The Proximity of Ideas

An Analysis of Patent Text Using Machine Learning

Fiona Sijie Feng (NYU Stern)

January 13, 2019

Motivation: Why do cities want Amazon's second HQ?

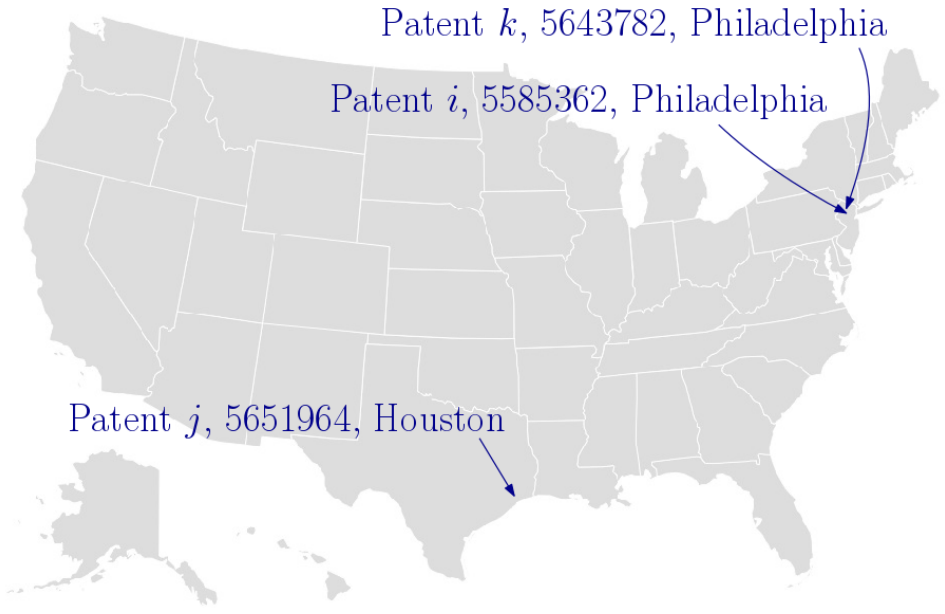
> *High-tech firms like Amazon create a "clustering effect"... whereby a company attracts workers with specialized knowledge in, say, software and data analysis. These workers are rare in other cities but reach a critical mass in a tech hub. And higher-skilled workers are more productive when they work in proximity to each other, sharing ideas and experiences.*

Chicago Tribune, 18 October 2017

Geography of Innovation and Knowledge Spillovers

- Innovation concentrated in top 10 U.S. cities:
  - 40% of new U.S. patents in 1975-85, increased to 46% for 2005-2015
- One reason is local knowledge spillovers (LKS):
  - Geographic proximity generate positive externalities or "spillovers" in knowledge transfers favoring local firms and inventors (Marshall, 1890).
- Cities want other local firms to learn from Amazon and also innovate in tech
  - Policymakers in NYC offered Amazon $1.3 bn in benefits; indicates knowledge spillovers have some perceived value
  - Not without criticism: "States also need to figure out the impact of any subsidy on other economic development plans, and not simply fall for name recognition over common sense." (TechCrunch, August 2018)

Standard Measure of Knowledge Spillovers: Citations



- Is Philadelphia patent ***k*** more likely to cite another Philadelphia patent ***i*** compared to Houston patent ***j***?
- Localization is large and significant: Jaffe, Trajtenberg, Henderson (JTH, 1993); Thompson (2006); Almeida and Kogut (1997, 1999)
- JTH (1993): Patents from same city more than twice as likely to cite each other as a patents from other cities

How well do citations reflect knowledge flows?

1. Many citations added by examiners and lawyers
    - Patent examiners responsible for two-thirds of forward citations; lawyers decide many citations (Alcacer and Gittelman, 2006; Alcacer, Gittelman, and Sampat, 2009; Moser, Ohmstedt and Rhode, 2017)
    - This paper: localization in citations may be driven by lawyer's knowledge of local patents
2. Citations related to knowledge flows are strategically omitted
    - 21-33% of relevant citations are strategically withheld; non-patent citations (e.g. scientific publications) more accurate reflection of knowledge flows (Lampe, 2012; Roach and Cohen, 2013)
    - Localization may be overstated if infringement discovery less likely for non-local patents (Ganguli, Lin and Reynolds, 2017)

Issues with citations well documented but no alternative until now

Measuring KS using Patent Text

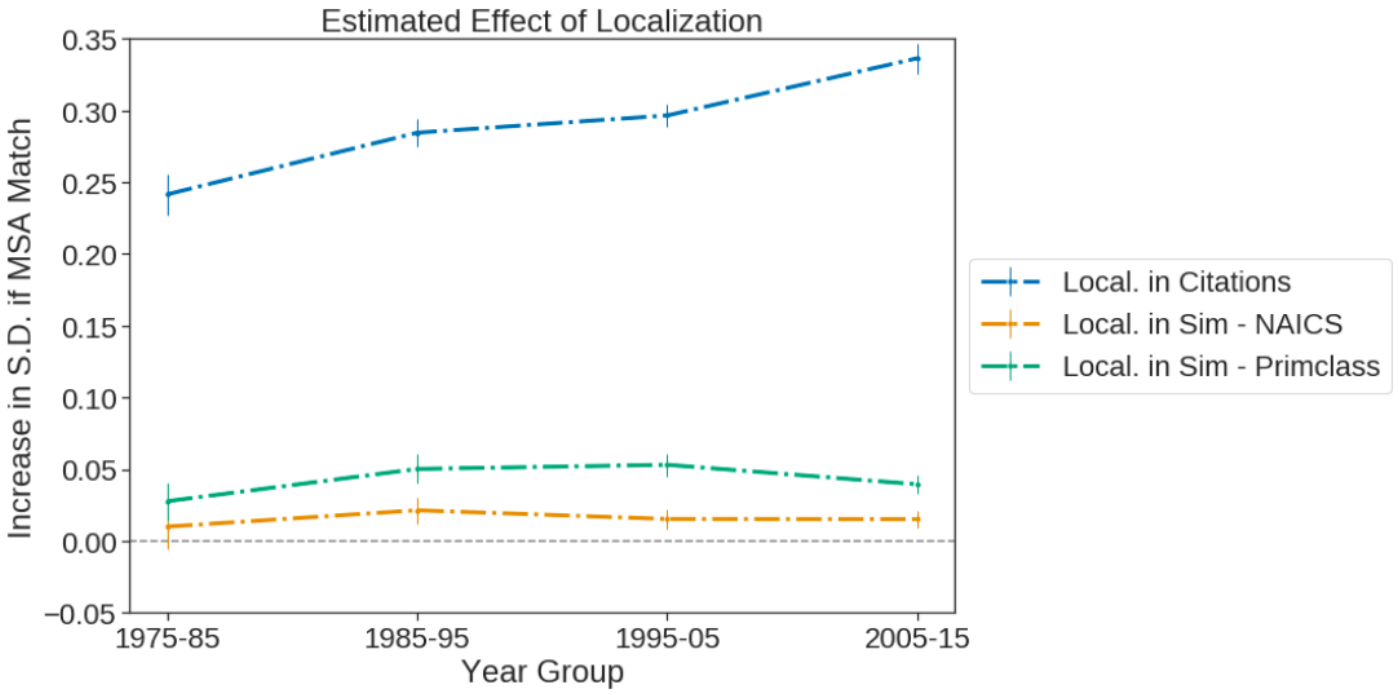
- Use patent abstract to identify the ideas of the invention
  - Machine learning algorithm converts patent abstracts to meaningful vector representations
  - Measure idea proximity using cosine similarity on patent vector representations
  - Captures knowledge relationships across patents driven by key technical terms used to describe invention
- Application of idea proximity: measuring localization
  - If there are LKS, then local inventors learning more from each other
  - Measuring localization: are local patents more textually similar to one another than compared to patents from other cities?

