Strategic Citations in Patents Using Machine Learning

Fiona Sijie Feng (NYU Stern)

May 1, 2019

Abstract

Patent citations are the most commonly used indicator of knowledge relationships across patents. However, citation may be prone to strategic behaviour: inventors and firms may over-cite to offset the likelihood of litigation, and under-cite to increase the scope of the patent. As an alternative to citations, I construct a measure of proximity in ideas expressed in patent text using unsupervised machine learning algorithm Doc2Vec. Based on the similarity across patent texts, I find evidence that (i) applicants may strategically omit citations to patents in different cities, as infringement discovery is less likely; (ii) applicants cite their own prior inventions *less* after they change firms. This implies that use of citations to measure knowledge flows are affected by strategic biases.

1. Introduction

I explore the relationship between patents and their citations using a measure of proximity in innovative ideas, as represented by patent abstract text. Abstract text is used as it summarizes the ideas of the invention. I derive vector space representations of patents using Document Vectors (Doc2Vec), an unsupervised machine learning algorithm, and use cosine similarity to measure their proximity in ideas space. Pairs of patents with high cosine similarity represent proximate inventions: for example, two patents US5651964 "Methods for the suppression of neu mediated tumors by the adenoviral EIA gene" and US5585362 "Adenovirus vectors for gene therapy" are have high cosine similarity as both are pharmaceutical patents that deal with adenoviral gene therapy, which is used in cancer treatment.

Given that patent text similarity can be derived for any pairs of patents, this represents a measure of relatedness that is not reliant on the reporting of citations, which is the most commonly used indicator of knowledge relationships across patents. 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 intellectual property 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 likelihood of citing local patents, i.e. other patents invented in the same city. 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.

I find evidence using patent text similarity there exists strategic incentives to leave out citations. Across all patents, I find that citations rate is not monotonically increasing relative to similarities, and in fact citation rate drops for patents with the highest similarity, but only if these patents are located in different cities. This suggests that incentives to leave out citations are weaker for patents located in the same city, as the likelihood of infringement discovery is stronger. Furthermore, I find that inventors cite their own prior patents less once they move firms. Given that inventors are less likely to be uninformed about their own work, this strongly suggests the presence of strategic omissions in patent citation behaviour. Since inventors are likely to innovate in highly similar technological areas,

the new firm may not want to cite similar prior patents of the inventor and limit the scope of the new patent. Another possible explanation, as explored in Feng (2019) is that different firms have different lawyers who cite from varying pools of prior patents.

I also find that there is evidence using patent text similarity that the relatedness between patents with citation relationships has declined slightly over time, corresponding to the rise in spurrious citations. Finally, I find mixed evidence for the effect of new inventors on new patents in the city: while citations to the new inventors' prior patents do increase, it may be the case that these citations are "perfunctory". A perfunctory citation is one made as a courtesy to a friend or colleague which may not indicate a knowledge influence on the patent. Watzinger et al. (2018) document this effect in citations to university professors who relocate. I find some evidence that this effect may also apply to inventors who relocate.

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.[1] Primary subclasses are nested in primary classes, which are in turn nested in

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

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 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 $N = 100$.[2] 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 [3] 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. Effect of external influences on citations

Evidence on the declining relevance of citations

I find evidence that external influences as discussed in 1 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

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

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

increased, more than doubling from 2.3 to 6.0 over the period 1985-2015 (figure B.1).[4] 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[5], 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 pairs that have the highest level of mutual similarity in patent text. 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

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

lack of incentive 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.

4. 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.

