PRELIMINARY DRAFT

The Proximity of Ideas: An Analysis of Patent Text Using Machine Learning

Fiona Sijie Feng (NYU Stern)

December 4, 2018

Click here for latest version

Abstract

This paper introduces a measure of proximity in ideas using unsupervised machine learning. I explore knowledge relationships in innovative ideas by deriving vector space representations of patent abstract text using Document Vectors (Doc2Vec), and using cosine similarity to measure their proximity in ideas space. I illustrate the potential uses of this method with an application to localization in knowledge spillovers. In the first case, I use the standard citations approach in measuring localization, but use text similarity to select a control case patent instead of USPC technology class. While this improves matching on unobserved technological differences, I find that local patents still receive about 0.9-1.4 times more local citations than the non-local control. This may partially be explained by the role of patent lawyers in determining the localization patterns of citations. As an alternative to citations, I calculate the localization in idea proximity using patent text similarity. I find less evidence of localization: within a technology field, patents within the same city are 0.02-0.06 times more similar to each other than patents from other cities. These findings indicate that localization effects may indeed be smaller than standard citations methods suggest. As ideas proximity provide a different lens into knowledge relationships, I also discuss some implications and potential limitations in the use of text similarity methods.

1. Introduction

This paper introduces a measure of proximity in ideas using unsupervised machine learning. I explore knowledge relationships in innovative ideas by deriving vector space representations of patent abstract text using Document Vectors (Doc2Vec), and using cosine similarity to measure their proximity in ideas space. Abstract text is used as it summarizes the ideas of the invention. I illustrate the potential uses of this method with an application to localization in knowledge spillovers.

One explanation for why innovation is concentrated in cities is that knowledge spillovers are geographically constrained. This means that knowledge transfers generate greater positive externalities for local firms and inventors, who learn from one another. A prominent literature of measuring knowledge spillovers has emerged from Jaffe et al. (1993) (henceforth JTH) that uses patent citations to study the "paper trail" left by the diffusion of innovative knowledge. The general consensus of this literature is that there are large and significant geographic localization effects for knowledge spillovers.

I apply similarity in two different ways. First, I use the standard citations methodology of measuring localization by examining the percentage of local forward citations made to local patents compared to a non-local control, across four decades of observations 1976-2015. Here, instead of selecting the control based on USPC[1] primary class, I select a control based on text similarity, which should provide a better proxy of underlying technological proximity between the two patents. While I do indeed find smaller localization effects with the similarity selected control, I still find that local patents receive local citations 0.9-1.4 times higher than that of the non-local control. Prior literature (Moser et al. (2017), Wagner et al. (2014)) has suggested that patent attorneys play a large role in determining patent citations. I control for lawyer effects by selecting a control from the same primary class *and* lawyer. I find that this reduces localization estimates substantially: local patents receive 0.3 times more local citations compared to the non-local control from the same lawyer. This provides a partial explanation for why better measures of technological proximity do not yield lower estimates of localization: citations may be localized in part because lawyers' knowledge of "citable" patents are geographically concentrated.

If citations may overstate localization of knowledge spillovers, a different approach may be useful. If local firms and inventors learn from each other's inventions, then within a technology field, patents the same city should express ideas more proximate to each other on average compared to patents

[1] United States Patent Classification. This is classification system of patents created by the United States Patent and Trademark Office (USPTO).

from different cities. Further, I find that patents that cite each other have much higher similarity on average, which suggests that a higher incidence of direct knowledge flows does imply greater proximity of ideas. Under this second approach, I find that patents within the same city are 0.02-0.06 times more similar to each other than patents from other cities. These findings provide further evidence that localization effects may be less than previously thought.

I address the concern that text similarity may be a noisy measure of idea proximity by validating its ability to find large effects on proximity: patents that are from the same primary class, or share a common backward citation, or share an inventor are found to have significantly higher text similarity. Thus, text similarity is not just noise: a noisy measure would find weak estimates across all dimensions. Another concern is that lawyers and examiners may also exert influence on patent text. I find that patents from the same lawyer and processed by the same examiner do have higher text similarity, but that this effect is attenuated when further controls for technology proximity across patent classes are included.

Citations and idea proximity provide different windows into knowledge relationships. Two potential explanations are discussed that may bridge the difference in localization of citations and localization of idea proximity. First, the number of local inventors that influence each other may be very small, which supports the microgeography literature (Feldman (2014); Catalini (2017)) that suggests disconnected clusters of innovation coexist even within the same city. Second, patent text will also reflect to influence of knowledge sources besides other patents, from scientific and other academic publications, to non-codified "tacit" knowledge. While Arora et al. (2018); Ganguli et al. (2017) discuss the importance of non-patent knowledge for innovation, patent text does indeed capture the influence of a broader range of potential knowledge flows, which may not be relevant for all studies. These discussions may provide some guidance for applied researchers seeking to understand best use and limitations of text similarity methods. Besides knowledge spillovers, proximity in ideas can potentially have a broad range of other applications. I discuss potential avenues for future research in the conclusion.

This paper provides a contribution to the literature on measures of knowledge and innovation through patent data. Alongside Jaffe et al. (1993), the prior literature using has found significant geographic localization in a variety of contexts: Murata et al. (2014) and Buzard et al. (2016) using spatial distance measures; Almeida and Kogut (1999), Agrawal et al. (2006), and Azoulay et al. (2011) using geographic mobility of inventors; Belenzon and Schankerman (2013) with university

patents and scientific publications. Only Thompson and Fox-Kean (2005) find that localization estimates are insignificant between extremely technologically proximate patents. Limitations of patent citations are well documented in the literature (Jaffe and De Rassenfosse (2017), Lerner and Seru (2015)). Three main critiques are raised: (i) the addition of citations from patent examiners as discussed in Alcacer et al. (2009); (ii) strategic considerations to add irrelevant citations to block potential infringement suits and to omit relevant citations to broaden the patent scope (Lampe (2012), Roach and Cohen (2013)); (iii) the influence of lawyers on applicant's citations (Moser et al. (2017), Wagner et al. (2014)). The evidence I find that localization may be less geographically constrained also contributes to the literature on agglomeration and urban economics, notably in support of Krugman (1991) and Cairncross (1997) who argued against spatial limitations to knowledge spillovers.

Recently, a small literature has bourgeoned around applying text analysis to patent text. Arts et al. (2018), Younge and Kuhn (2016) and Kelly et al. (2018) use a variation of term-document frequency to construct vector representations of patents, which uses the proportional counts of different terms within a patent. Kelly et al. (2018) adapts their measure to specifically account for the innovativeness of a patent, by overweighting *infrequent* terms up to the year of appearance of the patent. They also argue for the advantages of using patent text similarity against citations based measures, specifically that citations "given an incomplete representation of which predecessor technologies are important for a new patent." Additionally, Arts et al. (2018) validates the accuracy of text-based similarity measures with technology experts.

My contribution differs in that while these other measures use text-based frequency, they do not use machine learning methods that were devised to address shortfalls in frequency vector representations of documents. Primarily, frequency measures fail to account for terms with similar meanings (synonyms such as *software* and *program*) and terms with multiple meanings (polysemy such as *program*). Additionally, these measures do not utilize crucial information in semantic patterns such as the co-occurences in terms across documents. Thus, frequency based approaches may fail to capture the presence of similar ideas expressed through differing semantics and terminology. Finally, frequency measures results in extremely sparse and high-dimensional vectors, which may be computationally expensive. Unsupervised machine learning methods, including Latent Dirichlet Allocation (Blei et al. (2003)), Word2Vec (Mikolov et al. (2013b,a)), and Doc2Vec (Le and Mikolov (2014)) were devised precisely to address such concerns. Campr and Ježek (2015) specifically compares frequency measures against each of the unsupervised machine learning methods in how well they

each calculated document similarities compared to human annotators. Doc2Vec proved to be the most successful in their evaluation. However, the trade-off to using Doc2Vec is that its selection of text vectors is more black-box and less interpretable, as it utilizes a neural network structure.

Text analysis and unsupervised machine learning has also been applied to patents in non-similarity based contexts. Similar to Kelly et al. (2018), Kaplan and Vakili (2015) fit a Latent Dirichlet Allocation model on patent text to determine breakthrough innovation. They find that a topic-originating or breakthrough patent receives approximately 1.4 times more citations than the average patent. Bergeaud et al. (2017) also use relevant keywords found in patent abstracts to construct a semantic network classification system to map the technological taxonomy of patents. Bryan et al. (2018) use a machine learning approach to identify in-text citations within patent text, which they propose as a better measure of direct knowledge flows. Packalen and Battacharya (2015) use the appearance of new terms to examine how adoption of new technology varies by city size.

Roadmap The paper proceeds as follows. Section 2 discusses the data sources, and outlines the NLP methodology. Section 3 presents the estimation and results for the application of similarity to examine local knowledge spillovers. Results and further discussion of text similarity, including policy implications, are found in 4. I conclude with a discussion of potential avenues for other applications in 5. Further analyses, tables, and graphs can be found in the appendix.

2. Data and Methodology

2.1. Data Sources

Patent data is taken from PatentsView on all utility patents granted 1976-2016, containing data both on inventors (including unique identifiers and location) and patents (assignee, application date, grant date, primary class and subclass). Bibliographic text data is taken from the USPTO Bulk Data Products, which has all patent bibliographic text from 1976 to end of 2015. Patent abstracts are taken to be representative of the knowledge contained in patents, as they are a summary of the invention. Citations, lawyer, and examiners data for each patent are also taken from PatentsView. Following prior literature, the patent's location is determined as the MSA where the highest proportion of inventors are located.

Patent technology fields Each patent is assigned three technological *fields*, with each field being nested in the previous. At the broadest level, an NAICS-based industry classification is given using the USPC to NAICS concordance crosswalk, which delegates each patent to a NAICS category according to its USPTO 3-digit primary classification. Additionally, many patents are also assigned a primary *sub*class.[2] Primary subclasses are nested in primary classes, which are in turn nested in a NAICS industry label. There are over 150,000 subclass labels; 450 class labels, and 33 NAICS industry labels.

2.2. Patent Abstracts to Vector Space Representations: Document Vectors from Doc2Vec

Using patent abstract texts, I use procedures standard in the NLP literature to clean and convert text to vector representations (see section A.1 for details). I use the

The Doc2Vec algorithm was introduced by Le and Mikolov (2014) as a means to meaningfully summarize text contained within documents. It is a straightforward extension of the Word2Vec model of Mikolov et al. (2013b,a), which was developed to represent words meaningfully in a vector space (provide "word embeddings"). Word2Vec was found to be surprisingly powerful in capturing linguistic regularities and patterns, for example that $vec(\text{“}Madrid\text{”}) - vec(\text{“}Spain\text{”}) + vec(\text{“}France\text{”})$ is closer to $vec(\text{“}Paris\text{”})$ than any other word vector. The objective of Word2Vec is to situate words that have similar meanings close to one another. Similarly, Doc2Vec has the objective of situating similar *documents* close to one another by placing document vectors (DocVec) close to each other in vector space. To do this, the algorithm uses the "context" around each term in the document to derive a vector representation that maximizes the probability its the appearance. (See A.1.1 for more details on the algorithm; figure A.2 illustrates diagramatically the inputs and outputs of the algorithm) I implement the algorithm using the `gensim` package in Python (Řehůřek and Sojka (2010)).

For example, for the sentence "Provides for unattended file transfers", the central word "unattended" has the context ["Provides", "for", "file", "transfers"]. Different sentences will have different context and center words. Before the algorithm is implemented, common words or stop words such as "for" are removed and each word is stemmed to the root. "Provides" and "transfers" become "provid" and "transfer." The document identifier, in this case the patent number "US7502754," is treated as a

[2] Patents may also include other discretionary classifications, which are not used in my data.

context word for ever word in the patent. Thus, the context for "unattended" would become: ["provid", "file", "transfer", "US7502754"]. The goal of the algorithm is to select word vectors that maximise the probability of the center word, given the context words. In terms of document vectors, the algorithm will attempt to situate the patent document vector as close as possible to the words within the patent text.

Every word and document is assigned a vector of dimension $N = 100$.[3] The vectors are optimized using a neural network which maximises the log probability of the appearance of each central word. The resulting vector places words that arise in similar contexts close to each other, and documents that contain similar words close to each other.

2.3. Measuring Knowledge Spillovers: Cross Patent Similarity

Cosine similarity [4] has been used to measure technological proximity in Jaffe (1989) and Bloom et al. (2013), as well as being standard in the NLP literature (Mihalcea et al. (2006)). The prior literature used vectorizations of patent classes listed for each patent, which had the issues of being of varying lengths with unassigned weights for each class. The primary advantage of NLP patent vector outputs is that they are *jointly* determined, and position each patent vector *relative* to all other patents within the corpus. Thus, cross-patent comparisons using NLP vector outputs are much more internally consistent than using vectorizations of patent class selections.

For two patents, i and j, the cosine similarity between them is:

$$sim(i,j) = \frac{PV_i \cdot PV_j}{\|PV_i\| \left\|PV_j\right\|} \tag{2.1}$$

Where PV_i is the patent vector representation of i. This is preferred to Euclidean distance as it is factors in the "size" of the vector; a Euclidean distance measure would assign positive distance to two vectors that contained the exact same words, but of different quantities. Cosine similarity normalises all measures to be in the range $[-1, 1]$.

[3] This is a rule-of-thumb in the literature, according to Lin et al. (2015)

[4] Other measures, such as Hellinger distance, were also used but found to be very highly correlated with cosine similarity.

Number and proportion of common backward citations

Since the incidence of direct citation between two random patents are rare, a measure of "indirect" knowledge flow between the two patents would be the number or proportion of common backward citations between two patents:

$$ncc(i,j) = |\{citations_i\} \cap \{citations_j\}| \tag{2.2}$$

$$pcc(i,j) = \frac{|\{citations_i\} \cap \{citations_j\}|}{|\{citations_i\}|} \tag{2.3}$$

Where $ncc(i,j)$ represents the number of common backward citations between patents i, j and $pcc(i,j)$ the proportion of backward citations of $i$that were also made by j. For example, if patent i cites $\{A,B,C,D\}$, and j cites $\{D,E\}$, $ncc(i,j) = 1$ and $pcc(i,j) = 0.25$. In each case, self-citations are removed first. These variables can be thought of as measuring the degree of similitude in the patent knowledge sources of the two patents using a citations-based approach.

