patent and any technology subclass-based control patent can be even 25 times higher than the technological exposure of a technology class-based control patent to the cited patent (here, technological exposure is calculated relative to the cited patent). While similar in spirit, our analysis of heterogeneity within patent classes is distinct from this analysis, for our approach provides a direct measure of how much heterogeneity there is within patent classes.

As with most earlier studies of the localization of knowledge flows, we rely on patent citations. In addition to those already cited, such localization studies include Almeida & Kogut (1999), Verspagen and Schoenmakers (2004), Singh & Marx (2013), and Blit (2017). The use of citations to measure knowledge flows, however, is not uncontroversial. On the one hand, inventor surveys have shown that patent citations do capture knowledge flows, although not without noise (e.g. Jaffe et al. 2000; Duguet and MacGarvie 2005). On the other hand, the main purpose of citations is to limit the scope of the patent rather than identify idea inputs (e.g. Jaffe et al. 1993; Strumsky et al. 2012) and many citations are added by the patent examiner (Thompson 2005; Alcacer and Gittelman 2006). Consequently, citations may simply indicate patents that are similar to the citing invention rather than prior work that the patent builds upon.[1] While this is a worthwhile debate, we bypass it in this paper, and merely acknowledge that it presents an important caveat in the interpretation of our results and that of other papers that use patent citations to measure knowledge flow.

As discussed, we differ from most of the prior literature on knowledge spillovers in that we identify control patents based on patent texts, whereas JTH and almost all other papers in this literature identify control patents based on the patent classification system. One salient exception is Thompson (2005) who uses examiner-added citations as controls. We recreate his analysis as one benchmark for our results. To the best of our knowledge, only Arts et al. (2018) have employed a text-matching-based approach to study the localization of knowledge flows. But while they employ a bag of words approach to determine similar patents, we use a neural

[1] Other important papers contributing to this debate include Roach and Cohen (2013), Moser et al. (2017), and Arora et al. (2018). Arora et al. (2018) compare the localization of citations to patents with earlier priority dates with that of citations to patents with later priority dates (which they argue cannot represent knowledge flows), and find that since both sets of citations display the same degree of localization, citations are unlikely to be a useful tool for examining the localization of knowledge flows. This finding presents an important potential caveat interpreting our and other citation-based analyses as evidence of knowledge flows.

network algorithm to calculate the vector space positions of documents. This neural network algorithm simultaneously places every patent document and every word that appears in patents in a vector space. A further point of departure with Arts et al. (2018) is that in addition to replicating the JTH localization analysis with the new control patents, we also examine heterogeneity within technology classes and subclasses. We also show that when vector space coordinates are used to select control patents for case-control matching rate analyses, several additional exclusion restrictions must be employed relative to the exclusion restrictions employed in JTH. Both our analysis and Arts et al. (2018) complement other analyses that have advanced the use of textual information in patent analyses (e.g. Younge and Kuhn 2016; Packalen and Bhattacharya 2012, 2015).

The balance of the paper is organized as follows. The next section presents the data and the machine learning approach that we use to obtain a vector space representation of the technology space. Section 3 shows that proximity in this new technology space representation is linked with closeness in other dimensions such as whether a pair of patents shares a common technology class. Section 4 uses the new technology space representation to examine within patent class and patent subclass heterogeneity. Section 5 utilizes the new technology space representation to re-examine whether knowledge flows are localized, and section 6 concludes.

2. Data and the Machine Learning Approach

2.1 Patent Data

Our raw data were retrieved from the USPTO's PatentsView website and consist of all patents granted by the U.S. Patent and Trademark Office (USPTO) during 1975-2017. For each patent, we utilize the following fields: application year, location of each inventor, assignee, main technology class and subclass, citations to other U.S. patents, and the patent text (title, abstract, claims, summary, and descriptions of the drawings). We exclude from our sample the 14.2% of patents that, based on text comparisons, are likely to be continuation patents (Section 2.1.3 discusses this in further detail). The resulting sample consists of 5,058,820 patents.

2.1.1 Sample Periods and Sample Sizes

While we generate our technology space representation based on all patents granted during 1975-2017, our localization analysis focuses on citing patents with application years 2001-2009 and cited patents with application years 1975-2009. We limit the analysis to citing patents with application years 2001 and later because data on whether a citation was added by the inventor or an examiner is only available from that year forward. We limit to citing patents with application years no later than 2009 to allow bechmarking with the JTH analysis.[2]

Our sample of 19,595,315 citing-cited patent pairs has 1,493,726 unique citing patents and 2,598,828 unique cited patents. This sample is used in our analyses in sections 3 and 4. For our localization analysis (section 5) we further restrict the sample to U.S. cited patents (as in JTH), which allows us to conduct state and MSA-level localization analyses. This U.S. sample has 13,169,144 citing-cited patent pairs,

It is worth noting that in each of these two samples the same citing patent appears multiple times if it cites multiple cited patents. We refer to each such instance (corresponding to a cited-citing pair) as a *citing patent instance*.

2.1.2 Patent Location

We assign each patent to a unique location – country, state and Metropolitan Statistical Area (MSA) – based on the reported address of the inventors. When a patent has inventors from multiple distinct locations, the patent location is determined based on a majority rule (the country, state, or MSA with the most inventors listed on the patent residing there). When two or more locations are tied for the most inventors, we randomly assign the patent to one of the tied locations.

2.1.3 Patent Text

[2] The USPC patent classifications used by JTH were discontinued for patents granted after 2014. Therefore, due to the delay between patent application and grant, patents with application years beyond 2009 often do not have a USPC patent class.

We generate the text string representing the content of each patent by concatenating the following text fields in order: title, abstract, claims, summary, and descriptions of the drawings. We remove from the resulting text string very common non-technical words (such as "there" and "within") and words whose use is more likely to be linked to differences in patent jargon rather than differences in the technical nature of inventions (such words include "embodiment" and "apparatus"). We also apply a word stemming algorithm that changes words that are similar in terms of their meaning but have a different word ending. From the stemmed text strings, we then eliminate words that are the same as the preceding word and words that are mentioned fewer than 10 times in the corpus of patents. In the final text editing step, we select the first 1,000 remaining words to represent the textual content of each patent.[3]

For 14.2% of the patents in our patent data the first 50 characters of the constructed representative text string is identical to that of another patent with either an earlier application year or the same application year. When such "duplicates" are found, we exclude from the analysis all but the one with the earliest application year. In most such cases that we examined by hand, the duplicate patents were continuation patents.[4]

2.2 The Machine Learning Approach for Constructing the Vector Space Representation of the Technology Space

To estimate a vector space position for each patent document in the patent corpus we apply a recently developed machine learning approach. The algorithm that we employ is the distributed bag of words version of the Doc2Vec algorithm (Le and Mikolov 2014). We utilize the Gensim implementation of this algorithm (Rehurek and Petr 2010).

The main purpose of the Doc2Vec algorithm is to produce a low-dimensional vector space representation for each document in a given corpus in such a way that the vector space

[3] For example, the edited text string for patent number 9,785,496 starts with "semiconductor die chip wafer measur enabl fill cell wafer multipl enabl open process manufactur semiconductor wafer chip die util electr measur fill cell structur configur target expos varieti short leakag excess resist failur mode process evalu design experi multipl enabl fill cell variant target failur mode multipl enabl detect open" (the complete string has 1,000 words).

[4] The inclusion of continuation patents as controls for the initial patent is not in the spirit of the analysis since they are effectively the same patent. If they were included in the localization analysis, the power of the case-control localization test would be lower because the identifying variation of the case-control test comes from the cases when the citing and control patents are from different locations.

positions of similar documents are near one another. The algorithm calculates a vector space representation also for each word that appears in the corpus based on which documents contain the word. Any pair of words that appear in similar contexts are placed near one another in the vector space. Though in our analysis we utilize only the vector space positions of the patent documents, the algorithm simultaneously calculates a vector space position also for each word, and thus the vector space positions of documents are in part based on the estimated vector space positions of words.

While a corpus of texts may have millions of documents and tens of thousands or more distinct words, the Doc2Vec algorithm embeds each document in a low-dimensional vector space. In implementing the algorithm, the key choice parameter is the dimension of the vector space representation. Following other applications of the algorithm, we use it to estimate 300-dimensional document vectors. The algorithm thus reduces the number of dimensions in which the documents differ to 300. This dimension reduction is conducted so that words that are distinct but likely have a related meaning (as they appear in similar contexts) have a similar influence on the vector space position of a document.

The algorithm is a neural network with three layers. The first layer (the input layer) has one node representing each document in the corpus. The second layer (the hidden layer) has one node for each dimension of the vector space representation. Thus, in our application the number of nodes in the hidden layer is 300. The third layer (the output layer) has one node representing each distinct word in the corpus.

The matrix of all estimated document vectors forms the link between the input layer and the middle layer. Given a corpus with D documents, and a selected dimension of the vector space representation of 300, the size of this "input vector" matrix is $D \times 300$. The matrix of all estimated word vectors in turn forms the link between the middle layer and the output layer. Given a corpus with V distinct words in it, the size of this "output vector" matrix is then $300 \times V$.

The input and output vector matrices that contain the document and word vectors form the parameters of the model that are estimated by the algorithm. The estimation is done by iterating over the corpus multiple times. We estimate these parameters using 20 iterations over the corpus. During each iteration, the parameters are adjusted based on one document at a time. For each document, the algorithm calculates a conditional probability for all the words in the corpus that appear in the document and also for some words that do not appear in that particular

document but do appear in other documents in the corpus. The corresponding document vectors and word vectors are then adjusted so that the conditional probability is high for all those words that actually appear in the document and low for those words that do not appear in the document.

Following other applications that have employed the Doc2Vec algorithm, we measure the proximity of any two patents by the cosine similarity of their vector space representations. Cosine similarity of two vectors is calculated as the cosine of the angle between unit-normalized versions of the vectors. Cosine similarity is thus 0 for vectors that are orthogonal to one another, 1 for vectors that have the exact same direction, and -1 for vectors that are diametrically opposed to one another. Larger values of this proximity measure indicate greater similarity.

In the next section, we present some summary statistics of our proximity variable (cosine distance) and show that this variable correlates in expected ways with different patent pair characteristics.

3. Distribution and Correlates of Proximity in the Vector Space

3.1 Distribution of Proximity in the Vector Space

We pair up each of the 19,595,315 citing patent instances in our full sample with a randomly chosen patent that has the same application year as the citing patent. This yields 19,595,315 citing-random patent pairs. We then compute the proximity (cosine distance) of each pair. Figure 1 shows the distribution (having mean 0.100 and standard deviation 0.058) of these proximities. The figure shows that the technology space locations of patents is neither degenerate (as the distribution is not degenerate) nor randomly dispersed throughout the vector space (as the distribution is not centered on zero).

For each citing patent instance, we also determine which patent with the same application year is closest to it in the technology space, and then calculate the proximity (cosine distance) between the citing-closest patent pair. Figure 2 shows this distribution. There is considerable variation across citing patents in how near is their closest patent. The ability to determine not only the closest patent, but also *how close* it is, is an important feature of our approach, and one

that we will utilize in our localization analysis to show evidence of localization even in cases where close controls are available.

3.2 Correlates of Proximity in the Vector Space

We now examine the extent to which more proximate patents share characteristics that we might expect to be shared by patents that are similar. This analysis demonstrates that the vector space position of patents captures meaningful aspects of the technology space.

One characteristic that more proximate patents in the vector space would be expected to share more often is the technology class that USPTO examiners have assigned to each patent. Figure 3 presents the relationship between a patent pair's proximity (horizontal axis) and the likelihood that the two patents are in the same technology class (vertical axis). This figure is constructed based on the 19,595,315 citing patent instance-random patent pairs described above. While patent pairs that are far from one another are almost never in the same technology class, the fraction of patent pairs that are in the same technology class rises as the proximity of the patent pair increases. When the proximity of a patent pair in the vector space is above 0.4, the patents have the same technology class more than half the time. The vector space representation of the technology space thus has some overlap with the manually curated technology classification representation of the technology space; the machine learning algorithm is able to capture some of the characteristics of the technology space that human curators of the technology classification system have deemed important.