4.1. 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 4.1 and table C.2. 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 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

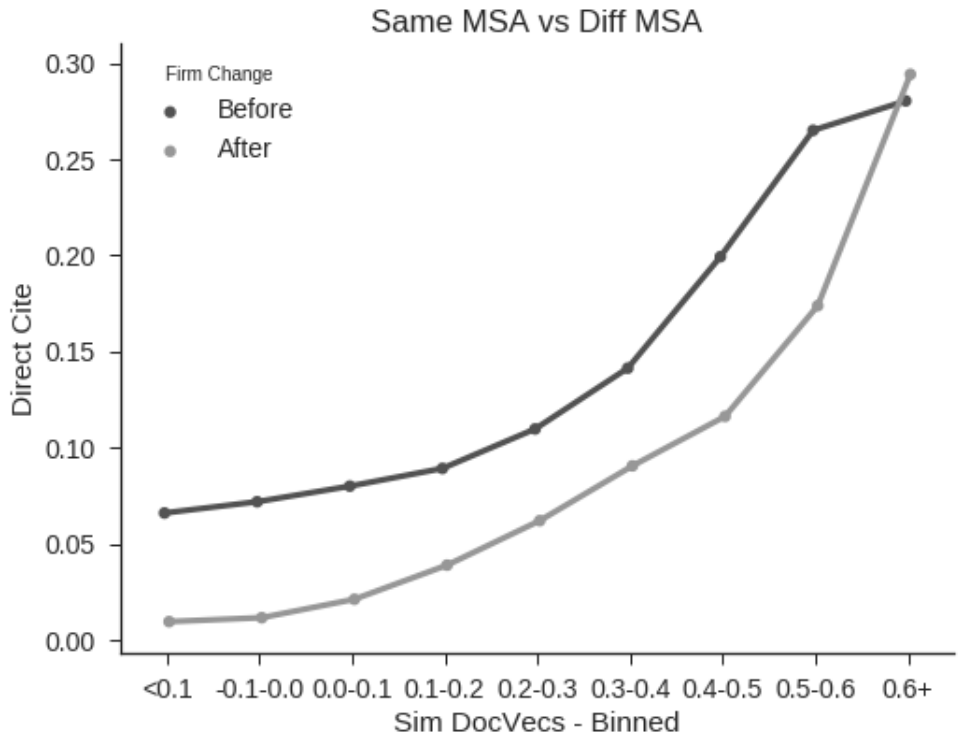


Figure 4.1: Rates of direct citation by DocVecs similarity. See C.2 for table of results.

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.

4.2. 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.[6] 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

[6] 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 4.1: Changes in number of common cited patents in inventor's own patents before and after firm change

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.1,C.3)

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. This follows evidence in Feng (2019) 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.

4.3. 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 4.1). 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. The rationale is to determine if the mobile inventor had an effect in changing the citing firm's direction of innovation. For example, suppose A is the mobile inventor's prior patent, and B is the new citing patent, and C is the control from the same firm and primary class as B. B cites A, but C does not. If the inventor's patent A exerted a significant influence on the firm, then B should be (i) more similar to A, (ii) be less similar to the citing firm's other patents compared to C.

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.1, 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 not 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 patents 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. In relation to the example given above, I find that B is indeed more similar to A, but not less similar to the citing firm's other patents compared to C, the control patent.

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 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.

5. Conclusion

In this paper, I examined the influence of strategy on citations. I measure textual similarity across pairs of patents as an alternative method of capturing knowledge relatedness across patents. In line with prior literature, I find evidence that citations may be biased by strategic considerations related to the likelihood of infringement discovery. While increasing overall citations provide protection from potential litigation, inventors and firms also do not want to over-cite as this reduces the scope of the patent and decreases its value. I find evidence that non-local patents (i.e. patents from different cities) are more likely to be left uncited if they are extremely similar textually to the applicant patent. More conclusively, I find evidence that inventors cite their own prior inventions *less* after they move firms. This points unequivocally to there being strategic omissions in the citation behaviour of firms. Additionally, using patents from inventors who move to different cities, I find that while citations to the new inventor, there is some evidence that these citations are “perfunctory”.