Main Results

1. Geographic localization has smaller impact on patent text similarity than citations
2. Text similarity less vulnerable to influence of patent examiners and lawyers
3. Remaining biases (e.g. collocation of technologically similar firms) likely overstates localization $\implies$ local knowledge spillovers may not drive agglomeration and innovation
    - Skilled labour pool and access to professional services more important
4. Location-based subsidies may yield smaller than expected agglomeration benefits

Different effect of geography on citations vs. text similarity

Local patents insig.-0.05 S.D.s more similar to each other; local citations are 0.24-0.30 S.D.s more likely for local patents.



Geographic proximity affects proximity of innovative ideas far less than citations; citations may not greatly affect innovation

Outline

Data & Methodology

Data Sources & NLP

Data

- PatentsView by USPTO (US Patent and Trademark Office): approx 2.1 million U.S. utility patents granted 1976-2015
  - Removed: patents granted to mostly foreign inventors, patents with duplicated text
  - Bibliographic data: application and grant date, assignee, PTO primary classification, abstract text
  - Complete data for citations to patents and non-patents
  - Lawyer and examiner of each patent
- Location of each patent is Metropolitan Statistical Area (MSA) where highest proportion of inventors are located
- PTO NAICS industry crosswalk: maps PTO class (450 total) to a NAICS industry code (33 industries)

Example

United States Patent **5,586,435**
Kokalis **December 24, 1996**

Hydraulic closed loop control system

Abstract

A closed loop hydraulic system especially adapted as a shot control system for a die casting machine. The control system includes a voice coil driven pilot servo valve mounted to a high flow proportional valve. The control system provides a highly controllable restriction of the outflow of hydraulic fluid from a hydraulic shot cylinder which drives a shot plunger of a die casting machine. Position transducers are provided for the pilot servo valve, proportional spool valve, and shot cylinder and a control system monitors these inputs to provide closed loop control over the shot process. A hydraulic accumulator is located in each of the pressure source and return lines of the pilot servo valve and a hydraulic accumulator tank is mounted directly to the proportional valve. The configuration of the elements of the control system along with features intended to optimize the hydraulic circuits provide a die cast shot control system with exceptionally high repeatability, frequency response and controllability coupled with low maintenance characteristics.

Inventors: **Kokalis; George P.** (Ann Arbor, MI)
Assignee: **Servo Kinetics** (Ann Arbor, MI)

Family ID: **22245571**
Appl. No.: **08/543,207**
Filed: **October 13, 1995**

Related U.S. Patent Documents

Application Number	Filing Date	Patent Number	Issue Date
389071	Jan 3, 1995		
94508	Jul 20, 1993		

Current U.S. Class: **60/416**; 164/154.2; 164/457; 60/413; 91/363R; 91/365

Why use Natural Language Processing (NLP)?

- Topic Modeling: Blei, Ng, Jordan (2003) introduced *unsupervised* machine learning (no prior labelling of data) algorithm, Latent Dirichlet Allocation, to determine underlying topics of bodies of texts
    - Widely applied in sociology, political science, legal studies, digital humanities
- Neural network and deep learning: advancements in recent years (e.g. Google Translate)
- Patents and NLP: Kelly et. al. (2018), Arts et. al. (2018), Kuhn & Younge (2016), Kaplan and Vakili (2015)
    - Mostly use count-based vectorization of patent text that does not apply machine learning freq

Unsupervised ML: Doc2Vec (Document Vectors)

How to map 2.1 million patents texts meaningfully into a vector space?

- Goal of Doc2Vec: situate documents that use similar text close to one another
- Based on algorithm Word2Vec: derives meaningful vector representations of words by situating words with similar meanings (e.g. *software, program*) close to one another in vector space
- Key idea: synonyms appear surrounded by similar context words (e.g. *"computer software tools", "computer program tools"*)
- Extension to documents: treat each document ID (patent number 558635) as a context word in each of its phrases
- Output: each word and each patent assigned vector $\in \mathbb{R}^{100}$

Document Vectors: Example

Each sentence fragment in each patent abstract form separate phrase

... controllable restriction of the outflow of hydraulic fluid ...

Document Vectors: Example

Remove common words and stem to root word

control restrict outflow hydraulic fluid

Document Vectors: Example

Identify current center word and current context words

$$\underbrace{\text{control restrict}}_{\substack{\text{context words} \\ m \text{ word window}}} \quad \underbrace{\text{outflow}}_{\substack{\text{central word} \\ \text{position } t}} \quad \underbrace{\text{hydraulic fluid}}_{\substack{\text{context words} \\ m \text{ word window}}}$$

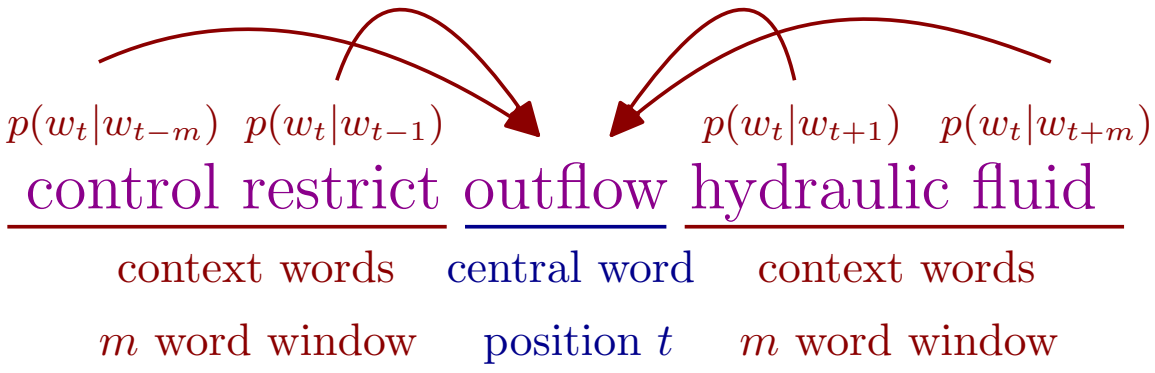
Other phrases have different center and context words