Figure 4 shows the corresponding analysis in relation to the likelihood that the patents share a technology subclass. The results show that patents that are more proximate in the vector space are more likely to share also a technology subclass. Together, Figures 3 and 4 confirm that the vector space positions of patents reflect technological characteristics of patents rather than pure random variations.

Figures 5 and 6 present the same relationship between patent pair proximity and likelihood of being in the same class/subclass where the sample is each citing patent instance paired with the closest patent having the same application year (as in Figure 2 above). Again, the frequency at which a patent pair is in the same technology class or technology subclass increases

with the patents' proximity in the vector space. Visual inspection of pairs with proximities above about .8, suggests that one of the patents is often a (non-identical) continuation of the other, and it is thus not surprising that such patents are frequently in the same patent class. However, for most citing-closest patent pairs, the likelihood of being in the same technology class is well below 100%. The two approaches to delineating the technology space – the machine learning approach and the manually curated technology classification – thus result in two related but distinct representations of the technology space. While we are not in a position to evaluate which approach yields a better representation of the technology space, our results below (Section 4) show that there is considerable systematic heterogeneity among patents within the same technology class and subclass.

To more formally evaluate the technology space mapping, we employ a regression analysis that examines whether patent pair proximity is correlated with a number of different characteristics that the pair of patents may share. In particular, for the 19,595,315 citing patent instance-random patent pairs described above, we regress the proximity variable on indicator variables that capture whether the two patents have a common technology class, technology subclass, assignee, inventor, backward citation, country, state, and MSA, and whether the two patents are linked by a cross citation between them.

Table 1 presents the results. As shown in column 1, patents that share a patent class have on average 0.056 higher proximity (56% higher proximity than the mean of 0.10, or roughly one standard deviation (which is 0.058)). If in addition they share the same subclass, their proximity is on average a further 0.042 higher, so that patents pairs having the same class and subclass are 1.7 standard deviations more proximate. The strongest relationships are with patent pairs having an inventor in common (0.105 higher proximity or 1.8 standard deviations) and having a cross-citation between the pair (0.108 higher proximity or 1.9 standard deviations). Patent pairs with a common assignee and making a common backward citation also have higher proximity (0.037 and 0.065, respectively).

With respect to geographical location, the results suggest that controlling for the other patent pair characteristics, patent pairs from the same location are on average 0.012 more proximate (0.2 standard deviations), which breaks down as 0.007 more proximate if they are in the same country, an additional 0.001 if they are in the same state, and a further 0.004 if they are also in the same MSA. Conditional on country and MSA, whether patent pairs are from the same

state does not seem to be strongly related to proximity and is only significant at the 10% level. This is perhaps due to most states having only one or two cities with significant innovative activity and suggests that an MSA-level analysis of knowledge flow localization might be more informative than a state-level analysis.

Columns 2 and 3 present the results when citing patent and random patent fixed effects, respectively, are added to the specification. The results are virtually unchanged.

Overall, the results in Table 1 not only show that our proximity measure is correlated with patent characteristics in an expected way, they also show that even after controlling for whether two patents are in the same technology class and in the same technology subclass, the proximity of a patent pair is correlated with the pair being in the same geographical location. This suggests that technology class and technology subclass control matching approaches may not be enough to adequately account for the agglomeration of technological activity in localization analyses. We examine this more closely in the section that follows.

4. Heterogeneity within Technology Classes and Subclasses

The principal challenge in examining knowledge flow localization stems from the dual facts that: (1) similar patents tend to be co-located, and (2) patents tend to cite patents that are similar to themselves. When both of these conditions hold, patents will disproportionately cite other patents in the same location, even if knowledge flows are not localized. JTH put forward the technology class-based approach as an attempt to eliminate this potential source of bias.

The extent to which technology class-based matching is successful in eliminating this potential source of bias depends on whether there is systematic heterogeneity within technology classes. TFK provided indirect evidence of significant heterogeneity within technology classes by showing that JTH results change if one selects control patents based on technology subclasses instead. By contrast, in this section we demonstrate directly that there is technological heterogeneity within technology classes (and subclasses). Specifically, we show that even when one compares patents in the same technology class (or in the same technology subclass), patents that are more proximate in the technology space are (1) more likely to be co-located, and (2) more likely to cite one another.

We begin by examining the extent to which more proximate patents are more likely to be in the same MSA, even when one limits comparisons to patents in the same technology class. We use the same sample of 19,595,315 citing patent instances as in the previous section. We first determine for each citing patent instance the distance to all other patents that have the same application year (as before) and main technology class, and order these by proximity. For each citing patent, we then select the closest and furthest such patents, and also patents that correspond to the 99th, 98th, 97th, 96th, 95th, 94th, 93rd, 92nd, 91st, 90th, 80th, 70th, 60th, 50th, 40th, 30th, 20th, and 10th percentiles in terms of their vector space proximity to the citing patent instance. For example, when there are 400 patents with the same technology class and the same application year as a given citing patent, the fifth closest patent is the patent on the 99th percentile in terms of proximity and the 393rd closest patent is the patent on the 2nd percentile. Each citing patent instance is thus paired with 20 potential control patents that all share the same application year and technology class, but differ in their proximity to the citing patent.

We next calculate, separately for each percentile, what fraction of citing patent and potential control pairs are in the same MSA. We refer to this frequency as the *Citing-Control Location Match Rate*. Figure 7 reports the *Citing-Control Location Match Rate* as a function of the technology space proximity percentile of the potential control patent. The horizontal axis thus reports the relative proximity between the citing patent and a potential control patent, and the vertical axis reports how often the potential control patent is from the same MSA as the citing patent.

As shown in Figure 7, we observe significant variation in the *Citing-Control Location Match Rate* across the proximity percentiles, suggesting geographical clustering within technology classes. While 42% of the closest potential control patents are in the same MSA as the citing patent and 15% of potential control patents on the 90th percentile are in the same MSA as the citing patent, only 10% of the median potential control patents (50th percentile) are in the same MSA. Hence, even within technology classes, different geographical regions tend to specialize in different technological areas. Thus, the computer processor example mentioned in the introduction seems to be part of a systematic pattern.

Figure 8 presents the corresponding analysis at the technology subclass level. That is, now all potential control patents for a citing patent must have the same application year and technology subclass as the citing patent. The pattern is similar but, perhaps not surprisingly, less

pronounced. For the closest potential control patents, 18% are in the same MSA as the citing patent, while the same fraction is 11% for the potential control patents on the 90th percentile, and only 7% for the median potential control patents. The geographic clustering of technology is evident even within technology subclasses.

However, this presence of agglomeration within technology classes and subclasses is only problematic for case-control analyses of localization if in addition patents are more likely to cite other patents that are more similar to themselves (even after controlling for technology class or subclass). To examine this possibility and the extent to which different control patent selection approaches resolve it, we compare the distribution of the vector space proximity between cited and citing patent pairs against four other distributions: (1) proximity between cited patents and random control patents with the same application year as the citing patent, (2) proximity between cited patents and random control patents with the same technology class and application year as the citing patent (the JTH control patent), (3) proximity between cited patents and random control patents with the same technology subclass and application year as the citing patent (the TFK control patent), and (4) proximity between cited patents and patents with the same application year as the citing patent that are most proximate to the citing patent (our machine learning-based control patent).

Figure 9 presents our results. It is evident from the difference between the citing-cited proximity and random control-cited proximity distributions that patents tend to cite patents that are similar to themselves (as citing-cited pairs are much closer than random control-cited pairs). Moreover, the difference between the citing-cited proximity distribution and technology class control-cited proximity distribution is large, indicating that even among patents in the same technology class, patents tend to cite patents that are more similar to themselves. The subclass control-cited proximity distribution is closer to that of the citing-cited distribution, but there remains a large difference between them. The only proximity distribution that in any way resembles that of the citing-cited is the closest control-cited proximity distribution. This, coupled with our earlier finding of agglomeration within technology classes and subclasses, highlights the fact that the machine learning approach that we utilize to identify the closest control has the potential to yield much more effective controls than even the technology subclass-based approach.

We further examine within technology class and subclass heterogeneity in a regression framework similar to that of section 3.2. The only distinction is that, whereas in that earlier analysis we paired each citing patent instance with a random patent with only the same application year as the citing patent, we now pair each citing patent instance with a random patent with the same application year and the same technology class as the citing patent. We regress the vector space proximity of these patent pairs on the same patent pair characteristics as before to determine whether even among patent pairs in the same class, proximity is correlated with variables such as whether there is a citation link between the patents and whether they are in the same location. We also conduct the corresponding subclass-level analysis, pairing each citing patent instance with a random patent with the same application year and technology subclass as the citing patent.

Table 2 shows the results. Columns 1-3 show that for patent pairs with the same application year and technology class, being more proximate in the vector space is correlated with having the same subclass, the same assignee, the same inventor, a cross-citation, a common citation, the same country, the same state, and the same MSA. Columns 4-6 in turn show that the same holds for patent pairs that are from the same technology subclass. Consistent with the results shown in Figures 7-9, there is systematic heterogeneity within both technology classes and subclasses.

This systematic heterogeneity within technology classes and subclasses raises the concern of possible bias in localization analyses that employ technology classes or subclasses to select control patents. Even when the null hypothesis of no localization holds, such analyses are likely to find that the cited and citing patents are from the same location more often than the cited and control patents. Hence, it is important that we re-examine whether knowledge flows are localized using our machine learning based approach to select control patents.

5. Re-Examination of Localization of Knowledge Flows

5.1 Localization Methodology

Our analysis of localization of knowledge flows utilizes the case-control methodology of JTH. As in JTH, we begin with a set of cited patents and find the set of patents that cite these. As discussed above in section 2.1.1, we restrict the set of cited patents to U.S. patents to obtain a sample of 13,169,144 citing-cited patent pairs. As in JTH, we further exclude those citing-cited pairs for which the assignees of the two patents are the same, as such patents do not reflect external knowledge flows. This yields a sample of 12,169,144 citing-cited patent pairs.

The most critical step in the case-control approach is the selection of the control patents. Our re-analysis of localization differs from JTH in how the control patent is selected. For each cited-citing pair, JTH identify the control patent based on the application year, technology class, and grant date of the citing patent. Specifically, in their approach, the control patent must (1) share the same 3-digit technology class and application year as the citing patent and (2) not cite the cited patent. If multiple such patents exist, the control is chosen from among these as the one whose grant date is closest to that of the citing patent.

Following JTH, we also consider as admissible control patents only those patents that share the same application year as the citing patent and do not cite the cited patent. From this set of admissible control patents, we select as the control patent the patent whose machine learning determined position in the vector space representation of the technology space is closest to the vector space position of the citing patent. This generates a set of cited-citing-control patent triplets that are used to calculate two location match rates: (1) the fraction of citing and cited patents that are in the same location, and (2) the fraction of control and cited patents that are in the same location. We refer to the former fraction as the *Citing-Cited Location Match Rate* and the latter fraction as the *ML Control-Cited Location Match Rate*. Whether knowledge flows are localized is revealed by a comparison of these two matching rates. Under the null hypothesis of no localization in knowledge flows, the two matching rates are equal, provided that the control patent properly accounts for technological agglomeration. By contrast, under the alternative hypothesis of localized knowledge flows, the location match rate is greater for the citing-cited pair than it is for the control-cited pair.[5]

[5] The identifying variation of this matching rate test comes from those cases for which the citing and control patents are in different locations. Hence, when geographical agglomeration of inventive activity is extreme, so that the control patent is always in the same location as the citing patent, the matching rate test has no statistical power.