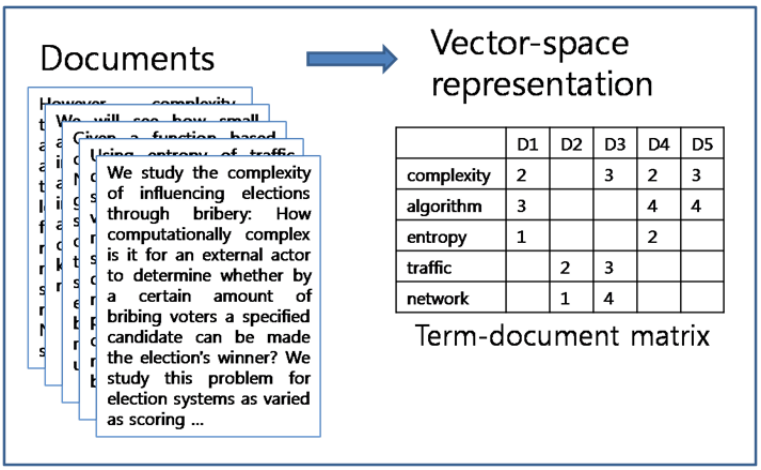
Appendix

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.[7] 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.

[7] 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

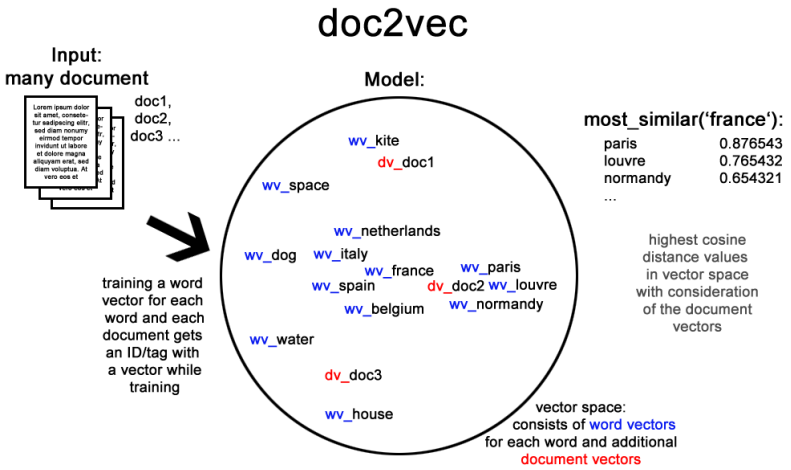


Figure A.2: Illustration of Document Vectors.

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.

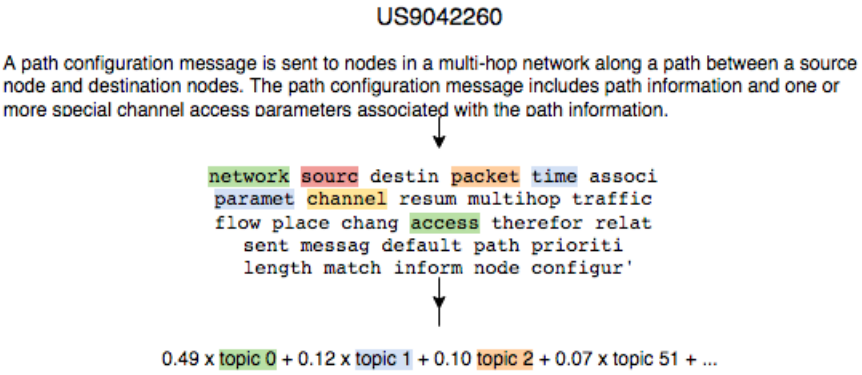


Figure A.3: Example of a patent converted into a distribution over topics.