Technological Field Proximity

Since each patent is assigned technology field labels in the form of primary classes, technological field proximity between two primary classes can be measured using the average similarity of a sample of patents in each primary class. For each year t, I take a sample of up to 1000 patents in each primary class pair pc_i, pc_j that were granted in the previous 5 years. I then calculate the mean of the pairwise similarities between all such pairs. Thus:

$$sim(pc_i, pc_j)_t = mean\left(\{sim(i,j)|i \in pc_i, j \in pc_j\}_{t-5,t}\right) \tag{2.4}$$

Intuitively, this represents the *expected* similarity between two patents if only their technology field was known. Cross field similarity are analogous to the technological proximity measures of Jaffe (1986); Bloom et al. (2013). Both papers, alongside other citations-based methods of measuring technological proximity, rely on the vectorization of PTO classes. These methods may lead to inconsistent results as each patent may have any number of non-primary classifications. The standard procedure has been to normalize or weight each of the classes listed, which discretizes the vector space and leads to discontinuities in the proximity measures.[5]

[5]A patent with one class would be represented by a vector with 1 in the class column and 0 elsewhere; two classes 0.5 in

Year Group	NAICS Match	S.D.	Primclass Match	S.D.	Inventor Match	S.D.	Direct Citation	S.D.	Year Match	S.D.
1975-85	0.126	0.137	0.187	0.145	0.301	0.148	0.328	0.148	0.126	0.138
1985-95	0.124	0.135	0.186	0.145	0.320	0.163	0.322	0.146	0.124	0.135
1995-05	0.129	0.134	0.196	0.147	0.312	0.158	0.302	0.147	0.129	0.134
2005-15	0.141	0.136	0.200	0.146	0.310	0.170	0.300	0.152	0.141	0.136

Table 2.1: Average DocVecs Similarity for pairs of patents that match on each column. The standard deviation of the similarity in that sample reported in the next column. Samples are partitioned by decade.

Validating Similarity Measures

Prior expectations about patent similarity can be used to validate the vectors generated by the Doc2Vec algorithm. In table 2.1, the baseline group average is the average pairwise similarity for patent pairs from within the same NAICS industry granted within 5 years of one another. We should expect that, on average, similarity between patent pairs of the same primary class should be *higher* than pairs within the same NAICS industry, since industry represents a broader definition of technology field. Table 2.1 shows that patents within the same primary class have average similarity around 1.5 times that of patents just within the same NAICS industry. Patent pairs sharing an inventor have 2.5 times the similarity of the baseline group. Patent pairs that have a direct citation relationship also have a comparable level of similarity to patent pairs sharing an inventor. On the other hand, we should also expect that patent pairs from the same grant year should not have average similarity higher than the baseline, since the time difference between 0 years and 1-5 years is not large enough to have a significant impact on technological difference. Table 2.1 shows there is virtually no difference between average similarity of patents granted in the same year and the baseline. In general, variance is higher in smaller samples such as patent pairs with matching inventors. Since DocVecs captures trends in similarity that matches prior expectations, it is unlikely that results are being driven by noise in the vectors generated by the algorithm.

3. Application of similarity: estimating geographic localization

The similarity measure can be applied in two ways to estimate geographic localization of knowledge spillovers. In the first case, I replicate and extend the work of JTH up to recent years. This standard methodology involves the selection of a control patent that is as close as possible in grant date to the

each class column and elsewhere; and so on.

"target" patent, within the same primary class. I then select different control patents using (i) patent text similarity; and (ii) a patent from the same primary class and the same lawyer. While selecting a control based on similarity does not drastically alter the estimates for localization, selecting on lawyer does significantly diminish localization estimates up to 2005.

In the second case, I look for evidence of localized knowledge spillovers by estimating whether within-technology field patents from the same MSA are more similar than if they are from different locations. The rationale is that if firms and inventors from the same "cluster" (defined as a technology field within an MSA, for example Pharmaceuticals in Philadelphia) are learning more from knowledge generated by each other, then the similarity of patents *within* a cluster should be higher than similarity of patents *across* clusters (for example, similarity of patents from Pharmaceuticals in Philadelphia to Pharmaceuticals in Boston). I find much less evidence of localization when examining patent text similarity.

3.1. First application: selecting different controls under standard citations methodology

I replicate and extend the work of JTH in order to have a baseline estimate of localization effects. JTH sampled patents in their control (target) group in the following manner: from the years 1975 to 1980, they select a random sample of Top Corporate (top 200 by R&D total expenditure measured by Compustat) and Other Corporate patents, and all patents granted to Universities. Their sample size is 950 for 1975 and 1450 for 1980 respectively. Then, for each "target" patent in the sample find a control patent that is as close as possible to the target in *grant date* in the *same patent primary class.* JTH claim that this accounts for the "existing distribution of technological activity," and thus if forward citations are more likely to be from the same geographical area as the target patent over the control, then it is evidence for the existence of localized knowledge spillovers.

I replicate this method using a larger sample of target patents granted 1976-2005,[6] and limit forward citations to be within 10 years of the target patent's grant date. Self-citations of patents granted to the same assignee are similarly excluded. The only point of departure is that due to lack of data, I do not use separate categories of patents by assignee "type", and pool all patents by grant year. Compared to the original JTH results (table III, p. 590), my results are fairly well aligned with their 1980 cohort

[6] 2005 is the last year that 10 year forward citations are available for

figures for top corporate patents. [7]

The next step is to select a different control patent. In the first substitution, I determine a control that is the patent with the *highest similarity* to the target from a different MSA and a proximate grant date. If we interpret patent text similarity as a better reflection of unobserved technological proximity, then this method may provide better control than merely selecting on PTO primary class, as challenged previously by Thompson and Fox-Kean (2005). This approach is in line with previous attempts such as Arts et al. (2018), who also use a text-based similarity measure to select better control patents.

The second substitution follows a separate line of concern. I select a control patent in the same primary class, different primary class, and different assignee to the target (same as in the JTH match), but also from *the same lawyer as the target*. Previous literature such as Moser et al. (2017) has drawn attention to the large effect that patent attorneys play in deciding citations for patent applications. Therefore, localization patterns may be overly influenced by the patent knowledge of lawyers rather than knowledge flows across inventions. If patent lawyers do not have an important role to play in determinizing the localization of patent citations, then further selecting a control that matches on both primary class *and* attorney should not yield very different results for localization, compared to the baseline replication. However, if these results prove significant, this may explain why better measures of technological proximity do not yield lower estimates of localization: citations may be localized in part because lawyers' knowledge of "citable" patents are geographically concentrated, not necessarily because knowledge flows across patents are.

Once a control in each case has been selected, I calculate the percentage of forward citations matching the target's MSA for both the target and control. Under this method, localization is significant if the target patent has more local citations compared to the control. Results for the percentage of forward citations matching the target's MSA under each control selection method is presented below in table 3.1.

3.1.1. Localization under different controls

Control selection using text similarity does improve in accounting for unobserved technological proximity, resulting in lower estimates for localization. However, these effects are still highly significant, with local citations being 0.9-1.4 times higher for the local target patent compared to the non-local

[7] 8.8% for target match and 3.6% for control match; compared to 9.09% and 3.77% for my results. Slight discrepancies may arise due to sample selection and slight differences in removing self-citations.

	1975-85	1985-95	1995-05
Control Selection: Standard JTH			
Target, Pct Cite in Target MSA	0.091	0.097	0.11
Control, Pct Cite in Target MSA	3.8	3.5	4.5
Ratio	2.4	2.8	2.4
p-value	0	0	0
N	58647	107358	185154
Control Selection: Similarity			
Target, Pct Cite in Target MSA	9.2	9.7	11.0
Control, Pct Cite in Target MSA	4.8	4.1	5.1
Ratio	1.9	2.4	2.0
p-value	0	0	0
N	36917	67332	117137
Control Selection: Lawyer			
Target, Pct Cite in Target MSA	9.4	10.0	11.5
Control, Pct Cite in Target MSA	7.5	7.9	8.9
Ratio	1.3	1.3	1.3
p-value	0	0	0
N	22914	51837	85855

Table 3.1: Baseline results for JTH replication under different control selection methods. Each column represents the average percentage of forward citations to the target or control in the target's MSA, for patents granted within a certain decade. Sample sizes vary due to the inability to find control patents under certain methods of selection.

control. Interestingly, once we select a control that is from the same lawyer as the target, localization estimates shrink dramatically. Local citations are now only 0.3 times higher for the local target. These findings confirm the important role of lawyers in determining how localized citations are, which "muddies the waters" of determining the size of local knowledge flows. Hypothetically, there are a number of mechanisms by which lawyers could bias citations towards localization: (i) lawyers operating in a select few cities may cite the patents of their clients, who are likely to be operating in similar technology fields to begin with; (ii) lawyers may cite the patents of other firms and inventors within these cities that they have encountered either through their own networks or other transactions.

The confounding factor is that many lawyers only operate within one city, representing a handful of firms. Thus, it may be difficult to disentangle the localization of the inventor's knowledge flows (which citations should proxy), from the localization of the lawyer's knowledge of patents (which add noise

and bias to the citations measure). Since I find that lawyers representing technologically similar firms across different cities cite significantly more patents in the target's city, this suggests that the size of the bias from lawyers towards localization of citations is not small.

3.1.2. Regression model for estimating localization

The above exercise can be represented as a regression model in the form:

$$pct\,cites\,in\,MSA_{T,i} = \beta_0 + \beta_1 I(MSA_i = MSA_T) + X_i + \epsilon \tag{3.1}$$

Where $i \in \{T = target, C = control\}$. Here, if patent i is the target patent, the indicator $I(MSA_T = MSA_T) = 1$, while for the control patents $I(MSA_C = MSA_T) = 0$.

To account for the potential effect of other variables on citations and localization, X_i represents further controls for the patent i, including year, primary class, MSA, lawyer, and examiner fixed effects.[8] The percentage of citations is normalized prior to the regression, so that$\hat{\beta}_1$ represents the increase in the (standardized) percentage of local citations when the patent is also local. For consistency, I normalize all three samples under each control selection regimes using the standard JTH control sample.

Estimates of localization for fixed effects including year and primary class are presented in table table D.1; for all fixed effects, results are presented in table D.2. The size of the localization estimates are not particularly sensitive to the inclusion of further controls. In the standard JTH control selection, local patents are found to receive 0.23-0.30 S.D.s more local citations compared to the non-local control. This is diminished to 0.20-0.27 S.D.s in the sample where similarity is used to select the control. Finally, selecting on the same lawyer reduces localization estimates to 0.08-0.12 S.D.s.

3.2. Second application: evidence of localization in the proximity of ideas

The findings from the previous section suggest that citations may overstate localization due to the influence of lawyers. But the bias towards localization of citations does not necessarily imply that knowledge spillovers themselves are not localized - it suggests that it may be useful to address the question of localization from a different approach. If knowledge spillovers localized, then this suggests

[8] Only the 100 largest in each category are included to reduce dimensionality in the covariates matrix.

that local firms and inventors learn from each other's inventions. Thus, patents from a particular cluster should express ideas more proximate to each other compared to patents from differing clusters. In table 2.1, patent pairs that have a direct citation relationship are found to have much higher similarity on average, which suggests that direct knowledge flows imply greater proximity of ideas.

Patent text similarity may also be particularly suited to picking up knowledge that were "in the air", tacit knowledge, or common knowledge inputs other than citations. While patent citations reflect knowledge flows from other patents, patent text should reflect the influence of patent, non-patent, and tacit knowledge flows. Ganguli et al. (2017) find a significant role for geographic distance acting as a barrier for such tacit knowledge. In their study of patent interferences, the simultaneous instances of identical invention by two or more independent parties, they find that interfering patents are much more likely to arise from the same geographic location. If, as proposed by Ganguli et al. (2017) and Audretsch and Feldman (1996), tacit knowledge flows are geographically bounded, then we should find even more proximate ideas within a cluster. For further discussion of the relationship between knowledge flows and idea proximity, see section §4.

Sample Construction For patents within the same technology field (either a NAICS industry or a PTO primary class), I sample patent *pairs* within the same MSA (i.e. patents from the same cluster), and patent pairs from different MSAs (across clusters). Patents from the same MSA are slightly over sampled to ensure a sizable number of patent pairs from the same MSA across a range of technological fields. Patent pairs are granted within 5 years of each other and are assigned to different firms. [9]

The use of both industry level and primary class level technology fields allows me to capture potential differences in the dynamics of knowledge spillovers. If knowledge spillovers are more localized for firms within the same industry, then patent text within the same industry cluster (i.e industry-MSA pair) should be more similar. Further, if we expect knowledge to be more specialized across clusters at the industry level, then patent text should be more similar within an industry cluster compared to within a primary class cluster.

[9] While some patent pairs may have the same target patent, the number of appearances made by multiples of the same patent is extremely small relative to the entire sample, thus curtailing the presence autocorrelation. Heteroskedastic-robust standard errors are used in regression estimates.

Proximity of ideas within cluster and across clusters The unconditional sample means for *within cluster* and *across clusters* patent text proximity are reported in table 3.2. While the similarity of patents within the same cluster are higher than patents across clusters, the effects are relatively modest: on average, within cluster patent pairs have text proximity 0.02-0.06 higher than patent pairs across clusters. These results are more closely aligned with the conjecture that citations over-estimate the localization of knowledge spillovers.

A number of important limitations should be considered. First, there is the concern that rather than there being limited localization in idea proximity across patent clusters, document vector similarity itself is a poor measure of idea proximity. Mirroring the analysis in table 2.1, I address this concern in section 2.2.2 below. Second, there may be other factors affecting patent text similarity besides knowledge spillovers. For example, both patent lawyers and patent examiners may affect the abstract text. This is further discussed in section 2.2.3. Finally, there is likely to be simultaneity bias between the location of the patent and the similarity of patent text, if firms from similar technology fields are also more likely to collocate. However, this bias is likely to be positive and implies that the effect of local knowledge spillovers should be even smaller than in the reported results. These concerns can be partially addressed through moving to a regression model framework and including suitable control variables.

