A History of Collaboration in U.S. Invention: Changing Patterns of Co-invention, Complexity and Geography

Frank van der Wouden*
Department of Geography
University of California, Los Angeles

1255 Bunche Hall
Los Angeles, California
United States of America

*fvanderwouden@ucla.edu

Author's manuscript

As published in:

Industrial and Corporate Change (2020)

Cite as:

Frank van der Wouden, **A history of collaboration in US invention: changing patterns of co-invention, complexity and geography**, *Industrial and Corporate Change*, Volume 29, Issue 3, June 2020, Pages 599- 619, https://doi.org/10.1093/icc/dtz058

Abstract

Research suggests that increasing collaboration in knowledge production is explained by rising complexity of knowledge. Yet, there is little long-run, systematic, empirical evidence on the relationship between complexity and collaboration. A new database is introduced that identifies all (co-)inventors on more than 3 million US patents between 1836-1975. Empirical analysis reveals (1) collaboration on U.S. patents began to increase in the 1940s; (2) there is a robust positive relationship between complexity & collaboration; and (3) increasing complexity is associated with local rather than non-local collaboration.

Keywords: collaboration, complexity, geography, inventors, patents

Introduction

In a recent paper, Wuchty et al. (2007) show that U.S. invention increasingly results from collaboration. Theories explaining scientific collaboration are extensive, ranging from resource optimization (Eaton, 1951), increased productivity (de Solla Price, 1986), access to ideas and resources (Wray, 2002), to intellectual or social linkages (Thorsteinsdottir, 2000). A broad argument that runs through much of this work is that knowledge production is becoming increasingly complex, requiring inputs that exceed the capacity of individual inventors. For example, to generate new knowledge on particle physics the Large Hadron Collider Project pools resources from over 100 countries, linking thousands of scientists and engineers, many of whom bring highly specialized and unique skillsets to the research team.

Political and scholarly interest in collaboration and complexity rests upon the notion that both have a positive relationship with the value, quality or quantity of output. Collaborative work is found to spur creativity (Uzzi & Spiro, 2005), receive more citations (Glanzel, 2002), have higher acceptance rates for academic publication (Presser, 1980), raise productivity (Lee & Bozeman, 2005) and enable researchers to engage with larger research questions (Thagard, 1997). In similar vein, complex knowledge is seen as a key source of competitive advantage (Kogut & Zander, 1992; Maskell & Malmberg, 1999; Sorenson, Rivkin, & Fleming, 2006) for firms and regions (Balland & Rigby, 2017). Access to complex knowledge allows inventors to engage with more advanced activities and capture the benefits that might arise from such engagement. These benefits may be pecuniary within firms or manifest as increased levels of economic development within society.

Surprisingly, empirical evidence on the relationship between complexity and collaboration is limited. There is no long-run, large scale, systematic evidence linking collaboration and complexity, primarily because empirical data on historical collaboration is lacking. Most research on collaboration focuses on relative short time-frames, specific fields or projects. The same is true for measures of complexity because these are difficult to design, construct and operationalize over time and space.

The primary aim of this paper is to examine patterns of co-invention on US patents along the axes of collaboration, complexity and geography. Using state-of-the-art machine learning and text mining algorithms, I construct a unique inventor-patent database that identifies all (co-)inventors and their geographical location(s) on more than 3 million patents between 1836 and 1975. The raw data originate in the United States Patent and Trademark Office (USPTO) and in the HISTPAT database of Petralia et al. (2016). I merge the inventor collaboration data with information on the technology classes of US patents, available from the USPTO, to construct measures of complexity for each patent. The resulting data reveals the characteristics of US inventor collaboration on patents of varying complexity, across geographies and at different moments in time. The key findings of this study are that (1) collaboration has increased sharply since the 1940s; (2) there is a positive and significant relationship between complexity and collaboration across US patents; and (3) the increasing complexity of patents promotes local rather than non-local collaboration.

The rest of this paper is structured as follows. In the second section I motivate the research and discuss recent work concerning collaboration, knowledge complexity, the geography of collaboration, and I situate this work in the context of shifts in U.S. economic history. The third section describes the construction of the historical long-run U.S. patent-inventor database. Section four presents the empirical analysis. The final section presents and discusses the key findings of this paper.

2. Literature

2.1 Economic and institutional context

This paper examines patterns of collaboration between inventors on U.S. patents between 1836 and 1975. This time-window covers important transformations that significantly impacted the U.S. economy. The period 1870-1970 is often referred to as the "Golden Age" of U.S. invention (Akcigit et al., 2017; Gordon, 2016). Researchers of American innovation have long stressed the importance of understanding shifts in the U.S. institutional environment across this period that altered the nature of invention.

First, the construction and enforcement of a patent system created a market for technologies which allowed inventors to appropriate, sell or license their ideas, providing incentives for inventors to specialize in technology (Lamoreaux et al., 2013; Lamoreaux & Sokoloff, 1996). Although the United States was not the first country to develop such a system, it is generally perceived as constructing the first *modern* patent system in 1836[1]. This system with certified examiners promoted the diffusion of technological knowledge through publication of patent records (Lamoreaux & Sokoloff, 1999; Lamoreaux & Sokoloff, 2005).

Importantly, under the 1836 US patent system, only individuals were able to apply for patents. Firms were not allowed to patent, even if the invention was created in their shop (Coriat & Weinstein, 2012). Fisk (1998) argues that part of this legislation rested upon the notion that an invention is the result of one person's 'genius'. To maintain their competitive position, firms purchased or licensed new technological ideas from inventors, thus securing the market for knowledge (Lamoreaux & Sokoloff, 2005). The construction of the US postal system significantly impacted technological progress. During the late 18th and early 19th century, the postal system rapidly extended communications throughout the country, providing access to information and long-distance communication (John, 2009). Acemoglu et al. (2016) provide empirical evidence that U.S. counties with a post office had significantly greater patenting rates than counties without a post office.

Transformations of the US capitalist system also significantly affected US knowledge production. At the turn of the 19th and 20th centuries the US transformed from entrepreneurial capitalism to corporate capitalism, dominated by large firms (Coriat & Weinstein, 2012). With the anti-trust Sherman Act of 1890, large corporations faced increased competition and relying on the acquisition of technologies from external inventors was an increasingly risky strategy. As an alternative, firms began to construct in-house R&D facilities to internalize knowledge production, complicating questions over the ownership of intellectual property. Although the 1836 Patent Act only allowed individuals to apply for patents, knowledge ownership rights began to change with the emergence of the 'shop right' doctrine in the 1880s. Shop rights' granted employees ownership of patents they invented, but employers were granted licenses to the patented technology if they funded the research producing the new ideas. By the early 1900s new technologies became increasingly complicated, demanding multiple and different skills and resources. Fisk (1998, p.1141) argues that "collective research and development had become the source of most inventions long before courts and the public finally realized it". This new reality led to a 1933 Supreme Court ruling that 'the respective rights and obligations of employer and employee, touching an invention conceived by the latter, spring from the contract of employment' (in Coriat & Weinstein, 2012). Thus, if employees were 'hired-to-invent', the product of their labor now belonged to the firm.

2.2 Collaboration

Scholars across different disciplines have recognized the increasing presence of collaboration in the production of knowledge. Since the 1950s, there is empirical evidence that knowledge production in general is no longer the product of the 'lone-wolf' (Patel, 1973) or individual genius (Merton, 1968), but increasingly the outcome of cooperation. This trend is observed in almost every field and across the globe (Crescenzi et al.,2016; Hoekman et al., 2010; Jones et al., 2008; Merton, 1973; Wuchty et al., 2007).

The increasing complexity of contemporary knowledge production requires inputs that exceed that of the individual. De Solla Price (1963) refers to this trend as the shift to 'big science', emphasizing the increasing importance of complex science-based technologies (Noble, 1979; Pavitt, 1984). Within such a system, knowledge producers collaborate to optimize resources (Eaton, 1951), increase productivity (de Solla Price, 1986), access complementary ideas and resources (de Solla Price, 1970; Wray, 2002) or for social reasons (Thorsteinsdottir, 2000). Jones (2009) argues that the cumulative nature of knowledge produces a 'knowledge burden' that makes subsequent innovation all the harder. This knowledge burden is dampened to some degree by collaboration (Uzzi & Spiro, 2005).

Collaboration is positively associated with the value, quality or quantity of output. Collaborative work receives more citations than individual work (Frenken et al., 2005; Glanzel, 2002), it has higher acceptance rates in academic publication (Presser, 1980), it raises productivity (Lee & Bozeman, 2005) and enables researchers to engage with more significant research questions (Thagard, 1997). On the team-level there is ample evidence that diversity among collaborators increases performance (Burt, 2004; Uzzi & Spiro, 2005). Faems et al. (2005) find that collaborating firms are more likely to produce commercially successful products than non-collaborating firms. Katz & Martin (1997) provide a critical review on research collaboration.

An important subset of collaboration in knowledge production is inventor collaboration. The collaboration of inventors has received increasing attention in the last few decades, ranging from analysis of the composition of collaborative teams (Bercovitz & Feldman, 2011; Crescenzi et al., 2016) to network formation in groups of inventors (Cowan, Jonard, & Zimmermann, 2007) to examination of co-invention and innovative productivity (Breschi & Lenzi, 2016; Giuliani, Martinelli, & Rabellotti, 2016; Van der Wouden & Rigby, 2019). In this work, collaborative inventive activity is seen as an important aspect of the knowledge creation process, because it is a key mechanism through which knowledge flows between agents (Strumsky & Thill, 2013).

The analysis of inventor collaboration has been restricted to relatively recent data. Using USPTO patent records from 1975 to 1995, Wuchty et al. (2007) report that the average number of inventors on patents has been increasing over time. Others report similar findings (Crescenzi et al., 2016; Fleming & Frenken, 2007; Lobo & Strumsky, 2008). While this research shows the rise of inventor collaboration over the last 30-40 years, we have little information about the longer-run history of co-invention. Collaboration on inventive activities before 1975 remains largely a black box.

2.3 Knowledge complexity

The production of knowledge is widely recognized as a critical source of competitive advantage and long-run economic growth (Romer, 1986; Solow, 1957). Combined with a growing awareness of the heterogeneity of firms (Nelson & Winter, 1982), this has prompted development of the knowledge-based view of the firm, in which coordination, recombination and integration of the (specialized) knowledge of individuals is seen as central to firm performance

(Grant, 1996; Almeida and Kogut, 1999). That not all knowledge is equal is clear from Polanyi (1958). Thus, some forms of knowledge impact the competitiveness of firms and regions more than others (Maskell & Malmberg, 1999; Asheim and Gertler, 2005).

There is growing awareness of the importance of complexity to the production and value of knowledge. Knowledge complexity is linked to the architecture of ideas by Simon (1962). Kogut & Zander (1992) argue that complexity is a critical dimension of what makes knowledge tacit. Different visions of knowledge complexity exist in the field. For Fleming & Sorenson (2001), complexity can be derived from the interaction between knowledge components and how easily these components can be combined. Others argue that knowledge is complex when it can surprise the observer and its characteristics cannot simply be linked to components (Axelrod & Cohen, 2000; Tsoukas, 2005). From both perspectives, complex knowledge is emergent and more likely to rely on tacit, non-ubiquitous knowledge than less complex knowledge. Like tacit knowledge, complex ideas tend not to flow readily over space (Balland & Rigby, 2017; Sorenson et al., 2006).