Current center word appears in different phrases

Goal: predict center word given context words

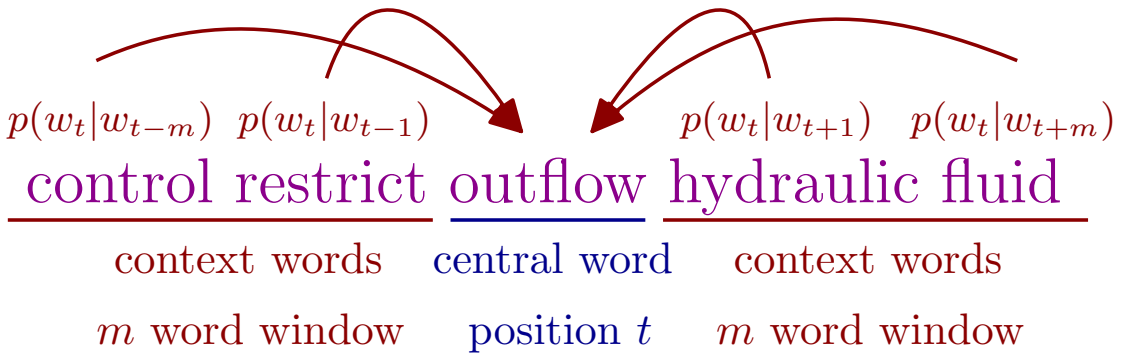
Document Vectors: Example

Choose word vectors to maximize probability of center word



Document Vectors: Example

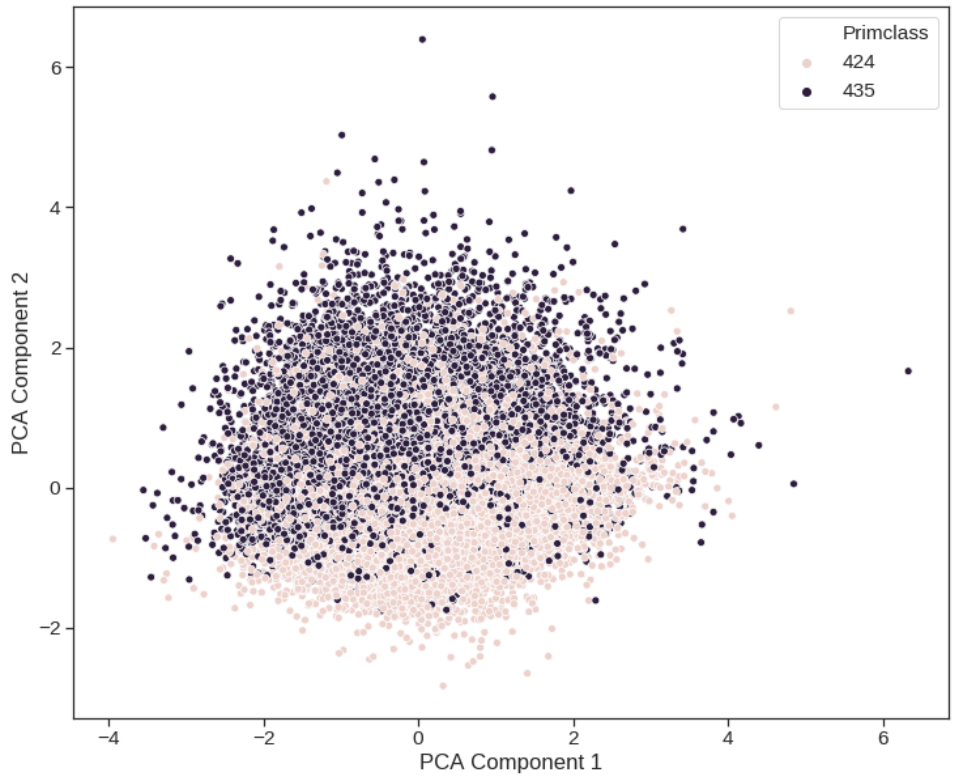
Patent ID treated as another context word in each phrase of the abstract



Patent vectors derived in same way as word vectors

Algorithm places patent vectors close to words used in abstract suppDV

Pharmaceutical Patents Principle Components Analysis



- Pharmaceutical patent DVs show some clustering across primary classes; picks up differences in terms used
  - 424 Drug, Bio-affecting and Body Treating Compositions
  - 435 Chemistry: Molecular Biology and Microbiology

Data & Methodology

Patent Similarity

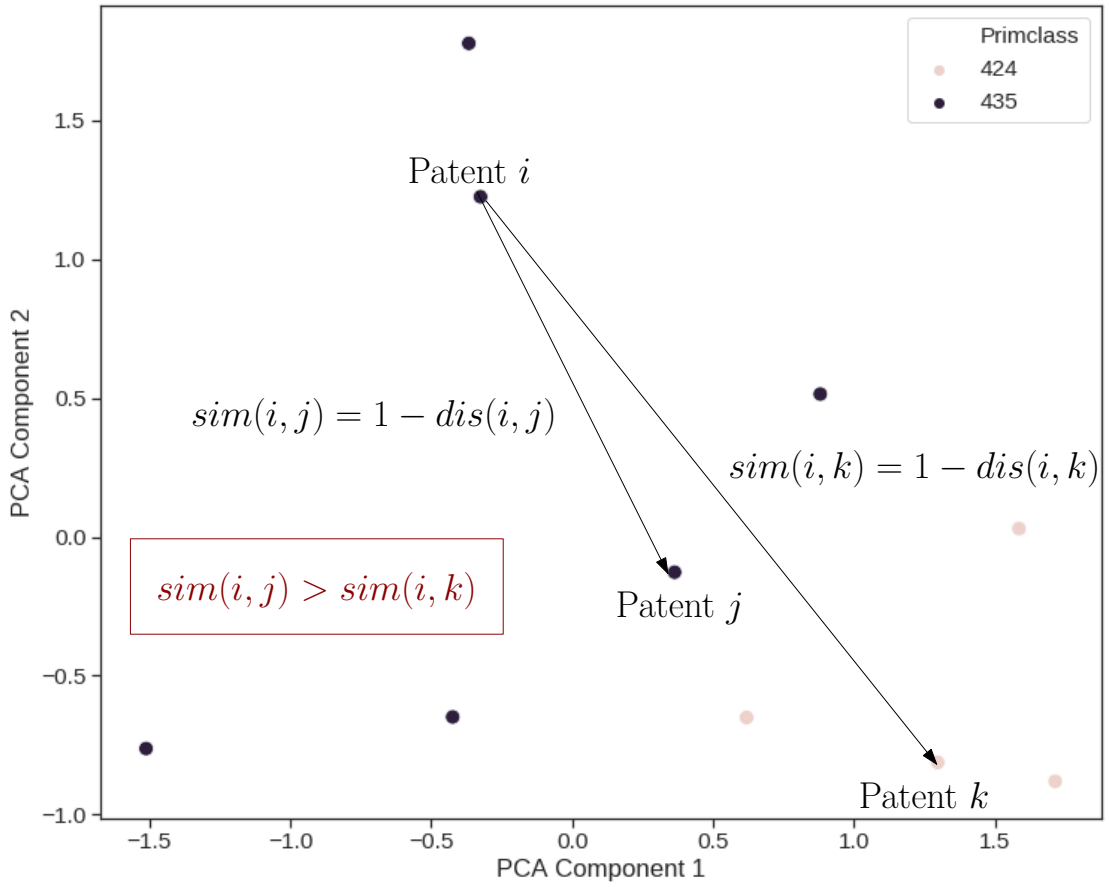
Similarity across Patent Vectors

Define idea proximity as cosine similarity between two patents i, j represented by patent vectors PV_i, PV_j:

$$sim(i,j) = \frac{PV_i \cdot PV_j}{\|PV_i\| \, \|PV_j\|} \in [-1, 1]$$

- Other similarity measures highly correlated with cosine similarity once normalized
- Cosine similarity is standard in NLP literature
- Jaffe (1989) and Bloom and Van Reenen (2008) also use cosine similarity to measure technological proximity using vectorization of patent classes

Similarity: Pharmaceutical Patents PCA Example



High Similarity Patents (0.33)

514 - Drug, Bio-Affecting, and Body Treating Compositions

- US5651964 (Houston): The method disclosed involved introduction of adenovirus early 1A gene (the E1A gene) products into affected cells. These products, which are preferably introduced by transfection of the E1A gene into affected cells, serve to suppress neu gene expression as measured by a reduction of p185 expression. ...
- US5585362 (Philadelphia): Vectors of the present invention preferably also include an additional deletion to accommodate a transgene and/or other mutations which result in reduced expression or over-expression of adenoviral protein and/or reduced viral replication.