Year Group	1975-85	1985-95	1995-05	2005-15
Technology field: NAICS Industry				
Within Cluster, $I(MSAMatch) = 1$	0.127	0.127	0.133	0.145
Across Clusters, $I(MSAMatch) = 0$	0.124	0.12	0.125	0.137
Ratio	1.02	1.059	1.062	1.056
p-value	0.001	0	0	0
N	194131	282112	443885	578056
Technology field: Primary Class				
Within Cluster, $I(MSAMatch) = 1$	0.197	0.193	0.195	0.197
Across Clusters, $I(MSAMatch) = 0$	0.19	0.182	0.184	0.188
Ratio	1.036	1.059	1.063	1.044
p-value	0	0	0	0
N	171893	252886	407176	537878

Table 3.2: Average similarity of patent text within and across clusters. Within cluster implies patent pairs in the sample are from the same MSA, as well as the same technology field. Across clusters implies patent pairs are from different MSAs. All patent pairs are granted within 5 years of each other, and are assigned to different firms.

3.2.1. Regression model for estimating localization in idea proximity

In regression form, I estimate for each technology field sample:

$$sim(i,j) = \beta_0 + \beta_1 I(MSAMatch_{i,j}) + X_{i,j} + \epsilon_{i,j} \tag{3.2}$$

Here $I(MSAMatch_{i,j}) = 1$ if patent i, j are from the same location (i.e. $MSA_i = MSA_j$). Similar to equation (3.1), X_i represents further controls[10] including year, primary class, MSA, lawyer, and examiner fixed effects. I also include other match controls for the patent pair, which may affect the similarity in their patent text: $I(Lawyer\,Match)$, if patents are assigned to firms that share the same lawyer; $I(Inventor\,Match)$, if patents share an inventor (after inventor relocates to different firm); $I(PrimclassMatch)$, if patents are from the same primary class (only for patent pairs within the same NAICS industry). Similarity is also normalized prior to regression. The estimated localization in idea proximity is given by $\hat{\beta}_1$ and represents how many S.D.s more similar patents are from within a cluster compared to across clusters.

	1975-85	1985-95	1995-05	2005-15
Technology Field: NAICS	0.0171***	0.0354***	0.0344***	0.0333***
	(0.0053)	(0.0043)	(0.0034)	(0.0030)
N	192773	280962	437405	563881
Adjusted R^2	0.08	0.09	0.10	0.06
Technology Field: Primary Class	0.0277***	0.0502***	0.0531***	0.0395***
	(0.0066)	(0.0052)	(0.0039)	(0.0033)
N	170564	251218	400729	518334
Adjusted R^2	0.07	0.08	0.08	0.06
Controls: Year, PC, MSA, Examiner, Lawyer Match and FEs				

Table 3.3: Regression estimates for the localization of idea proximity, which represents how many S.D.s more similar patents within cluster (location-technology field) are compared to across clusters. Standard errors of estimates are reported in parenthesese below. A separate sample is computed for each definition of technology field at the industry and primary class level.

3.2.2. Validation: is similarity a noisy estimate for idea proximity?

One concern may be that instead of geographic proximity being a weak determinant of patent proximity, in fact similarity is not a good measure of patent proximity due to noise. I address this concern by seeing if other match variables are strong determinants of similarity across patents. If similarity

[10] Fixed effects are for patent i only, to reduce dimensionality. Match effects depend on both patents.

is a poor indicator of patent proximity, then we should expect matching on these dimensions to also produce small and possibly imprecise estimates of their effects. Using prior expectations, I expect the following match variables to have a significant effect on patent similarity for patents within the same NAICS industry:$I(InventorMatch)$, if patents share an inventor (after inventor relocates to different firm); $I(PrimclassMatch)$, if patents are from the same primary class; and $I(CommonCited \geq 1)$ to indicate the presence of at least one common cited patent between the pair.

Results for the estimated effect of matching on other variables are reported in table 3.4. I find that estimates for these other match effects are large and significant. The estimated effect of sharing an inventor increases similarity by 1.27-1.52 S.D.s Sharing a common cited patent increases text similarity by 0.84-1.51 S.D.s. Note that the rise in citation rates in recent decades has meant that sharing a common cited patent has declined in effect on DocVecs similarity. Patents from the same primary class have 0.40-0.44 S.D.s higher text similarity. Thus, it is reasonable conclude that similarity is not a noisy estimate for idea proximity, as it is able to pick up on the proximity of patent text across differing dimensions. The effect of matching on the same location cluster may indeed be small.

	1975-85	1985-95	1995-05	2005-15
$I(InvMatch)$	1.2789***	1.5206***	1.3817***	1.2664***
	(0.1042)	(0.0713)	(0.0559)	(0.0563)
N	192841	281222	437685	569252
Adjusted R^2	0.07	0.07	0.08	0.06
$I(CommonCited \geq 1)$	1.5084***	1.2870***	1.1149***	0.8388***
	(0.0963)	(0.0637)	(0.0398)	(0.0254)
N	192841	281222	437685	569252
Adjusted R^2	0.07	0.07	0.08	0.06
$I(PrimclassMatch)$	0.4413***	0.4449***	0.4291***	0.4045***
	(0.0085)	(0.0066)	(0.0050)	(0.0042)
N	192841	281222	437685	569252
Adjusted R^2	0.08	0.09	0.10	0.07
Controls: Year and PC FEs				

Table 3.4: Effect of matching on other variables for patents within the same NAICS industry. Estimates are the increase in S.D.s of similarity when matching on each variable is true. Standard errors are reported below in parentheses.

3.2.3. Validation: how do lawyers and examiners affect patent text?

Another related concern is that patent text is also subject to the external influence of lawyers and examiners. Related to the exercise in the above section, I check for their effect by seeing how much text similarity increases when two patents have the same attorney or were processed by the same examiner. However, lawyers and examiners are not assigned randomly: they are both either selected or assigned in a manner correlated with the technology field of the patent. Therefore, omitting the effect of technological proximity across patents may overstate the influence of lawyers and examiners on patent text. I control for technological proximity using equation (2.4), that is, by calculating the mean similarity of prior patents in each patent's respective primary class. Note that in the regression results presented in table 3.3, matching on lawyer and examiner has already been included as a control. These estimates are to ascertain how large of a role examiners and lawyers have to play in shaping patent text.

Results are presented in table 3.5. Prior to controlling for technology proximity and all other fixed effects, patent pairs from the same lawyers and examiners do indeed have substantially higher text similarity. However, after controlling for the proximity in the patents' respective technology fields, the effect of lawyers and examiners on patent text similarity does decrease substantially. Lawyers, echoing results in section 3.1, have a much larger effect on similarity: patents from the same lawyer have 0.14-0.24 S.D.s more similar text than patents from different lawyers, even after controlling for all other effects. The text of patents processed by the same examiner have 0.08-0.16 S.D.s higher similarity; it appears that their effect has declined in the last three decades.

These results indicate that including technology proximity in primary class would attenuate some biases towards higher similarity in patent text due to the influence of external parties.

	1975-85	1985-95	1995-05	2005-15
Controls: Year and PC FEs				
$I(Lawyer\ Match)$	0.2720***	0.3721***	0.4429***	0.3694***
S.E.	(0.0366)	(0.0281)	(0.0299)	(0.0276)
N	192773	280962	437405	563881
Adjusted R^2	0.06	0.07	0.08	0.05
$I(Examiner\ Match)$	0.4571***	0.4867***	0.4170***	0.4158***
S.E.	(0.0184)	(0.0165)	(0.0186)	(0.0213)
N	192773	280962	437405	563881
Adjusted R^2	0.07	0.07	0.08	0.05
Controls: Tech proximity and all other FEs				
$I(Lawyer\ Match)$	0.1884***	0.1460***	0.2362***	0.1811***
S.E.	(0.0468)	(0.0256)	(0.0271)	(0.0260)
N	102330	280954	437386	563865
Adjusted R^2	0.11	0.12	0.12	0.08
$I(Examiner\ Match)$	0.1334***	0.1555***	0.0897***	0.0769***
S.E.	(0.0249)	(0.0161)	(0.0181)	(0.0205)
N	102330	280954	437386	563865
Adjusted R^2	0.11	0.12	0.12	0.08

Table 3.5: Effect of matching on lawyer and examiner for patents within the same NAICS industry. Estimates are the increase in S.D.s of similarity when matching on each variable is true. Standard errors are reported below in parentheses.

3.2.4. Including technology proximity in estimating localization of idea proximity

For patents within the same NAICS industry, including prior similarity across primary classes may be able to address some concerns about other factors that may raise text similarity, besides knowledge spillovers. It may also partially address the simultaneity bias between text similarity and collocation. In table 3.6, the estimate of localization diminishes further with the inclusion of primary class similarity as a control for technological proximity: the estimate of localization ranges from insignificant to 0.04 S.D.s above the mean; the interaction effect is insignificant at the 5% level across all decades.

	(1)				(2)			
	1980-85	1985-95	1995-05	2005-15	1980-85	1985-95	1995-05	2005-15
$I(MSAMatch)$	0.0170	0.0390***	0.0300***	0.0274***	0.0178	0.0415***	0.0277***	0.0222***
	(0.0120)	(0.0070)	(0.0048)	(0.0038)	(0.0120)	(0.0069)	(0.0048)	(0.0043)
$I_{MSA} * sim_{DV}(pc_i, pc_j)$					-0.0012	-0.0045	0.0050	0.0085*
					(0.0116)	(0.0067)	(0.0047)	(0.0045)
$sim_{DV}(pc_i, pc_j)$	0.2805***	0.2913***	0.2816***	0.2932***	0.2808***	0.2925***	0.2804***	0.2908***
	(0.0090)	(0.0055)	(0.0040)	(0.0036)	(0.0095)	(0.0057)	(0.0042)	(0.0039)
N	40323	110982	215861	344313	40323	110982	215861	344313
Adjusted R^2	0.12	0.13	0.13	0.08	0.12	0.13	0.13	0.08
Controls: technology proximity and all other controls								

Table 3.6: Estimates of localization in ideas proximity including technology proximity and all other controls for patent pairs in the same NAICS industry. Due to a lack of patent data prior to 1976, technology proximity is only available 1980 onwards. Equation (1) uses primary class similarity as a separate control; (2) includes interaction effects with the location match indicator.

3.3. Discussion of findings

I applied similarity to the examination of localized knowledge spillovers in two ways. In the first application, I replicate JTH's original methodology and use text similarity to find a control patent. While estimates of localization in citations do diminish, suggesting that selecting a control using similarity does a better job of addressing unobserved technological differences in the target and control patent, I also find that localization in citations diminish substantially more when selecting a control from the same lawyer. This complicates the validity of the experiment in identifying the localization in knowledge spillovers, as it indicates that lawyers' knowledge of patents may drive the geographic concentration in citations. As an alternative test, I investigate the localization in idea proximity of patent text: examining whether patents from within a cluster (the same technology field and city) are more similar than patents across clusters. I find that localization estimates are weak, suggesting that citations may in fact overstate localization in knowledge spillovers.

4. Discussion of patent text similarity

The results from the previous applications showed that knowledge relationships may appear quite different under examination using the proximity of ideas versus standard citations measures. In light of those findings, I discuss some limitations and implications of these results. The discussion in this section may also help applied researchers who want to further understand what text and text similarity

may reflect.

4.1. Relationship between text similarity and knowledge

High similarity does not reflect direct knowledge flows The results from the previous section relied on the argument that if a group of patents have a high incidence of shared knowledge flows, then the similarity of text *within such a group* should be high. For example, patent pairs that have a direct citation relationship will have higher similarity *on average* (table 3.4). However, an *individual* patent pair with high similarity does not imply that a direct knowledge flow has taken place. Inferring the presence of localized knowledge spillovers using text similarity is appropriate as it relies on comparison across aggregate or group means. If accurate indications of direct knowledge flows for individual patent pairs are required for the research question, for example if the focus was on patent and inventor networks, then citations may still be the more appropriate measure. Because text vectors are determined algorithmically, similarity of individual pairs may not all be well-aligned with human judgment. Thus, similarity may be better suited to assess differences in samples of similarity. From the citation replication exercise in section 3.1, the controls that were selected improved on the standard method, which shows that similarity in patent text vectors can be useful for some matching exercises.

Can high local knowledge spillovers lead to weak local idea proximity? Is it possible that local knowledge flows may still lead to dissimilar local ideas? This may be true in some cases where learning about the innovation agenda of local rivals may lead others to differentiate their inventions. However, it is still the case that *on average*, a patents that have a direct citation relationship have high similarity, which is to say that patents taking knowledge from other patents should still express more proximate ideas. This does imply that it may not be possible to determine whether low similarity patent pairs have differentiated their inventions, or are simply unrelated.

Interpreting weak local idea proximity One way to interpret the findings of weak local idea proximity is that inventors may be directly influenced by very few other local inventors. This may support the literature on microgeography, which suggests that "what appears to be a cluster at the county level may indeed be several geographically (and often technologically) distinct clusters, each with different social relationships and unique needs." (Feldman (2014)) The question is whether or not this

implies localization at the city level is high or low. As a simple example, consider inventor i from city A that has 100 inventors all operating in the same technology field. Suppose inventor i exchanges knowledge with 5 other inventors local to city A and 5 inventors from other cities. If we were to examine the proportion of inventor i's influences in city A, then localization from this perspective would be large at 0.5. However, if we were to examine the proportion of inventor i's influences relative to all other possible influences in city A, localization would be much smaller at 0.05. Thus, the choice of denominator may decide the magnitude of localization effects.

Another way to look at it would be whether or not "counter-evidence" of knowledge spillovers are important to consider, that is, the lack of knowledge flows where they should exist. Patent citations and other approaches such as interference are more or less silent about this counter-evidence, whereas similarity (which can be generated for any two patents) captures the presence of low proximity in local ideas. Ultimately, it may be left up to researchers to decide whether or not this counter-evidence is important. Even in the case when it is not, similarity can be useful in identifying which knowledge relationships can be disregarded. In terms of findings on the localized knowledge spillovers, even if similarity is considered too "broad" of a measure to accurately capture relevant knowledge flows, the evidence in section 3.1 still suggests that localization is likely overestimated using citations.

Patent texts reflect non-patent knowledge influences Continuing with the discussion from section 3.2, another reason why localization in idea proximity may be weak is that inventors are highly influenced by other sources knowledge, which may be commonly available and thus non-local (for example, knowledge acquired from the internet). Other sources of knowledge include: academic and scientific publications, textbooks, technical journals, and less concrete examples such as tacit knowledge, background knowledge, and knowledge "in the air". Recent literature by Arora et al. (2018); Ganguli et al. (2017) and Packalen and Battacharya (2015) emphasize the role that external knowledge sources also play in the innovation process; which is to say, other patents are by no means exhaustive of relevant knowledge required for new inventions. For example, even in the writing of this paper, multiple other concurrent papers appeared that also utilized text analysis methods which were made widely available through internet-based learning tools and resources such as Coursera and Stack Overflow.