Many scholars have argued that knowledge is becoming increasingly complex over time, witnessed in the shift to big science and the importance of science-based technologies (Pavitt, 1984; de Solla Price, 1963). Some stress this is troublesome as producing new knowledge becomes increasingly difficult and costly, slowing economic growth (Jones, 2009). Surprisingly, there is limited empirical evidence of the growth of knowledge complexity over time, though Balland et al. (2018) explore this question. This paper explores long-run shifts in knowledge complexity within U.S. patents and links collaboration to these shifts.

2.4 Geography of collaboration

For many of the reasons discussed above, the production of complex knowledge is more often the result of collaboration than is the production of less complex knowledge. At this time, we know relatively little about the characteristics of co-inventors on patents that exhibit different degrees of complexity. Sorenson et al. (2006) and Balland & Rigby (2017) use patent citation data to report that more complex knowledge does not flow as readily as less complex knowledge. However, it is unclear whether this means that there is an inverse relationship between the knowledge complexity of patents and the distance separating the co-inventors of these patents.

Within the economic geography literature debate is raging around the nature of knowledge flow and the extent to which local and non-local forms of interaction are more important for aggregate regional performance. On the one hand, it is argued that knowledge production depends heavily on face-to-face interactions provided by local buzz, the dynamics and interactions that arise from the co-location of economic agents (Storper & Venables, 2004). On the other hand, Bathelt et al., (2004) claim that knowledge pipelines linking clusters of agents across space have been given insufficient attention. There is plenty of empirical work in support of both sides of this debate, and growing awareness of the costs as well as the benefits of interaction (Esposito & Rigby, 2018).

To the extent that knowledge varies in quality, and that more complex knowledge tends to be more valuable, examining the geography of co-inventorship across patents characterized by different levels of complexity would add to these debates. This question is explored in the historical analysis of collaboration and complexity presented below.

3. Data

3.1 Searching, matching and recording

The data generated for this work extends the publicly available HistPat database (Petralia et al., 2016). HistPat contains geographical information for historical patents provided by the USPTO between 1790 and 1975. The authors of this work scraped text from digitized historical patent files available from Google and EspaceNet, and recorded the first inventor and a geographical location for each patent. Unfortunately, the early versions of HistPat provided no information for any possible additional co-inventors.

I contribute to these data by identifying all inventors and their geographical locations for each USPTO patent generated within the U.S. between 1836 and 1975[2]. I use the raw scraped text files for the 4,125,734 patents and examine whether each word in these text files is part of a first or family name, or a geographical location in the US. Figure 1 shows an example of the data for USPTO patent 1. The upper part of the image is the scanned copy of the original document. The lower part is the digitized version of this document, published by Google Patents. Note that the quality of the original patent document may have deteriorated over time, resulting in digitized documents that can be difficult to read by Optical Character Recognition software and ultimately misspelled texts. The data for the lists of first and family names comes from the digital USPTO patents from 1975 up to 2005 (Lai et al., 2012), inventor names in HistPat, and the U.S. Census. The data for the geographical locations come from the same sources as well as the U.S. Bureau of Economic Analysis.

Once a (fuzzy) match between a word on the patent text with one of the lists with names and/or location occurs, a series of complex algorithms is run. These algorithms can broadly be placed in two groups. The first set examines the words before and after the matched word to determine whether the matched word is a name or geographical location. For example, the word "EDISON" can be matched to inventor "THOMAS EDISON", but also to the location "EDISON, NEW JERSEY". Examining the text before and after the matched word helps to distinguish between name and location. At the end of this stage, more than 8 million names on the 4.1 million patents are recorded. For about 60% of these observed names, the algorithms also associate a location. All the words that are categorized as geographical locations but can't be linked to a name on the patent are ignored.

The second set of algorithms record a series of more than 30 statistics for each matched word. These statistics are used for the machine learning exercise, discussed below, to determine whether an observed name is truly an inventor and not a witness, examiner, corporation or reference. For instance, once a name is observed I count how often it occurs in the text, how *far* the name is from the word "inventor", how many words are between the observed name and the top and bottom of the patent, if it is adjacent to corporate identifications, and so on. Similar operations are undertaken for an observed geographical location that is linked to a name, but those statistics are only used to generate a likelihood measure of a correct name-location link. Table 1 shows a truncated snapshot of the resulting database. Note that the algorithms have been able to match words from the digitized text in Figure 1 to actual names and locations. At this stage it is still unclear whether the observed name corresponds to an inventor.

Figure 1. Photo-copied and digitized records of USPTO patent US1A

UNITED STATES PATENT OFFICE.

JOHN RUGGLES, OF THOMASTON, MAINE.

LOCOMOTIVE STEAM-ENGINE FOR RAIL AND OTHER ROADS.

Specification of Letters Patent No. 1, dated July 13, 1836.

To all whom it may concern:

Be it known that I, JOHN RUGGLES, of Thomaston, in the State of Maine, have invented a new and useful improvement or improvements on locomotive-engines used on railroads and common roads by which inclined planes and hills may be ascended and heavy loads drawn up the same with more facility and economy than heretofore, and by which the evil effects of frost, ice, snows, and mud on the rail causing the wheels to slide are obviated.

their heads with sufficient force to project them outward easily when pressed up into their sockets, the springs react against the top of a cap, or case made to inclose, and protect them from mud or other impediments to their easy action, the case is in form of the section of a cone, and may be seen at W, W, Fig. 1, it is fitted, and screwed firmly to the rim, the upper end being supported by braces *d*, *d*, which are fastened to the spokes, attached to the cogs is a rod about half an inch diameter passing up through the spiral

DESCRIPTION (OCR text may contain errors) CLAIMS available in

UNITED STATES PATENT UFFICE.

JOHN miennes, or Tnomnmit. imma.

LOGOMQTIVE BTEAM-ENGINE FOR BAIL .AINE OTHER ROADS.

To alt ltvmm it muy confirm `13e tt. known that I, Joni: Hochmut, of 'lhomnstoin in the Stute of Moine, have n vented n new sind useful improvement or tltprovcments on locomotive-engines used on railroads and common roads by which inclined planes and hills muy he ascended and heavy loads drawn up the suine with more facility and eeonomv than heretofore, ncl hywhich the evil eflects of frost. ice, snows.. and nnn'l yon the ruil causing the wheels to slide nre obVinted.

Table 1. Observed names and geographical locations for USPTO patents (truncated)

Patent	Year	Possible Inventor	City	County	State	A1	A_x
1	1836	JOHN RUGGLES	THOMASTON	KNOX	MAINE	3	…
2	1836	GEORGE SULLIVAN				0	…
2	1836	FRANKLIN BROWN				0	…
3	1838	FOWLER M RAY	CATSKILL		NEW YORK	-48	…
4	1853	BENJAMIN IRVING	GREENPOINT	KINGS	NEW YORK	0	…

3.2 Supervised machine learning to identify inventors

The next step is to distinguish between inventors and non-inventors in the data. Non-inventor names can correspond to other entities that have name-like characteristics. I use supervised machine learning techniques to classify each of the 8 million observed names as either an inventor (1) or a non-inventor (0). Each observed name on a patent is called an event.

To distinguish between inventors and non-inventors, I use supervised binary classification machine learning algorithms. These techniques typically make use of three different datasets. The training set contains verified or *known* information on events and the corresponding characteristics. Machine learning algorithms can be deployed to this dataset to model and *learn* which characteristics and structure among the characteristics can best classify events. A validation dataset is used to tune the parameters of the candidate models and prevent the models from over-fitting to the training data. A test set is used to evaluate the performance of the candidate models and select the final model. The final model is used to predict for the events in the mined data whether each event is an inventor.

To construct the co-inventor data I follow this approach, but have to overcome a key issue. The only training data that can be built for this project is the first inventor data recorded in HistPat. If the observed names generated by my algorithms match with the inventor listed on the patent recorded in HistPat, I have a *known* event for an inventor with corresponding characteristics[3]. However, I have no information on *known* events for non-inventors. Such situations call for a one-class classifier approach in which only characteristics for one class can be learned. Events are classified into that class if their characteristics fit into the class given certain thresholds derived from distributions in the training data. The quality and accuracy of one-class classifiers often underperform multi-class classifiers because they can only learn from one class. Based on this underperformance, I follow a multi-class classifying approach[4].

To be able to continue with the multi-class classifier approach, I generate artificial *known* event data for non-inventors. From the work of Wuchty et al. (2007) it is reasonable to assume that the average number of inventors on a patent is well below two before 1975. Given the fact that there are more than 8 million events for 4 million patents a fair share of events are non-inventors. I take a random sample of 30% of all the events that are not *known* inventors and artificially classify them in my training data as *known* non-inventors[5]. The remaining 70% of not *known* inventors are excluded from the training data. The training data now consists of over 3 million events with *known* inventors and more than 1 million artificial *known* non-inventors. Although there might be false-negative events in the latter group (non-inventors that are actually inventors), the majority of events provide data to learn from.

There is an expansive battery of machine learning algorithms available that can be used to train a supervised binary classification model. It is uncertain which algorithm will generate the best performance. Therefore, it is best practice to explore a variety of algorithms and examine fitness criteria. Some of the algorithms have multiple parameters that can be optimized to increase the performance of the models. This leads to extensive searches through parameter space and requires considerable computing power. The final trained model is fit using gradient boosting machine algorithms.

The next step is to disambiguate the inventors that are assigned to patents and assign each unique inventor with a classifying ID. This means that we want to know if Josephine Bruin living in Los Angeles in 1919 is the same inventor as Josephine Bruin living in Los Angeles in 1921. To deal with such issues, many disambiguation approaches have been developed. I loosely follow Ventura et al. (2015) and construct a supervised machine learning approach. This approach involves a training, validation and test dataset. My training and validation dataset is the

Lai et al. (2012) database that holds information on disambiguated inventors on US patents between 1975 and 2012. The test database is the event data I have mined. All databases hold event-level information with a series of similar characteristics. Importantly, the training database also holds information about which inventors are the same.

The disambiguation approach involves the following steps:

1. Select characteristics on which pairs of inventors are to be compared. These characteristics need to occur in all databases. I select: first name, middle name, last name, year, city, county, state, technology class (1:5).
2. Generate for each of the characteristics a method to compare similarity between two strings (i.e. 'Josephine' and 'Josphine')
3. Pair-wise compare all inventors in the training database and find the similarity score for each of the characteristics. This results in 12 similarity scores.
4. For each pair-wise comparison in the training database we know whether the pair of inventors are the same unique inventor. We can classify the vector of similarity scores with a 1 if the compared inventors are the same and 0 if they are not.
5. Train a series of supervised machine learning algorithms to 'learn' which combination of similarity characteristics correspond to pair-wise comparison of identical (1) and different inventors (0). The trained model can be used to predict whether a vector of similarity characteristics correspond to the same inventor or not.
6. Repeat step 3 and 4 for the newly mined event data. This results in hundreds of millions of vectors with similarity characteristics.
7. Apply the model generated in step 5 to the data generated in step 6.
8. Assign the inventors that are identified as the same individuals with the same ID.