B. Application of similarity: assessing the quality of citations

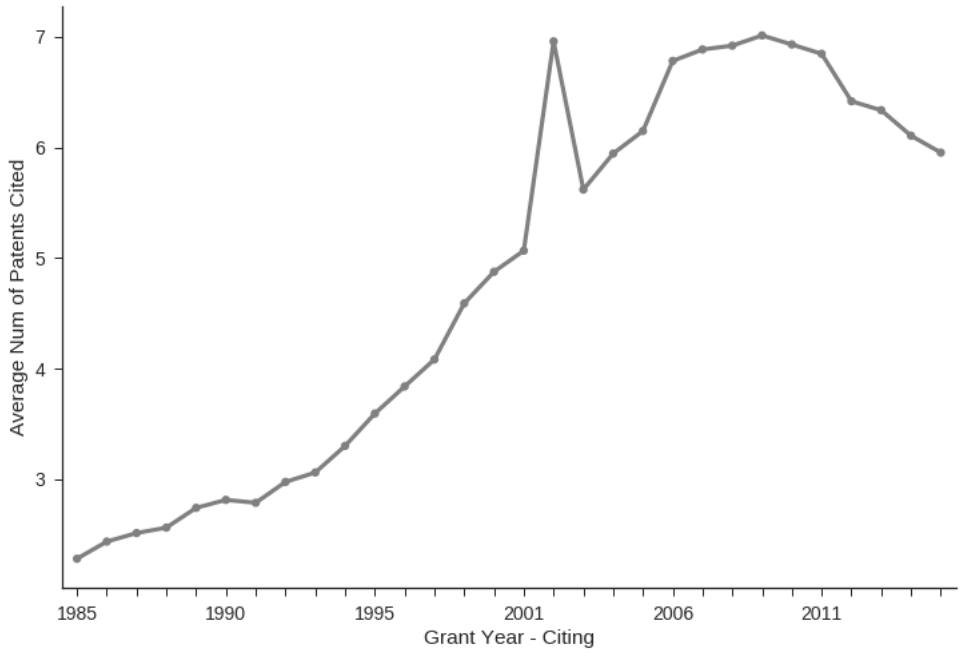


Figure B.1: Averate number of patents cited over time

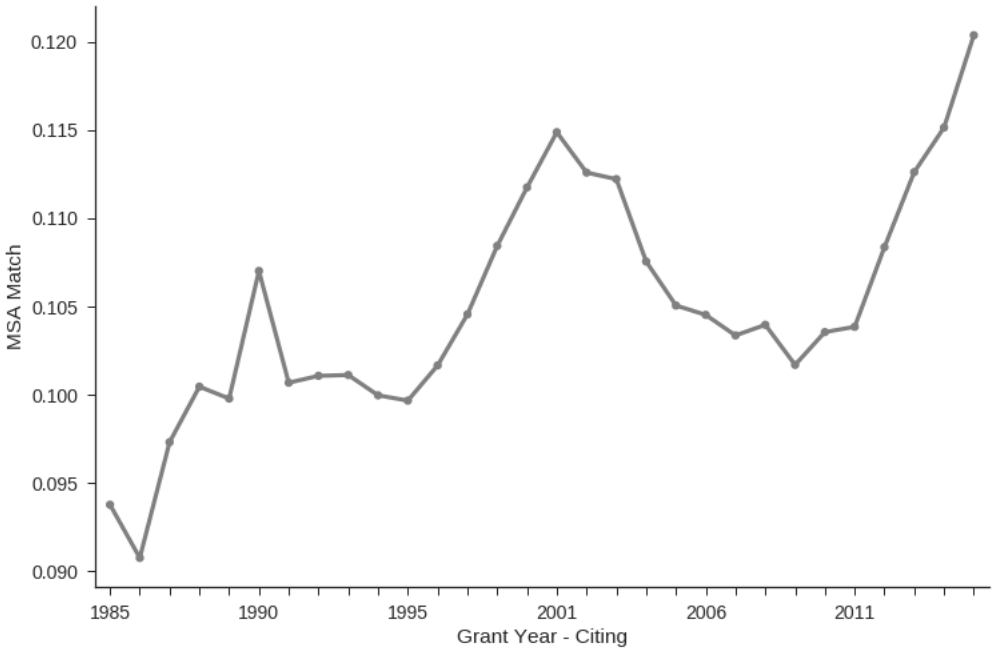