One possibility is that inventors rely extensively on external knowledge in their ideas, which citations-based studies cannot capture, but is reflected in patent text. As more knowledge becomes easily and

Average Similarity				
	1975-85	1985-95	1995-05	2005-15
$I(Common\,NPC \geq 1) = F$	0.293	0.280	0.258	0.250
$I(Common\,NPC \geq 1) = T$	0.426	0.401	0.382	0.548
p-value	0.091	0.000	0.000	0.000
t-value	1.692	5.217	12.092	189.720

Table 4.1: Average text similarity of patents that share at least one common non-patent citation ($I(Common\,NPC \geq 1) = T$), compared with patents that do not ($I(Common\,NPC \geq 1) = F$). Because random patent pairs have very sparse citation relationships, this uses a sample of patents that already share at least one common patent backward citation, which is why the baseline comparison group already has a high level of similarity.

commonly accessible through the internet, we may see that the use of external knowledge "homogenizes" innovation across locations. Marshall (1920) even suggested that geographic location would matter less through "cheapening... means of communication." This effect may be significant even if inventor networks are local or microgeographic clusters exist. In table 4.1, I find that similarity for patents that share non-patent citations (available from PatentsView) are much higher, which supports the claim that patent text reflects the knowledge flows from external sources.

Another example is introduction of new terms into the patent corpus. As new technology are developed, references make their way into patents. Patents that are the first to contain the new term are assumed to share some external sources of knowledge about the new technology. table E.1 show that patents introducing new terms rarely cite any backward citations in common, but do exhibit some similarity in their text. For example, "Adenovirus" is a term for a virus that causes many common infections, particularly respiratory illness. With the development of gene therapy technology in the early 1990s, the first patent applications containing the term adenovirus appeared in 1993. Gene therapy delivers "correct" genes inside affected cells, and adenoviruses are often used as carriers for the corrected genes. In 1993, thirteen adenovirus patents were applied for that were later granted.[11] While all adenovirus patents apparently utilised some common external knowledge sources, the average number of backward citations that was shared was 0.03, which represented an average of 0.0% of backward citations made.

However, if the influence of external knowledge is not desirable in all cases. It is up to the researcher to decide if incorporating these influences is appropriate for the research question. Patent text provides a new window into knowledge relationships, but one that may be too wide for certain

[11] Failed applications are not accessible via the USPTO.

applications.

Drawbacks of patent abstract text The main drawback with using the abstracts of patents may be written to be intentionally general or vague, and may not express the core content of the patent, as examiners focus on assessing claims text when deciding patentability. However, this may also be something of a "reverse advantage" for abstracts, as they are not overly scrutinized and potentially altered by both lawyers and examiners. Future work will incorporate the use of top claims text as well as abstracts. If claims texts are more specialized in the language that is used, then we might expect slightly different patterns of knowledge relationships to emerge.

4.2. Policy Implications

My findings may also have important implications for R&D policy. Broadly speaking, there are three avenues for public support of private research: 1. funding for regional clusters; 2. funding for nation-wide industry; 3. funding specific firms or institutions. In recent years, funding for regional clusters received a signficant boon in support from the Obama administration, who spent $225 million on regional cluster projects as of April 2012 Chatterji et al. (2014). The largest of these was the Energy Regional Innovation Cluster in Philadelphia, which received $122 million in funds from the Department of Energy. Local efforts include notably the Boston Waterfront Innovation District, launched in 2010 by Mayor Thomas Menino (Mehta et al. (2012)). Nation-wide industry R&D funding include support for renewable energy and nanotechnology (the National Nanotechnology Initiative was launched in 2001). While both efforts are underpinned by some economic rationale of generating positive externalities, the *kind* of externality differs: positive geographic externalities leading to better local economic outcomes for regional clusters, and positive social externalities leading to better population health for industry R&D funding.

The implications from my findings may be that, from the perspective of generating new innovation, the two forms of R&D funding are roughly equivalent. If there are less positive externalities generated by geographic proximity, then research boosted by local funds could be more easily appropriated by firms in the same technology field in other locations. On the other hand, this may weaken the argument for regional policy if it is intended to boost its own relative advantage in innovation. How much this evidence hampers enthusiasm for regional cluster policy depends on whether or not policymakers consider innovation-embodied knowledge spillovers to be a first order concern. Indications are

that it is not; policymakers are usually much more focused on outcomes such as employment growth and rates of entrepreneurship.

The lower rate of geographically localized knowledge spillovers may be one possible contributing factor to the difficulty in "governmentally inducing" innovation. While notable successes such as SIlicon Valley and Route 128 in Boston receive considerable attention, many failed cases do not (Lerner and Seru (2015)). My findings show that firms are less inclined to share commercially appropriable innovative knowledge with local competitors. This may particularly affect the innovative capacity of firms in artificial "hubs" which lack the benefits from other forms of shared knowledge. If local knowledge spillovers are not particularly conducive towards innovation, then other agglomeration externalities such as shared labour pools and shared inputs may have particular importance in ensuring the survival of regional hubs. Unfortunately, these other factors may be much more difficult to engineer through policy interventions. Aritificial innovation hubs which rely heavily on knowledge spillovers generated through the presence of proximate firms may face challenges to success without the presence of other agglomerative benefits.

A separate but related literature examines the contribution of public funding for university research on innovation. I conjecture that if universities lack incentives to commercially appropriate their own research, it may indeed provide positive externalities to local firms engaged in similar innovation areas. Thus, one way that regional cluster policy could boost their relative advantage could be to also allocate funds for local university research.

5. Conclusion

This paper focuses on knowledge dynamics of ideas embodied by patent text. I contribute methodologically to the literature on text analysis of patents by using an unsupervised machine learning approach in generating vector representations of patent text. I also offer an alternative lens to examining knowledge and find that different pattens of localization of knowledge spillovers are uncovered when examining patent citations vs idea proximity. These findings add nuance to our understanding of localized knowledge flows and highlight the possible importance of common and non-patent knowledge in generating innovation.

In further research, I address the question of whether Marshall-Arrow-Romer spillovers (the concentration of technologically similar firms within a city) facilitates greater innovation compared to

Jacobs spillovers (the presence of technologically diverse firms within a city). Patent text may provide a unique contribution to assessing the diversity of innovative knowledge, which may otherwise be difficult to measure. Text similarity (or rather its reverse, text distance) may also be particularly suited in identifying "novel" ideas (similar to Kaplan and Vakili (2015) and Kelly et al. (2018)). This would be a useful tool for questions assessing incremental vs radical innovation, as text similarity can easily identify inventions that are very close to prior inventions, and those that are distinct.

As shown in section 3.1 and discussed in Arts et al. (2018), patent vectorizations provide a powerful alternative to USPC classification in assessing technological relatedness across patents. This could generate more accurate technological clusters or neighbourhoods of patents. Such an application would be beneficial in analyzing the effect of intellectual property rights on cumulative innnovation, as in Galasso and Schankerman (2014), in providing more precise indicators of growth or decline within a technological cluster. Another application could be a different angle on a similar question related to agglomeration and microgeography, by identifying the extent to which similar patents cluster within a location, related to Catalini (2017). This methodology and much of the discussion could also be easily applied to scientific and academic texts, in the examination of collaboration (Johri et al. (2011)), the effect of patents on science (Murray and Stern (2007); Azoulay et al. (2009)), and the evolution of scientific knowledge (Azoulay et al. (2011)). I believe the tools and methods discussed in this paper would be beneficial to any researcher of knowledge and innovation.

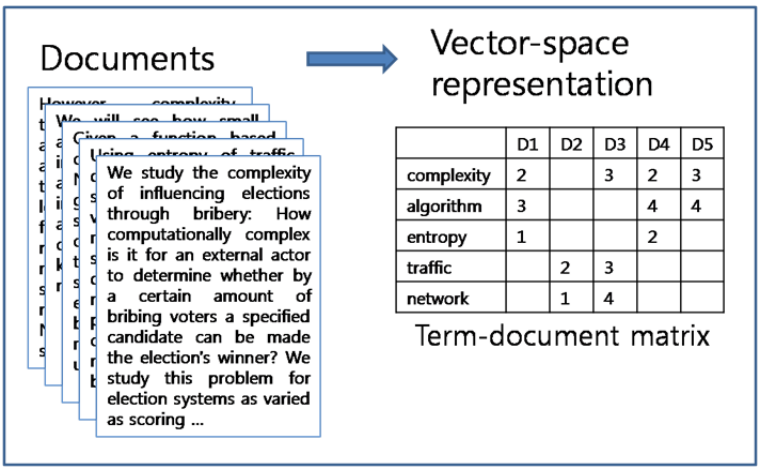
Appendix

A. Text to Data

A.1. Text cleaning

Each abstract is stemmed to the root word (for example, computer to comput), and stop words (such as “and”, “the”) are removed. The first step in converting text to data is to represent words and documents in their simplest vector forms. For all algorithms besides Document Vectors, input into the algorithms involve the construction of a document-term matrix from all patents; each row is indexed by the document ID and each column represents a word in the vocabulary. A document row vector represents the count of the number of times the term appears in the document. For the terms, I drop all terms that appear in more than 10% of all patents, and those that appear in fewer than 20.[12] Of the resulting terms, I keep the most common 40,000, in order to maintain a manageable matrix dimensionality. Once all 2,306,041 patents have been transformed into a document-term matrix of dimension 2306041×40000, I proceed to transforming patents into a smaller dimensional vector representation using the methods described below. This procedure is commonly called the *bag-of-words* representation of text data.

[12] Including very common and very infrequent terms may introduce noise and considerable increases in computation times.



	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

Figure A.1: Example of Document Term Matrix

A.1.1. Paragraph Vectors (Doc2Vec)

One recent advance in NLP which utilises neural networks is Paragraph Vectors, introduced by Le and Mikolov (2014). This is a straightforward extension of the word2vec model of Mikolov et al. (2013b,a). The word2vec model attempts to rectify one of the well-known problems of NLP: the inability of "one-hot" word vectors to account for word similarity. Typically, word vectors are represented as sparse vectors. For example, in a complete vocabulary of ["good", "fair", "fine"], the word *good* would be represented as the vector $[1,0,0]$, *fair* as $[0,1,0]$ and *fine* as $[0,0,1]$. Clearly, each of these vectors are orthogonal to each other and have a similarity of 0. Instead of using this class of word vectors, word2vec tries to represent words as dense vectors that encode such similarities; a word2vec vector for each of the three words ["good", "fair", "fine"] will have a *high* similarity.

The way that this is done is through looking at the *context* of a word. For example, for the sentence "Provides for unattended file transfers", the word "unattended" has the context ["Provides", "for", "file", "transfers"]. We want to represent each of these words as a vector of arbitrary dimension n. One way to account for context is to predict the context words given the target (Skip-gram); while another way is to predict the target word given the context (Continuous Bag-of-Words). Under Skip-gram, the optimization problem is to maximise the probability of any context word given the current center word. So the objective function is given by:

$$J(\theta) = -\frac{1}{V}\sum_{t=1}^{V}\sum_{-m\leq j\leq m}\log p(w_{t+j}|w_t) \tag{A.1}$$

Where θ represents all parameters: input vector ("one-hot") representation of each word, and the output word2vec representation of each word. m represents the length of the context window; for example $m=1$ gives the context for "unattended" as ["for", "file"]. The objective function is minimized using stochastic gradient descent.

Paragraph Vectors, or Doc2Vec, extends word2vec merely by adding an additional variable, which will be treated as an additional context vector: paragraph ID. For my data, this will be the patent number, which uniquely identifies every abstract document. Thus, including paragraph ID as an additional word for each context generated from that paragraph will also generate a unique vector associated with the paragraph, as well as the word vectors. Intuitively, the paragraph vector will represent what was learned in other context windows belonging to the paragraph, outside of the present context window: that is, it "acts as a memory that remembers what is missing from the

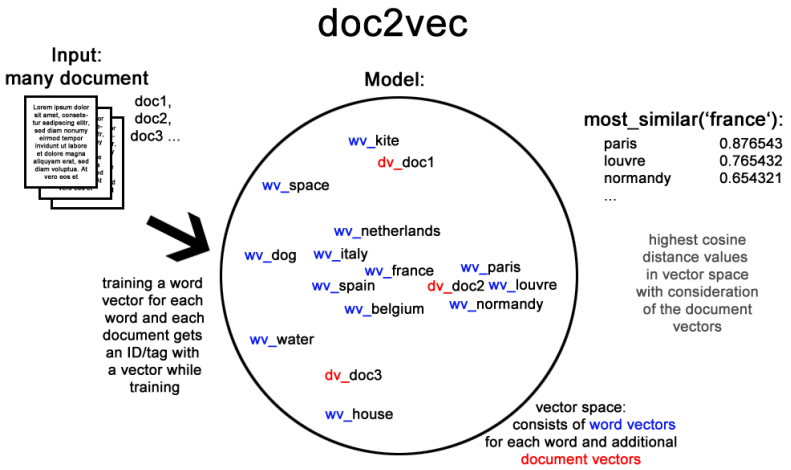


Figure A.2: Illustration of Document Vectors.

current context." (Le and Mikolov (2014))

Such an approach has been shown to be extremely powerful in accurately capturing cross-word and cross-document similarity (papers?), which is why it is the main focus of my analysis. Other vector representations of patents that I use do not specifically optimize to capture such similarity using contexts.