On average, does similarity aligns with expectations?

- Similarity may not align with human judgment for some individual pairs
- This paper: sample average similarity more relevant. Can validate using what we know and expect about patents.
- Use baseline sample of patent pairs in same industry, compare sample average similarity to subsamples:
    1. Within same primary class
    2. Sharing an inventor
    3. With direct citation relationship
    4. Granted in the same year
- If these similarities align with expectations, then similarity contains useful information about how patent texts are related

Check 1: Similarity higher when expected

Expectation: similarity between patent pairs (i) within the same primary class; (ii) sharing an inventor; (iii) with direct citation relationship should all have higher similarity on average than baseline

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

Patent similarity captures higher knowledge relatedness when expected

Check 2: Similarity no different when expected

Expectation: similarity between patent pairs from the same grant year is no different from pairs granted within 5 years

Year Group	1975-85	1985-95	1995-05	2005-15
NAICS Match	0.126	0.124	0.129	0.141
S.D.	0.137	0.135	0.134	0.136
Year Match	0.126	0.124	0.129	0.141
S.D.	0.138	0.135	0.134	0.136

Patent pairs have no difference in knowledge relatedness when expected

Check 3: Greater knowledge relatedness with non-patent cite

Roach and Cohen (2013): non-patent citations are more accurate reflections of knowledge spillovers. Validate that knowledge relatedness higher when patents share a non-patent citation compared to a patent citation.

Year Group	1975-85	1985-95	1995-05	2005-15
Primclass Match	0.191	0.184	0.186	0.190
Common Pat Cite	0.322	0.304	0.287	0.272
Common Non-Pat Cite	0.355	0.503	0.394	0.418
p-value	0.761	0.000	0.000	0.000

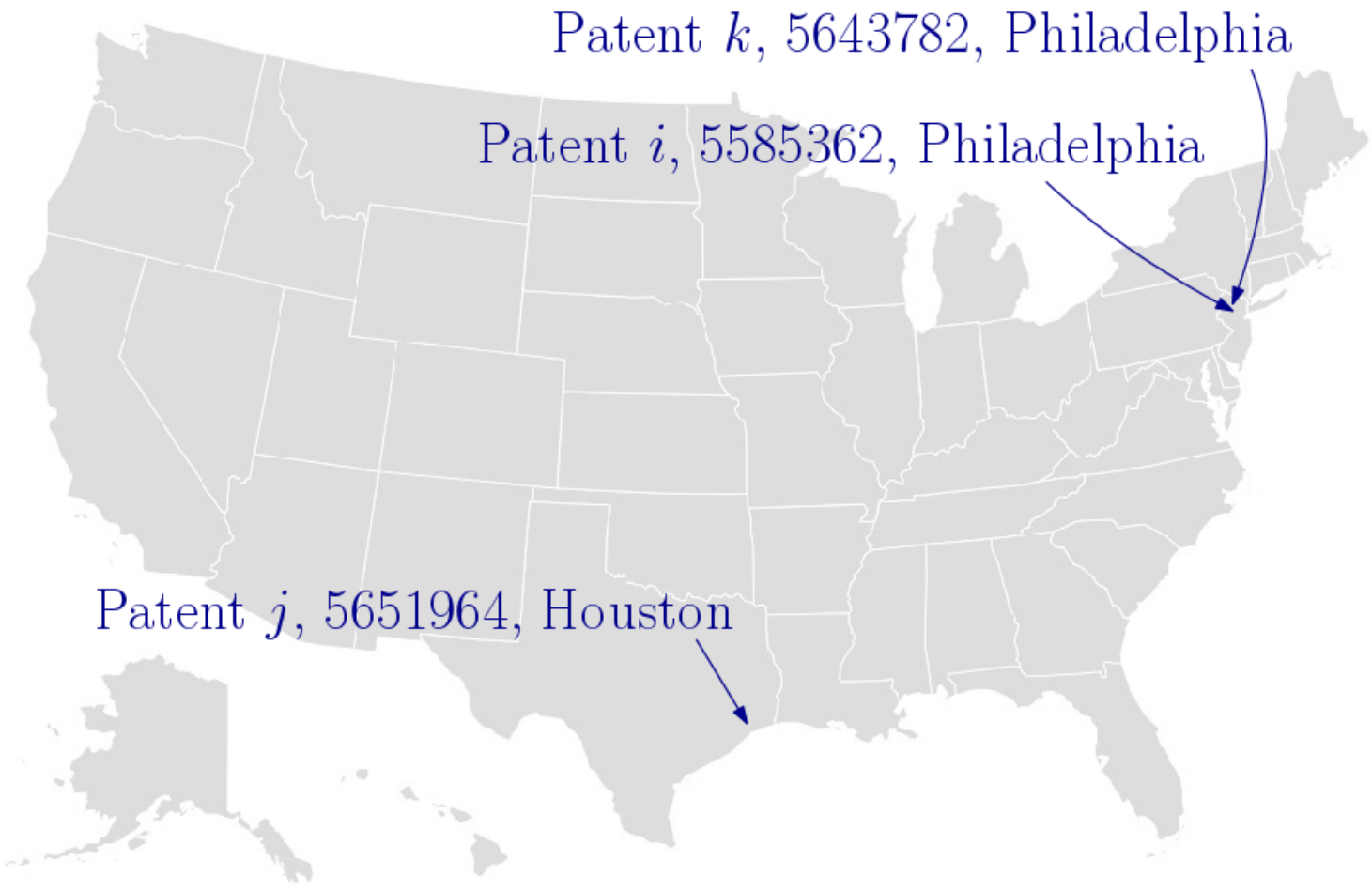
Patent pairs sharing a common non-patent citation have greater knowledge relationship compared to baseline, as well as patents sharing a patent citation

Estimation and Results

Measuring Effect of Geographic Localization

How to measure effect of geographic localization?