Figure B.2: Proportion of cited patents in the same MSA over time

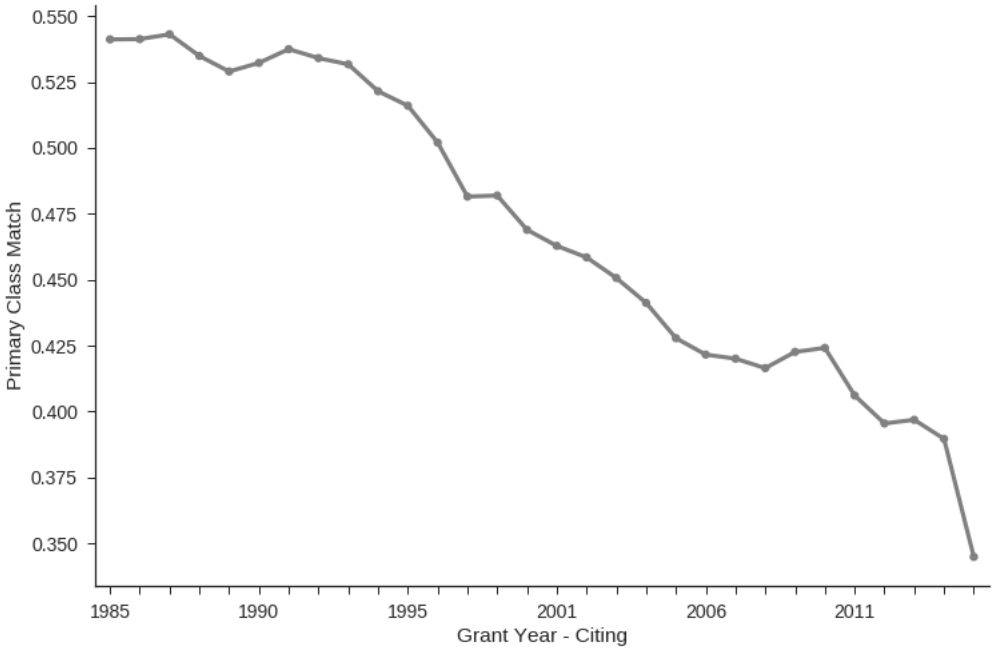


Figure B.3: Proportion of cited patents in the same primary class over time