A.1.2. Latent Dirichlet Allocation

Latent Dirichlet Allocation, first introduced by Blei et al. (2003), is a method of Topic Modelling that assumes that a document can be represented as a linear distribution hidden variables called *topics*. It is a Hierarchical Bayesian hidden variables model. The Data Generating Process assumes that each topic is a linear distribution over terms in the corpus. For each document, which is a distribution over topics, each term is assumed to be generated by first drawing a topic, then drawing a term from that topic. Because this is an unsueprvised method, the algorithm then jointly determines the topics distribution over terms and each document's distribution over topics. See A.1.2 for more details on the assumptions of the LDA model. table A.1 shows a breakdown of selected topics' distribution over terms. figure A.3 provides an example of the input and outputs of the algorithm

The number of topics K is a parameter that is determined ex-ante; as per Hoffman et al. (2010), the recommendation is that the model with the lowest log perplexity be selected, although there is not a universally agreed upon procedure. I fit a LDA model on a training subset of the same document-term matrix representing all patent abstracts with $20, 30, ..., 120$ topics. Then, the model was fit on the test

Topic	Distribution over terms	Description
0	0.040*"network" + 0.039*"inform" + 0.033*"comput" + 0.031*"communic" + 0.028*"user" + 0.027*"memori"	Networks & Coding
2	0.066*"time" + 0.057*"sensor" + 0.040*"detect" + 0.032*"event" + 0.031*"paramet" + 0.027*"level"	Monitoring & Coding
11	0.116*"power" + 0.068*"voltag" + 0.049*"output" + 0.045*"circuit" + 0.026*"suppli" + 0.026*"transistor"	Electronics
36	'0.071*"composit" + 0.059*"polym" + 0.049*"weight" + 0.041*"coat" + 0.018*"resin" + 0.016*"c"	Polymers, Chemicals
53	'0.065*"metal" + 0.065*"solut" + 0.037*"ion" + 0.036*"carbon" + 0.032*"concentr" + 0.023*"reaction"	Metals, Chemicals

Table A.1: Selected Topics as outputted by LDA. Description added post hoc.

set and the log-perplexity calculated. I selected $K = 60$ as it had the lowest log perplexity across the models.

A snippet from the resulting topics is shown in A.1, alongside the six highest probability terms in teach topic. The output I am interested in is the probability across each of the 60 topics of each patent document. I take this as the Topic Model vector representation of each patent.

Data generating process With probabilistic models, treat observations as outcomes of a data generating model and infer the hidden parameters of that model using posterior inference. Define a "topic" as a discrete distribution over a fixed vocabulary. Assume each topic is generated by drawing a distribution over terms in the vocabulary represented by the vector: $\boldsymbol{\beta_k} = (\beta_{k,1}, ..., \beta_{k,V}) \sim Dir(\eta)$. Additionally, asssume that each document d is generated by the following process:

1. Draw a vector distribution over topics: $\boldsymbol{\theta}_d = (\theta_{d,1}, ..., \theta_{d,K}) \sim Dir(\boldsymbol{\alpha})$
2. For each word $w_{d,n}$:
    a) Draw a topic $k_{d,n} \sim Multinomial(\boldsymbol{\theta}_d)$
    b) Draw a word based on that topic's distribution over the vocabulary $w_{d,n} \sim Multinomial(\boldsymbol{\beta}_{z_{d,n}})$

Then the posterior of the hidden variables, conditional on the observed words in each document, is given by:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D} | w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})} \tag{A.2}$$

An inference algorithm is used to approximate the posterior. Thus, from the observed set of V vocabulary terms $w \in 1, ..., V$, the hidden topics $k \in 1, ..., K$ (a distribution over words in the vocabulary), and each document's distribution over topics $(\theta_{d,1}, ..., \theta_{d,K})$ are derived.

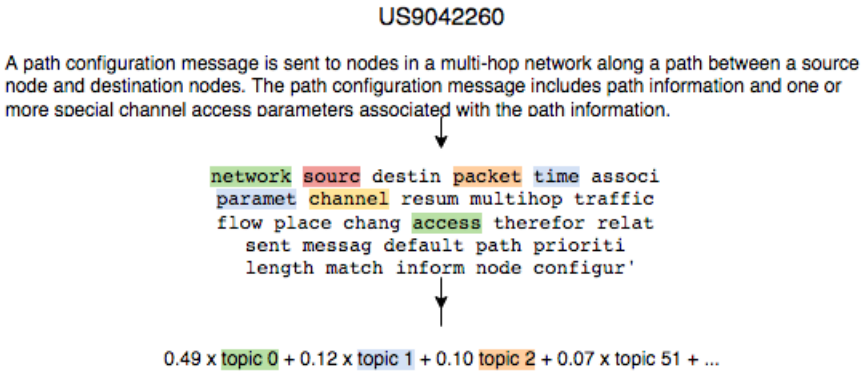


Figure A.3: Example of a patent converted into a distribution over topics.

B. Application of similarity: assessing the quality of citations

The validity of citations as a measure of knowledge spillovers are challenged by the existing literature. It has been widely used in practice because, until now, another such measure has not been available. The problems with using patent citations to proxy for knowledge flows have been well documented. The two dominant concerns are: (i) many citations added by external agents (either law firms or patent examiners), which obfuscates the relationship between the patent and citation as a direct knowledge "flow"; (ii) there are strategic reasons for withholding relevant citations. Namely, citing patents that are closely proximate to the invention limits the scope of the patent and thus reduces the value of the intellectual property. These effects can result in substantial measurement error: Alcacer and Gittelman (2006) find that on the average patent, two-thirds of citations are added by the examiner, while Cotropia et al. (2013) find that applicant citations are often ignored by examiners who conduct their own search of prior art. Citations are also strategic in that, according to Jaffe and De Rassenfosse (2017), "although applicants at the USPTO have a duty to disclose what they know, they have no duty to search for prior art and may be better off by remaining ignorant." Inventors seeking to maximise the value of their IP may be inclined to leave out the most relevant citations; Lampe (2012) finds that applicants withhold between 21% to 33% of relevant citations, as determined by the applicant firm's previous citations. Using a survey of lab managers, Roach and Cohen (2013) also find that patent citations are more reflective of a firm's appropriability strategies in ways that are not revealing of "true" knowledge flows.

In addition, concern over the possibility of patent litigation can potentially lead to a rise in spurrious citations. Lerner and Seru (2015) discuss tactics used by practitioners to offset the likelihood of

lawsuits: "...patent lawyers sometimes urge weak applicants to employ the "kitchen sink" approach to citations: to cite a wide variety of prior art, burying the relevant stuff under a mountain of irrelevant prior art in the hopes that the time-pressed examiner will not discover it." The combination both the incentive to omit highly relevant citations through either wilfull ignorance or strategy and the inclusion of irrelevant citations further casts doubts on the ability of citations to accurately proxy for knowledge flows.

It is also possible that these incentives drive up the measure of localization using citations. If firms are concerned that the probability of infringement discovery by rivals in the same city are more likely, this may induce a greater rate of citation for local firms. Lin et al. (2014) indeed find that patent interference claims occur more frequently between inventors located close together. The omission of relevant patents located elsewhere may further be defensible through both the defense of plausible ignorance and the lower probability of infringement discovery.

Patent vector similarity may not be subject to the same criticism. Because patent abstracts must be accurate summaries of the invention at hand, this limits the ability of applicants to omit important technological terms in order to hide the relevance of previous knowledge. Legal considerations could still play a role in determining how inventions are described: it is likely that applicants may choose words to distance their inventions from a handful of closely related patents. However, since similarity can be determined for *any* pair of patents, the ability for applicants to internalize their choice of terms relative to the entire patent corpus is limited. On the other hand, applicants have complete choice over their list of relevant prior art, which are difficult to hold accountable to an external criteria of accuracy. The authority of the patent examiner to make additions to citations list is precisely a measure enacted to counteract this problem.

B.1. Effect of external influences on citations

Evidence on the declining relevance of citations

I find evidence that such external influences do play a role in determining both the level of relevance of backward citations (i.e. patents cited by the applicants) and the potential omission of relevant citations. It has been well documented that patent litigation has been rising over time Marco et al. (2017). The number of backward citations (excluding self-cites) made by new patents has also in-

creased, more than doubling from 2.3 to 6.0 over the period 1985-2015 (B.1).[13] Meanwhile, the average similarity of patents to their backward citations has declined (from 0.28 to 0.25, B.4) as well as the percentage of citations made to patents in the same primary class (54.1% to 34.4%, B.3). The decline in similarity to citations is robust across citations from (i) the same and different primary class; (ii) the same and different cities (see B.5,B.6). Taken together, these trends would indicate that the relevance of citations have been diluted by the addition of less related citations. However those made to patents within the same MSA has increased, although not consistently over the period: the share of local backward citations rose from 9.3% in 1985 to 12% in 2015.

Evidence of external influences on rate of local citations

Sample construction I examine the possibility of strategic omission of relevant citations using a dataset of "potentially citeable" patent pairs. I sample of set of *target* patents and find a complete list of their backward citations. For each backward citation, I find all their forward citations: each target patent is then matched with another such forward citation, granted *after* the target. Thus, each target is matched with a patent that has a backward citation in common, so that the target is "potentially citable" by the matched patent. I then calculate cross-patent similarity for each pair. To prevent noise from bins with few observations[14], the lowest bin includes all values below, and the highest bin includes all values above. Over 2.4 million pairs of similarities are calculated.

Evidence of strategic omissions In the absence of strategic motives, the rate of citation should be increasing monotonically with similarity between patents. Greater similarity between the texts of two patents should indicate greater potential relevance. Overall, I find that the rate of citation is *not* increasing monotonically with similarity; the rate of citation in fact declines for patent pairs that have the highest level of mutual similarity. While 6.3% of target patents are directly cited when their similarity ranges between 0.5-0.6, only 4.2% are cited for similarity 0.6+. To account for technology differences, I find that this trend also holds for patent pairs within the same primary class: 7.8% of target patents are directly cited when the patent pairs have similarity between 0.5-0.6, and only 4.6% when similarity is 0.6+. (See B.7,B.9,B.1) In fact, the only sample group for which the rate of citation *does* increase monotonically is for patent pairs in the same city, which confirms the lack of incentive

[13] To avoid truncation bias, only citations granted within the previous 10 years of the new patent were counted.
[14] Below the 1st percentile and above the 99th

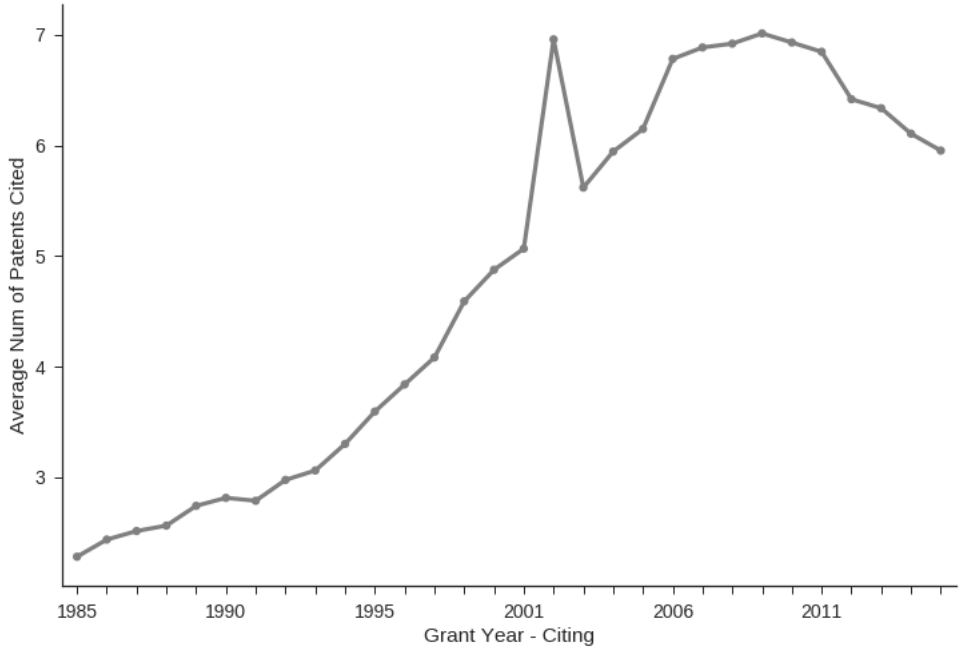


Figure B.1: Averate number of patents cited over time

for strategic omission (B.8). This is contrasted by the stark decline in the rate of citation for patent pairs from different cities with the highest similarity: while 6.3% of target patents are cited when similarity is 0.5-0.6, only 2.2% are cited when similarity is 0.6+. For patent pairs in the same city, the rate of citation increases from 6.3% to 7.5%. Interestingly, the convergence of the rate in citation up to the 0.5-0.6 bracket might indicate diminishing strategic incentives to omit non-local patents as patents become more similar, but the divergence in their citation rates for patents with the highest similarity strongly indicates that firms are strategically leaving out the most relevant citations to patents from other cities. Local patents also over-represent less relevant citations, as the citation rate for pairs with lower similarity are consistently higher for local patent pairs. These findings taken together provide evidence that external influences on the selection of citations tends to favour local citations overall.