The above approach for choosing the control patent based on the citing and control patents' technology space coordinates is our baseline specification. While this is the most straightforward recreation of the JTH methodology in the technology space context, our preferred specification makes three additional restrictions on the set of admissible control patents. First, to ensure that continuation patents of the citing patent are not being selected as the control, we exclude from the set of admissible control patents those patents that have at least one inventor in common with the citing patent. Second, we exclude from the set of admissible control patents those patents that have the same assignee as the cited patent. Since citing-cited patent pairs with a common assignee were removed from the analysis from the outset (so as to be able to interpret the results as evidence of localization of knowledge flows external to the firm), this additional exclusion ensures that the two location matching rates retain similar distributions under the null hypothesis of no localization in knowledge flow. Third, we exclude from the set of admissible control patents any patents that have the same assignee as the citing patent. Not only will such patents almost always be in the same location as the citing patent (reducing the power of the test), but even in cases where they are in different locations choosing same-firm controls is problematic due to the presence of intra-firm knowledge flows. For example, as Blit (2017) shows, the citation (knowledge flow) could occur not because the citing patent is in the same location as the cited patent (i.e. due to knowledge flow localization), but because the citing firm has a satellite R&D centre in the location of the cited patent (i.e. due to intra-firm knowledge flows). In the extreme, if knowledge flows perfectly within the firm, any patent of the firm is just as likely to be citing the cited patent regardless of location and if we allowed controls to have the same assignee as the citing patent, we may not find knowledge localization even when it in fact exists.

In summary, in this preferred specification for selecting control patents, patents in the admissible control patent set satisfy the following five conditions: (1) they have the same application year as the citing patent, (2) they do not cite the cited patent, (3) they do not have any inventors in common with the citing patent, (4) they do not have the same assignee as the cited patent, and (5) they do not have the same assignee as the citing patent. By contrast, in the baseline specification, we only employ restrictions (1) and (2).

Having identified the admissible control patent set, in this preferred specification we then again select as the control the patent whose technology space position is closest to the technology space position of the citing patent.

To benchmark our results, we also calculate the location matching rates when the control patent is chosen based on the technology class of the citing patent (the JTH approach) and when the control patent is chosen based on the technology subclass of the citing patent (the TFK approach). In addition, we calculate location match rates also for the case where the control patent is chosen randomly from among all patents with the same application year as the citing patent; this approach corresponds to the case when one does not try to control for geographic agglomeration of technological activity. Finally, we also estimate the degree of localization by comparing the fraction of inventor-added citing patents that are in the same location as the cited patent with the frequency of examiner-added citing patents that are in the same location as the cited patent; this approach follows the methodology put forward in Thompson (2006).

5.2 Localization Results

We begin with the results averaged across all citing-cited patent pairs. Further below, we then take advantage of the fact that our approach allows us to distinguish between those citing-cited patent pairs for which a close control patent is found versus those citing-cited patent pairs for which even the closest control patent is far from the citing patent in technology space.

The first two columns of Table 3 show the results for our baseline and preferred specifications, respectively. Across the three panels, the *Citing-Cited Location Match Rate* and the *Control-Cited Location Match Rate* as well as their difference and statistical significance are shown for three levels of location: Country (top panel), State (middle panel), and MSA (bottom panel). For both the baseline and preferred specifications (columns 1 and 2, respectively) the results suggest some degree of knowledge flow localization at all three levels of location.

To benchmark our results, columns 3 to 5 show the corresponding results for the TFK approach in which the control patent is chosen to match the technology subclass of the citing patent (column 3), for the JTH approach in which the control patent is chosen to match the technology class of the citing patent (column 4), and for the approach in which the control patent

is chosen randomly from patents that have the same application year as the citing patent (column 5). As expected, the *Citing-Cited Location Match Rate* is relatively stable across the columns, with the small differences being due to slightly different samples due to cases where no appropriate control was available. The sample is most different (about 10% smaller) for the TFK (subclass) controls in column 3, and consistent with the aforementioned sample selection concern, the *Citing-Cited Location Match Rate* is somewhat larger across all three panels.

As expected, across all three panels, the *Control-Cited Location Match Rate* drops significantly from left to right, suggesting that our proximity controls do best at controlling for agglomeration, followed by the TFK (subclass) controls, and finally the JTH (class) controls, which do better than not controlling at all.

Column 6 presents the results for the Thompson (2006) approach that compares the location match rate of inventor-added citations to that of examiner-added citations. Here again we find evidence of localization since the inventor-added citations are significantly more likely to be in the same location as the patent they are citing, than are the examiner-added citations. Not surprisingly, the reported %*Citing matching* (which more accurately for this column is the inventor-added citations location match rate) is higher than the same row in other columns, since for the other columns the reported number is the fraction of all citing patents (both inventor and examiner added) that are in the same location as the cited patent. That is, for each panel, the first row of the other columns is a mix of the match rates for the inventor-added and examiner-added citations (the first two rows) in column 6.

If indeed only inventor-added citations represent actual knowledge flows and the data on the source of the citation is accurate, we may want to perform the entire analysis solely on the set of inventor-added citations. Table 4 presents the corresponding results for this case when the analysis is restricted to cited-citing patent pairs where the citation was added by the inventor. As would be expected since these represent actual knowledge flows, the *Citing-Cited Location Match Rate* and the computed location matching rate differences are now somewhat larger. However, qualitatively the results remain similar to those reported in Table 3. Since inventor-added citations are more likely to represent actual knowledge flows, the remainder of the analysis focuses on the subsample of citations that were made by inventors.

In both Tables 3 and 4, in all cases the difference between the *Citing-Cited Location Match Rate* and the *Control-Cited Location Match Rate* is positive and significant, providing

support for the localization hypothesis, at least for those approaches for which one believes that the selected control patents properly account for agglomeration. However, there is cause to be concerned that even for our proximity controls, this result could potentially be driven by the subset of citing-cited pairs where a close control was not available. Fortunately, our approach allows us to not just identify the nearest control patent but to also compute how close that nearest control patent is to the citing patent. This enables us to perform separate location matching rate comparisons for those cases where the nearest control patent is relatively close to the citing patent (and hence appropriate controls for the agglomeration of technological activity) and for those cases where the nearest control patent is relatively far from the citing patents.

Figure 10 reports the results of such an analysis for our baseline specification and focusing only on inventor-added citations to U.S. patents. The figure shows both *Citing-Cited Location Match Rate* and the *ML Control-Cited Location Match Rate* as a function of the technology space proximity between the citing patent and the control patent. As a benchmark, the figure shows also the location match rate between the cited patent and the JTH control patent as well as the location match rate between the cited patent and a control patent chosen randomly from patents with the same application year. Because the technology space proximity is calculated using the cosine distance metric, larger values of the proximity measure indicate patent pairs that are closer to one another in the technology space.

The results shown in Figure 10 suggest that when the control patent better accounts for technological agglomeration (in that the control patent is close to the citing patent in the technology space) the hypothesis of localized knowledge flows is no longer supported, as there is no discernible difference between the *Citing-Cited Location Match Rate* and the *ML Control-Cited Location Match Rate* for values of proximity above 0.6. However, the significance of this finding is limited for two reasons. First, particularly for proximity values of 0.7 and greater, many of the control patents are slightly modified continuation patents of the citing patent, and thus almost by definition they will be in the same location as the citing patent. In addition, many control patents will also have the same assignee as the citing patent and thus will also mostly be in the same location as the citing patent. It is thus not surprising that the two matching rates would be practically the same in this range; the result is merely due to the low power of the test when the citing and control patent tend to be in the same location often. Second, while we drop citing-cited pairs for which the assignee is the same in the citing and cited patents, in our

baseline specification we do not exclude from the set of admissible controls patents that have the same assignee as the cited patent. This likely generates upward bias in the *ML Control-Cited Location Match Rate* relative to the *Citing-Cited Location Match Rate*.

Before proceeding to the corresponding figure for our preferred specification, we note that Figure 10 also shows that except in the relatively rare cases where the most proximate control patent is far to the citing patent (proximity values less than 0.4), the location match rate between the cited patent and a technology class-based control patent (JTH approach) is lower than that for our proximity-based control. Lowest of all, is the location match rate between the cited patent and the random control. This pattern is consistent with our machine learning determined control patents generally being more effectively controls than the JTH control patents which only partially account for technological agglomeration.

Figure 11 shows the matching rates as a function of the citing-control proximity for the preferred specification. The preferred specification addresses these aforementioned issues by excluding from the set of admissible control patents three types of patents: (1) any patent that has an inventor in common with the citing patent, (2) any patent that has the same assignee as the cited patent, and (3) any patent that has the same assignee as the citing patent. For low and moderate proximities (proximities below 0.6) the result is clear: *Citing-Cited Location Match Rate* is higher than the *ML Control-Cited Location Match Rate*, indicating that there is localization in knowledge flow.

For high proximities (proximities above 0.6) it is difficult to see much difference between the two matching rates. This is due in part to the matching rates becoming noisy due to the relatively small number of citing instances for which an admissible control with that high proximity was available (the number of citing instances with proximity of 0.6 is 4953, with proximity 0.7 is 963, and with proximity 0.8 is 245). In addition, because cases with high citing-control proximities represent a tiny fraction of the 7,889,133 total observations we should be concerned about selection. A closer look at the ML control-citing patent pairs for proximities of .6 or higher reveals that more than 60% of control patents are in the same MSA as their associated citing patent. Visual inspection of the most proximate citing and control patent assignee names reveals that more than half of citing and controls have the same effective assignee (in spite of patents with the same assignee as the citing patent not being admissible

controls in our preferred specification) due to imperfect disambiguation of assignee names (misspellings, short forms of names, etc.) and subsidiary-parent company relationships.

For robustness, since one cannot rely on the disambiguated assignee id alone, we repeat the analysis, imposing an additional restriction that control patents cannot have more than a 100 character text fragment overlap with the citing patent. This restriction takes advantage of the fact that assignees often use similar legal jargon in their patents (especially in the background section). Figure 12 shows the matching rates as a function of the citing-control proximity for the preferred specification with this additional restriction. The *Citing-Cited Location Match Rate* is now clearly higher than the *ML Control-Cited Location Match Rate* even when the control patent is close to the citing patent in technology space.

Tables 5 and 6 present a similar analysis to Figures 11 and 12, respectively, but for the country, state, and MSA level of analysis. Given that the cases where the control patent is closest to the citing patent most effectively account for the technological agglomeration of technological activity, we present the *Citing-Cited Location Match Rate* and the *ML Control-Cited Location Match Rate* for different percentiles of cases with the most proximate citing and control patents.

Table 5 presents the results for our preferred specification and for only the inventor-added citations to U.S. patents. As would be expected, we generally find that the more proximate the controls (the higher the citing-control proximity subsample), the higher fraction of controls that match the location of the cited patent. The difference between *Citing-Cited Location Match Rate* and the *ML Control-Cited Location Match Rate* decreases as we analyse higher percentiles but remains significant in all cases but the 0.1% of the sample with most proximate controls for the country and state level analyses. While it is tempting to conclude that there is no evidence for localization at the level of the country and the state, since for the cases with the best controls to account for technological agglomeration we can no longer reject the null that there is no localization, a more conservative interpretation is that these results are likely due to selection issues and/or the aforementioned loss of power due to imperfect assignee disambiguation and subsidiary-parent company relationships. The large jumps in the *Citing-Cited Location Match Rate* across subsamples indeed suggests that the smaller samples may not be directly comparable to the overall sample. Moreover, when we introduce the additional restriction that control patents that have more than a 100 character text fragment overlap with the citing patent are not admissible controls, our analysis strongly rejects the null hypothesis of no localization even for

the sample of 0.1% most proximate control patents, and at each of the country, state, and MSA level (see Table 6).