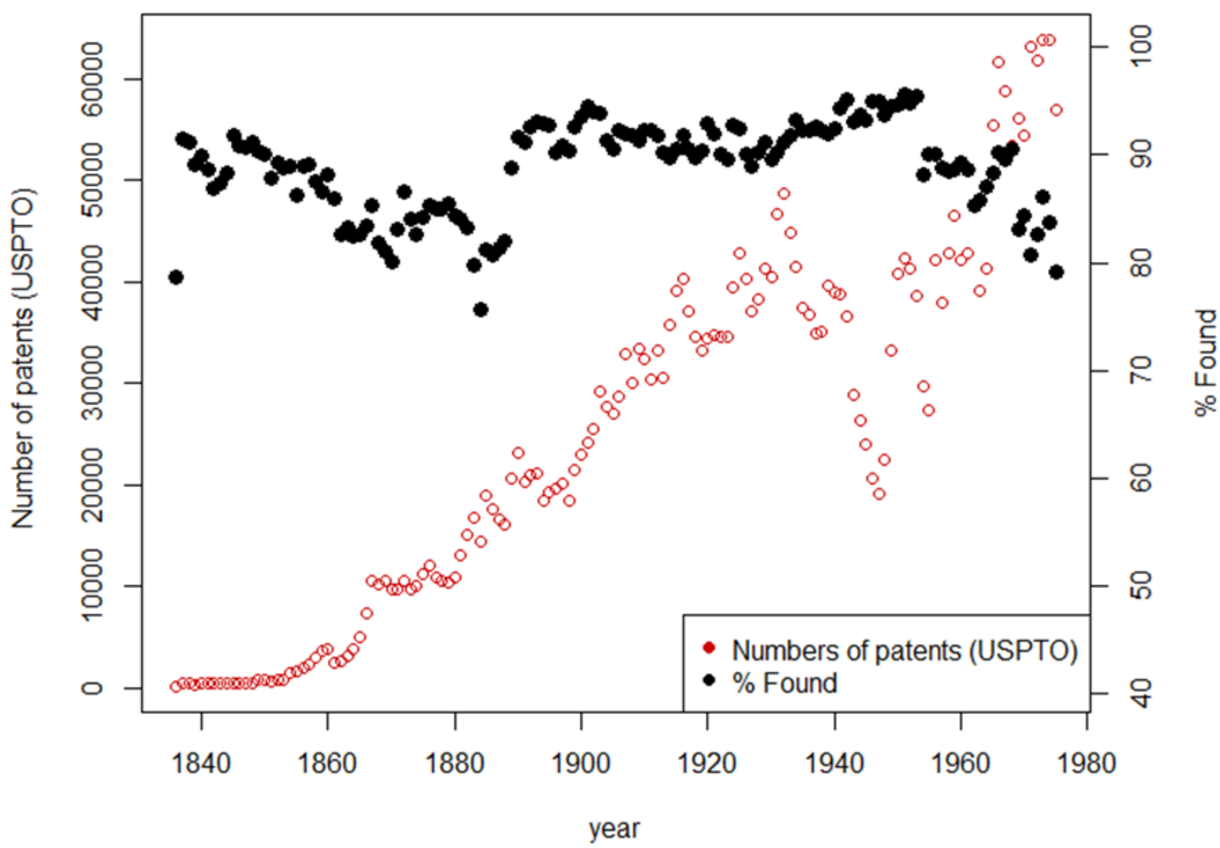
The resulting inventor-patent database identifies 1,922,754 inventors, with 4,437,960 observations on 3,365,253 unique US patents between 1836 and 1975. Figure 2 shows the annual number of patents granted by the USPTO (left axis) and the percentage of the patents for which at least one inventor is identified. On average approximately 80-90% of the patents produced each year are captured in the new inventor-patent dataset. The wide coverage of the annual number of patents and lack of theoretical or empirical motivations to expect systematic bias in the unobserved patents suggests that this database is a representative sample of the historical USPTO patents granted between 1836 and 1975.

Table 2 shows the top 10 inventors identified by these exercises. All names in this table are well known inventors. Surprisingly, Thomas A. Edison is not the most observed inventor in our dataset, although he is generally regarded as the most inventive individual in the late 19th and early 20th century. Time Magazine assigns 1,093 patents to his name, while I only find 536 patents on which he is an inventor. There are at least two reasons why Edison might not be identified on the other 500 or so patents. First, the quality of the photo-copy may have been too poor to read. Additionally, Edison's name occurs on a disproportionately large numbers of patents as a reference and later as a firm name or assignee. The machine learning algorithms have *learned* this and require a large number of positive scores to identify "Thomas A. Edison" as an inventor. If the algorithms can't construct these scores in the patent document, the probability that "Thomas A. Edison" is truly an inventor of the patent is assessed as low by the learning algorithm and is not regarded as such. This 'Edison effect' is not likely to affect other inventors because they do not occur so often on other patents as a reference, firm name or assignee. The number of patents associated with other top-inventors in the database closely reflects the counts other data-sources associate to these inventors.

Table 2. Top 10 inventors in database

Number	Name	Patents
1	Francis H. Richard	738
2	Thomas A. Edison	536
3	Elihu Thomson	516
4	John F. O'Connor	511
5	Edwin H. Land	465
6	Clyde C. Farmer	458
7	George Albert Lyon	453
8	Carleton Ellis	481
9	Louis H. Morin	406
10	Thomas E. Murray	388

Figure 2. Annual number of patents granted and percentage of patents in database



4. Empirical results

4.1 Collaboration on US patents

Collaboration on US patents has increased over time. Figure 3 indicates that up until the 1920s, the annual share of patents generated by co-inventors remained below 20 percent. After the 1920s a rapid increase in rates of co-invention occurs and by the 1960s about 30 percent of patents result from collaboration. By the 1970s, approximately 40 percent of patents result from co-invention.

The average number of inventors on patents has also expanded over time. Figure 4 shows the annual average number of inventors on U.S. patents. The number of co-inventors on patents remained relatively stable until the 1940s. However, after 1940, a steady increase in the number of inventors on patents is observed. By 1970 the average team size is 1.6 inventors per patent compared with a value of about 1.2 inventors at the end of the 19th century. The same trend is observed if only collaborative patents are considered. The in-set in Figure 4 shows that in the 1940s the average team-size increases rapidly. This indicates that the increase in the average number of inventors on a patent is not produced by fluctuations in the annual number of single inventors, but can be attributed to growing average team-sizes. Note that the figures presented here for the late 1970s align closely with the average numbers documented for 1975 in Wuchty et al. (2007).

Figure 3. Share of collaborative patents between 1836 and 1975

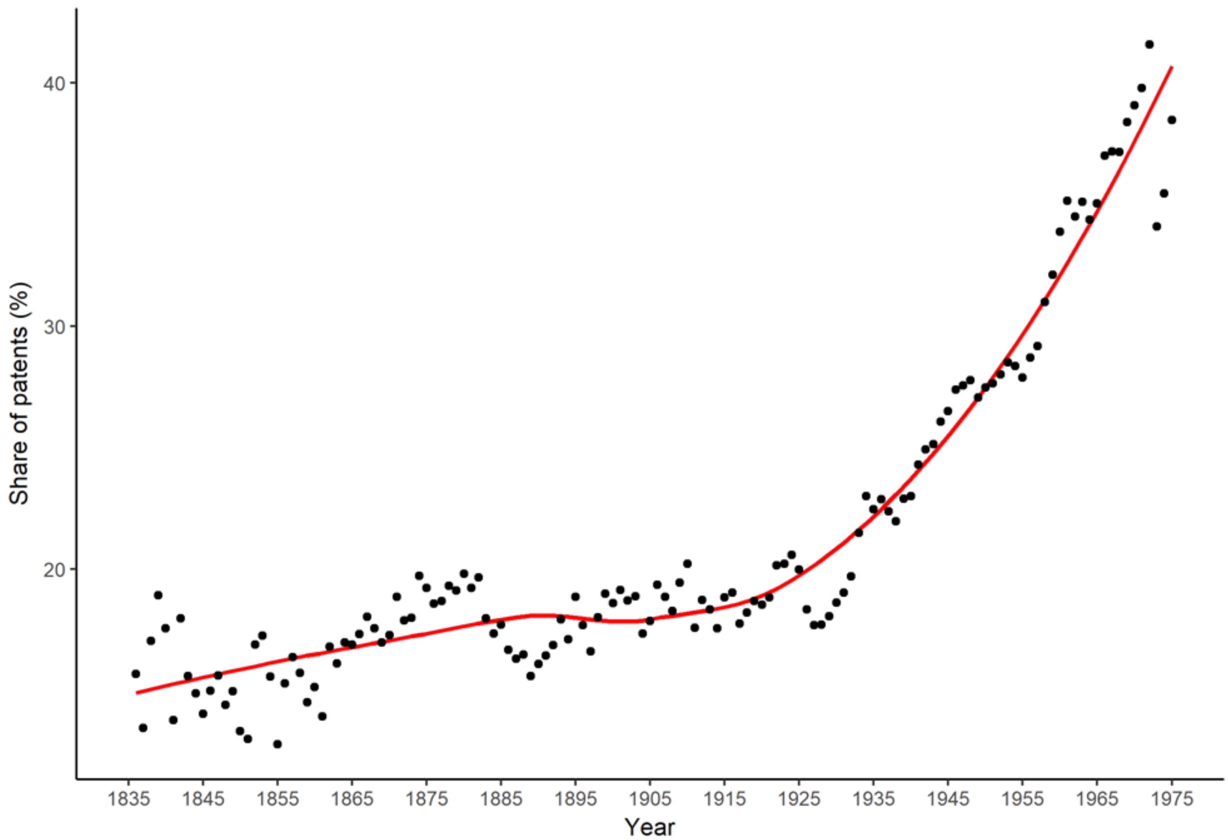
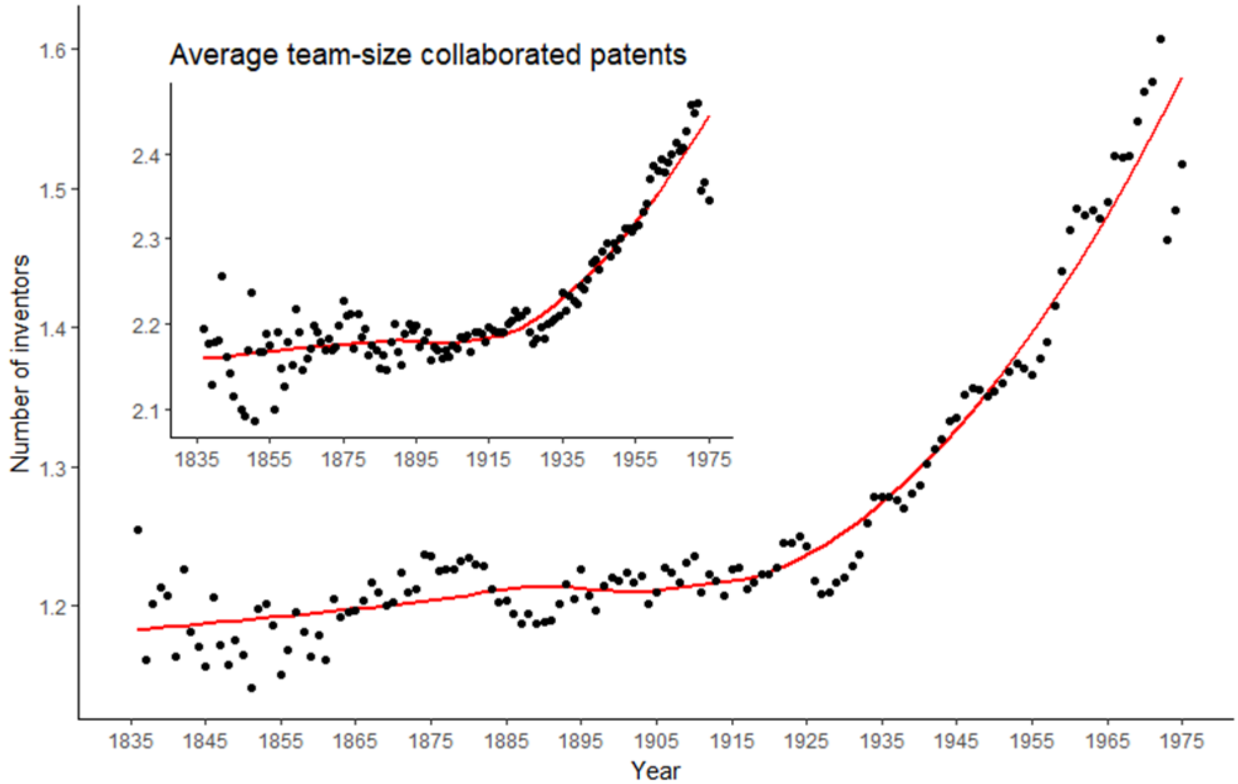