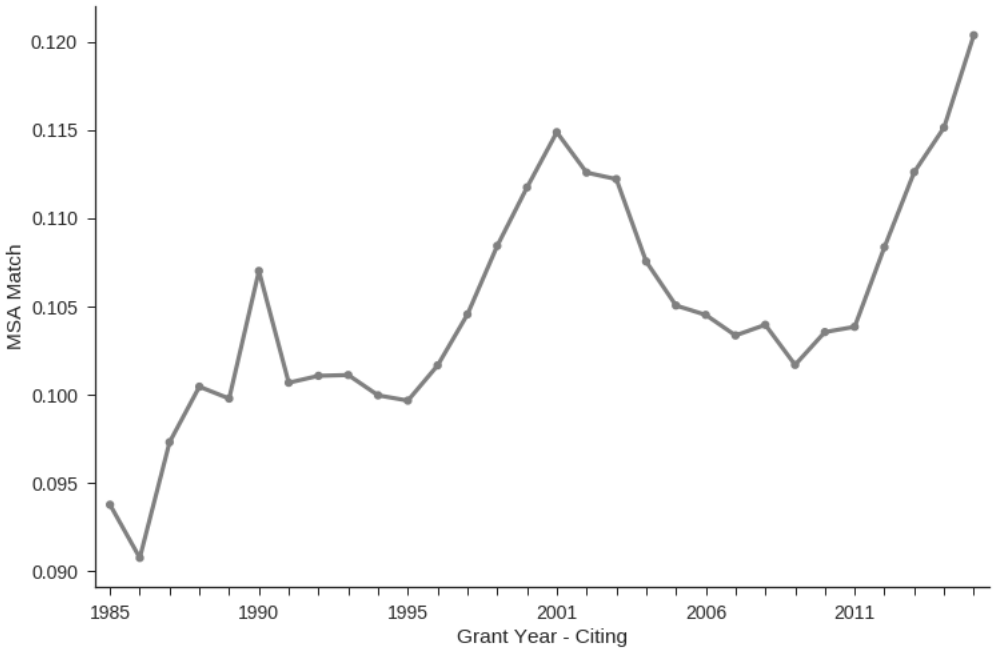


Figure B.2: Proportion of cited patents in the same MSA over time

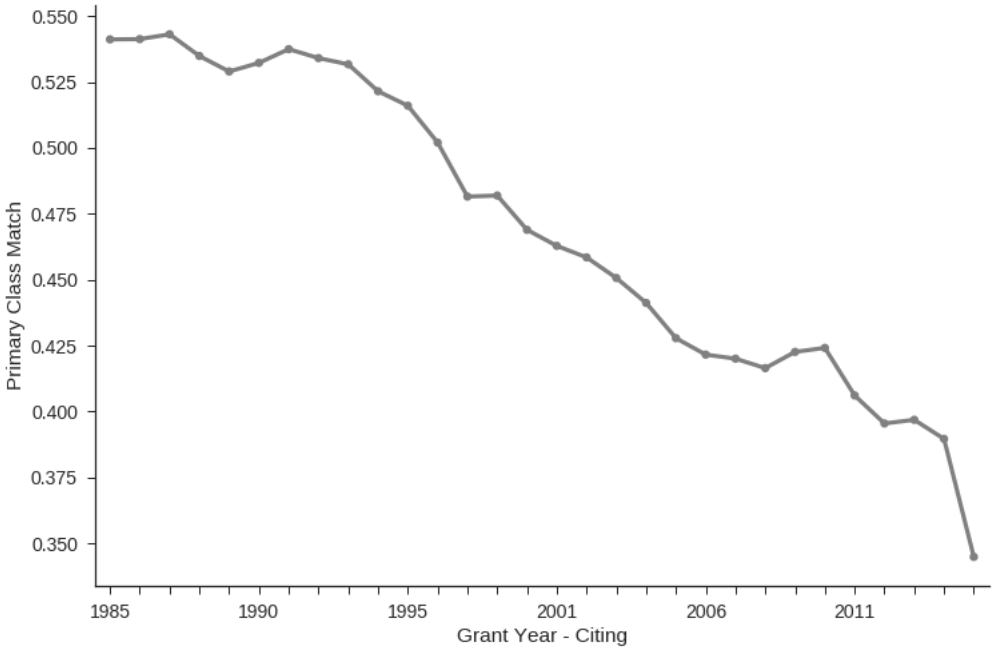


Figure B.3: Proportion of cited patents in the same primary class over time

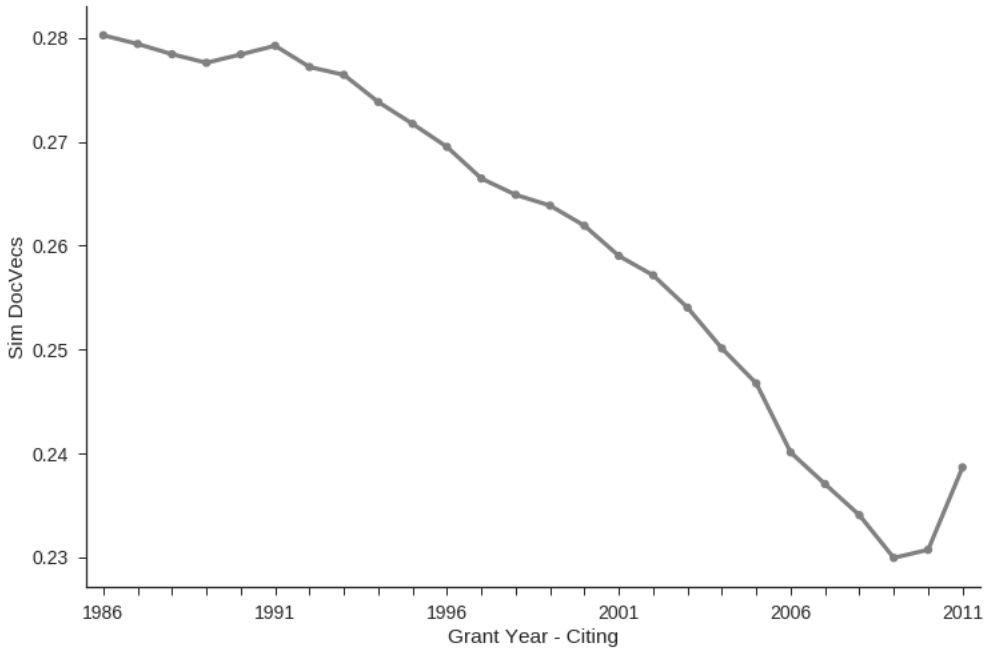


Figure B.4: Average DocVecs similarity to cited patents

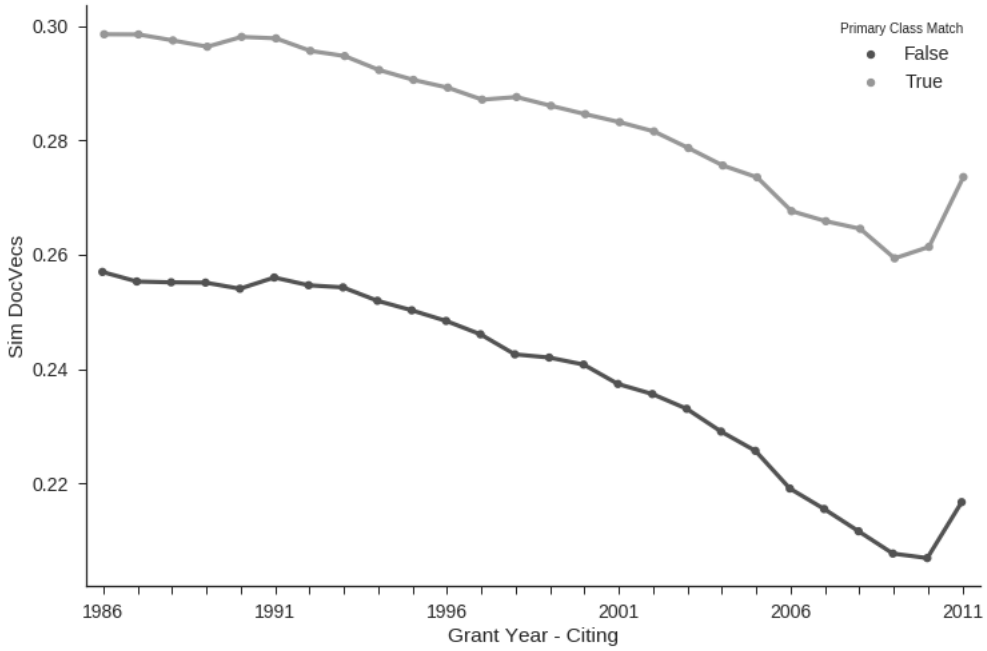


Figure B.5: Average DocVecs similarity to cited patents in the same primary class

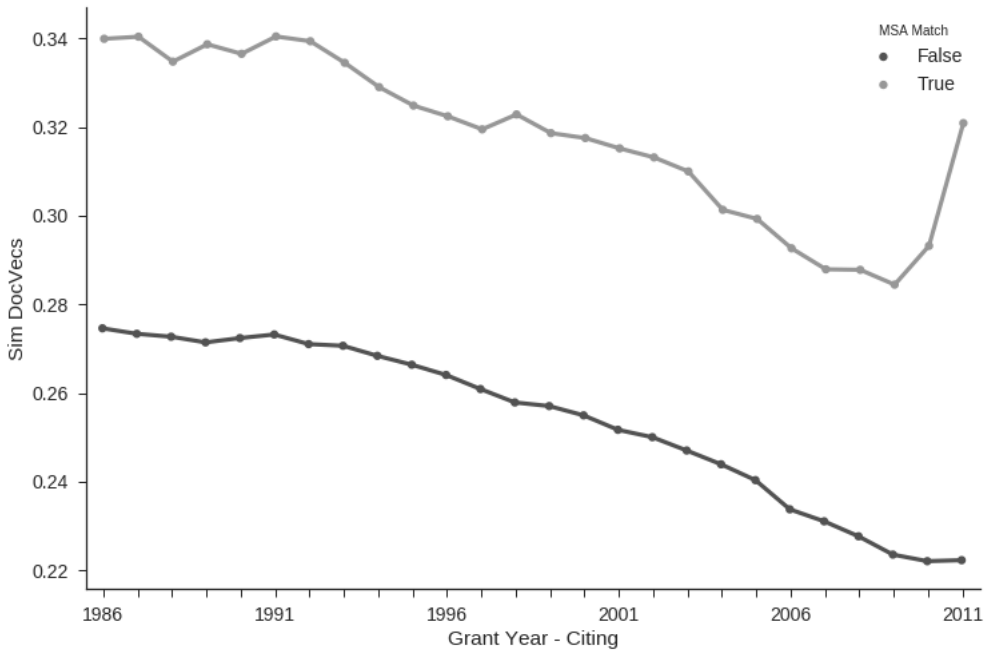


Figure B.6: Average DocVecs similarity to cited patents in the same MSA

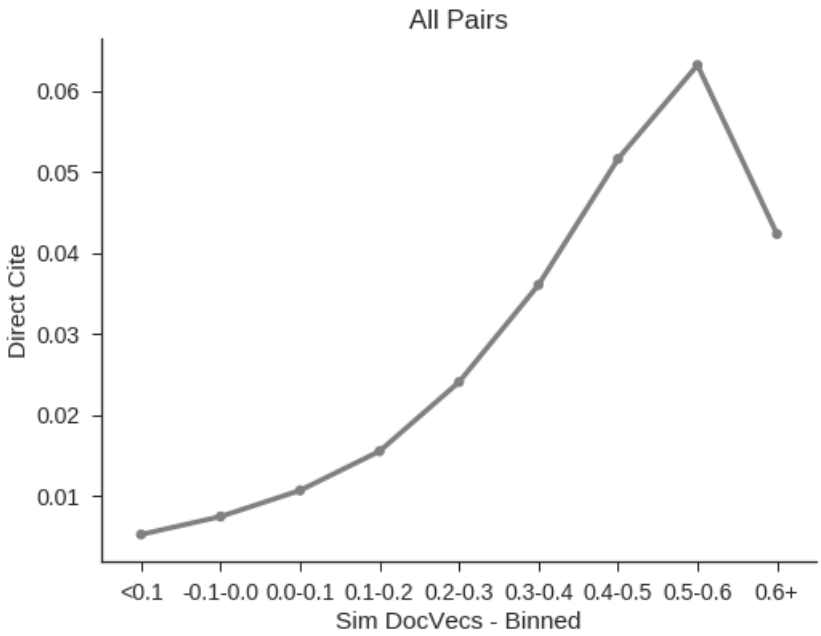


Figure B.7: Rate of direct citation conditional on level of DocVecs Similarity, All Pairs

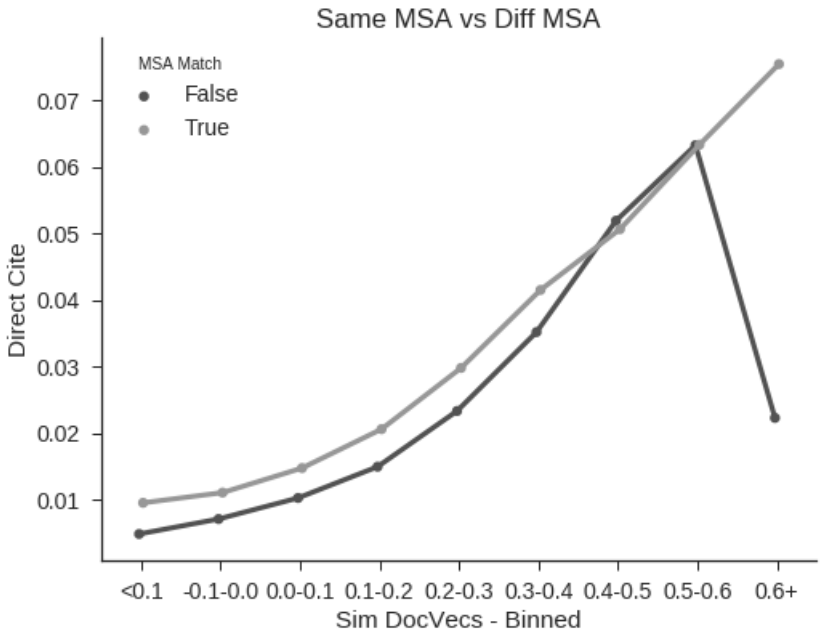


Figure B.8: Rate of direct citation conditional on level of DocVecs Similarity, Same MSA vs Different MSA

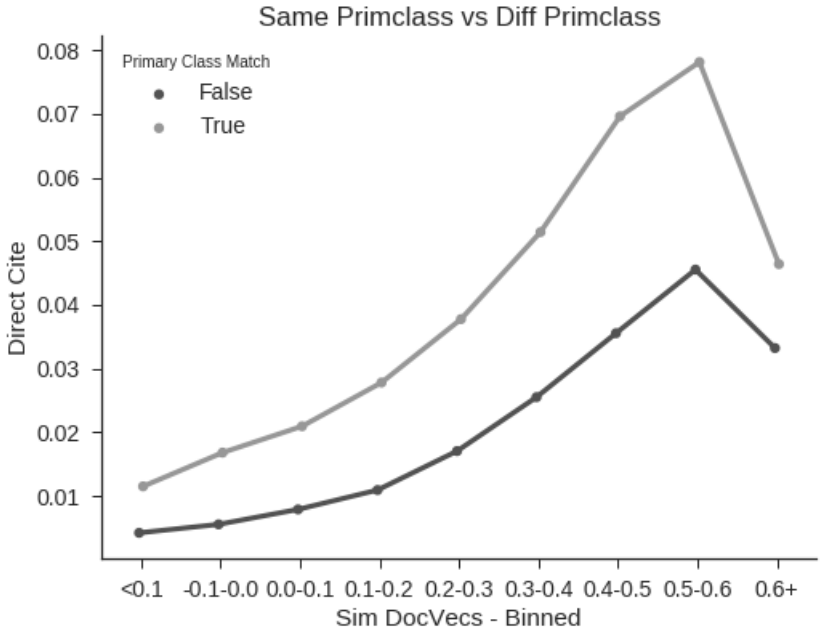


Figure B.9: Rate of direct citation conditional on level of DocVecs Similarity, Same Primary Class vs Different Primary Class

	<0.1	-0.1-0.0	0.0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6+
All Pairs, N	51355	205500	498927	685041	562529	293309	102019	27531	18958
All Pairs, Prop Cited	0.005	0.007	0.011	0.016	0.024	0.036	0.052	0.063	0.042
Same MSA, N	3768	16056	42380	65643	63246	41573	19994	8327	7163
Same MSA, Prop Cited	0.01	0.011	0.015	0.021	0.03	0.041	0.051	0.063	0.075
Diff MSA, N	47587	189444	456547	619398	499283	251736	82025	19204	11795
Diff MSA, Prop Cited	0.005	0.007	0.01	0.015	0.023	0.035	0.052	0.063	0.022
p-value	0	0	0	0	0	0	0.466	0.982	0
Same NAICS, N	21756	92343	239948	354306	313789	175065	64802	18362	13949
Same NAICS, Prop Cited	0.006	0.009	0.013	0.019	0.028	0.042	0.059	0.072	0.047
Diff NAICS, N	29599	113157	258979	330735	248740	118244	37217	9169	5009
Diff NAICS, Prop Cited	0.005	0.006	0.009	0.012	0.019	0.028	0.039	0.046	0.029
p-value	0.355	0	0	0	0	0	0	0	0
Same Primclass, N	7035	34445	105794	185777	190332	119502	48211	14968	13148
Same Primclass, Prop Cited	0.012	0.017	0.021	0.028	0.038	0.051	0.07	0.078	0.046
Diff Primclass, N	44320	171055	393133	499264	372197	173807	53808	12563	5810
Diff Primclass, Prop Cited	0.004	0.006	0.008	0.011	0.017	0.026	0.036	0.046	0.033
p-value	0	0	0	0	0	0	0	0	0

Table B.1: Summary table of rates of direct citation by DocVecs similarity

C. Application of similarity: examining knowledge flows through inventor mobility

Inventors are expected to be consistent in their knowledge of their own prior patents and prior citations. This fact can be exploited to further explore the nuances of the citation measure of knowledge flows. I examine the rate of citation for their own previous work, to see if citations may "miss" existing knowledge flows due to strategic motives after the inventor changes firms. Further, I compare the lists of citations made by inventors before and after they change firms to see how much of a difference this makes in their reported knowledge flows.