The balance of evidence thus leads us to reject the null hypothesis that there no is localization in knowledge flow. Though this finding comes with three important qualifications. First, when utilizing the case-control approach for studying knowledge flow localization, the results are always dependent on the assumption that the chosen control patents properly account for agglomeration. We, of course, cannot be certain that our approach fully controls for this. Our analysis has, however, moved the literature forward in this respect by demonstrating that the substantive results obtained in earlier, technology classification-based analyses, are robust to using our alternative (and arguably better), machine learning-based delineation of the technology space for the analysis of localization.

Second, while our results are very clear when averaged across all citing patent instances, as the average results indicate that there is localization, our results are slightly murkier for the cases where a very close control patent for the citing patent instances can be found, particularly for the country and state level analyses. We do not, however, believe these results of no localization to be particularly robust, as it is a very small subsample and in addition we find strong evidence of localization when we impose the additional restriction of no 100-character string overlap between the control and citing patents. Clearly, however, further research is warranted. Perhaps more advanced machine learning tools would be able to find even closer control patents for a broader set of patents which would help bring certainty to this important question.

Third, as mentioned in the introduction, much doubt remains about the suitability of using patent citations to study knowledge flows (e.g. Arora et al 2018) and we sidestepped this important debate. Instead, our focus was on constructing the new machine learning delineated vector space representation of the technology space and on applying it to re-examine some of the highest profile contributions in the economics of innovation literature.

6. Conclusion

Localization of knowledge spillovers are a key potential driver of agglomeration and remain a key focus in innovation policy. Many cities and regions strive to create innovation clusters where inventors have the opportunity to learn from each other, with the hopes of enhancing the region's economic prosperity. Given their potential importance, knowledge spillovers have deservedly received a lot of attention from innovation scholars.

Previous empirical studies have generally supported the conclusion that knowledge flows are localized to a significant degree. However, concerns over the methodology used in these analyses has cast a shadow over their substantive implications and practical relevance. The fundamental assumption in the case-control methodology employed in these analyses is that one is able to identify which patents are comparable to one another in order to properly account for geographic agglomeration of different types of inventive activities. However, the existing technology classification-based approach for delineating the technology space has been seen as inadequate in this respect. For if there is technological heterogeneity within technology classes and subclasses and this heterogeneity is correlated with the geographical locus of innovation, the case-control approach will yield biased estimates of localization.

In this paper, we have explored a new approach for delineating the technology spaced. We first showed that the position of each patent in the technology space can be determined using the recently developed Doc2Vec machine learning algorithm. We applied this algorithm to the text of more than 5 million patents to compute the position of each patent in a 300-dimensional vector space. We showed that the proximity of a pair of patents in this vector space is correlated with other measures of patent similarity, including having a common technology class, subclass, cited patent, inventor, assignee and location. Having shown that the machine learning approach captures meaningful aspects of the technology space, we utilized the new technology space representation to evaluate the extent of technological heterogeneity within patent classes and patent subclasses. Our findings showed that technology is geographically agglomerated even within patent classes and patent subclasses. An important implication of such systematic heterogeneity within patent classes and patent subclasses is that technology class and technology subclass-based analyses of knowledge flows can yield spurious evidence of localization.

Our re-examination of the knowledge flow localization hypothesis utilized the vector space representation of the technology space to identify a control patent for each citing patent in order to account for agglomeration. The results continue to provide support for the localization

hypothesis. Thus, the bias inherent in technology class- and subclass-based localization analyses was not the sole driver of the results. Our machine learning based approach places these findings on a firmer footing and thereby contributes to the important academic and policy discussion on knowledge flow localization. We hope that our exploration of the new machine learning approach sparks further research that utilizes this and related tools in the study of knowledge flow localization and the economics of innovation more generally.

References

Alcacer, J., and M. Gittelman, 2006, "Patent Citations as a Measure of Knowledge Flows: the Influence of Examiner Citations," *The Review of Economics and Statistics* **88**(4) 774-779.

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

Arora, A., S. Belenzon, and H. Lee. 2018, "Reversed Citations and the Localization of Knowledge Spillovers," *Journal of Economic Geography* **18**(3) 495-521.

Arts, S., Cassiman, B. and J. C. Gomes, 2018, "Text Matching to Measure Patent Similarity," *Strategic Management Journal* **39**(1) 62-84.

Blit, J. 2017, "Learning Remotely: R&D Satellites, Intra-Firm Linkages, and Knowledge Sourcing," *Journal of Economics and Management Strategy* **26**(4) 757-781.

Duguet, E. and M. MacGarvie, 2005, "How Well Do Patent Citations Measure Flows of Technology? Evidence from French Innovation Surveys," *Economics of Innovation and New Technology* **14**(5) 375-393.

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

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

Jaffe, A. B., M. Trajtenberg and R. Henderson, 1993, "Geographic localization of knowledge spillovers as evidenced by patent citations," *Quarterly Journal of Economics* **108**(3) 577-598.

Jaffe, A.B., Trajtenberg, M. and M. S. Fogarty, 2000, "The Meaning of Patent Citations: Report on the NBER/Case-Western Reserve Survey of Patentees," *American Economic Review* **90**(2) 215-218.

Jaffe, A.B. and M. Trajtenberg, 2002, *Patents, Citations, and Innovations: A Window on the Knowledge Economy*. MIT Press, Cambridge.

Le, Q. and T. Mikolov, 2014, "Distributed Representations of Sentences and Documents," *Proceedings of the 31st International Conference on Machine Learning*.

Marshall, A, 1890, *Principles of economics*. MacMillan, London.

Moser, P., Ohmstedt, J. and P. W. Rohde, 2017, "Patent Citations-An Analysis of Quality Differences and Practices in Hybrid Corn," *Management Science* **64**(4) 1926-1940

Murata, Y., R. Nakajima, R. Okamoto and R. Tamura, 2014, "Localized Knowledge Spillovers and Patent Citations: A Distance-Based Approach," *Review of Economics and Statistics* **96**(5) 967-985.

Packalen, M. and J. Bhattacharya, 2012, "Words in Patents: Research Inputs and the Value of Innovativeness in Invention," National Bureau of Economic Research Working Paper No. 18494.

Packalen, M. and J. Bhattacharya, 2015, "Cities and Ideas," National Bureau of Economic Research Working Paper No 20921.

Rehurek R. and S. Petr, 2010, Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.*

Roach, M and W. M. Cohen, 2013, "Lens or Prism? Patent Citations as a Measure of Knowledge Flows from Public Research," *Management Science* **59**(2) 504-525.

Singh, J. and M. Marx, 2013, "Geographic Constraints on Knowledge Spillovers: Political Borders vs. Spatial Proximity," *Management Science* **59**(9) 2056-2078.

Strumsky, D., Lobo, J. and S. van der Leeuw, 2012, "Using Patent Technology Codes to Study Technological Change," *Economics of Innovation and New Technology* **21**(3) 267-286.

Thompson, P., 2006, "Patent Citations and the Geography of Knowledge Spillovers: Evidence from Inventor- and Examiner-Added Citations," *Review of Economics and Statistics* **88**(2) 383-389.

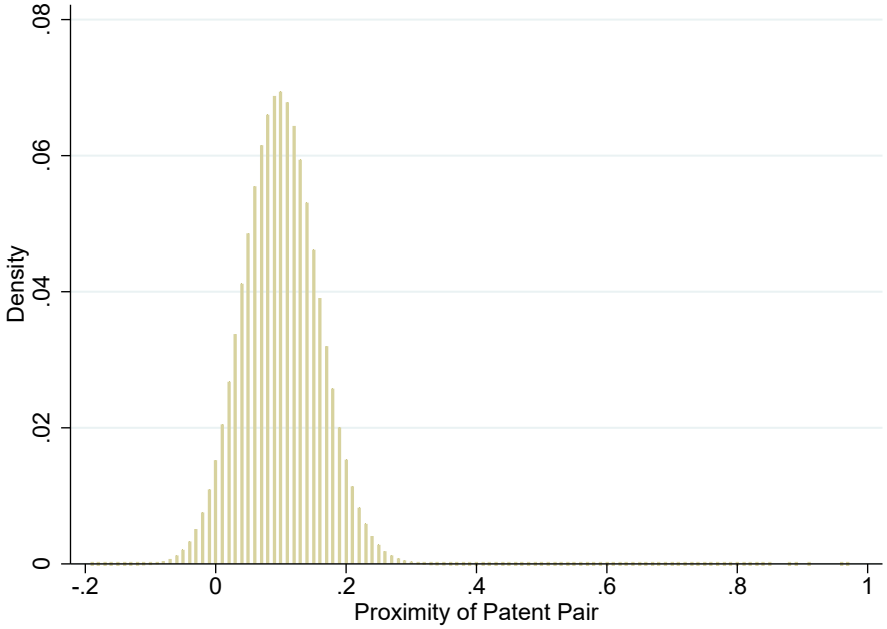
Thompson, P. and M. Fox-Kean, 2005, "Patent citations and the geography of knowledge spillovers: a reassessment," *American Economic Review* **95**(1) 450-460.

Verspagen, B. and W. Schoenmakers, 2004, "The Spatial Dimension of Patenting by Multinational Firms in Europe," *Journal of Economic Geography* **4**(1) 23-42.

Younge, K. A. and J. M. Kuhn, 2016, "Patent to Patent Similarity: A Vector Space Model," Manuscript.

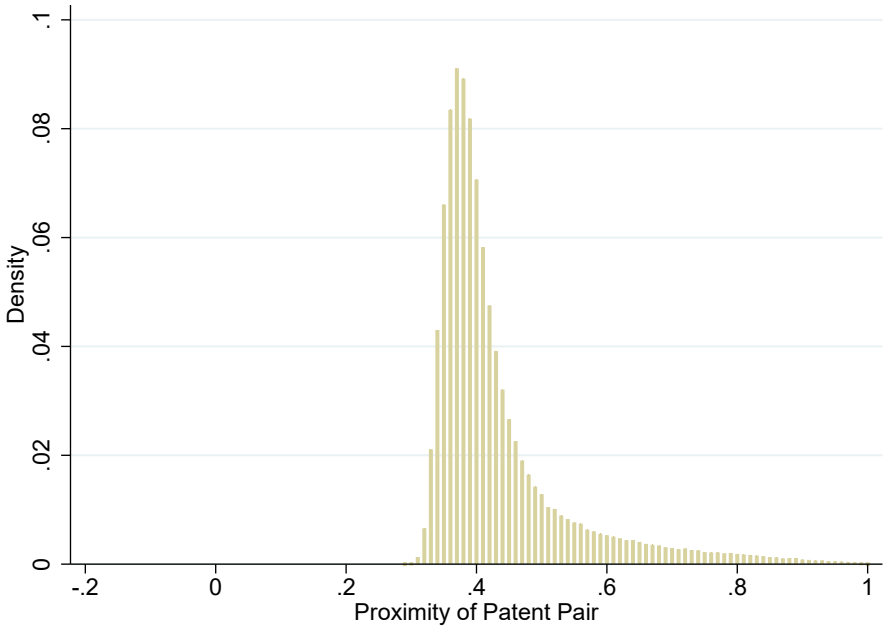
Tables and Figures

Figure 1: Distribution of proximity for citing and random patent pairs.