Figure 4. Average number of inventors on a patent between 1836 and 1975



4.2 Complexity and collaboration

Did the complexity of patents increase over time, and is there a positive relationship between complexity and collaboration? The *complexity* measure of a patent originates in Fleming & Sorenson (2001) and is based on the *NK* model of Kauffman (1993). Patent examiners classify each patent into technological classes. The patents examined here are classified in 438 primary classes that can be disaggregated to 10,562 mainline classes and 153,305 sub-classes. The proposed complexity measure is based on the ease with which each mainline class can be recombined with other mainline classes on a single patent. That is, if a mainline class co-occurs relatively often with other mainline classes on a patent, the ease of recombination of the technology class is relatively high. Formally, the ease of recombination of mainline class i is defined as

$$E_i = \frac{\text{Count of Mainline Classes Previously Combined with Mainline Class } i}{\text{Count of Previous Patents in mainline class } i} \tag{1}$$

However, the number of mainline classes increases rapidly over time and has an influence on the ease of recombination measure. Patents in 1840 are classified in 372 mainline classes, while the 1970 patents are classified across 7,722 mainline classes. This means that the ease of recombination of technology classes in early year patents is relative low compared to later year patents because there are fewer classes in which a patent can be located. As a control I construct a standardizing coefficient S defined as

$$S = \frac{\text{Count of unique mainline classes in 1975}}{\text{Count of unique mainline classes in year } t} \tag{2}$$

This standardizing coefficient S is used to get an adjusted ease of recombination score

$$EA_i = E_i * S \tag{3}$$

The complexity of patent l is defined as the number of mainline classes divided by the sum of the ease of recombination of these classes. Formally,

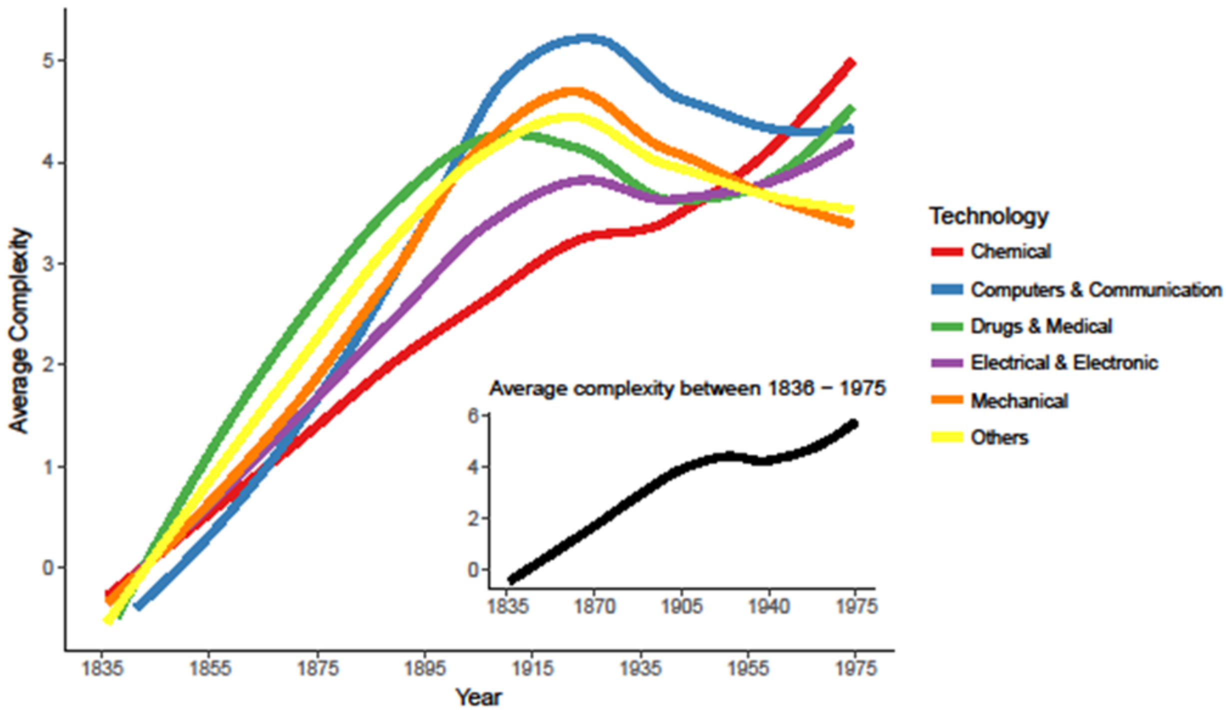
$$Complexity_l = \frac{\text{Count of Mainline Classes on Patent } l}{\sum_{i \in l} EA_i} \tag{4}$$

The ease of recombination for a patent is calculated using a ten-year window of patent recombinations recorded up to the grant year date of the focal patents. Thus, the ease of recombination for technology classes shifts over time.

Figure 5 shows that the average complexity of patents has been increasing over time. The inset shows that the average complexity increases through the nineteenth century, falls from the 1920s to the mid-1940s and then increases once more. The period of initial increase matches the Second Industrial Revolution (Gordon, 2016) and the 'golden age' of US invention (Akcigit et al., 2017). The drop in complexity after the 1920s aligns with the Great Depression. The increase in complexity after the 1940s is consistent with the literature documenting the shift to 'big science' in knowledge production.

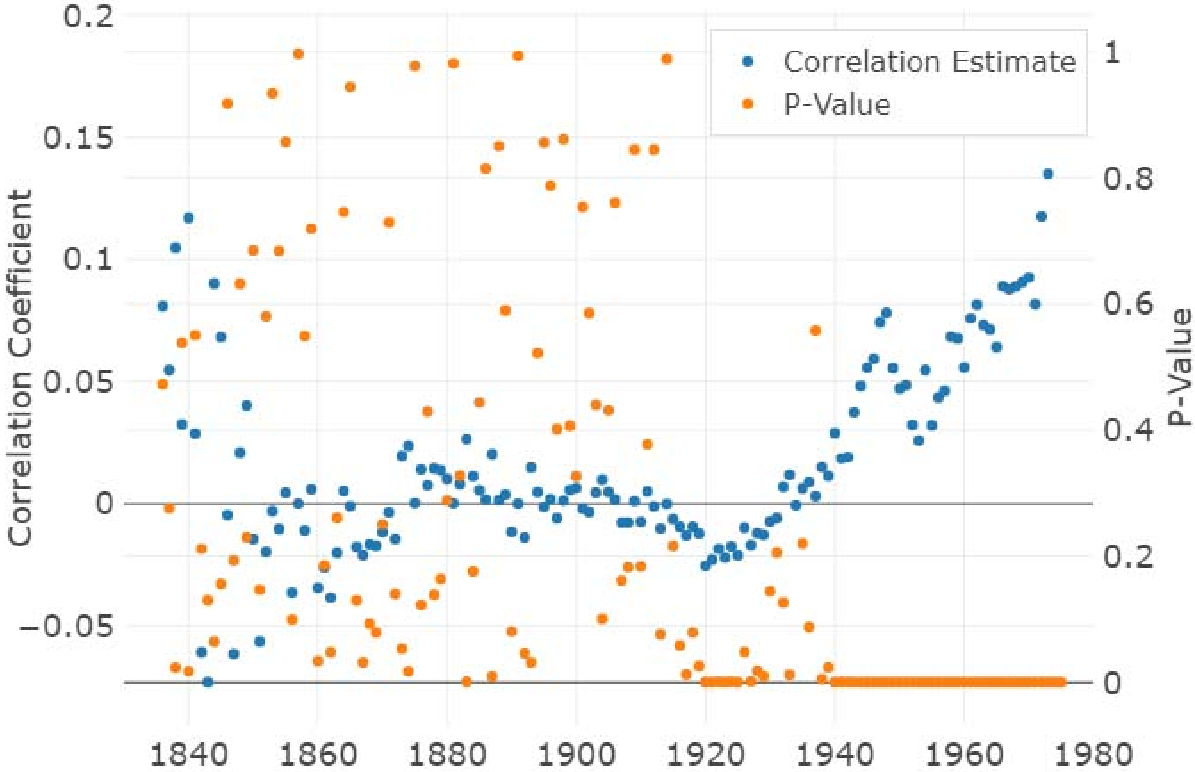
The average complexity of patents varies across the six broad technological categories defined by Hall et al. (2001). Up until the 1900s the 'Drugs & Medical' patents were most complex. Throughout the first half of the 20th century, patents in the 'Computers & Communication' category are most complex, on average. This period marks the development of the telephone. The decrease after the 1920s might be the result of the 'exhaustion' of this generation of communication technologies. Indeed, complexity rises again in the 1950s when computer related technologies begin to emerge. In the 1960s 'Chemicals' patents are on average the most complex. Patents in the 'Drugs & Medical' category increased in complexity when the production of drugs became integrated with synthetic organic chemistry post World War II (Drews, 2000). Remarkably, the 'Mechanical' category has long been among the most complex categories, but falls to the bottom of the complexity ranking by 1975.