Pharmaceutical Patents



Two ways to measure localization

1. Localization with standard citations: are the patents from Philadelphia more likely to cite each other than the patent from Houston, conditional on being in same technology field?
    - Rationale: Geographic localization facilitates learning between inventors, reflected in reference to each other's patents
2. Localization with idea proximity: are the Philadelphia patents more similar to each other than to the Houston patent?
    - Rationale: If geographic localization facilitates learning amongst local inventors, then local inventions should be more proximate to each other

Estimation and Results

Localization with Citations

1. Localization with Citations

- Replicate Jaffe, Trajtenberg, Henderson (1993) original "treatment"/control methodology up to 2005 (+10 years forward citations)
- Dataset creation:
  1. Target patents: select all patents with forward citations 1975-2005
  2. Select control patent under different methods
  3. Remove self-citations (patents with same assignee) in each case
- Compare rate of local citations for local target patent vs a non-local control that is technologically as close as possible
  - Remove differences in citations driven by technological differences
- Standard method: select control from same USPTO primary technology class
  - Concern: some primary classes cover broad range of patents (e.g. all pharmaceutical patents fall into one of two classes)

1. Localization with Citations: Control selection

1. Standard method: same primary class, different MSA, as close in grant date as possible.
    - Target: "Card reading terminal having protective shield for input port" (312 - Supports: Cabinet structure)
    - Control 1: "Operator work station"
2. Using idea proximity: granted in same year, different MSA, with highest text similarity
    - Control 2: "Uniport interface for a bar code reading instrument"
    - May be better at addressing unobserved technological differences between target and control
3. Effect of lawyers: same primary class, different MSA, granted in same year, shares a lawyer with the target
    - Control 3: "Method of merchandising cutter bits and display case"
    - Check if influence of external parties have significant effect on localization

1. Localization with Citations: Identification

- Regression equation:

$$\textit{norm pct cites in } MSA_T = \beta_0 + \beta_1 I(MSA_i = MSA_T) + X_i + \epsilon_i$$

- Localization is measured by $\hat{\beta}_1$: How many more local citations (in S.D.s) does the local target patent receive compared to the control?
- X_i represents control variables: year and primary class fixed effects
- MSA_T: MSA of target patent
- Normalize and repeat cross sections for decades 1975-85, 1985-95, 1995-05, 2005-15

Results: Standard Primclass vs Idea Proximity

Localization in Citations, Standard Control			
	1975-85	1985-95	1995-05
Standard Primclass	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
Idea Proximity	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 using idea proximity reduces localization estimates; improves in accounting for unobserved technological differences in target and control

Results: Standard Primclass vs Lawyers

Localization in Citations, Standard Control			
	1975-85	1985-95	1995-05
Standard Primclass	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
Same 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

Localization estimates sharply declines when control is also from same lawyer; shows lawyers have considerable confounding influence

Localization with Citations: Results

- Contrary to expectations, localization estimates more affected by lawyers than by text similarity
  - Differences in citations driven not just by unobserved technological differences but also by external parties who have large influence
- Highly sensitive to control selection; source of noise for estimate
- Localization with standard method shows upward time trend co-occuring with drastic changes in citation behavior:
  - Average number of (non-self) citations to patents in prior 10 yrs rises from 3 in 1990 to 7 in 2010

Alternative approach may be helpful in circumventing some of these concerns

Estimation and Results

Localization with Similarity

2. Localization with Similarity

- Measure similarity of random patent pairs within the same technological field:
    1. Sample random pairs of patents granted within 5 years, belonging to same (i) NAICS industry; (ii) primary class
    2. Remove pairs assigned to same firm
- Use both NAICS industry and primary class to allow for intra-industry spillovers as primary class definitions can be narrow (710 Computers: Input/Output vs 711 Computers: Processing)
- No need for control selection; examines knowledge relationships across *any* pair of patents

2. Localization with Similarity: Identification

- Regression equation:

$$norm\,sim(i,j) = \beta_0 + \beta_1 I(MSA_i = MSA_j) + X_{i,j} + \epsilon_{i,j}$$

- Localization is measured by $\hat{\beta}_1$: How much more similar (in S.D.s) are local patents to each other than compared to patents from different MSAs?
- $X_{i,j}$ represents control variables: year and primary class fixed effects; more added later
- Normalize and repeat cross sections for decades 1975-85, 1985-95, 1995-05, 2005-15

2. Localization with Similarity: Results

Similarity: Within-NAICS				
	1975-85	1985-95	1995-05	2005-15
I(*MSA Match*)	0.0323***	0.0605***	0.0610***	0.0571***
	(0.0053)	(0.0043)	(0.0034)	(0.0030)
N	192841	281222	437685	569252
Adjusted R^2	0.07	0.07	0.08	0.06
Controls: Year and PC FEs				

Local patents in same NAICS industry 0.03-0.06 S.D.s more similar to each other than compared to non-local patents

2. Localization with Similarity: Results

Similarity: Within-Primclass				
	1975-85	1985-95	1995-05	2005-15
I(*MSA Match*)	0.0527***	0.0798***	0.0787***	0.0573***
	(0.0064)	(0.0050)	(0.0038)	(0.0032)
N	170882	252174	401623	529686
Adjusted R^2	0.07	0.07	0.08	0.07
Controls: Year and PC FEs				

Local patents in same primary class 0.05-0.08 S.D.s more similar to each other than to non-local patents

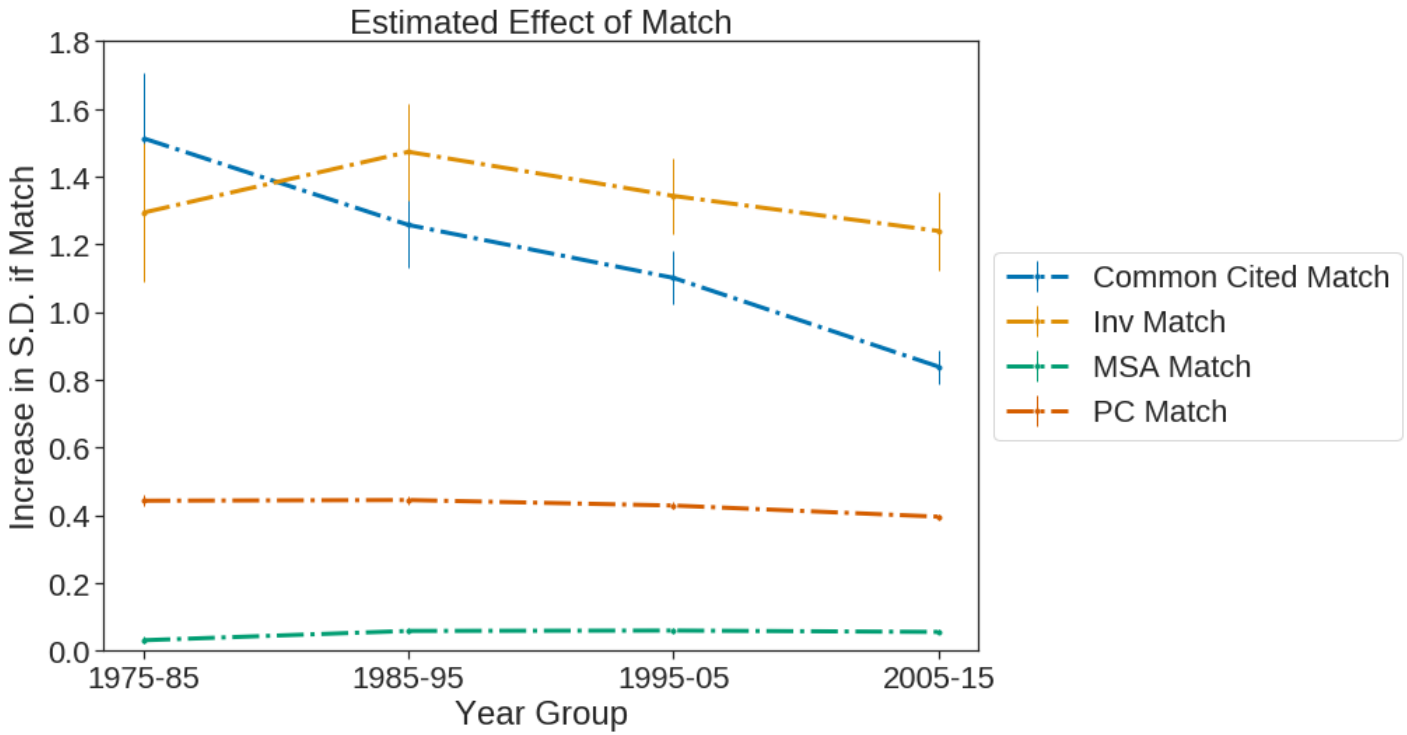
Estimation and Results

Validation Exercises

1. Are localization estimates low because similarity is noisy?

Matching on any dimension that positively affects idea proximity should produce small and imprecise estimates