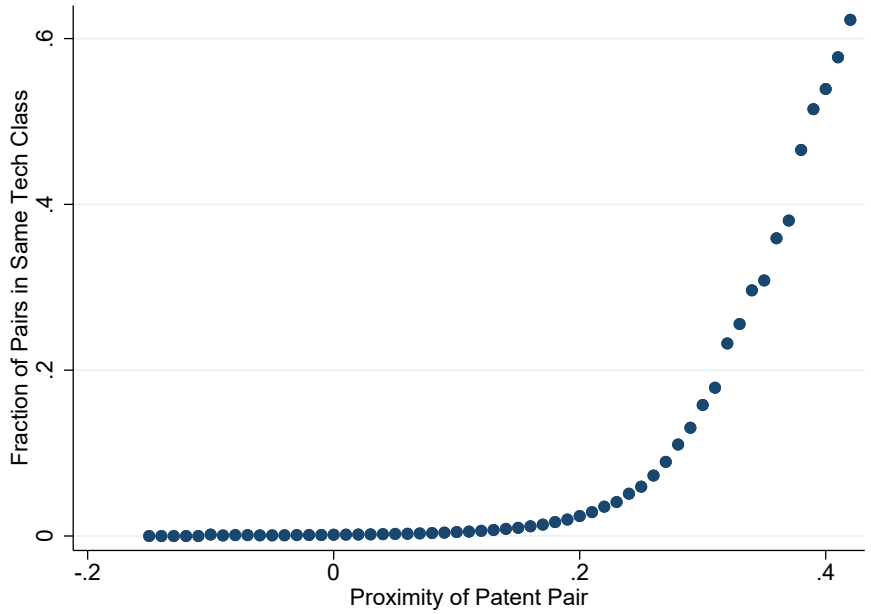
Notes: Each of the 19,595,315 citing patent instances in our sample is paired with a random patent with the same application year as the citing patent. Proximity is measured as the cosine distance (rounded to the nearest 0.01) between the citing-random patent pairs.

Figure 2: Distribution of proximity for citing and closest patent pairs.



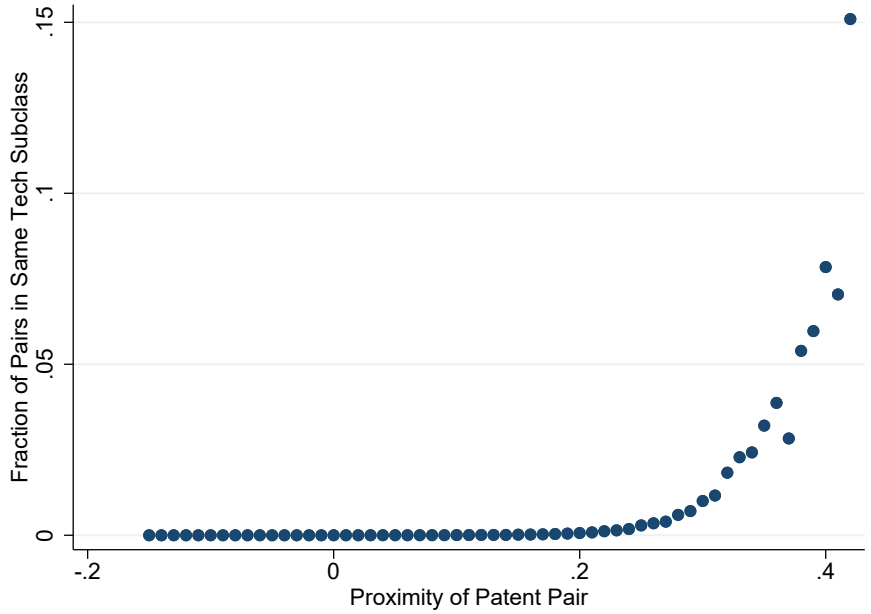
Notes: Each of the 19,595,315 citing patent instances in our sample is paired with the closest patent having the same application year as the citing patent. Proximity is measured as the cosine distance (rounded to the nearest 0.01) between the citing-closest patent pairs.

Figure 3: Probability that citing patent and random patent are in the same technology class, by patent pair proximity.



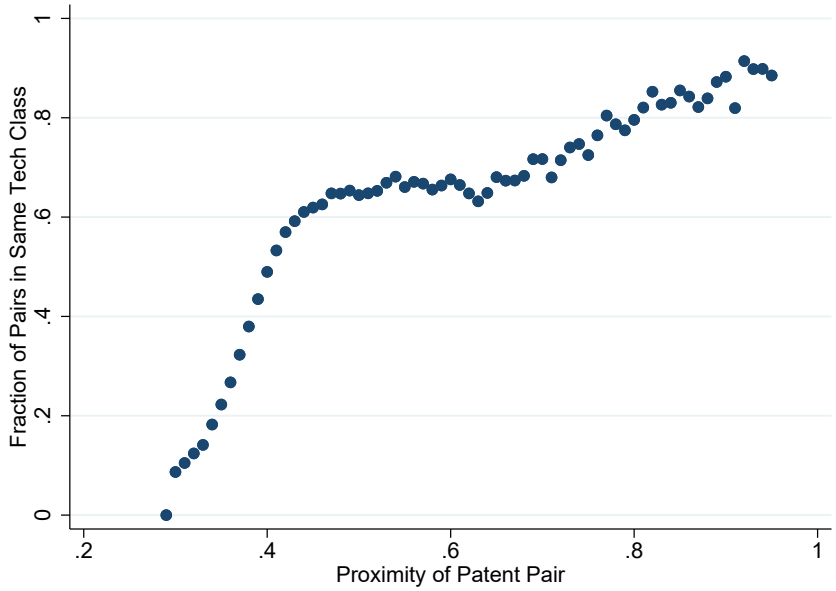
Notes: Fraction of citing-random patent pairs having a given proximity (as measured by cosine distance) that share the same technology class. Proximity values with at least 50 patent pairs are reported. Sample is the same 19,595,315 citing-random patent pairs as in Figure 1.

Figure 4: Probability that citing patent and random patent are in the same technology subclass, by patent pair proximity.



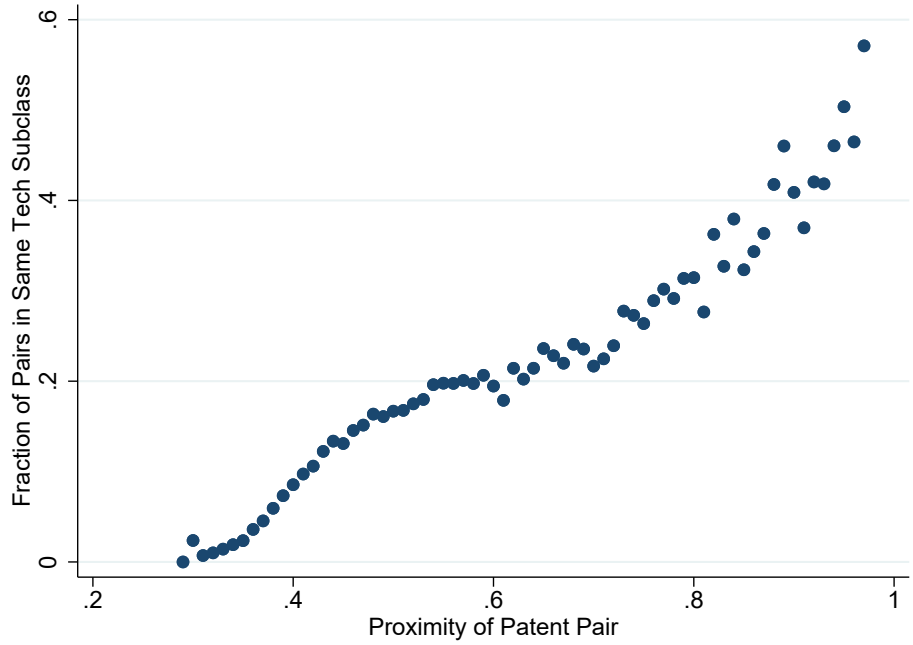
Notes: Fraction of citing-random patent pairs having a given proximity (as measured by cosine distance) that share the same technology class and subclass. Proximity values with at least 50 patent pairs are reported. Sample is the same 19,595,315 citing-random patent pairs as in Figure 1.

Figure 5: Probability that citing patent and closest patent are in the same technology class, by patent pair proximity.



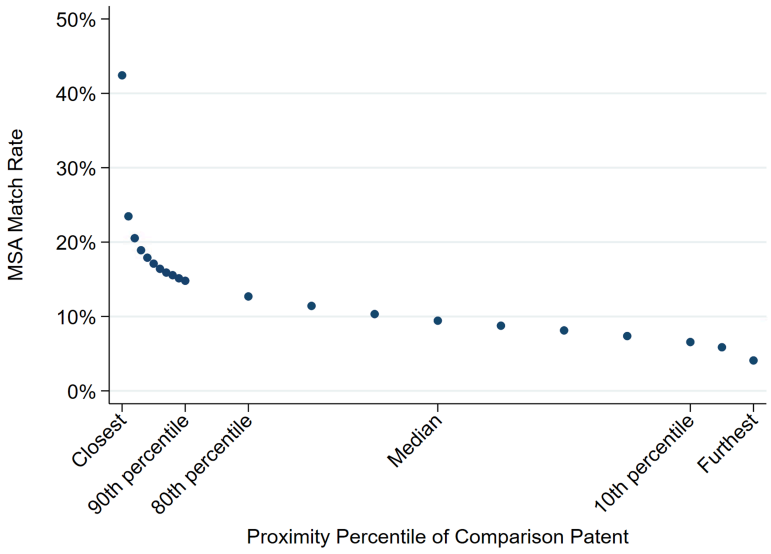
Notes: Fraction of citing-closest patent pairs having a given proximity (as measured by cosine distance) that share the same technology class. Sample is the same 19,595,315 citing-closest patent pairs as in Figure 2.

Figure 6: Probability that citing patent and closest patent are in the same technology subclass, by patent pair proximity.



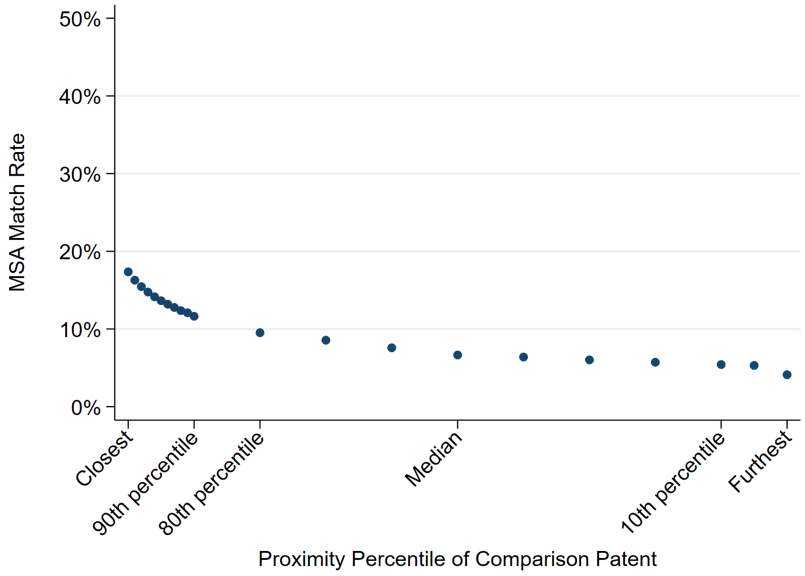
Notes: Fraction of citing-closest patent pairs having a given proximity (as measured by cosine distance) that share the same technology class and subclass. Sample is the same 19,595,315 citing-closest patent pairs as in Figure 2.