Finally, previous research such as Almeida and Kogut (1999); Azoulay et al. (2011) have used changes in the citation rate once inventors move cities to argue for localization. I compare the similarity of the mobile inventor's patents to their new citations to determine if it impacted their firm's new innovation ouputs.

C.1. Rate of self-citation before and after firm change

A clear example of where strategic non-citation might emerge is in the rate of citation for inventor's own patents, once they move to a different firm. Since inventors cannot reasonably claim to be ignorant of their own inventions, we can safely assume that any discrepancies in the rate of citation must be attributed to strategic withholding on the part of the inventor or new firm. I compare the rate of inventor self-citation when they are at their first firm, to the rate of self-citation of patents at their second firm to the patents at their first firm.

Sample construction

Suppose inventor i has patents A, B, C at firm 1, and D, E at firm 2. Then I will compare the self-citation rates in the set AB, AC, BC before their firm change, and AD, AE, BD, BE, CD, CE after the change. Since inventors often work in slightly different areas at their new firm, it is also crucial to condition on pairwise similarity in order to ascertain the appropriate benchmark citation rate. I use all 12,377 inventors who have changed firms and their complete patents at their first and second firm to construct my sample. They account for 8.7% of the 141,583 total inventors in the data. I calculate the complete set of pairwise similarities in the resulting sample, which after removing outliers, results in a sample size of almost 3.3 million pairs.

Evidence of strategic omission

If, conditional on similarity, the rate of citation is lower for inventors after they change firms compared to before, then this indicates that the inventor or the new firm is more or less knowingly concealing relevant citations in order to enlarge the scope of the new invention. I find evidence to support this claim in figure C.1 and table C.3. On average, prior to the move, inventors cite their own inventions at the first firm in 12.5% of the observations, while after the move this drops in more than half to 5.8%. To allow for the possibility that inventors switch firms in order to work in different technology areas, I then condition the rate of self-citation on the similarity between the inventor's own patents. While this rate increases sharply with similarity between the two inventions, there is gap in the rate before and after changing firms that is consistent and statistically significant at almost every level of similarity. The difference grows with similarity up to the 0.5-0.6 bracket, after which the two measures converge for the highest levels of pairwise similarity. In the similarity range 0.5-0.6, inventors prior to their move

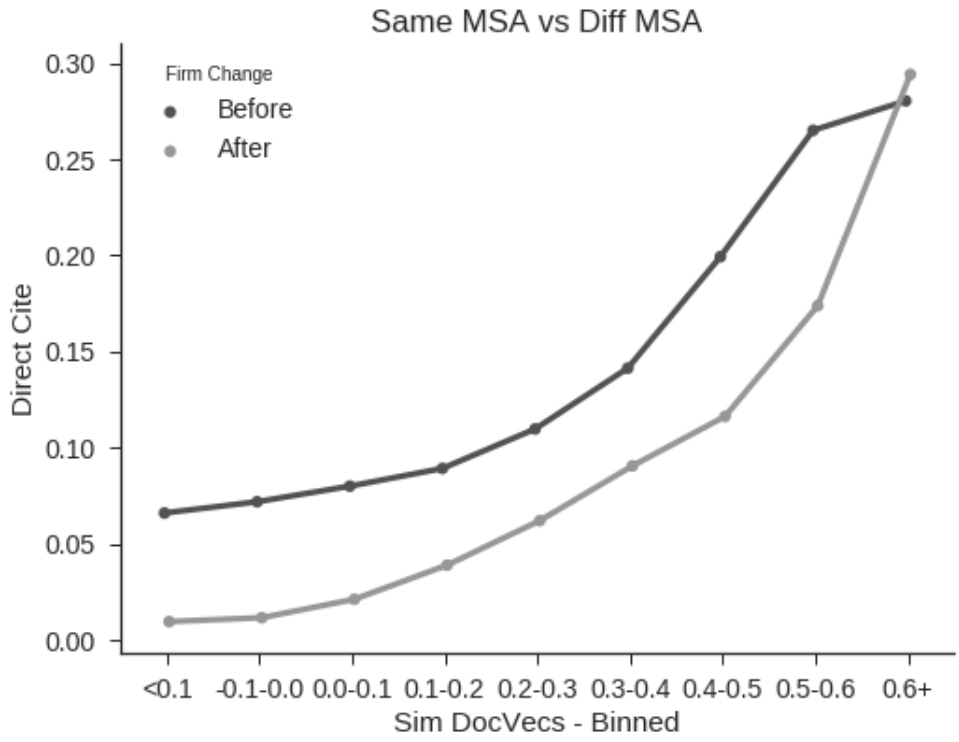


Figure C.1: Rates of direct citation by DocVecs similarity. See C.3 for table of results.

across firms self-cited at a rate of 26.5%, while after the change it becomes 17.4%, a difference of almost 9%. Interestingly, the difference is not statistically significant at the highest level of similarity, largely due to the tapering off of *within* firm self-citation. One explanation is that there is no risk of patent infringement lawsuits from yourself, and so firms can expand the scope of their new patents by not listing their own highly similar previous inventions. Finally, if inventors cited themselves at their new firms at the same rate as before their move, then the projected number of total citations would be 11,875, compared to the actual number of 6,118. This implies that there are almost as many "missing" self-citations as actual self-citations.

C.2. Changes in citations made before and after firm change

An implication of this finding may be that firms (that is, the assignee of the patent who "owns" the intellectual property), not inventors, determine which patents are cited in the application. Using the same sample, I then examine how many citations are shared before and after the inventor changes firms using the number and percentage of common citations as described in 2.3. Citations made to

other patents assigned to the same firm are excluded prior to the analysis.[15] I find that changing firms significantly reduces the both the number and proportion of common citations made by the same inventor. Prior to the move, inventors on average shared 13% of backward citations with their own other patents. After the move, this drops to 5% overall: 6% if the inventor changed to a different firm in the same city, 3% if the inventor relocated to a different city. To account for the inventor changing innovation agendas once they switch firms, I also condition on similarity and find that a gap in the percentage of common citations exists for all similarity levels, and is particularly high when new patents have similarity of 0.5-0.6 to prior inventions. (C.2,C.4)

Looking just at the overlap backward citations, it would indicate that inventors are utilizing vastly different knowledge sources. But this is not accompanied by a drastic shift in their fields of interest. The change in similarity to their previous inventions is small, although significant: from a mean of 0.29 to 0.25 after changing firms. The number of pairs from the same primary class also reflect a smaller change in the inventor's output: from 35% prior to 30% after changing firms. There is a larger change for inventors who move cities as well, which indicates that inventors who switch firms and cities are altering their innovation agenda more drastically. While evidence suggests that inventors do change firms to produce different innovations, this change is slight compared to what is suggested by the change in their citations lists.

The discrepancy in the amount of overlap in the inventor's citation list and the similarity to their own previous inventions suggests that citations may be determined more by firm specific factors than the inventors themselves. An inventor may contribute a couple of citations they know and used before, and the rest is selected from a pool of citations that the firm uses, also likely influenced by their choice in lawyers. I find evidence in section 3.1 that lawyers play a large role in determining a patent's citations. This is consistent with Wagner et al. (2014), who also show that firms who rely on professional service firms are more likely to cite patents that are part of the law firm's knowledge repository. These findings suggest that there is a further gap between what citations represent and the knowledge flows likely used by the inventor for their invention.

[15] Because outliers have an outsized effect in determining the average number of common cited patents, I drop observations with number of common cited patents above the 99th percentile. This drops approximately 3,264 observations.

	Num Common Cited	Pct Common Cited	Num Common Cited from Prev MSA	Pct Common Cited from Prev MSA	Sim DocVecs	Primclass Match
Before Firm Change, Mean	15.61	0.13	0.94	0.05	0.29	0.35
After Firm Change, Mean	1.92	0.05	0.05	0.02	0.25	0.30
p-value	0.00	0.00	0.00	0.00	0.00	0.00
After Firm Change - Same MSA, Mean	2.27	0.06	0.07	0.03	0.26	0.31
p-value	0.00	0.00	0.00	0.00	0.00	0.00
After Firm Change - Diff MSA, Mean	1.23	0.03	0.02	0.02	0.22	0.28
p-value	0.00	0.00	0.00	0.00	0.00	0.00

Table C.1: Changes in number of common cited patents in inventor's own patents before and after firm change

C.3. Effect of inventor mobility on patents in their new city

Following Almeida and Kogut (1999); Azoulay et al. (2011); Agrawal et al. (2006), I examine the changes in knowledge flows when inventors move cities. Of the 66,790 inventors I observe who changed firms in the previous subsections, about 12,846 inventors (19.2% of total) also moved cities. One key challenge with using mobility is that inventors often move cities to work in slightly different technology fields (as we saw above in C.1). Thus, there may be an appearance in higher "knowledge flows" when in fact what is picked up is the inventor moving to a different city to work in a technology area that is concentrated in the new city. Adapted from Azoulay et al. (2011), who focus on academic citations made to mobile scientists, one way to partially control for this is to focus on knowledge flows from the inventor's patents *prior* to the move.

In Azoulay et al. (2011), they find that article-to-article citations from the scientists' new location increases markedly after the move. My findings corroborate this pattern. I focus on the 6,497 prior patents from inventors who moved cities. Each patent has received at least one (non-self) forward citation. On average, 2.91% of these citations matched the new location before the inventor moved. Afterwards, this rate jumps to 7.75% (p-value$= 0.00$). Once again, citations provides unequivocal evidence for localization.

However, does the citation represent a knowledge flow from the mobile inventor's patent, or merely that the firm now knows the mobile inventor? That is, is there evidence that firms in the mobile inventor's new city who cite their prior patents are actually influenced by these patents, or is the citation in some ways "perfunctory", reflecting the inventor's reputation or part in the local inventors' network rather than a knowledge spillover from the inventor to the firm.

Sample construction

For the prior patents of mobile inventors, I gather all citations that were made by firms in the new city that had *not* cited the inventor before. Prior to their move, 4,316 assignees from the new location had cited their past work. After the move, this number jumps to 10,578, with new citing firms accounting for 80.3% of the total. I focus on these firms as it is somewhat plausible that they have newly "discovered" the mobile inventor's work due to the inventor's presence in the city. These new citing firms make 27,817 forward citations to the mobile inventor's prior patents. For each forward citation, I try to select a control patent from the same primary class and firm, granted as close as possible in date, that does *not* cite the same prior patent. I only succeed in finding a control patent in 8,951 cases as not all firms have prior patents in the same primary class, or any prior patents at all.

Evidence of knowledge flow from citation to newly citing firm vs "perfunctory" citations

I attempt to gauge the relevance of the mobile inventor's prior patent on the newly citing firm in two ways. First, I compare if the citing patent is more similar to the prior patent compared to the control. Then, I compare the average similarity of the citing patent and the control to their firm's own prior patents in the previous 5 years (i) overall; (ii) within the same primary class. This is to determine whether or not the citing patent represents a departure from the firm's usual innovation agenda, due to the influence of the new inventor's knowledge flow to the firm.

I find that the citing patent is more similar on average to the cited prior patent compared to the control. In C.2, the mean similarity of the citing patent is 0.278, while mean similarity is 0.234 for the control. The citing patent is about 18.8% more similar to the cited patent. However, I also find evidence that the citing patent is in any ways a "departure" from the firm's usual inventive activities. The citing patent and the control have identical average similarity to their own firm's prior patents, both across all primary classes and within the same primary class. When I rank the similarity of citing firm's prior parents to the new inventor's patent, I find that the citing patent was most similar in approximately 30% of cases. The median rank of the citing patent is 2.

These results suggest that while the new inventor may have influenced the citing patent, this patent was produced within the existing agenda of the citing firm. It is consistent with the explanation that firms in the new city were already working within the new inventor's technology field, and are citing the inventor who has become a peer. This suggests that those with the "absorptive capacity" (Cohen and

	Sim DocVecs to Cited	Mean Sim Docvecs, Own Prior Pats	Mean Sim Docvecs, Own Prior Pats in Citing PC
Citing	0.278	0.281	0.328
Control	0.234	0.28	0.327
t-value	26.637	1.096	1.458
p-value	0.00	0.273	0.145
N	8951	6407	6338

Table C.2: Changes in number of common cited patents in inventor's own patents before and after firm change. Differences in the number of observations arise due to a lack of other prior patents for citing patents' firms in different categories.