Within-NAICS



Similarity can capture very large and significant effects; low estimates for geographic match not just due to noise in similarity

2. Apples-to-apples comparison of magnitudes

- For patent pairs in same primary class, compare (unconditional) ratios of:
  - Probability of being cited if patents from the same MSA vs not (due to sparsity, use sample size $\sim$ 3 mil)
  - Similarity if patents from the same MSA vs not (use subsample of size $\sim$ 1 mil)

	Pr Direct Cite, %				Similarity			
Year Group	1975-85	1985-95	1995-05	2005-15	1975-85	1985-95	1995-05	2005-15
MSA Match = T	0.456	0.663	0.427	0.206	0.196	0.193	0.195	0.197
MSA Match = F	0.087	0.104	0.072	0.033	0.190	0.182	0.184	0.188
Ratio	5.271	6.398	5.960	6.146	1.033	1.060	1.064	1.043

Localization measured by Pr Direct Cite much greater than patent similarity; consistent with prior results examiners

3. Effect of lawyers and examiners on text vs citations?

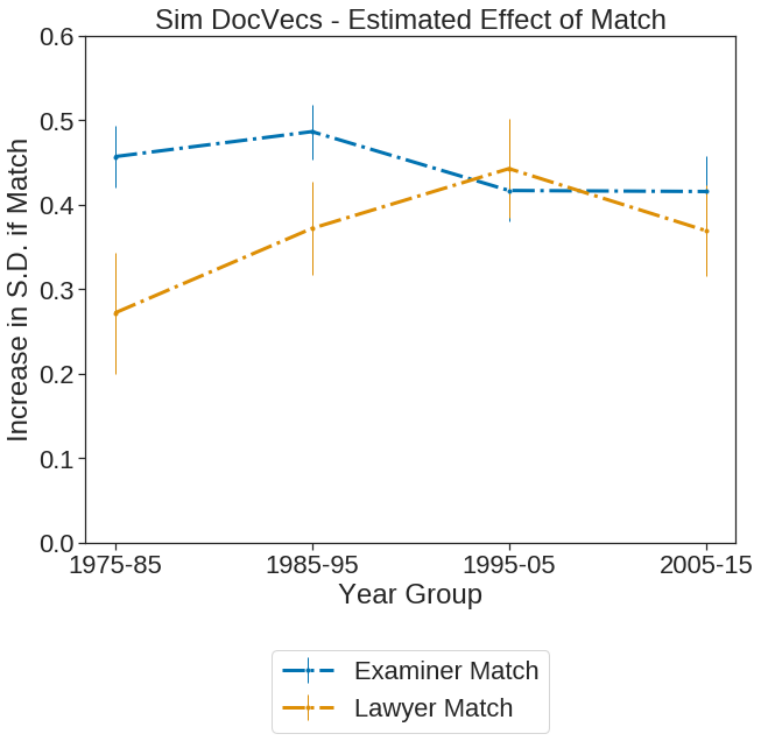
- How much does patent text similarity increase when two patents share a lawyer or examiner compared to patents that do not?
- How much does shared citations to prior patents increase?

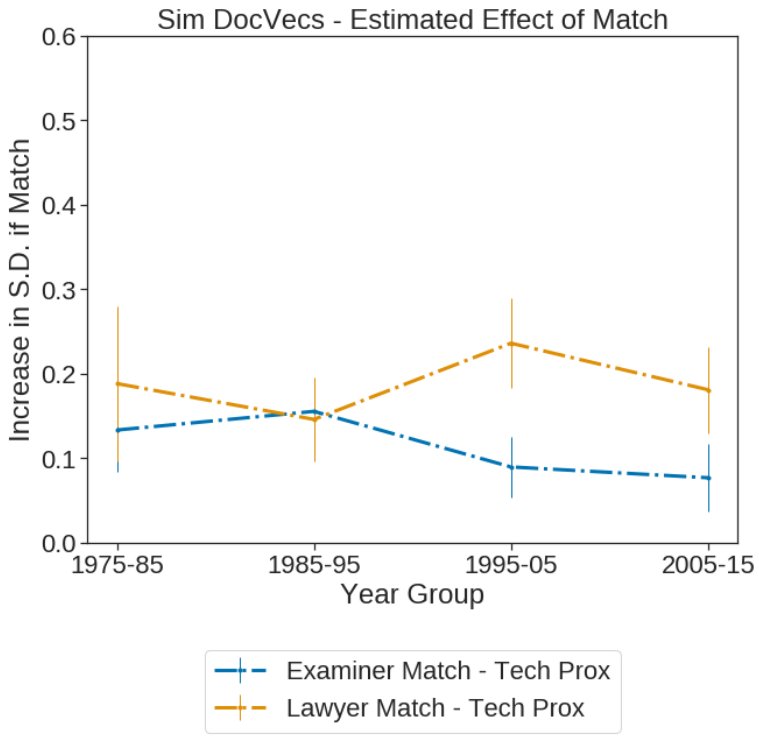
$$\text{Pct Common Cited} = pcc(i,j) = \frac{num\ citations\ by\ i,j}{num\ citations\ by\ i}$$

- Lawyers/examiners are selected/assigned in manner correlated with technology field of patent
  - Include control for technology field proximity using similarity in primary classes:

    $sim(pc_i, pc_j) = mean\left(\{sim(i,j)|i \in pc_i, j \in pc_j\}_{t-5,t}\right)$
  - Average similarity of prior patents in each of their respective primary classes in the past 5 years

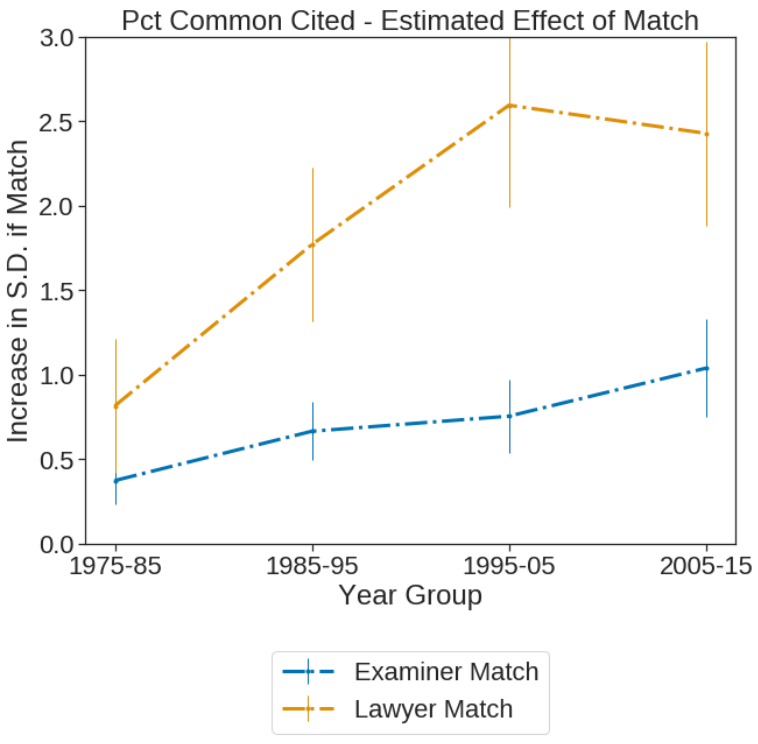
3. Effect of lawyers and examiners: Similarity