Figure 7: Probability that citing patent and control patent are in the same location when control patent is chosen from the same technology class, by proximity of control patent



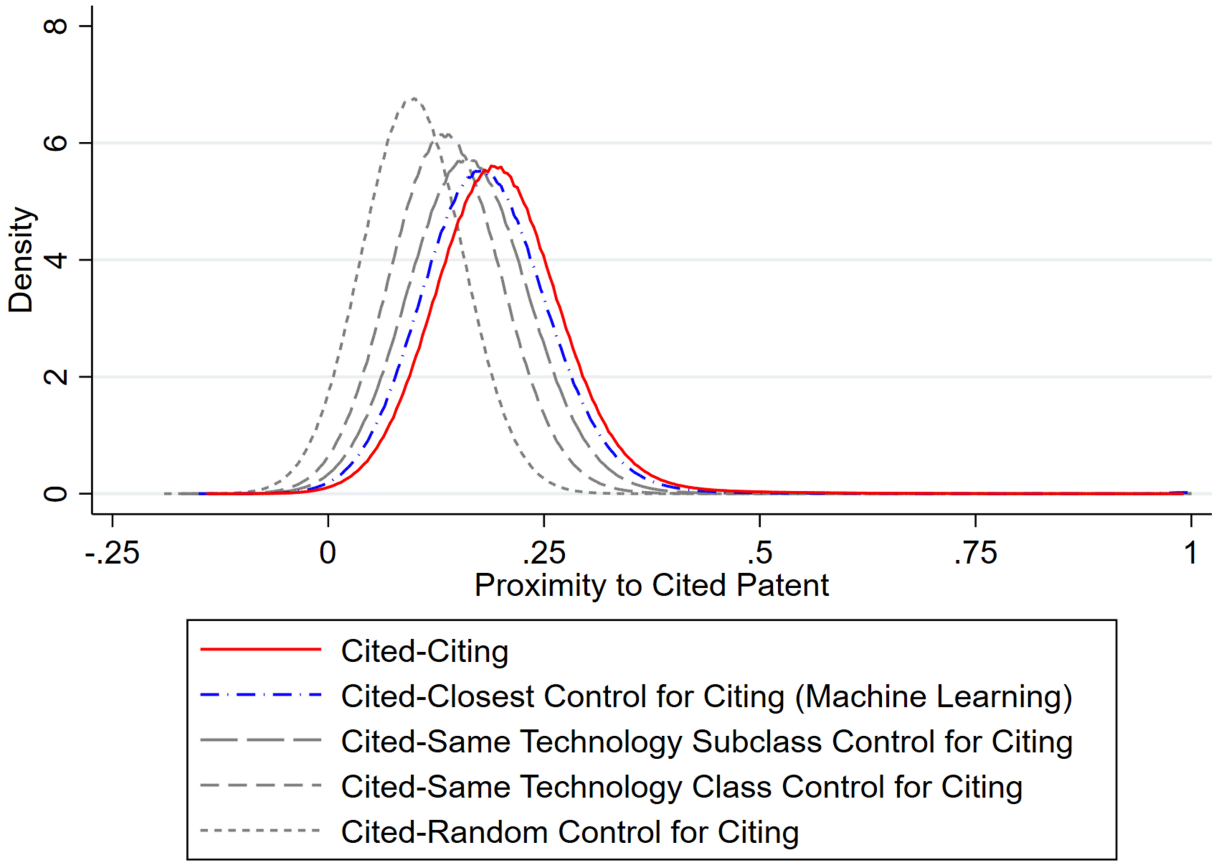
Notes: Horizontal axis captures proximity of citing patent instance and a potential control patent. Vertical axis captures fraction of citing and potential control patents that are from the same MSA. Each citing patent instance is paired with the closest patent, the furthest patent, and the patent at each reported proximity percentile, from among all patents sharing the same application year and same technology class as the citing patent. Sample consists of 19,595,315 citing patent instances and their potential controls at each of the computed percentiles.

Figure 8: Probability that citing patent and control patent are in the same location when control patent is chosen from the same technology subclass, by proximity of control patent



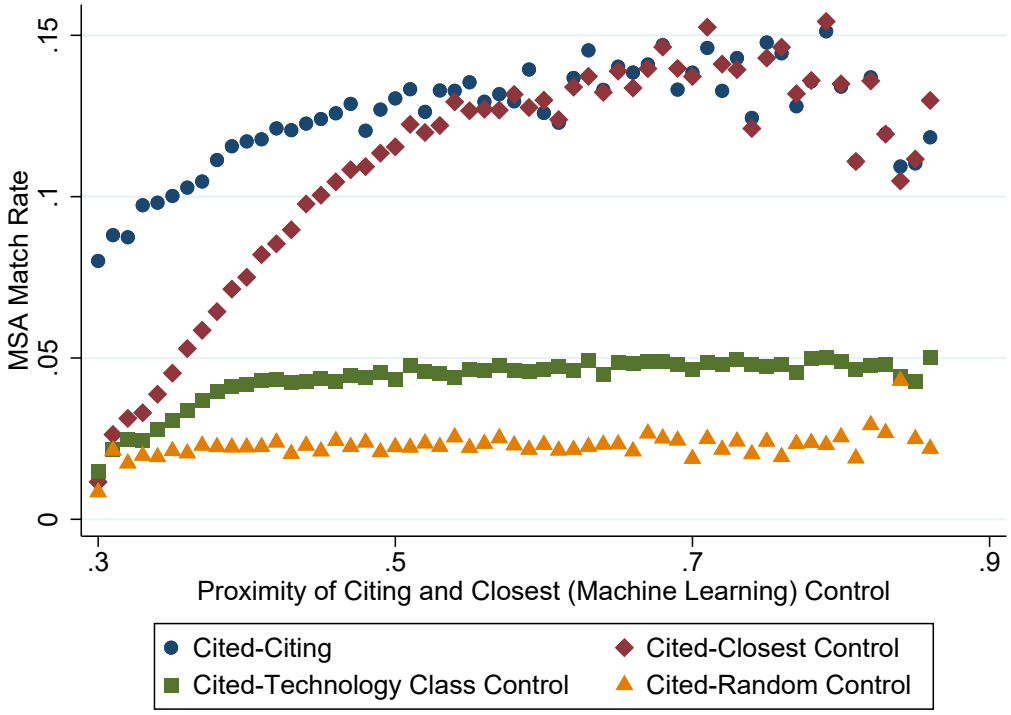
Notes: See notes to Figure 7. In contrast with analyses for Figure 8, where control patents have the same technology class as the citing patent, here all control patents have the same technology class and subclass as the citing patent.

Figure 9: Distribution of proximity to cited patent for citing and four different control patents.



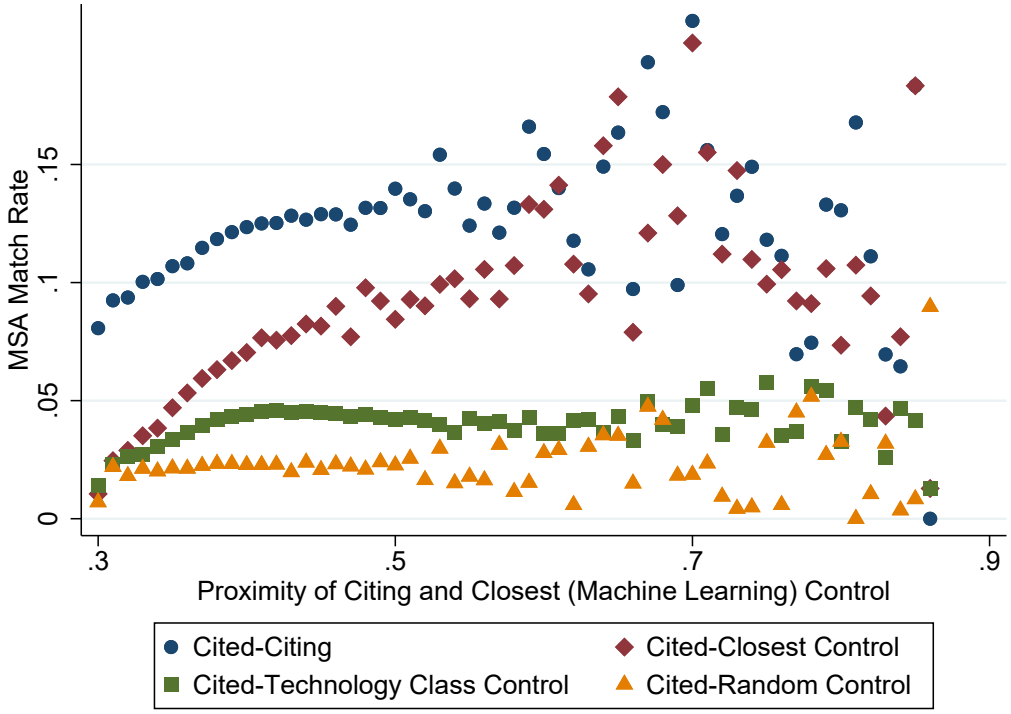
Notes: Distribution of proximity (cosine distance) is reported for five different patent pairs: cited-citing, cited-closest control, cited-technology subclass control, cited-technology class control, and cited-random control. Sample consists of 19,595,315 citing patent instances and their controls.

Figure 10: Location match rate at MSA-level for the baseline specification, by proximity of citing patent and closest control patent.



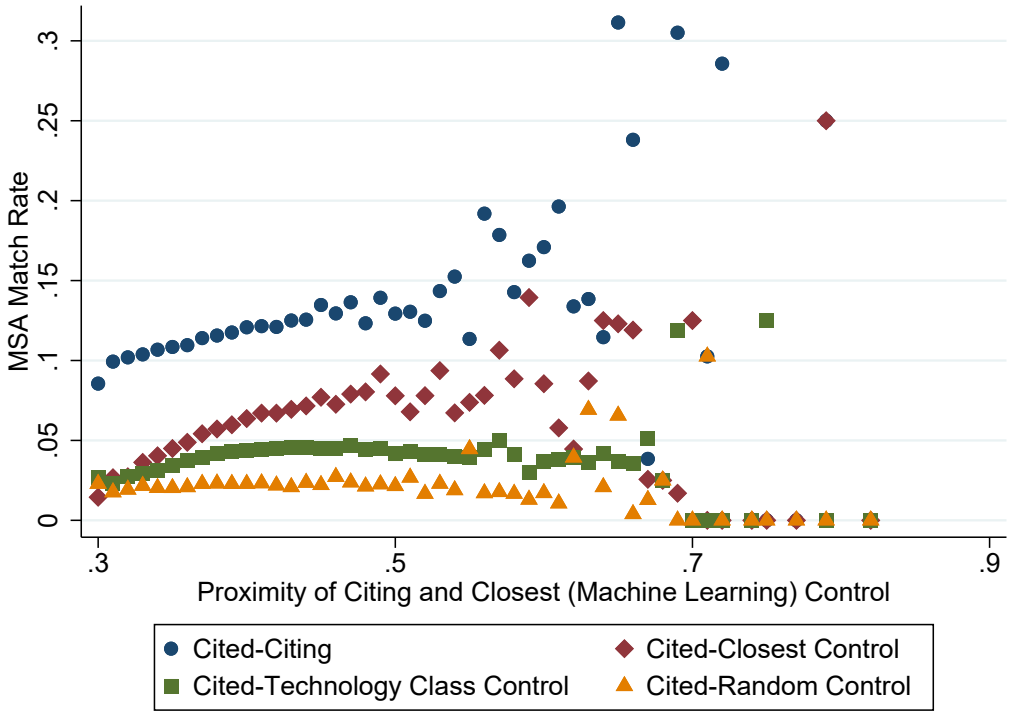
Notes: Vertical axis captures fraction of patent pairs (citing-cited, closest control-cited, technology class control-cited, and random control-cited) that are in the same MSA. Horizontal axis captures proximity between the citing patent and the closest control patent. Proximity is rounded to the nearest 0.01. Blue circles represent the MSA match rate between the citing and cited patents. Red diamonds represent MSA match rate between cited and closest control patent. Green squares represent MSA match rate cited and control chosen randomly from patents with the same technology class as citing patent. Orange triangles represent MSA match rate between cited patent and control chosen randomly from patents with same application year as citing patent. Sample consists of 7,806,934 inventor-added citations to U.S. patents and their associated cited and control patents.

Figure 11: Location match rate at MSA-level in the preferred specification, by proximity of citing patent and closest control patent.



Notes: See notes to Figure 10. In contrast with the baseline specification analyses shown in Figure 10, in the preferred specification three additional restrictions are imposed when the closest control patent is determined: the control patent cannot have the same inventor as the citing patent, the control patent cannot have the same assignee as the cited patent, and the control patent cannot have the same assignee as the citing patent. Sample consists of 7,889,133 inventor-added citations to U.S. patents and their associated cited and control patents.