Figure 5. Average complexity per technological category between 1836 and 1975



A key question in this paper is whether there is a positive relationship between complexity and collaboration at the patent level. Figure 6 plots the annual correlation between complexity and collaboration on all patents by year. There is no consistent and systematic relationship between complexity and collaboration before the 1920s. During the 1920s there is a significant but weak negative relationship between complexity and collaboration. Interestingly, from 1940s onwards there is an increasing, significant and positive relationship, suggesting that increasing complexity is associated with a higher probability of a patent being produced through collaboration.

Figure 6. Correlation between complexity and collaboration on a patent between 1836 and 1975



These ideas are extended in models examining the relationship between complexity and collaboration. The dependent variable is a binary variable indicating whether the patent is the result of a collaboration (1) or not (0). The key independent variable of interest is the complexity of the patent. The second independent variable is a time dummy that indicates whether the patent is produced after 1940 or not. This cut-off point is chosen because there is a significant structural break in the trend (Bai & Perron, 2003; Zeileis et al., 2003). Figure 6 clearly indicates there is a structural relationship between complexity and collaboration after 1940. Given the evidence on increasing collaboration and complexity over time presented above, a positive coefficient is expected. To explore whether the nature of the relationship between complexity and collaboration shifts after 1940 an interaction terms is added as the product of the time dummy and complexity. City, year and technology fixed effects are added to the model. It is plausible that specific city-level characteristics can promote or hinder collaboration. Year fixed effects control for time-specific shocks in collaboration across the sample. Technology fixed effects control for the possible heterogeneity in the propensity to collaborate across technology classes. The sample size is restricted to patents for which the primary inventor (first listed on patent) lives in an MSA to satisfy the requirements for the city fixed effects variable. The unit of analysis is the patent. Each patent is assigned to the MSA of the first inventor.

The results presented in Table 3 indicate that there is indeed a positive and significant relationship between complexity and collaboration on a patent, regardless of the model specification. Model 1 is the baseline model and reports a positive and significant relationship between complexity and collaboration. A coefficient of β_1 = 0.15 means that a 2.71 factor[6] increase in complexity is associated with a change in the *odds* of a patent being a collaboration by a factor of 1.16. In models 2-6, technology, city and year fixed effects are introduced. The estimated coefficient for complexity remains roughly the same value and is always significant.

In model 7 the time dummy and its interaction with complexity are introduced. The coefficient for complexity now corresponds to patents produced before 1940 (dummy variable is zero). For these patents, controlling for city and technology fixed effects, a coefficient of β_1 = 0.06 means that a 2.71 factor increase in complexity is associated with a change in the odds that a patent is produced by more than one inventor by a factor of 1.06. The significant coefficient for the interaction term between the time dummy and complexity indicates that post-1940, an increase of one unit of complexity raises the log odds of a patent being generated through collaboration by about 50% over the pre-1940 log odds ratio. The results presented here provide the first long-run, empirical evidence of a positive and significant relationship between the complexity of ideas and co-inventor collaboration.

Table 3. Results of Logistic Regression with fixed effects

Dependent variable	Is patent collaborated (0/1)?						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Complexity (log)	0.15***	0.10***	0.15***	0.10***	0.10***	0.03***	0.06***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)
Dummy > 1940							0.54***
							(0.01)
Complexity (log) * Dummy > 1940							0.03***
							(0.004)
Fixed Effects:							
- Technology		✓		✓		✓	✓
- City			✓	✓		✓	✓
- Year					✓	✓	
Constant	-1.23***	-1.78***	-1.38***	-1.78***	-1.09***	-1.88***	-1.95***
	(0.003)	(0.02)	(0.12)	(0.13)	(0.42)	(0.44)	(0.13)
N	1,764,733	1,764,733	1,764,733	1,764,733	1,764,733	1,764,733	1,764,733
Log Likelihood	-997,311	-976,428	-993,265	-974,236	-972,388	-960,603	-964,249
AIC	1,994,627	1,953,710	1,987,254	1,950,044	1,945,059	1,923,055	1,930,074

*p < .1; **p < .05; ***p < .01

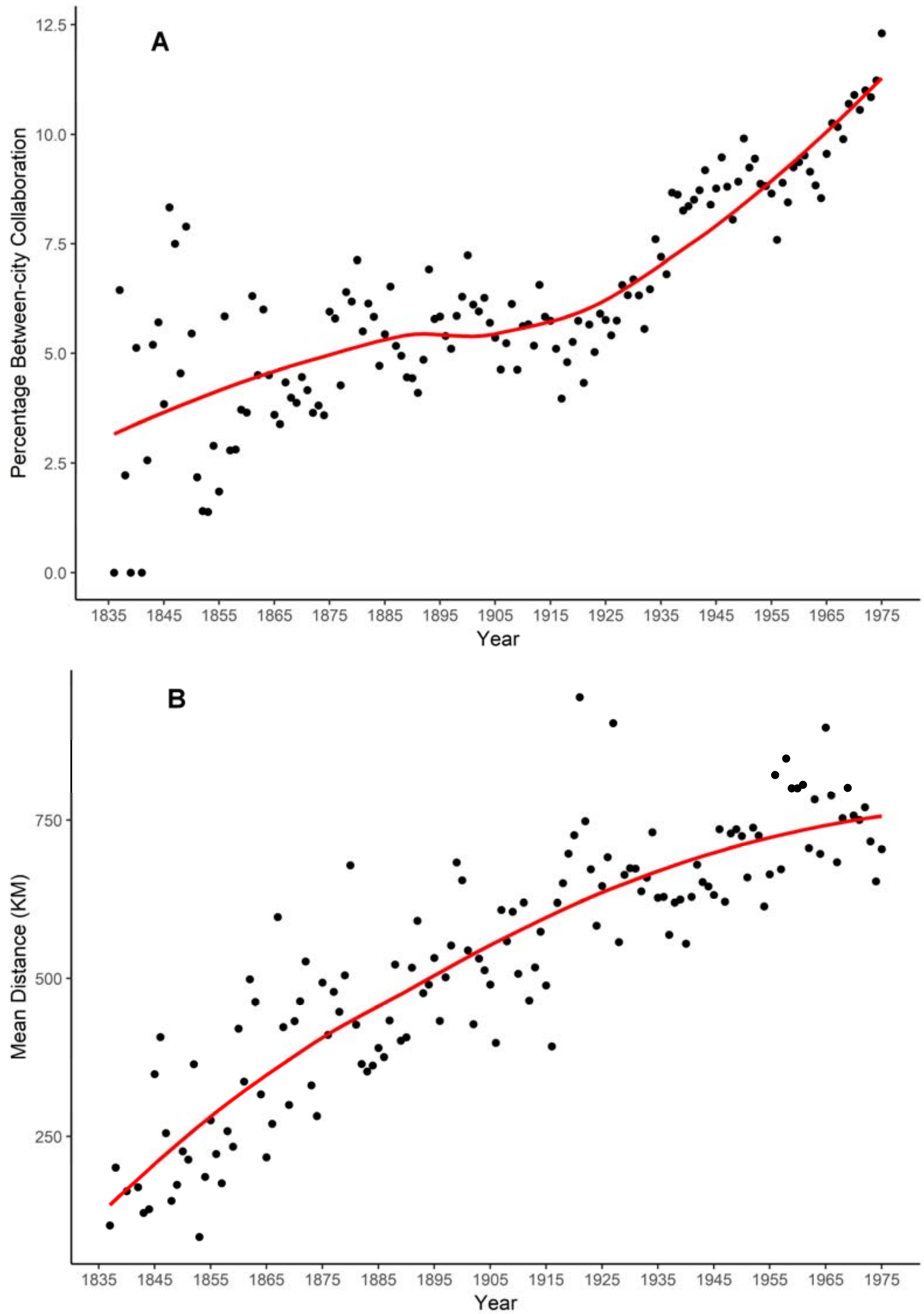
4.3 Geography of co-invention

This final section of analysis explores the geography of collaboration in U.S. patent production since 1836. The data used for this analysis comprise only those patents for which all inventors reside in a metropolitan area.

Geographical distance of co-invention

Geographic patterns of co-invention have changed over time. Figure 7A reveals that there is an increase in the share of between-city co-invention. Between-city collaboration increased from roughly 5 percent throughout the 19th century to over 10 percent in 1975. In addition, Figure 7B shows that the average geographic distance between co-inventors from multiple cities doubled from about 250 kilometers in the early-19th century to around 750 kilometers by 1975. Interestingly, the increase in average distance of between-city co-inventors seems to stagnate since the 1940s. This might have to do with the dynamics in the spatial distribution of the US population and inventors as the US expanded to the West. Such expansion might have spurred coast-to-coast collaboration in the initial years of settlement in the West, but leveled off as the population on the west-coast started to grow and provided an increasing local pool of possible (co-)inventors, limiting the need for long-distance coast-to-coast collaboration. Indeed, Hicks et al. (2001) show the increasing importance of west-coast patenting in the US innovative activity from 1983 to 1997. By the mid-1990s the Pacific census region is producing the most patents in the US.

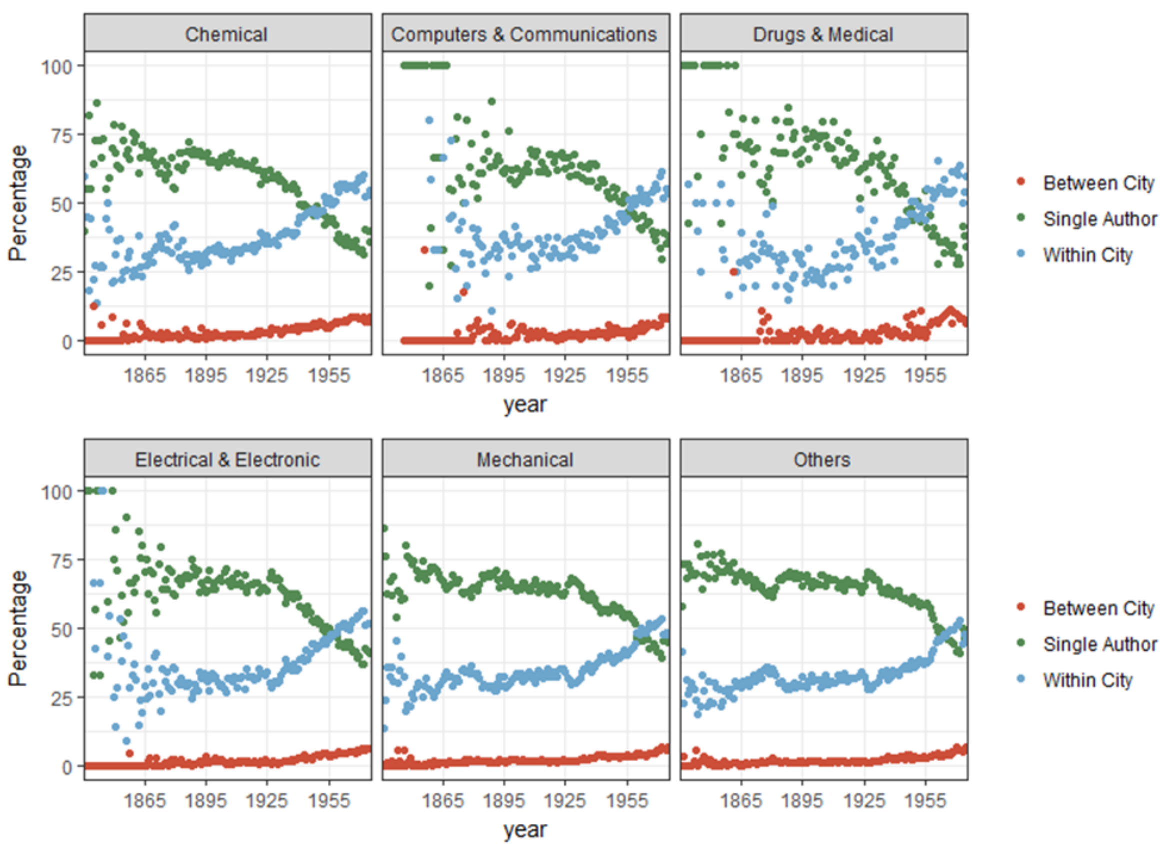
Figure 7. (A) Percentage of between-city co-invented patents and (B) average distance between-city co-inventors over time