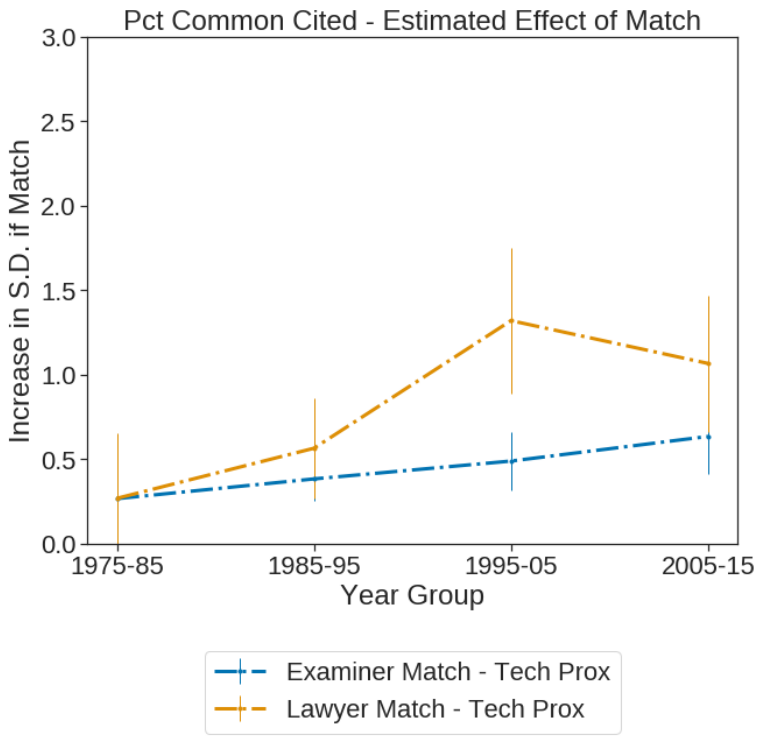




- Patents sharing a lawyer have about 0.2 S.D.s higher text similarity after including tech. prox.; 0.1-0.2 S.D.s for examiners
- Effects are fairly stable over time, declining at least in 2005-15

3. Effect of lawyers and examiners: Citations





- Patents sharing a lawyer have about insig.-1.3 S.D.s higher shared citations; 0.3-0.6 S.D.s for examiners
- Effect of lawyers and examiners on citations is much larger; trend is rising over time moreapples

4. Localization estimates after including other controls

- Further add control variables that may increase patent text similarity unrelated to possible knowledge spillovers: inventor mobility, lawyers, examiners match indicators and fixed effects
- Simultaneity bias may remain: firms from similar technology fields may choose to collocate in same MSA $\implies$ positive bias, localization still overestimated
    - For within-NAICS sample, partially address through inclusion of technology field proximity control

4. Further controls: Localization with Similarity

Similarity: Within-NAICS				
	1980-85	1985-95	1995-05	2005-15
I(*MSA Match*)	0.0170	0.0390***	0.0300***	0.0274***
	(0.0120)	(0.0070)	(0.0048)	(0.0038)
N	40323	110982	215861	344313
Adjusted R^2	0.12	0.13	0.13	0.08
Controls: Year, PC, MSA, Examiner, Lawyer FEs, Tech. Prox.				

Localization estimates decrease further to insig.-0.04 S.D.s when other controls included

4. Further controls: Localization with Similarity

Similarity: Within-Primclass				
	1980-85	1985-95	1995-05	2005-15
I(*MSA Match*)	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 FEs				

Localization estimates decrease further to 0.03-0.05 S.D.s

Discussion of Results

Why do results differ for citations vs similarity?

1. Citations may overstate localization:
    - Reflects lawyer's knowledge of patents from their local cities
2. Citations may not greatly influence innovation:
    - Highly similar inventions arise in other cities through very different citation patterns
    - Innovation may be "homogenized" across locations through use of external knowledge (e.g. scientific publications)
        - Inventors learn from outside knowledge sources not just other inventors and their patents
    - Inventors' citations change drastically after switching firms policy

concl

Citations may overstate localization

- Possible mechanism of overestimation: citations reflect lawyer's knowledge of potentially relevant patents
  - Use sample of target patents from standard citations exercise
  - Effect of lawyers' location on location of forward citations: what pct of target's forward citations come from cities that their lawyer also operates?

	1975-85	1985-95	1995-05
Pct Cites in Lawyer's MSAs	32.55	39.84	47.03
S.D.	31.93	33.52	37.4
N	31546	54754	90243

Lawyers' knowledge of related patents in their local cities may be mechanism through which citations overstates localization

Many highly similar inventions exist in other cities

- For each patent in prior sample, find top 50 most textually similar patents (i.e. "nearest neighbours"; remove those from same firm). Are neighbours from the same MSA?

	1975-85	1985-95	1995-05	2005-15
Pct Neighbours in Same MSA	3	3.22	4.02	4.41
S.D.	5.09	5.29	6.39	6.48
Pct Cites in Same MSA	9.16	9.72	10.96	12.99
S.D.	23.62	22.93	23.34	27.92
N	24124	41341	69699	61273

Pct of neighbours in same MSA far smaller than pct of citations in same MSA; pct of closest neighbours in same MSA slightly higher than pct of random patent pairs in same MSA (2.4%)

Many similar inventions exist in other cities

- For patents and their nearest neighbours, find pct common citations ($pcc(i,j)$)

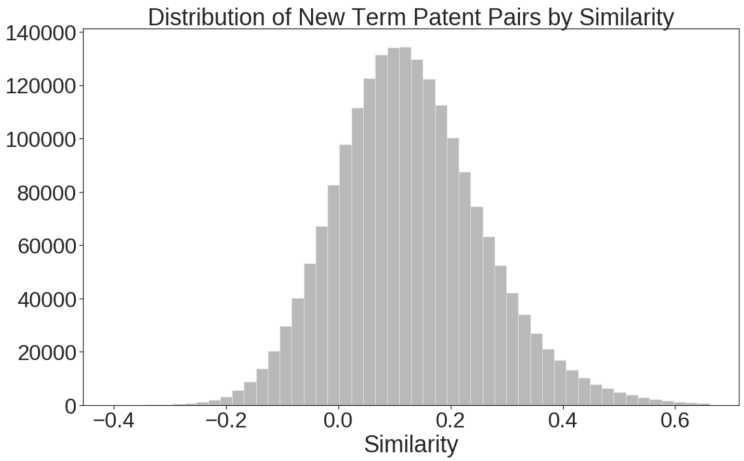
	1975-85	1985-95	1995-05	2005-15
$pcc(i,j)$	0.35	0.44	0.56	0.55
S.D.	0.83	1.08	1.96	2.36
N	23439	40294	58529	59154