Figure 12: Location match rate at MSA-level for the preferred specification with no 100-character overlap, by proximity of citing patent and closest control patent.



Notes: See notes to Figure 10. Figure is for the preferred specification with the additional restriction that control patents that have more than a 100 character text fragment overlap with the citing patent are not admissible controls. Sample consists of 7,050,019 inventor-added citations to U.S. patents and their associated cited and control patents.

Table 1: Relationship between patent pair proximity and patent pair characteristics

Dependent Variable: Proximity of patent pair

	(1)	(2)	(3)
Same technology class	0.056***	0.055***	0.055***
	(0.003)	(0.003)	(0.002)
Same technology subclass	0.042***	0.039***	0.039***
	(0.004)	(0.004)	(0.004)
Same assignee	0.037***	0.033***	0.034***
	(0.001)	(0.001)	(0.001)
Common inventor	0.105***	0.108***	0.106***
	(0.008)	(0.008)	(0.008)
Citation between pair	0.108***	0.109***	0.109***
	(0.018)	(0.019)	(0.017)
Common citation	0.065***	0.066***	0.065***
	(0.001)	(0.001)	(0.002)
Same country	0.007***	0.009***	0.009***
	(0.000)	(0.000)	(0.000)
Same State	0.001*	0.001***	0.000
	(0.000)	(0.000)	(0.000)
Same MSA	0.004***	0.004***	0.004***
	(0.001)	(0.000)	(0.000)
Citing patent FE	No	Yes	No
Random patent FE	No	No	Yes
R-squared	0.0148	0.1202	0.1280
Number of observations	19,332,013	19,332,013	19,332,013

Notes: Ordinary least squares regressions with robust standard errors clustered by citing patent technology class. An observation is a citing patent instance that is paired with a random patent with the same application year as the citing patent. The dependent variable is the proximity (cosine distance) between the two patents. Sample is constructed from citations made from patents filed between 2001 and 2009 to patents granted between 1976 and 2017 that have application year 1976-2009. Asterisks indicate statistical significance: * = p<0.1, ** = p<0.05, *** = p<0.01.

Table 2: Regression analysis of patent pair proximity and patent pair characteristics, for patent pairs with same application year and either the same technology class (columns 1-3) or the same technology subclass (columns 4-6).

Dependent Variable: Proximity of patent pair

	Same Technology Class			Same Technology Subclass		
	(1)	(2)	(3)	(4)	(5)	(6)
Same technology subclass	0.043***	0.041***	0.043***			
	(0.002)	(0.002)	(0.002)			
Same assignee	0.047***	0.039***	0.047***	0.068***	0.059***	0.068***
	(0.003)	(0.002)	(0.003)	(0.004)	(0.005)	(0.004)
Common inventor	0.171***	0.176***	0.171***	0.228***	0.212***	0.228***
	(0.014)	(0.013)	(0.014)	(0.006)	(0.006)	(0.006)
Citation between pair	0.074***	0.074***	0.074***	0.047***	0.045***	0.047***
	(0.018)	(0.019)	(0.017)	(0.008)	(0.007)	(0.008)
Common citation	0.061***	0.053***	0.062***	0.060***	0.040***	0.060***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.002)
Same country	0.013***	0.015***	0.013***	0.013***	0.016***	0.013***
	(0.001)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)
Same State	0.001*	0.002***	0.001*	0.004**	0.005***	0.004**
	(0.001)	(0.000)	(0.001)	(0.002)	(0.002)	(0.002)
Same MSA	0.006***	0.005***	0.006***	0.021***	0.016***	0.021***
	(0.001)	(0.001)	(0.001)	(0.003)	(0.002)	(0.003)
Citing patent FE	No	Yes	No	No	Yes	No
Random patent FE	No	No	Yes	No	No	Yes
R-squared	0.1159	0.3465	0.1893	0.4788	0.7333	0.5267
Number of obs.	19,331,144	19,331,144	19,331,144	17,447,618	17,447,618	17,447,618

Notes: See notes to Table 1. In contrast with analyses reported in Table 1, where each citing patent was paired with random patent with the same application year as the citing patent, here each citing patent is paired with a random patent that also has the same technology class (columns 1-3) or same technology subclass (columns 4-6) as the citing patent.

Table 3. Geographic matching frequencies for different choices of control patents.

	(1) Proximity: Baseline Specification	(2) Proximity: Preferred Specification	(3) Technology Subclass	(4) Technology Class	(5) Random	(6) Examiner Citations
Country Matches						
% Citing Matching	77.2%	77.3%	77.6%	77.3%	77.3%	84.7%
% Controls Matching	70.5%	67.2%	59.4%	53.6%	47.2%	59.8%
Difference	6.7%	10.1%	18.2%	23.7%	30.1%	24.9%
t-statistic	374.43	555.72	921.77	1254.47	1598.86	868.67
N	11,866,932	11,987,380	10,682,467	11,910,921	11,985,058	12,052,531
State Matches						
% Citing Matching	13.7%	13.8%	14.0%	13.8%	13.8%	15.7%
% Controls Matching	10.4%	8.8%	8.0%	5.9%	4.0%	9.5%
Difference	3.2%	5.0%	6.0%	7.9%	9.8%	6.2%
t-statistic	243.37	388.86	447.08	652.78	857.07	311.60
N	11,866,932	11,987,380	10,682,467	11,910,921	11,985,058	12,052,531
MSA Matches						
% Citing Matching	10.2%	10.3%	10.5%	10.3%	10.3%	11.7%
% Controls Matching	7.3%	5.9%	5.4%	3.7%	2.2%	6.9%
Difference	2.9%	4.4%	5.1%	6.6%	8.2%	4.8%
t-statistic	238.22	383.36	421.15	612.09	811.22	265.92
N	11,092,075	11,207,673	10,021,286	11,133,581	11,205,411	11,269,142

Notes: Matching percentage refers to the fraction of citing-cited and control-cited patent pairs that are in the same country (top panel), state (middle panel), or MSA (bottom panel). Sample consists of citations made from patents filed between 2001 and 2009 to patents with a U.S. inventor that were granted between 1976 and 2017 and have application year 1976-2009. In each column each citing patent instance is paired with a different control patent: closest patent from our baseline specification (column 1), closest patent from our preferred specification, which imposes the additional restrictions that control patent cannot have same inventor as citing patent or same assignee as cited or citing patents (column 2), random patent with the same technology subclass as citing patent (column 3), random patent with same the same technology class as citing patent (column 4), and random patent with the same application year as the citing patent (column 5). Column (6) compares inventor and examiner added citations as in Thompson (2006).

Table 4: Geographic matching frequencies for different choices of control patents: inventor citations only.

	(1) Proximity: Baseline Specification	(2) Proximity: Preferred Specification	(3) Technology Subclass	(4) Technology Class	(5) Random	(6) Examiner Citations
Country Matches						
% Citing Matching	84.7%	84.7%	84.9%	84.7%	84.7%	84.7%
% Controls Matching	75.7%	71.5%	61.0%	54.3%	47.1%	59.8%
Difference	9.0%	13.2%	23.8%	30.4%	37.6%	24.9%
t-statistic	463.04	663.86	1083.49	1432.56	1772.37	868.67
N	8,325,110	8,410,745	7,572,145	8,365,222	8,413,445	12,052,531
State Matches						
% Citing Matching	15.5%	15.7%	15.8%	15.7%	15.7%	15.7%
% Controls Matching	11.5%	9.5%	8.3%	6.0%	4.0%	9.5%
Difference	4.0%	6.2%	7.4%	9.6%	11.7%	6.2%
t-statistic	241.56	381.74	447.59	641.21	817.35	311.60
N	8,325,110	8,410,745	7,572,145	8,365,222	8,413,445	12,052,531
MSA Matches						
% Citing Matching	11.6%	11.7%	11.8%	11.7%	11.7%	11.7%
% Controls Matching	8.1%	6.4%	5.6%	3.8%	2.2%	6.9%
Difference	3.5%	5.4%	6.2%	7.9%	9.6%	4.8%
t-statistic	234.08	373.04	415.96	593.37	759.26	265.92
N	7,806,934	7,889,133	7,121,193	7,844,948	7,891,549	11,269,142

Notes: See notes to Table 3. In contrast with the analyses in Table 3, the sample is now limited to inventor-added citations only (except in column 6 which is reproduced from Table 3 for comparison).

Table 5: Geographic matching frequencies for preferred specification, inventor citations, and subsamples with closest controls

	(1)	(2)	(3)	(4)	(5)	(6)
Percentage of Sample	50%	10%	1%	0.7%	0.4%	0.1%
Proximity Values	≥ 0.38	≥ 0.44	≥ 0.54	≥ 0.57	≥ 0.61	≥ 0.74
Country Matches						
% Citing Matching	85.3%	84.7%	88.1%	89.6%	90.1%	91.7%
% Controls Matching	74.1%	75.7%	83.3%	87.3%	88.8%	91.2%
Difference	11.2%	8.9%	4.8%	2.3%	1.3%	0.5%
t-statistic	392.62	136.59	30.73	11.78	5.55	1.14
N	3,908,601	731,194	100,443	53,296	34,549	8636
State Matches						
% Citing Matching	16.7%	17.6%	17.8%	17.9%	17.2%	13.0%
% Controls Matching	11.0%	12.6%	14.4%	15.6%	15.7%	12.3%
Difference	5.7%	5.0%	3.4%	2.3%	1.5%	0.8%
t-statistic	230.56	85.51	20.47	10.08	5.37	1.49
N	3,908,601	731,194	100,443	53,296	34,549	8636
MSA Matches						
% Citing Matching	12.5%	13.1%	13.4%	13.7%	12.9%	10.2%
% Controls Matching	7.6%	9.0%	11.2%	12.3%	12.3%	8.9%
Difference	4.9%	4.1%	2.3%	1.4%	0.6%	1.4%
t-statistic	222.96	76.83	15.04	6.53	2.48	2.92
N	3,684,170	691,281	95,241	50,448	32,633	8033

Notes: Sample consists of citations made by inventors (examiner-added citations are excluded) on patents filed between 2001 and 2009 to patents with a U.S. inventor that were granted between 1976 and 2017 and have application year 1976-2009. The "% Controls Matching" reports the fraction of control-cited patent pairs that are in the same country (top panel), state (middle panel), or MSA (bottom panel), where the control patent is chosen as the closest patent to the citing patent in our preferred specification (as in Column 2 in Tables 3 and 4). The different columns correspond to different subsamples. In particular, we report the results for the 50%, 10%, 1%, 0.7%, 0.4%, and 0.1% of the subsample with most proximate controls to the citing patent, which correspond to proximity values of 0.38, 0.44, 0.54, 0.57, 0.61, 0.74, and above, respectively.