Within and between city collaboration

Collaboration occurs primarily by inventors from the same city. Figure 8 shows that within-city collaboration is growing at a faster rate than between-city collaboration, even as the cost of communication between cities has declined. Single authorship of patents is dropping rapidly across all categories. Although these general trends are observed across the six aggregate technology categories, patterns of collaboration differ across categories. For instance, since the 1940s the majority of patents in the 'Chemicals' category result from collaboration, whereas it takes until the 1960s for patents in the 'Mechanicals' and 'Others' category to reach this point.[7]

Figure 8. Collaboration across technology categories



Geography, collaboration and complexity

Is complexity associated with within-city rather than between-city collaboration? Collaboration on complex patents might tend to be local, because complex, tacit forms of knowledge are known to be difficult to move over space. At the same time, Van der Wouden and Rigby (2019) report that cities with more specialized knowledge cores build longer co-inventor pipelines than more diversified cities. Thus, the geographical pattern of collaboration and its relationship with complexity is still unclear.

To examine the geographical patterns of collaboration three separate analyses are performed. First, logistic regression models are estimated. The unit of analysis is the patent. Only collaborative patents are selected for which all inventors resided in an MSA. The dependent variable indicates whether a collaborative patent is the result of a within-city (0) or between-city (1) partnership. The key independent variable is, once more, the *complexity* of the patent. An additional independent variable controls for the number of inventors on a patent. Larger teams might be more likely to be non-local. If a patent becomes complex to the point that inputs from a larger number of inventors is needed, the probability this input is not found locally might increase too. City, year and technology fixed effects are added to the model.

To isolate the effect of complexity on collaboration, a matching procedure has been undertaken in which between-city collaborated patents are matched to within-city collaborated patents with identical scores on the number of inventors, city, year and primary technology class. Unmatched observations are dropped. The outcome is a sample with balanced covariates for within and between-city collaborated patents, except for the *complexity* score of the patent.

The results presented in Table 4 demonstrate a negative and significant relationship between complexity and between-city collaboration. In model 1 the estimated coefficient for *complexity* (β_1 = -0.08) indicates that a 2.71 factor increase in the complexity score is associated with an expected change in the odds of between-city collaboration by a factor of 0.92. Similar results are found for models with different specifications. In model 6, controlling for *Number of Inventors* and year, city and technology fixed effects, the estimated coefficient for *Complexity* (β_1 = -0.16) suggest that a 2.71 factor increase in the complexity score is associated with an expected change in the odds of between-city collaboration by a factor of 0.85. The *Number of Inventors* on a patent has a positive and significant relationship with between-city collaboration in all models. In model 6, controlling for *Complexity* and year, city and technology fixed effects, the estimated coefficient for *Number of Inventors* (β_2 = 0.43) indicates that for every additional inventor on the patent, the odds of the patent resulting from inter-city collaboration increases by 1.54. Together, these results indicate that, controlling for the number of inventors, complex patents are more likely to be produced by inventors in the same city.

A second analytical model explores the relationship between the average distance observed among co-inventors from different cities and the complexity of the patent[1]. Figure 9A plots these variables and a non-linear concave relationship is observed. Each point is a co-invented patent by inventors from different cities. The colors indicate the decade the patent was granted. The histograms on top and next to Figure 9A indicate the density of data-points on the complexity and distance axes, respectively. The figure suggests that initially the average geographical distance between inventors rises with the complexity of a patent. Perhaps this increase results from the lack of motivation for inventors to collaborate with non-local individuals on patents of rather low complexity – the perceived benefits of collaboration might not outweigh the extra costs of collaborating across distance. When complexity increases the

[1] Note that co-invented patents produced by inventors from the same city are excluded, because it is impossible to measure the geographical distance between these inventors in the historical patent data

benefits of collaboration may increasingly outweigh these costs, giving rise to an upward trend. However, when patents become increasingly complex, collaboration becomes gradually more local as indicated by the negative trend-line. This suggests that inventors of complex patents rely more strongly on the geographical co-location of inventors than inventors of less complex patents do.

The third piece of analysis focuses on within-city dynamics and links these dynamics to the average complexity of patents in a city. Operationalizing within-city collaboration with a metric distance of zero is problematic, because it assumes that the average distance between inventors is the same for all cities and over time. To allow for variation on the city-level, a measure of inventor density is constructed as the annual number of inventors in the city divided by the area of the city measured in square kilometers. In cities where density is higher the average geographical distance between inventors in that city is lower. In such cities, inventors will have greater opportunities for coordinated and unexpected interactions, enhancing knowledge spillovers between inventors (Carlino et al., 2001; Lobo & Strumsky, 2008; van der Wouden & Rigby, 2019).

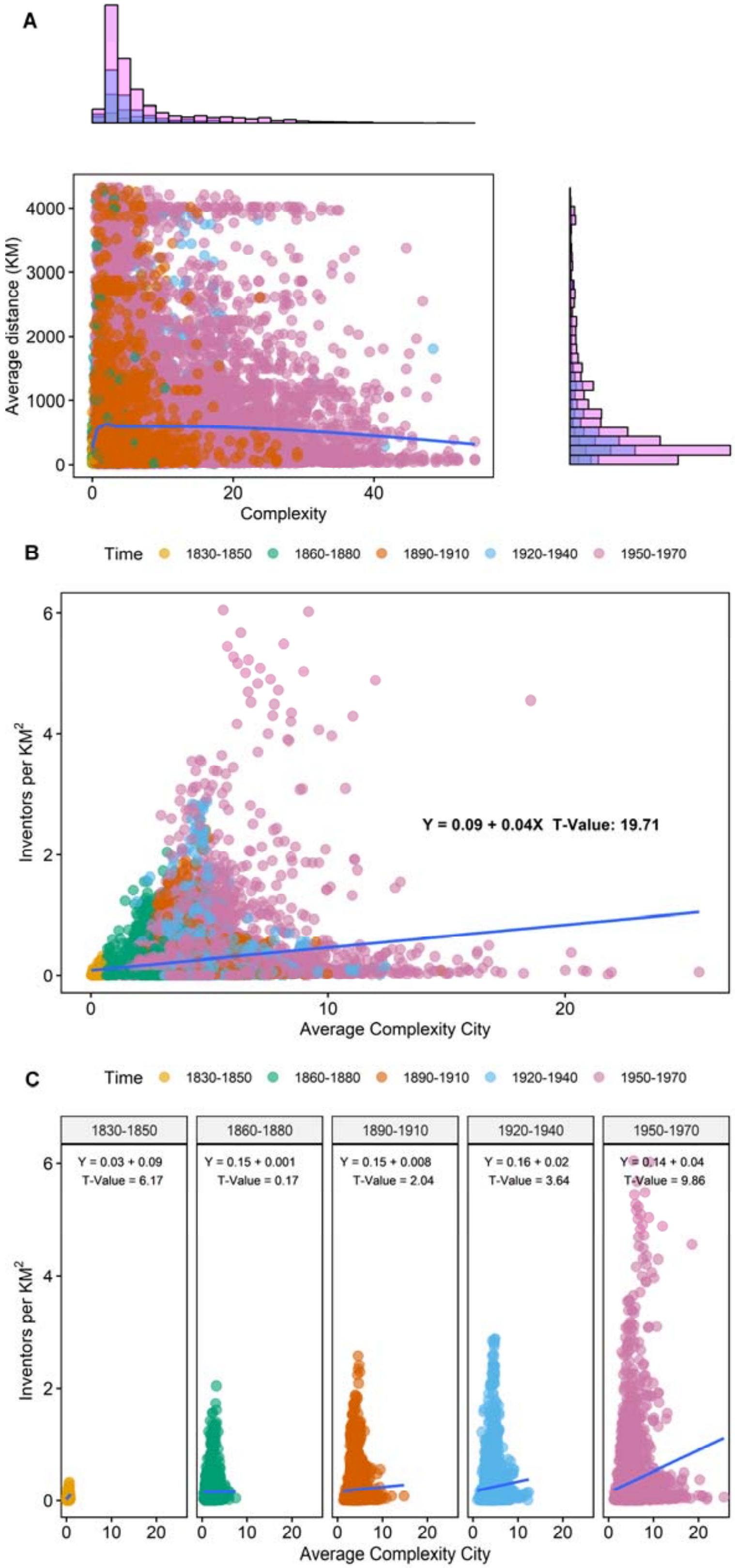
Figure 9B plots the average complexity of all patents in a city in a given year by the inventor density of this city at that time. The color of the points indicates the year of the observations, grouped 30 year time-windows. The figure shows a significant and positive relationship between the inventor density of a city and the average complexity of the patents in a city in a given year. This suggests that in cities where the average distance between inventors is smaller, the complexity of patents is on average greater. Figure 9C plots the same relationship but for the five 30 year time-windows separately. The average complexity and inventor density of cities are increasing over time. For all time-windows the relationship between average complexity and inventor density is positive and significant, except for the observations in 1860-1880. Interestingly, the slope of the relationship is getting steeper over time, indicating that city-level inventor density and the average complexity of the patent stock become more strongly tied together in recent decades. Together, the results of the three analyses indicate that increasing complexity on patents is associated with within-city rather than between-city collaboration.

Table 4. Statistical models on within- and between-city collaboration

Dependent variable:	Patent is collaborated within-city (0) or between-city (1)					
	(1)	(2)	(3)	(4)	(5)	(6)
Complexity (log)	-0.08***	-0.07***	-0.32***	-0.18***	-0.15***	-0.16***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Number of Inventors	0.34***	0.38***	0.36***	0.38***	0.40***	0.43***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Fixed effects:						
- City		✓				✓
- Year			✓		✓	✓
- Technology				✓	✓	✓
Constant	-2.14***	-1.80***	-1.78***	-2.29***	-2.15***	-1.69***
	(0.03)	(0.07)	(0.68)	(0.14)	(0.75)	(0.75)
N	160,484	160,484	160,484	160,484	160,484	160,484
Log Likelihood	-76,919	-75,809	-76,768	-76,177	-76,066	-75,098
AIC	153,844	151,983	153,775	153,131	153,143	151,566

*p < .1; **p < .05; ***p < .01

Figure 9. Relationships between (A) average distance and complexity,, (B) inventor density and average complexity of city



5. Conclusion

This paper examined historical U.S. inventor collaboration for the years 1836 to 1975 and presented a series of stylized facts on the collaboration, complexity and geography of co-invention for this largely unexplored era of patent production. Recent research has shown the increasing importance of collaboration in the production of knowledge. The key underlying assumption is that the rising complexity of knowledge requires resources that exceed the inputs of individuals. However, little empirical evidence on the relationship between collaboration and complexity has been available. The key findings of the empirical analysis in this paper are that (1) collaboration has increased sharply since the 1940s; (2) there is a positive relationship between complexity and collaboration on US patents; and (3) increasing complexity on patents promotes within-city collaboration.