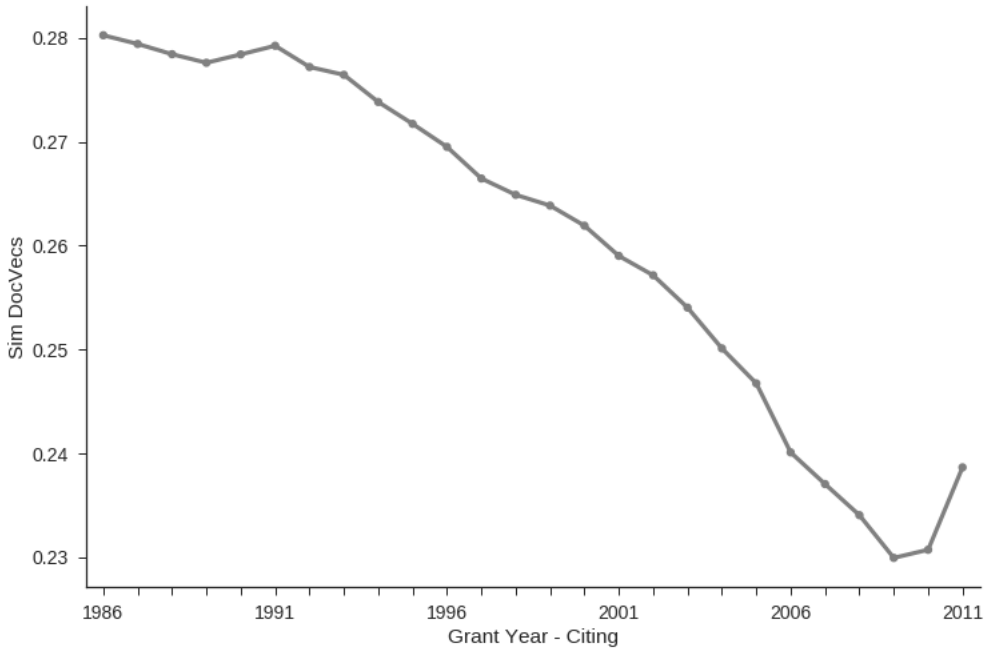


Figure B.4: Average DocVecs similarity to cited patents

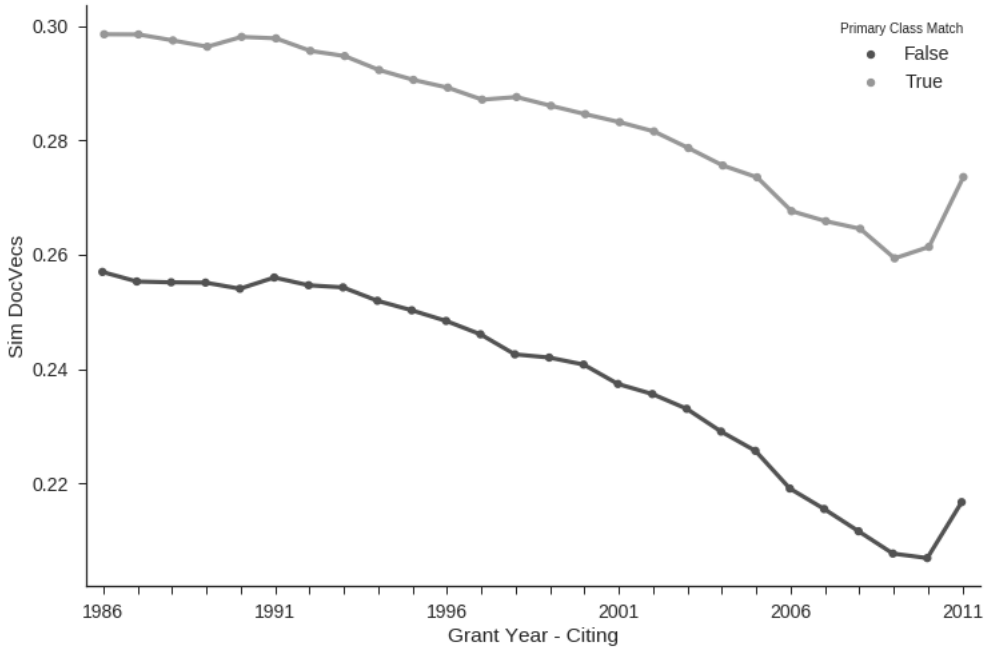


Figure B.5: Average DocVecs similarity to cited patents in the same primary class