Table 6: Geographic matching frequencies for preferred specification with no 100-character overlap, inventor citations, and subsamples with closest controls

	(1)	(2)	(3)	(4)	(5)	(6)
Percentage of Sample	50%	10%	1%	0.7%	0.4%	0.1%
Proximity Values	≥ 0.38	≥ 0.43	≥ 0.50	≥ 0.51	≥ 0.53	≥ 0.58
Country Matches						
% Citing Matching	84.9%	83.9%	78.7%	76.6%	74.7%	76.8%
% Controls Matching	71.0%	70.3%	65.9%	65.3%	63.7%	63.1%
Difference	13.9%	13.6%	12.7%	11.4%	10.9%	13.8%
t-statistic	438.09	224.11	58.88	38.14	26.80	16.19
N	3,301,023	933,164	84,225	45,711	25,270	5676
State Matches						
% Citing Matching	16.3%	17.0%	17.0%	17.1%	17.5%	20.7%
% Controls Matching	9.7%	10.5%	10.2%	10.2%	10.1%	12.9%
Difference	6.6%	6.5%	6.8%	6.9%	7.4%	7.8%
t-statistic	252.68	128.75	41.08	30.48	24.39	11.11
N	3,301,023	933,164	84,225	45,711	25,270	5676
MSA Matches						
% Citing Matching	12.3%	12.9%	13.8%	14.6%	15.6%	17.1%
% Controls Matching	6.7%	7.4%	7.9%	8.4%	8.3%	9.8%
Difference	5.6%	5.5%	5.9%	6.2%	7.3%	7.3%
t-statistic	239.55	121.84	38.12	28.81	24.89	11.14
N	3,112,406	883,272	80.040	43,454	24,073	5344

Notes: See notes for Table 5. Control patents are chosen according to the preferred specification with the additional restriction that control patents that have more than a 100 character text fragment overlap with the citing patent are not admissible controls. The different columns correspond to different subsamples. In particular, we report the results for the 50%, 10%, 1%, 0.7%, 0.4%, and 0.1% of the subsample with most proximate controls to the citing patent, which correspond to proximity values of 0.38, 0.43, 0.50, 0.51, 0.53, 0.58, and above, respectively.

Appendix A: Sample Patents and Computed Proximity

By way of example, we present a patent and then three other patents that are similar, somewhat similar, and dissimilar. Our patent of interest (#5,215,930) is a method for etching an integrated circuit. The second patent below (#5,607,543) also covers a process for etching an integrated circuit and not surprisingly we compute the proximity between the first and second patents to be 0.96. The third patent (#5,338,750) is also a method for fabricating an integrated circuit, though it is not specific to etching. Its proximity to the first patent is 0.72. The last patent (#5,624,563) is a process for the treatment of sludge water and is not at all related to the first patent. Our computed proximity to the first patent is 0.14.

US005215930A

United States Patent [19]

Lee et al.

[11] **Patent Number: 5,215,930**

[45] **Date of Patent: Jun. 1, 1993**

[54] **INTEGRATED CIRCUIT ETCHING OF SILICON NITRIDE AND POLYSILICON USING PHOSPHORIC ACID**

[75] Inventors: **Kuo-Hua Lee**, Lower Macungie Township, Lehigh County; **Chen-Hua D. Yu**, Allentown, both of Pa.

[73] Assignee: **AT&T Bell Laboratories**, Murray Hill, N.J.

[21] Appl. No.: **781,463**

[22] Filed: **Oct. 23, 1991**

[51] **Int. Cl.**[5] **H01L 21/76**

[52] **U.S. Cl.** **437/40**; 437/69; 437/228

[58] **Field of Search** ..... 437/69, 228, 40, 233; 148/DIG. 51, DIG. 85, DIG. 86, DIG. 117

[56] **References Cited**

U.S. PATENT DOCUMENTS

5,002,898 3/1991 Fritzenger et al. 437/69

FOREIGN PATENT DOCUMENTS

63-307743 12/1988 Japan .

OTHER PUBLICATIONS

"A New Deglaze Process for Doped Polysilicon," L. Lowell, *Solid State Technology*, Apr. 1991, pp. 149–153.

Poponiak, M., "Forming Dielectric Isolation", *IBM Technical Disclosure Bulletin*, vol. 20, No. 4, Sep. 1977, p. 1405.

Primary Examiner—George Fourson

Attorney, Agent, or Firm—John T. Rehberg

[57] **ABSTRACT**

A process for removing both the silicon nitride layer and polysilicon layer in a poly-buffered LOCOS process which utilizes hot phosphoric acid is disclosed.

11 Claims, 1 Drawing Sheet

US005607543A

United States Patent [19]

Eisenberg et al.

[11] Patent Number: 5,607,543

[45] Date of Patent: Mar. 4, 1997

[54] **INTEGRATED CIRCUIT ETCHING**

[75] Inventors: **Juli H. Eisenberg**, Allentown, Pa.; **Susan C. Vitkavage**, Orlando, Fla.

[73] Assignee: **Lucent Technologies Inc.**, Murray Hill, N.J.

[21] Appl. No.: **431,341**

[22] Filed: **Apr. 28, 1995**

[51] **Int. Cl.**[6] **H01L 21/30**

[52] **U.S. Cl.** **156/662.1**; 156/651.1; 156/652.1; 156/653.1; 437/228

[58] **Field of Search** 156/662.1, 651.1, 156/653.1, 657.1; 437/228; 252/79.1, 79.2

[56] **References Cited**

U.S. PATENT DOCUMENTS

3,715,249	2/1973	Panousis .	
5,002,898	3/1991	Fritzinger et al. .	
5,215,930	6/1993	Lee et al.	437/40
5,310,457	5/1994	Ziger	156/657
5,437,765	8/1995	Loewenstein	216/51

FOREIGN PATENT DOCUMENTS

OTHER PUBLICATIONS

"Silicon Processing For The VLSI Era–vol. 2–Process Integration"; Wolf; Lattice Press, Sunset Beach, Ca.; ©1990; pp. 17–32.

A New Deglase Process for Doped Polysilicon Larry, Lowell, Polaroid Corporation, Microelectronics Laboratory, Cambridge, Massachusetts. Apr. 1991, Solid State Technology, pp. 149–153.

Thin Film Processes, John L. Vossen and Werner Kern, RCA Laboratories, David Sarnoff Research Center, Princeton, New Jersey.

Characterization of Poly–Buffered LOCOS in Manufacturing Environment, R. L. Guldi, B. McKee, G. M. Damminga, C. Y. Yong and M. A. Beals, Texas Instruments Incorporated, Logic Operations, Semiconductor Group, Dallas, Texas. J. Electrochem 1967, p. 423.

Primary Examiner—R. Bruce Breneman
Assistant Examiner—George Goudreau
Attorney, Agent, or Firm—John T. Rehberg

[57] **ABSTRACT**

A process for removing both the silicon nitride layer and polysilicon layer in a poly-buffered LOCOS process which utilizes hot phosphoric acid and nitric acid is disclosed.

US005338750A

United States Patent [19]

Tuan et al.

[11] Patent Number: 5,338,750

[45] Date of Patent: Aug. 16, 1994

[54] **FABRICATION METHOD TO PRODUCE PIT-FREE POLYSILICON BUFFER LOCAL OXIDATION ISOLATION**

[75] Inventors: **Hsiao-Chin Tuan; Hu H. Chao**, both of Hsinchu, Taiwan

[73] Assignee: **Industrial Technology Research Institute**, Hsinchu, Taiwan

[21] Appl. No.: **982,708**

[22] Filed: **Nov. 27, 1992**

[51] **Int. Cl.**[5] **H01L 21/76**

[52] **U.S. Cl.** **437/70**; 437/69; 437/926; 437/968

[58] **Field of Search** 437/69, 70, 926, 968

[56] **References Cited**

U.S. PATENT DOCUMENTS

4,407,696	10/1983	Han et al.	437/69
4,897,364	1/1990	Nguyen et al.	437/69
5,002,898	3/1991	Fritzinger et al.	437/69
5,196,367	3/1993	Lu et al.	437/69
5,215,930	6/1993	Lee et al.	437/69

FOREIGN PATENT DOCUMENTS

0053957	3/1982	Japan	437/968
0208156	12/1982	Japan	437/968
0074350	4/1986	Japan	437/69
0204746	8/1988	Japan	437/69
0214142	8/1989	Japan	437/69
0068930	3/1990	Japan	437/69
0097038	4/1990	Japan	437/69

Primary Examiner—Tom Thomas
Assistant Examiner—Trung Dang
Attorney, Agent, or Firm—George O. Saile; Stephan B. Ackerman

[57] **ABSTRACT**

A method of forming a silicon oxide isolation region on the surface of a silicon wafer consisting of a thin layer of silicon oxide on the wafer, a layer of impurity-doped polysilicon, and a layer of silicon nitride. The oxidation mask is formed by patterning the silicon nitride layer and at least a portion of the doped polysilicon layer. The silicon oxide field isolation region is formed by subjecting the structure to a thermal oxidation ambient. The oxidation mask is removed in one continuous etching step using a single etchant, such as phosphoric acid which etches the silicon nitride and polysilicon layers at substantially the same rate to complete the formation of the isolation region without pitting the monocrystalline substrate.

United States Patent [19]

Hawkins

[11] **Patent Number: 5,624,563**

[45] **Date of Patent: Apr. 29, 1997**

[54] **PROCESS AND APPARATUS FOR AN ACTIVATED SLUDGE TREATMENT OF WASTEWATER**

[76] Inventor: **John C. Hawkins**, P.O. Box 566, Murrells Inlet, S.C. 29576

[21] Appl. No.: **519,423**

[22] Filed: **Aug. 25, 1995**

[51] **Int. Cl.**[6] **C02F 3/12**

[52] **U.S. Cl.** **210/607**; 210/624; 210/195.3; 210/253; 210/903

[58] **Field of Search** 210/605, 607, 210/621, 624, 626, 903, 195.1, 195.3, 202, 253, 258, 259

[56] **References Cited**

U.S. PATENT DOCUMENTS

3,468,795	9/1969	Bye-Jorgensen et al.	210/7
3,622,507	11/1971	Pasveer	210/6
3,964,998	6/1976	Barnard	210/7
4,243,522	1/1981	Ter-Borch et al.	210/774
4,376,275	3/1983	Raper	210/605
4,479,876	10/1984	Fuchs	210/605
4,537,682	8/1985	Wong-Chong	210/611
4,563,282	1/1986	Wittmann et al.	210/619
4,568,457	2/1986	Sullivan	210/151
4,604,206	8/1986	Sullivan	210/603
4,624,788	11/1986	Repin	210/624
4,663,044	5/1987	Goronszy	210/610
4,780,208	10/1988	Bohnke et al.	210/605
4,793,930	12/1988	Soeder et al.	210/614
4,867,883	9/1989	Daigger et al.	210/605
4,910,541	3/1990	Wade et al.	210/241
4,948,510	8/1990	Todd et al.	210/624
4,952,316	8/1990	Cooley	210/626
4,961,854	10/1990	Wittmann et al.	210/621
5,192,441	3/1993	Sibony et al.	210/603
5,192,442	3/1993	Piccinillo et al.	210/605
5,196,111	3/1993	Nicol et al.	210/96.1
5,205,936	4/1993	Topnik	210/614
5,211,847	5/1993	Kanow	210/610
5,213,681	5/1993	Kos	210/605
5,228,996	7/1993	Lansdell	210/605
5,234,595	8/1993	DeGregorio et al.	210/605
5,248,422	9/1993	Neu	210/605
5,252,214	10/1993	Lorenz et al.	210/605
5,266,200	11/1993	Reid	210/605
5,275,722	1/1994	Beard	210/195.1
5,288,407	2/1994	Bodwell et al.	210/617
5,326,459	7/1994	Hlavach et al.	210/150
5,342,522	8/1994	Marsman et al.	210/605
5,342,523	8/1994	Kuwashima	210/607
5,348,653	9/1994	Rovel	210/605
5,354,458	10/1994	Wang et al.	210/180
5,354,471	11/1994	Timpany et al.	210/607

OTHER PUBLICATIONS

EPA Manual for Nitrogen Control, pp. 272–274; Advanced Environmental Systems Leaflets.

Primary Examiner—Christopher Upton
Attorney, Agent, or Firm—Michael E. Mauney

[57] **ABSTRACT**

A process and apparatus for biological purification of wastewater resulting in reduction of biological oxygen demand, reduction of suspended solids, and for nitrification, wherein the process involves aerating wastewater in a treatment zone for reduction of biological oxygen demand, transferring the mixed liquor from this treatment zone on an alternate basis to a second or third treatment zone which undergoes an aeration/settle, fill and decant cycle. As the second or third zone is receiving influent, it is being decanted in a modified plug-flow fashion. The alternate treatment zones allow for a continuous discharge of treated effluent during the process. During aeration of the second or third treatment zone, mixed