Using the raw text files of the HistPat database I utilize search, match and machine learning techniques to identify and disambiguate all (co-)inventors including their geographical location as recorded on US patents between 1836 and 1975. The resulting inventor-patent database contains about 80-90% of the patents granted by the USPTO and identifies 1,922,754 inventors, from 4,437,960 observations on 3,365,253 unique patents. These data are used to examine co-invention on U.S. patents.

The research finds evidence for a growing tendency to collaborate on US patents between 1836 and 1975. This finding is in line with other research on inventor collaboration on US patents (Wuchty et al., 2007) and scientific collaboration on papers (Jones et al., 2008) that both extend back to 1975. The observed increase in the percentage of patents collaborated, along with the growth in average team-size after the 1930s might be linked to the 1933 US Supreme Court ruling to assign the rights of technologies developed by inventors hired-to-invent directly to the firm. This provided an incentive for firms to recruit and hire inventors, instead of buying licenses to patented technologies by inventors operating privately. In doing so, the firm becomes a platform in which inventors can collaborate instead of being direct competitors. More general, infrastructure and transportation developments throughout the 19th and 20th century have made it easier for inventors to collaborate across increasing distance. By the 1920s almost all major US cities were connected by the railroads; in 1930 about 40% of US homes had a telephone connection and 20% of Americans had access to an automobile.

The observed robust, significant positive relationship between complexity and collaboration on US patents confirms the theoretical claim that complexity and collaboration are significantly correlated. In addition, the evidence presented in this paper suggests that the impact of complexity on the odds of a patent being collaborated becomes markedly stronger after the 1940s. This supports a shift to 'big science' after World War II, in which the complexity of knowledge production accelerated. However, the directions of possible causal links between complexity and collaboration are not investigated and remain unclear.

Co-invention on US patents has a distinct geography. Most metropolitan collaboration is mostly between inventors from the same city. Moreover, as complexity increases, the odds of within-city collaboration increases and the average distance between inventors on patents decreases. Moreover, cities with higher inventor density – a crude measure of within-city distance of inventors – produce on average more complex patents than cities with lower inventor density do. These findings suggest that the production of complex knowledge relies more strongly on the geographical co-location of inventors than does less complex knowledge. Local collaboration allows repeated face-to-face meetings, spontaneous encounters and other interactions facilitated by the *local buzz,* than non-local collaboration. These interactions might be particularly important for the production of complex knowledge, because such knowledge relies difficult to diffuse tacit knowledge.

These findings have a number of implications. The role of teamwork in the production of knowledge has been shown to increase following a rather linear trend between 1955 and 2000 (Wuchty et al., 2007). With regards to patent data, we now know that this trend starts around the 1940s – before the 1940s there is no clear trend. Although not explicitly tested in this research, this change in patterns of collaboration correspond with the 1933 Supreme Court ruling to assign the patent rights directly to the firm if the inventors are on their pay-roll. Organizing the production of patents on the level of the firm, inventors become collaborators instead of competitors. This finding is in line with the general research findings put forward by Lamoreaux & Sokoloff (1996; 1999; 2005) and Lamoreaux et al. (2011).

The significant and positive relationship between the complexity of a patent and collaboration can be linked to the 'burden of knowledge' debate. Following the arguments put forward in Jones (2009), the accumulative nature of knowledge requires its subsequent producers to be knowledgeable of increasing stocks of knowledge. Since the resources of economic agents are limited, individual specialization and teamwork is an interesting strategy to produce knowledge. To some extent, one can think of collaboration and patent complexity this way. Collaborative patents tend to be more complex because they combine the specialized sets of knowledge of individual inventors. Indeed, highly specialized inventors only have expertise in a very limited number of technology classes and when combined on patents provide very complex patents. Tracking the average complexity of patents over time, the plateau it reaches in the first half of the 20th century perhaps indicates a technological paradigm shift in which the burden of knowledge decreases because of the emergence of radical new technologies lacking burdening stocks of knowledge.

The local nature of collaboration has important implications for the diffusion of knowledge across space. Although the localized nature of knowledge flows has been long recognized (see Almeida & Kogut, 1999; Audretsch & Feldman, 1996; Breschi & Lissoni, 2009; Feldman & Kogler, 2010; Jaffe, Trajtenberg, & Henderson, 1993; Malmberg & Maskell, 2002), how complexity mediates the relationship between collaboration and geography has largely remained unexplored. While Balland & Rigby (2017) report an uneven spatial distribution of knowledge complexity across US cities and Sorenson et al. (2006) illustrate how knowledge complexity affects the diffusion of knowledge across agents, the findings in this paper highlights that inventors are more likely to collaborate with local inventors when knowledge becomes increasingly complex. This suggests that complex knowledge is *stickier* than less complex knowledge because localized patterns of collaboration facilitate geographical lock-in of knowledge.

This *stickiness* of complex knowledge speaks to the literature on urban agglomeration and the geographic location of firms and industries. Economic agents sourcing non-local complex knowledge might face barriers not present sourcing less complex knowledge, providing incentives to co-locate with the specific complex knowledge. The processes of co-location set in motion important agglomerative dynamics, because it provides the current local economic agents with greater opportunities to collaborate local, enhancing the attractiveness of co-location even further. Indeed, from this perspective the local nature of collaboration on complex knowledge gives rise to forces of agglomeration and shapes uneven distributions of complex knowledge across space. The tendency of industries to co-locate in space might relate to the complexity of their knowledge (see Sorenson et al., 2006).

The paper presented here has at least two major short-comings. First, this paper has focused on inventors located in US cities. In doing so, the geography of collaboration as presented in this work is biased because it doesn't include rural-urban and foreign collaboration. Second, the poor quality of the text files makes it difficult to retrieve all (co-)inventors and geographical locations on patents. As a consequence, about 5-20% of patents per year are missing in the database.

Future research could focus on the understanding of patterns of rural-urban and foreign collaboration on US patents and examine the direction of causality between complexity and collaboration. Moreover, little is known on the mechanisms that structure tie formation between co-inventors and to what extent these structures have evolved over time. In addition, much remains unclear on how the movement of inventors in space affects patterns of collaboration and impact urban networks of invention. Finally, the role of firms has been neglected and remains unclear in collaboration on historical US patents.

Literature

Acemoglu, D., Moscona, J., & Robinson, J. A. (2016). State capacity and American technology: evidence from the nineteenth century. *The American Economic Review*, *106*(5), 61–67.

Akcigit, U., Grigsby, J., & Nicholas, T. (2017). The Rise of American Ingenuity: Innovation and Inventors of the Golden Age. https://doi.org/10.3386/w23047

Almeida, P., & Kogut, B. (1999). Localization of knowledge and the mobility of engineers in regional networks. *Management Science*, *45*(7), 905–917.

Audretsch, D. B., & Feldman, M. P. (1996). R & D Spillovers and the Geography of Innovation and Production. *The American Economic Review*, *86*(3), 630–640.

Axelrod, R. M., & Cohen, M. D. (2000). *Harnessing complexity*. Basic Books.

Bai, J., & Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, *18*(1), 1–22.

Balland, P.-A., & Rigby, D. (2017). The Geography of Complex Knowledge. *Economic Geography*, *93*(1), 1–23. https://doi.org/10.1080/00130095.2016.1205947

Balland, P., Jara-figueroa, C., Petralia, S., Steijn, M., Rigby, D., & Hidalgo, C. (2018). Complex Economic Activities Concentrate in Large Cities. *Working Paper*, 1–10.

Bathelt, H., Malmberg, A., & Maskell, P. (2004). Clusters and knowledge: local buzz, global pipelines and the process of knowledge creation. *Progress in Human Geography*, *28*(1), 31–56. https://doi.org/10.1191/0309132504ph469oa

Bercovitz, J., & Feldman, M. (2011). The mechanisms of collaboration in inventive teams: Composition, social networks, and geography. *Research Policy*, *40*(1), 81–93.

Breschi, S., & Lenzi, C. (2016). Co-invention networks and inventive productivity in US cities. *Journal of Urban Economics*, *92*, 66–75. https://doi.org/10.1016/j.jue.2015.12.003

Breschi, S., & Lissoni, F. (2009). Mobility of skilled workers and co-invention networks: an anatomy of localized knowledge flows. *Journal of Economic Geography*, *9*(4), 439–468.

Burt, R. S. (2004). Structural holes and good ideas. *American Journal of Sociology*, *110*(2), 349–399.

Carlino, G. A., Chatterjee, S., & Hunt, R. M. (2001). *Knowledge spillovers and the new economy of cities*. Economic Research Division, Federal Reserve Bank of Philadelphia.

Coriat, B., & Weinstein, O. (2012). Patent Regimes, Firms and the Commodification of knowledge. *Socio-Economic Review*, *10*, 267–292.

Cowan, R., Jonard, N., & Zimmermann, J.-B. (2007). Bilateral Collaboration and the Emergence of Innovation Networks. *Management Science*, *53*(7), 1051–1067. https://doi.org/10.1287/mnsc.1060.0618

Crescenzi, R., Nathan, M., & Rodríguez-Pose, A. (2016). Do inventors talk to strangers? on proximity and collaborative knowledge creation. *Research Policy*, *45*(1), 177–194. https://doi.org/10.1016/j.respol.2015.07.003

de Solla Price, D. J. (1970). Citation measures of hard science, soft science, technology, and nonscience. In *Communication Among Scientists and Engineers* (pp. 3–22). Lexington Books, D.C. Heath and Company.

Drews, J. (2000). Drug discovery: a historical perspective. *Science*, *287*(5460), 1960–1964.

Eaton, J. W. (1951). Social Processes of Professional Teamwork. *American Sociological Review*, *16*(5), 707–713.

Esposito, C. R., & Rigby, D. L. (2018). Buzz and pipelines: the costs and benefits of local and nonlocal interaction. *Journal of Economic Geography*.