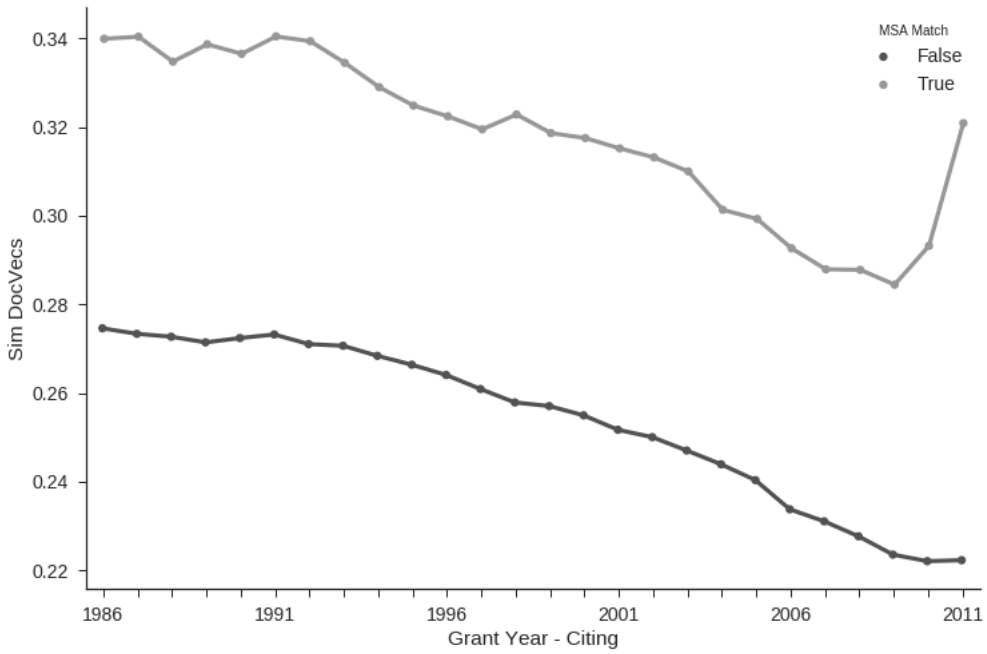


Figure B.6: Average DocVecs similarity to cited patents in the same MSA

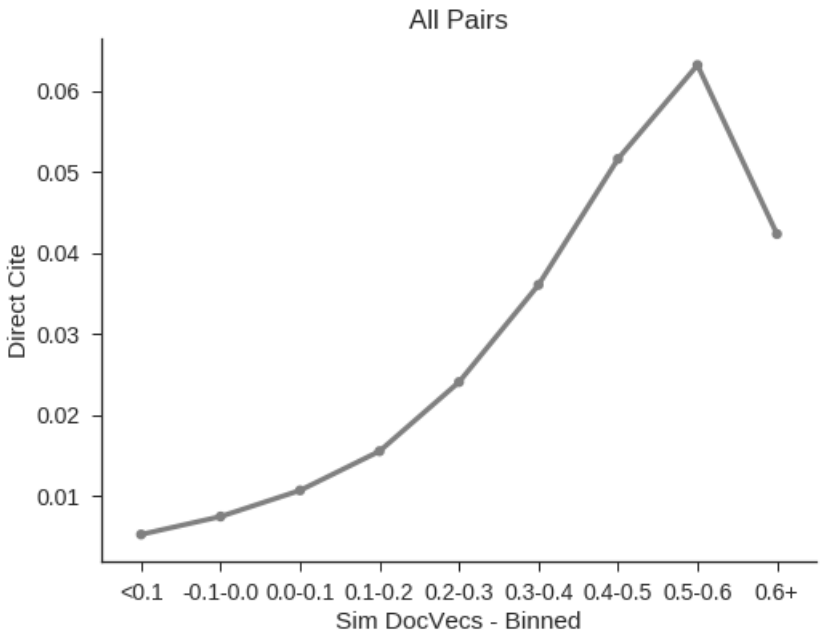


Figure B.7: Rate of direct citation conditional on level of DocVecs Similarity, All Pairs

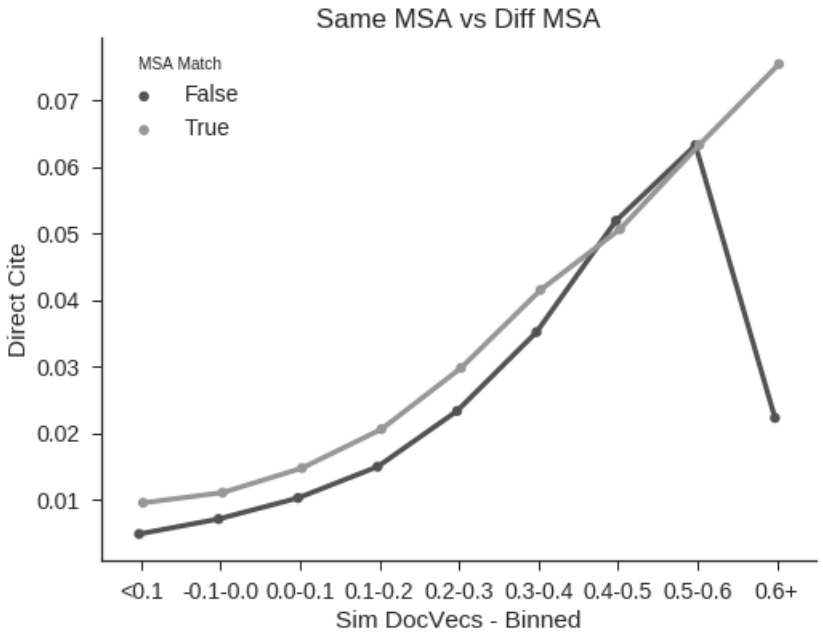


Figure B.8: Rate of direct citation conditional on level of DocVecs Similarity, Same MSA vs Different MSA