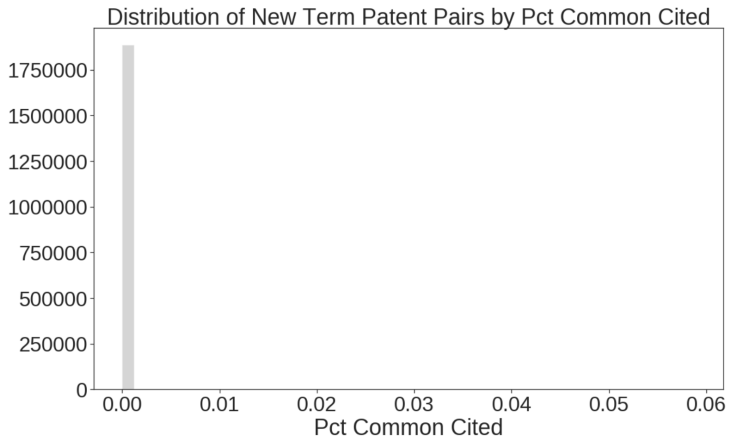
Similar inventions arise through very different citation patterns

- Example: US7086158 "Method for forming a dual metal gate structure" and US7531398 "Methods and devices employing metal layers in gates to introduce channel strain" has $sim(i,j) = 0.60$ but no common citations

Knowledge from external sources: new terms

- Citations reflect only patent references while patent text reflect influence of many sources of knowledge
  - Examine sample of patents first to use new technological terms (e.g. *bluetooth, adenovirus*) learned from external sources
  - Compare similarity and pct common citations to see if knowledge relatedness is captured





Similarity captures some knowledge relatedness in patents that are first to apply new technology; citations indicate no knowledge relationship

Knowledge from external sources: new terms

Two of the first "adenovirus" patents:

- US5651964 (Houston): The method disclosed involved introduction of adenovirus early 1A gene (the E1A gene) products into affected cells. These products, which are preferably introduced by transfection of the E1A gene into affected cells, serve to suppress neu gene expression as measured by a reduction of p185 expression. ...
- US5585362 (Philadelphia): Vectors of the present invention preferably also include an additional deletion to accommodate a transgene and/or other mutations which result in reduced expression or over-expression of adenoviral protein and/or reduced viral replication.

Similarity = 0.33; no common patent citations. Indicates external knowledge sources affects innovation when no citation "paper-trail" present.

Inventors' citations change after switching firms

- Inventors who move firms still work in highly similar areas
- Using sample of patents by mobile inventors:
  - Compare pct shared citations to prior patents before and after moving firms, conditional on inventions being highly proximate

$sim(i,j) > 0.3$	$prior\ inv_i, prior\ inv_j$	$prior\ inv_i, new\ inv_j$
$pcc(i,j)$	23.66	10.00
S.D.	37.74	26.39
N	91191	34260

Patents cited after moving firms drastically different even when inventor's new patents highly similar to prior patents

Inventors' citations change after switching firms

Example:

- Prior to changing firms inventor had two patents:
  - US4549561 'Coin handling machine'
  - US4564036 'Coin sorting system with controllable stop'
  - $sim(prior\ inv_i, prior\ inv_j) = 0.57$; $pct\ common\ cited = 100\%$
- After changing firms:
  - US4549561 'Coin handling machine'
  - US5297986 'Coin sorting apparatus with rotating disc'
  - $sim(prior\ inv_i, post\ inv_j) = 0.60$; $pct\ common\ cited = 12.5\%$

Citation patterns can vary widely even for the same inventor

Policy Implications

- Citations-based literature: important to have firms and inventors geographically close to one another to learn from each other
  - Policymakers should incentivize geographic proximity, e.g. Kendall Square Initiative in Boston or the Research Triangle in North Carolina
- My results: unclear if geographic proximity facilitates much knowledge exchange
  - Inventors can acquire relevant knowledge for innovation through external sources
- For local government: not much of a case for incentivizing geographic proximity from perspective of knowledge spillovers
- For non-local firms and inventors: may also benefit if localization effects not as strong

Conclusion

- Local knowledge spillovers may not greatly affect innovation and agglomeration
- Other applications of idea proximity:
  - Patent text: Is specialisation or diversity in knowledge is better for innovation and regional growth? Do innovation clusters grow or decline after entry of large firm?
  - Academic text: How did availability of Department of Defense funding shape the trajectory of computer scientific research and development?
  - Job review ratings and text: Does workers' subjective well-being affect firm performance?

Further discussion of Doc2Vec

- Punctuation is removed: including punctuation shown in literature to perform worse
- Continuous Bag of Words (CBOW): ordering of context words irrelevant
    - Probability of center word (m) given context word (c) given by "softmax" function:

$$p(m|c) = \frac{\exp(u_c^T v_m)}{\sum_{w=1}^{V} \exp(u_w^T v_m)}$$

- Optimization problem: maximize log probability of all center and context words
    - Algorithm solved using stochastic gradient descent
- Other NLP methods (Latent Dirichlet Allocation) used as robustness check, results are aligned with DocVecs DVEx

Effect of lawyers on Pr Direct Cite vs Similarity

	Pr Direct Cite, %				Similarity			
Year Group	1975-85	1985-95	1995-05	2005-15	1975-85	1985-95	1995-05	2005-15
Within-NAICS								
Lawyer Match = T	1.825	1.786	2.009	1.390	0.152	0.170	0.192	0.191
Lawyer Match = F	0.016	0.023	0.020	0.011	0.125	0.121	0.127	0.139
Ratio	116.862	78.594	101.838	131.096	1.219	1.400	1.514	1.376
Within-Primclass								
Lawyer Match = T	5.304	6.488	5.880	3.989	0.274	0.260	0.268	0.258
Lawyer Match = F	0.123	0.164	0.120	0.057	0.191	0.184	0.186	0.190
Ratio	42.992	39.640	48.935	70.545	1.440	1.418	1.442	1.356

Pr Direct Cite between patents much higher if both patents share a lawyer

Effect of examiners on Pr Direct Cite vs Similarity

	Pr Direct Cite, %				Similarity			
Year Group	1975-85	1985-95	1995-05	2005-15	1975-85	1985-95	1995-05	2005-15
Within-NAICS								
Examiner Match = T	0.486	0.548	0.845	0.688	0.201	0.202	0.201	0.205
Examiner Match = F	0.016	0.025	0.021	0.012	0.124	0.120	0.127	0.139
Ratio	29.604	21.601	40.107	57.685	1.623	1.679	1.582	1.477
Within-Primclass								
Examiner Match = T	0.347	0.476	0.515	0.388	0.208	0.198	0.199	0.205
Examiner Match = F	0.122	0.176	0.126	0.060	0.188	0.182	0.185	0.190
Ratio	2.850	2.709	4.092	6.423	1.105	1.089	1.076	1.079

Pr Direct Cite between patents much higher if both patents share an examiner; smaller effect if in same primary class apples further

Frequency vectorizations vs Machine Learning

- Frequency vectorizations of words in abstracts cannot account for synonyms (e.g. *software, program, algorithm*)
- Patent abstracts also repeat many terms such as *system, method, control* which may overstate similarity in some texts
- Raw text more affected by lawyers' "style" of writing
- Campr and Jezek (2015) finds Doc2Vec closest to human judgment of document similarity Doc2Vec