Faems, D., Van Looy, B., & Debackere, K. (2005). Interorganizational Collaboration and

Innovation: Toward a Portfolio Approach. *Journal of Product Innovation Management*, *22*, 238–250.
Feldman, M. P., & Kogler, D. F. (2010). *Stylized facts in the geography of innovation*. *Handbook of the Economics of Innovation* (1st ed., Vol. 1). Elsevier BV. https://doi.org/10.1016/S0169-7218(10)01008-7
Fisk, C. L. (1998). Removing the "Fuel of Interest" from the "Fire of Genius": Law and the Employee Inventor, 1830-1930. *University of Chicago Law Review*.
Fleming, L., & Frenken, K. (2007). the Evolution of Inventor Networks in the Silicon Valley and Boston Regions. *Advances in Complex Systems*, *10*(01), 53–71. https://doi.org/10.1142/S0219525907000921
Fleming, L., & Sorenson, O. (2001). Technology as a complex adaptive system: Evidence from patent data. *Research Policy*, *30*(7), 1019–1039. https://doi.org/10.1016/S0048-7333(00)00135-9
Freeman, L. C. (1979). Centrality in Social Networks Conceptual Clarification. *Social Networks*, *1*(1968), 215–239.
Frenken, K., Hölzl, W., & Vor, F. de. (2005). The citation impact of research collaborations: the case of European biotechnology and applied microbiology (1988–2002). *Journal of Engineering and Technology Management*, *22*(1–2), 9–30. https://doi.org/10.1016/j.jengtecman.2004.11.002
Giuliani, E., Martinelli, A., & Rabellotti, R. (2016). Is Co-Invention Expediting Technological Catch Up ? A Study of Collaboration between Emerging Country Firms and EU Inventors. *World Development*, *77*, 192–205. https://doi.org/10.1016/j.worlddev.2015.08.019
Glanzel, W. (2002). Coauthorship patterns and trends in the sciences (1980-1998): A bibliometric study with implications for database indexing and search strategies. *Library Trends*, *50*(3), 461–473. https://doi.org/10.1017/CBO9781107415324.004
Gordon, R. J. (2016). *The rise and fall of American growth: The US standard of living since the civil war*. Princeton University Press.
Grant, R. M. (1996). Toward a Knowledge-Based Theory of the Firm. *Strategic Management Journal*, *17*(Winter Special Issue), 109–122.
Hall, B. H., Jaffe, A. B., & Trajtenberg, M. (2001). The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools. Retrieved from http://www.nber.org/papers/w8498
Hicks, D., Breitzman, T., Olivastro, D., & Hamilton, K. (2001). The changing composition of innovative activity in the US — a portrait based on patent analysis. *Research Policy*, *30*(4), 681–703. https://doi.org/https://doi.org/10.1016/S0048-7333(00)00147-5
Hoekman, J., Frenken, K., & Tijssen, R. J. W. (2010). Research collaboration at a distance: Changing spatial patterns of scientific collaboration within Europe. *Research Policy*, *39*(5), 662–673. https://doi.org/http://dx.doi.org/10.1016/j.respol.2010.01.012
Jaffe, A. B., Trajtenberg, M., & Henderson, R. (1993). Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations. *The Quarterly Journal of Economics*, *108*(3), 577–598. Retrieved from http://www.nber.org/papers/w3993
John, R. R. (2009). *Spreading the news: The American postal system from Franklin to Morse*. Harvard University Press.
Jones, B. F., Wuchty, S., & Uzzi, B. (2008). Multi-University Research Teams: Shifting Impact, Geography, and Stratification in Science. *Science*, *322*(5905), 1259–1262. https://doi.org/10.1126/science.1158357

Jones, Benjamin F. (2009). The Burden of Knowledge and the Death of the Renaissance Man: Is Innovation Getting Harder? *Review of Economic Studies*, *76*(1), 283–317. https://doi.org/10.1111/j.1467-937X.2008.00531.x

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy*, *26*(1), 1–18. https://doi.org/10.1016/S0048-7333(96)00917-1

Kauffman, S. A. (1993). *The origins of order: Self organization and selection in evolution*. Oxford university press.

Kogut, B., & Zander, U. (1992). Knowledge of the firm, combinative capabilities, and the replication of technology. *Organization Science*, *3*(3), 383–397.

Lai, R., D'Amour, A., Yu, A., Sun, Y., & Fleming, L. (2012). Disambiguation and Co-authorship Networks of the U.S. Patent Inventor Database (1975 - 2010). Harvard Dataverse. https://doi.org/hdl/1902.1/15705

Lamoreaux, N R, & Sokoloff, K. L. (1996). Long-term change in the organization of inventive activity. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(November), 12686–12692. https://doi.org/10.1073/pnas.93.23.12686

Lamoreaux, N R, & Sokoloff, K. L. (1999). Inventors, firms, and the Market for Technology in the Late Nineteenth and Early Twentieth Centuries. In Naomi R. Lamoreaux, D. M. G. Raff, & P. Temin (Eds.), *Learning by Doing in Markets, Firms, and Countries* (pp. 19–60). University of Chicago Press.

Lamoreaux, Naomi R., Sokoloff, K. L., & Sutthiphisal, D. (2013). Patent Alchemy: The Market for Technology in US History. *Business History Review*, *87*(01), 3–38. https://doi.org/10.1017/S0007680513000123

Lamoreaux, Naomi R, & Sokoloff, K. L. (2005). *The decline of the independent inventor: a schumpterian story?* National Bureau of Economic Research.

Lamoreaux, Naomi R, Sokoloff, K. L., & Sutthiphisal, D. (2011). The Reorganization of Inventive Activity in the United States during the Early Twentieth Century. In *Understanding Long-Run Economic Growth: Geography, Institutions, and the Knowledge Economy* (pp. 235–274).

Lee, S., & Bozeman, B. (2005). The Impact of Research Collaboration on Scientific Productivity. *Social Studies of Science*, *35*(5), 673–702. https://doi.org/10.1177/0306312705052359

Lobo, J., & Strumsky, D. (2008). Metropolitan patenting, inventor agglomeration and social networks: A tale of two effects. *Journal of Urban Economics*, *63*(3), 871–884. https://doi.org/10.1016/j.jue.2007.07.005

Lucas, R. E. (1988). On the mechanics of economic development. *Journal of Monetary Economics*, *22*(1), 3–42. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-45549121614&partnerID=tZOtx3y1

Malmberg, A., & Maskell, P. (2002). The elusive concept of localization economies: Towards a knowledge-based theory of spatial clustering. *Environment and Planning A*, *34*(3), 429–449. https://doi.org/10.1068/a3457

Maskell, P., & Malmberg, a. (1999). The Competitiveness of Firms and Regions: "Ubiquitification" and the Importance of Localized Learning. *European Urban and Regional Studies*, *6*(199901), 9–25. https://doi.org/10.1177/096977649900600102

Merton, R. K. (1968). The Matthew effect in science. *Science*, *159*(3810), 56–63.

Merton, R. K. (1973). The sociology of science: Theoretical and empirical investigations. University of Chicago press.

Nelson, R. R., & Winter, S. G. (1982). *An Evolutionary Theory of Economic Change*. Harvard University Press. Retrieved from http://books.google.com/books?hl=en&lr=&id=6Kx7s_HXxrkC&pgis=1

Noble, D. F. (1979). *America by design: Science, technology, and the rise of corporate capitalism*. Oxford University Press, USA.

Patel, N. (1973). Collaboration in the Professional Growth of American Sociology. *Social Science Information*, *12*(6), 77–92. https://doi.org/10.1177/053901847301200604

Pavitt, K. (1984). Sectoral patterns of technical change: Towards a taxonomy and a theory. *Research Policy*, *13*(6), 343–373. https://doi.org/10.1016/0048-7333(84)90018-0

Presser, S. (1980). Collaboration and the quality of research. *Social Studies of Science*, *10*(1), 95–101.

Price, D. de S. (1963). Big science, little science. *Columbia University, New York*, 119.

Price, D. J. de S. (1986). Little Science, Big Science and Beyond. Columbia University Press.

Romer, P. M. (1986). Increasing Returns and Long-Run Growth. *Journal of Political Economy*, *94*(5), 1002–1037. https://doi.org/10.2307/1833190

Simon, H. A. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society*, *106*(6), 457–476.

Sokoloff, K. L. (1988). Inventive Activity in Early Industrial America: Evidence From Patent Records, 1790–1846. *The Journal of Economic History*, *48*(04), 813. https://doi.org/10.1017/S002205070000663X

Solow, R. (1957). Technical Change and the Aggregate Production Function. *The Review of Economics and Statistics*, *39*(3), 312–320. https://doi.org/10.2307/1926047

Sorenson, O., Rivkin, J. W., & Fleming, L. (2006). Complexity, networks and knowledge flow. *Research Policy*, *35*(7), 994–1017. https://doi.org/10.1016/j.respol.2006.05.002

Storper, M., & Venables, A. J. (2004). Buzz: Face-to-face contact and the urban economy. *Journal of Economic Geography*, *4*(4), 351–370. https://doi.org/10.1093/jnlecg/lbh027

Strumsky, D., & Thill, J.-C. (2013). Profiling Us Metropolitan Regions By Their Social Research Networks and Regional Economic Performance. *Journal of Regional Science*, *53*(5), 813–833. https://doi.org/10.1111/jors.12048

Thagard, P. (1997). Collaborative knowledge. *Noûs*, *31*(2), 242–261.

Thorsteinsdottir, O. (2000). External research collaboration in two small science systems. *Scientometrics*, *49*(1), 145–160.

Tsoukas, H. (2005). *Complex knowledge: Studies in organizational epistemology*. Oxford University Press.

Uzzi, B., & Spiro, J. (2005). Collaboration and Creativity: The Small World Problem. *American Journal of Sociology*, *111*(2), 447–504. https://doi.org/10.1086/432782

van der Wouden, F., & Rigby, D. L. (2019). Co-inventor networks and knowledge production in specialized and diversified cities. *Papers in Regional Science*, *0*(0). https://doi.org/10.1111/pirs.12432

Ventura, S. L., Nugent, R., & Fuchs, E. R. H. (2015). Seeing the non-stars: (Some) sources of bias in past disambiguation approaches and a new public tool leveraging labeled records. *Research Policy*, *44*(9), 1672–1701. https://doi.org/10.1016/j.respol.2014.12.010

Wray, K. B. (2002). The Epistemic Significance of Collaborative Research. *Philosophy of Science*, *69*(1), 150–168. https://doi.org/10.1086/338946

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, *316*(5827), 1036–1039. https://doi.org/10.1126/science.1136099

Zeileis, A., Kleiber, C., Krämer, W., & Hornik, K. (2003). Testing and dating of structural changes in practice. *Computational Statistics & Data Analysis*, *44*(1–2), 109–123.

Footnotes

[1] US Congress, "An Act to Promote the Progress of Useful Arts", July 1836

[2] The analysis is limited to 1836 because before 1836 the USPTO did not make use of patent examiners. It is generally accepted that this institutional change has significantly impacted US patenting activity (Lamoreaux et al., 2011; Sokoloff, 1988).

[3] This assumes that the scraped inventor names recorded in the HistPat data are correct.

[4] Several extensive one-class classifier algorithms are ran. None came close to the performance of multi-class classifiers.

[5] In larger samples the quality of my trained models decreases, indicates that I falsely artificially assign inventors to the non-inventor class, obscuring the learning process. In smaller samples there are not enough correct non-inventors artificially assigned as non-inventor in the training data to learn from.

[6] The complexity score is log transformed using natural logarithms, for which the base is approximately 2.71.

[7] Figure 8 slightly deviates from Figure 1. The data used in Figure 1 includes patents with inventors who reside in non-MSA locations. These are excluded from Figure 8.