	<0.1	-0.1-0.0	0.0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6+
Before Firm Change, N	3437	12802	32168	49481	50733	36190	19608	10204	9301
Before Firm Change, Prop Cites	0.066	0.072	0.08	0.089	0.11	0.142	0.199	0.265	0.28
After Firm Change, N	3018	8594	18736	24880	22741	15429	7683	2702	1452
After Firm Change, Prop Cites	0.01	0.012	0.022	0.039	0.062	0.091	0.116	0.174	0.294
Diff, p-value	0	0	0	0	0	0	0	0	0.278

Table C.3: Rate of self-citation before and after firm change

	<0.1	-0.1-0.0	0.0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6+
Before Firm Change, N	3620	12775	32021	49163	50138	35447	18975	9756	9033
Before Firm Change, Pct Common Cites	0.039	0.034	0.044	0.062	0.086	0.135	0.235	0.419	0.658
After Firm Change, N	3205	8593	18731	24868	22734	15419	7678	2701	1452
After Firm Change, Pct Common Cites	0.007	0.009	0.015	0.028	0.046	0.071	0.107	0.211	0.52
Diff, p-value	0	0	0	0	0	0	0	0	0

Table C.4: Rate of self-citation before and after firm change

Levinthal (2000)) to appropriate the knowledge brought by the new inventor are largely working within the same domain. As to whether or not the existing firm would have made the same invention *but for* the knowledge flows from the new inventor, the evidence is unclear. Some influence is suggested, but perhaps the contribution is not significant enough to drastically alter the invention.

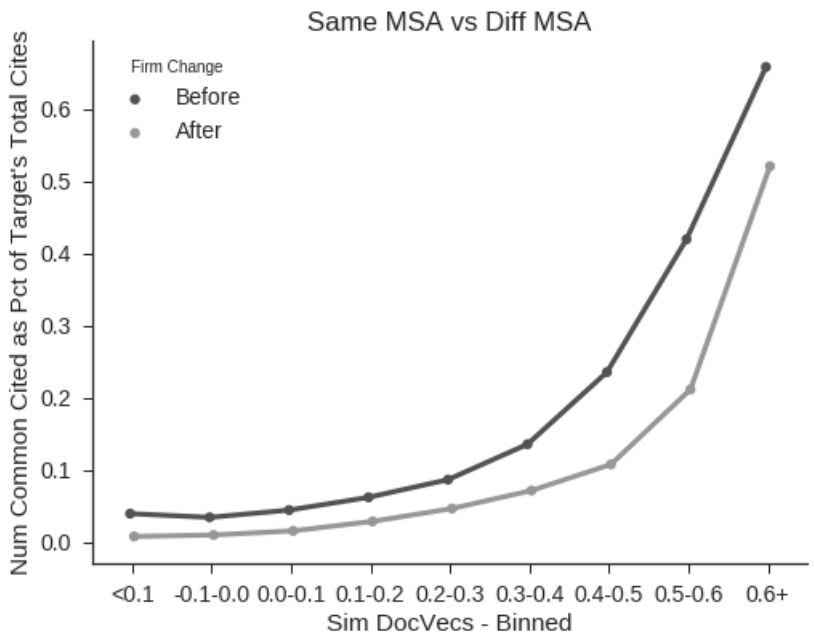


Figure C.2: Rate of direct citation conditional on level of DocVecs Similarity, Same Primary Class vs Different Primary Class

D. Application of similarity: estimating geographic localization

	1975-85	1985-95	1995-05
Control Selection: Standard JTH	0.2422***	0.2849***	0.2968***
	(0.0074)	(0.0051)	(0.0041)
N	58647	107358	185154
Adjusted R^2	0.03	0.05	0.05
Control Selection: Similarity	0.2029***	0.2681***	0.2621***
	(0.0100)	(0.0067)	(0.0054)
N	36917	67332	117137
Adjusted R^2	0.02	0.03	0.04
Control Selection: Lawyer	0.0806***	0.0935***	0.1150***
	(0.0136)	(0.0086)	(0.0069)
N	22914	51837	85855
Adjusted R^2	0.02	0.03	0.03
Controls: Year and Primary Class FE			

Table D.1: Regression results for JTH replication under different control selection methods. Localization estimates ($\hat{\beta}_1$) are the row values with standard errors in parentheses beneath. Sample sizes vary due to the inability to find control patents under certain methods of selection. Controls in the regression model include year and primary class fixed effects.

	1975-85	1985-95	1995-05
Control Selection: Standard JTH	0.2340***	0.2783***	0.2864***
	(0.0073)	(0.0050)	(0.0040)
N	58647	107358	185154
Adjusted R^2	0.04	0.06	0.08
Control Selection: Similarity	0.1983***	0.2656***	0.2633***
	(0.0100)	(0.0067)	(0.0053)
N	36917	67332	117137
Adjusted R^2	0.04	0.06	0.08
Control Selection: Lawyer	0.0842***	0.0916***	0.1119***
	(0.0136)	(0.0086)	(0.0068)
N	22914	51837	85855
Adjusted R^2	0.05	0.06	0.08
Controls: Year, Primary Class, MSA, Lawyer, Examiner FE			

Table D.2: Regression results for JTH replication under different control selection methods. Localization estimates ($\hat{\beta}_1$) are the row values with standard errors in parentheses beneath. Sample sizes vary due to the inability to find control patents under certain methods of selection. Controls in the regression model include year and primary class fixed effects.

E. Discussion

	Sim DocVecs	Num Common Cited	Pct Common Cited, Target's Citations	First Year	First Year, Num Pats	First Year, Num Pairs
gui	0.09	0.00	0.00	1992	1348	796727
lun	0.12	0.00	0.00	1995	415	63044
asic	0.11	0.00	0.00	1987	198	16716
url	0.10	0.01	0.00	1995	111	4847
serd	0.12	0.02	0.00	1998	75	1929
chat	0.10	0.00	0.00	1992	9	1299
bist	0.12	0.00	0.00	1990	42	810
femto	0.15	0.04	0.00	2007	27	563
angst	0.16	0.00	0.00	1994	40	549
mcm	0.10	0.00	0.00	1991	32	440
www	0.15	0.01	0.01	1995	9	291
efus	0.12	0.02	0.00	2000	22	201
femtocel	0.22	0.09	0.00	2007	8	198
adenovir	0.26	0.03	0.00	1993	15	98
cyclin	0.18	0.00	0.00	1991	13	80
n 1	0.13	0.00	0.00	1998	13	63
dvd	0.16	0.00	0.00	1996	10	44
websit	0.18	0.00	0.00	1996	10	42
gpu	0.05	0.00	0.00	2001	9	36
pcie	0.21	0.00	0.00	2004	9	35

Table E.1: Comparison of similarity and backward citation overlap for new patents using new terms.

References

Ajay Agrawal, Iain Cockburn, and John McHale. Gone but not forgotten: knowledge flows, labor mobility, and enduring social relationships. *Journal of Economic Geography*, 6(5): 571–591, 2006.

Juan Alcacer and Michelle Gittelman. Patent citations as a measure of knowledge flows: The influence of examiner citations. *The Review of Economics and Statistics*, 88(4): 774–779, 2006.

Juan Alcacer, Michelle Gittelman, and Bhaven Sampat. Applicant and examiner citations in us patents: An overview and analysis. *Research Policy*, 38(2):415–427, 2009.

Paul Almeida and Bruce Kogut. Localization of knowledge and the mobility of engineers in regional networks. *Management science*, 45(7):905–917, 1999.

Ashish Arora, Sharon Belenzon, and Honggi Lee. Reversed citations and the localization of knowledge spillovers. *Journal of Economic Geography*, 18(3):495–521, 2018.

Sam Arts, Bruno Cassiman, and Juan Carlos Gomez. Text matching to measure patent similarity. *Strategic Management Journal*, 39(1):62–84, 2018.

David B Audretsch and Maryann P Feldman. R&d spillovers and the geography of innovation and production. *The American economic review*, 86(3):630–640, 1996.

Pierre Azoulay, Waverly Ding, and Toby Stuart. The impact of academic patenting on the rate, quality and direction of (public) research output. *The Journal of Industrial Economics*, 57(4):637–676, 2009.

Pierre Azoulay, Joshua S Graff Zivin, and Bhaven N Sampat. The diffusion of scientific knowledge across time and space: Evidence from professional transitions for the superstars of medicine. Technical report, National Bureau of Economic Research, 2011.

Sharon Belenzon and Mark Schankerman. Spreading the word: Geography, policy, and knowledge spillovers. *Review of Economics and Statistics*, 95(3):884–903, 2013.

Antonin Bergeaud, Yoann Potiron, and Juste Raimbault. Classifying patents based on their semantic content. *PloS one*, 12(4):e0176310, 2017.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Nicholas Bloom, Mark Schankerman, and John Van Reenen. Identifying technology spillovers and product market rivalry. *Econometrica*, 81(4):1347–1393, 2013.

Kevin Bryan, Yasin Ozcan, and Bhaven Sampat. A user's guide to in-text citations, 2018.

Kristy Buzard, Gerald A Carlino, Robert M Hunt, Jake K Carr, and Tony E Smith. Localized knowledge spillovers: Evidence from the agglomeration of american r&d labs and patent data. 2016.

Frances Cairncross. The death of distance: How the communications revolution will change our lives. 1997.

Michal Campr and Karel Ježek. Comparing semantic models for evaluating automatic document summarization. In *International Conference on Text, Speech, and Dialogue*, pages 252–260. Springer, 2015.

Christian Catalini. Microgeography and the direction of inventive activity. *Management Science*, 2017.

Aaron Chatterji, Edward Glaeser, and William Kerr. Clusters of entrepreneurship and innovation. *Innovation Policy and the Economy*, 14(1):129–166, 2014.

Wesley M Cohen and Daniel A Levinthal. Absorptive capacity: A new perspective on learning and innovation. In *Strategic Learning in a Knowledge economy*, pages 39–67. Elsevier, 2000.

Christopher A Cotropia, Mark A Lemley, and Bhaven Sampat. Do applicant patent citations matter? *Research Policy*, 42(4):844–854, 2013.

Maryann P Feldman. The character of innovative places: entrepreneurial strategy, economic development, and prosperity. *Small Business Economics*, 43(1):9–20, 2014.

Alberto Galasso and Mark Schankerman. Patents and cumulative innovation: Causal evidence from the courts. *The Quarterly Journal of Economics*, 130(1):317–369, 2014.

Ina Ganguli, Jeffrey Lin, and Nicholas Reynolds. The paper trail of knowledge spillovers: Evidence from patent interferences. 2017.

Matthew Hoffman, Francis R Bach, and David M Blei. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864, 2010.

Adam B Jaffe. Technological opportunity and spillovers of r&d: Evidence from firms' patents, profits, and market value. *American Economic Review*, 76(5):984–1001, 1986.

Adam B Jaffe. Real effects of academic research. *The American economic review*, pages 957–970, 1989.

Adam B Jaffe and Gaétan De Rassenfosse. Patent citation data in social science research: Overview and best practices. *Journal of the Association for Information Science and Technology*, 68(6):1360–1374, 2017.

Adam B Jaffe, Manuel Trajtenberg, and Rebecca Henderson. Geographic localization of knowledge spillovers as evidenced by patent citations. *the Quarterly journal of Economics*, 108(3):577–598, 1993.

Nikhil Johri, Daniel Ramage, Daniel A. McFarland, and Daniel Jurafsky. A study of academic collaborations in computational linguistics using a latent mixture of authors model. In *Association for Computational Linguistics (ACL) Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Portland, OR, June 2011. URL pubs/acl-latech2011.pdf.

Sarah Kaplan and Keyvan Vakili. The double-edged sword of recombination in breakthrough innovation. *Strategic Management Journal*, 36(10):1435–1457, 2015.

B Kelly, D Papanikolaou, A Seru, and M Taddy. Measuring technological innovation over the long run. Technical report, Working Paper, 2018.

Paul R Krugman. *Geography and trade*. MIT press, 1991.

Ryan Lampe. Strategic citation. *Review of Economics and Statistics*, 94(1):320–333, 2012.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196, 2014.

Josh Lerner and Amit Seru. The use and misuse of patent data: Issues for corporate finance and beyond. *Booth/Harvard Business School Working Paper*, 2015.

Chu-Cheng Lin, Waleed Ammar, Chris Dyer, and Lori Levin. Unsupervised pos induction with word embeddings. *arXiv preprint arXiv:1503.06760*, 2015.

Jeffrey Lin et al. The paper trail of knowledge transfers. *Federal Reserve Bank of Philadelphia Business Review. Second Quarter*, 2014.

Alan C Marco, Asrat Tesfayesus, and Andrew A Toole. Patent litigation data from us district court electronic records (1963-2015). 2017.

Alfred Marshall. *The economics of industry*. Macmillan and Company, 1920.

Aditi Mehta, Jessica K Martin, and Charlotte B Kahn. City of ideas: Reinventing boston's innovation economy: The boston indicators report 2012. 2012.

Rada Mihalcea, Courtney Corley, Carlo Strapparava, et al. Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, volume 6, pages 775–780, 2006.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013b.

Petra Moser, Joerg Ohmstedt, and Paul W Rhode. Patent citations - an analysis of quality differences and citing practices in hybrid corn. *Management Science*, 64(4):1926–1940, 2017.

Yasusada Murata, Ryo Nakajima, Ryosuke Okamoto, and Ryuichi Tamura. Localized knowledge spillovers and patent citations: A distance-based approach. *Review of Economics and Statistics*, 96(5):967–985, 2014.

Fiona Murray and Scott Stern. Do formal intellectual property rights hinder the free flow of scientific knowledge?: An empirical test of the anti-commons hypothesis. *Journal of Economic Behavior & Organization*, 63(4):648–687, 2007.

Mikko Packalen and Jay Battacharya. Cities and ideas, 2015.

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

Michael Roach and Wesley M Cohen. Lens or prism? patent citations as a measure of knowledge flows from public research. *Management Science*, 59(2):504–525, 2013.

Peter Thompson and Melanie Fox-Kean. Patent citations and the geography of knowledge spillovers: A reassessment. *American Economic Review*, pages 450–460, 2005.

Stefan Wagner, Karin Hoisl, and Grid Thoma. Overcoming localization of knowledge - the role of professional service firms. *Strategic management journal*, 35(11):1671–1688, 2014.

Kenneth Younge and Jeffrey Kuhn. Patent-to-patent similarity: a vector space model. 2016.