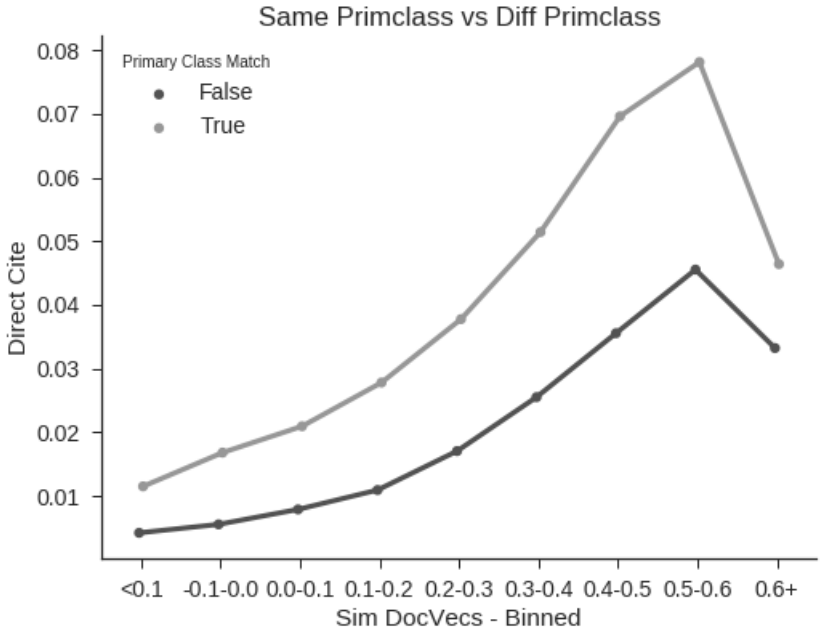


Figure B.9: Rate of direct citation conditional on level of DocVecs Similarity, Same Primary Class vs Different Primary Class

	<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. Inventor mobility

	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.1: 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.2: 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.3: Rate of self-citation before and after firm change

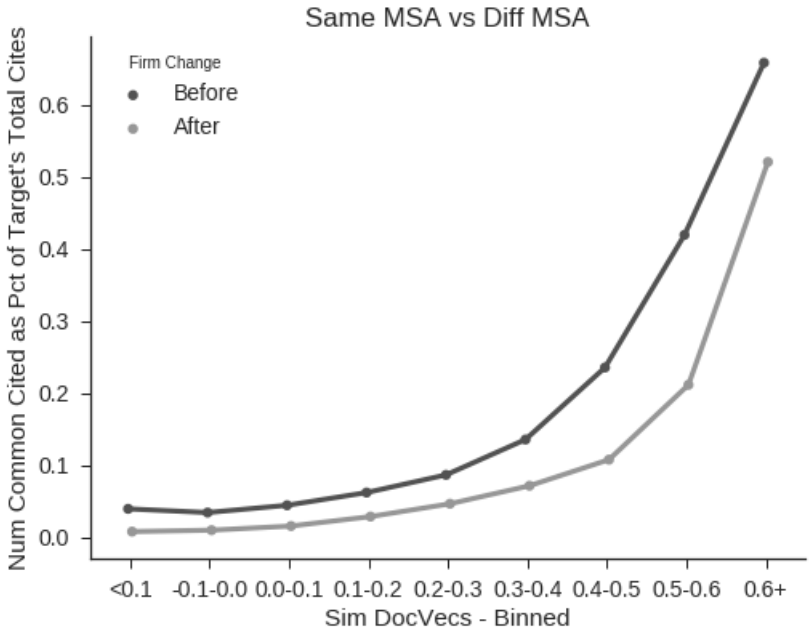


Figure C.1: Rate of direct citation conditional on level of DocVecs Similarity, Same Primary Class vs Different Primary Class

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.

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

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.

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.

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.

Fiona Sijie Feng. The proximity of ideas: An analysis of patent text using machine learning. Technical report, Working Paper, 2019. URL http://https://fsfeng.github.io/academic/assets/current_draft.pdf.

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. 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.

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.

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.

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.

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.

Martin Watzinger, Lukas Treber, and Monika Schnitzer. Universities and science-based innovation in the private sector. Technical report, mimeo, 2018.