CERGE
Center for Economics Research and Graduate Education
Charles University Prague

Essays on Economics of Innovation

Taras Hrendash

Dissertation

Prague, November 2022

Taras Hrendash

Essays on Economics of Innovation

Dissertation

Prague, November 2022

Dissertation Committee

PATRICK GAULÉ (University of Bristol; chair)

ŠTĚPÁN JURAJDA (CERGE-EI; local chair)

ANDREAS MENZEL (CERGE-EI)

MARTIN SRHOLEC (CERGE-EI)

JAROMÍR KOVÁŘÍK (University of the Basque Country)

Referees

1ST READER (1st Reader affiliation)

2ST READER (2st Reader affiliation)

To my parents.

Table of Contents

Abstract

This thesis consists of three chapters.

The first chapter proposes a novel multi-layer clustering algorithm aimed to identify technology clusters from the network of collaboration ties among innovators with geo-coded locations. Using this novel algorithm, I identify innovation clusters in the U.S. Patent Inventor Database by simultaneously exploring two dimensions: the spatial distribution of inventors and the patterns of interconnections among them. Based on the clusters identified, I show that a combination of proximity and interconnectedness of inventors within the cluster boundaries is related to higher quality of innovations than those produced outside the clusters.

In the second chapter, I exploit the introduction of the USPTO's Prioritized Examination (Track One) Program to capture the impact of shortened pendency on the likelihood that a pending or granted patent will be commercialized via the transfer of property rights. I find that the Track One program significantly increased the probability of commercial reassignment of applications that were more likely to be prioritized.

In the third chapter, joint with Christian Fons-Rosen and Patrick Gaulé, we investigate causes of the ageing of the U.S. scientific workforce. Using novel data on the population of U.S. chemistry faculty members between 1960 and 2010, we find that the secular increase in the age of the academic workforce has mainly been driven by changes in the numbers of new faculty hires over time.

Acknowledgments

This research would not have been possible without the top-quality training in advanced economics and financial support toward living costs I received from CERGE-EI.

CERGE-EI was a life-changing place for me, where I had a chance to meet my scientific supervisor, Patrick Gaulé, who inspired my deep interest in the field of Economics of Innovation and has constantly encouraged my commitment to reaching the final stage of the long journey towards a doctoral dissertation.

I am extremely grateful to Professor Gaulé for his patience and excellent supervision throughout all ten years of my PhD studies. My thesis would not have been possible without the possibility to co-author with Patrick Gaulé and Christian Fons-Rosen.

I would like to thank Jaromír Kovářík, Štěpán Jurajda, Andreas Menzel, and Martin Srholec for enriching my thesis with priceless feedback and numerous suggestions.

When studying on mobility, I encountered productive atmosphere and openness of researchers at the Max Plank Institute for Innovation and Competition, especially thanks to Professor Dietmar Harhoff and Fabian Gaessler. This helped me to make major progress on my job market paper and thesis chapter on "The Impact of Prioritized Examination on Commercialization of Patents".

Many helpful comments and suggestions related to different parts of this thesis were received from conference participants at the fourth Conference of Geography of Innovation in Barcelona, twelfth Annual Northwestern/USPTO Conference on Innovation Economics in Chicago, and tenth SEI Faculty Workshop & Doctoral Consortium in Bologna.

Financial support from the Czech Science Foundation (GAČR) projects 16-05082S and 17-09265S, as well as Charles University Grant Agency (GAUK) project 508718 is gratefully acknowledged.

A fruitful collaboration and a unique job opportunity offered to me by Daniel Münich and Martin Srholec at CERGE-EI's IDEA Think Tank have provided me with a solid financial ground along the major part of my PhD path.

Last but not least, CERGE-EI has seen me married not only to the field of Economics of Innovation, but also to my wife Alona, who has mightily strengthened my faith in reaching the goal of thesis defence. Thank you, my dear.

Czech Republic, Prague — Taras Hrendash
November 2022

Introduction

This dissertation presents an empirical analysis of three different topics in the field of Economics of Innovation. Each of the three dissertation chapters is self-contained and lacks direct links to the other chapters. All three chapters are based on different data sources, but each one includes data that was originally collected in the United States.

The first chapter focuses on innovation clusters. There are three most common contexts in which innovation clusters are discussed in the literature. First, a geographical space of clustering, proximity, and spatial concentration of inventors are considered to be essential features of clusters in the geography of innovation studies (Verspagen & Schoenmakers, 2004; Giuliani, 2007; Huber, 2012; Ter Wal, 2013; Nomaler & Verspagen, 2016). Second, clustering as a particular pattern of connections among economic actors has been studied in the literature on social network analysis. In the context of innovations, one example of such connections is a collaboration network of innovators. At the intersection of social network analysis and the geography of innovation, several studies have considered the role of collaboration networks in combination with geographic proximity in determining the innovation and economic performance of regions (Lobo & Strumsky, 2008; Strumsky & Thill, 2013; Coffano, Foray, & Pezzoni, 2017; Hazır, LeSage, & Autant-Bernard, 2018), and have explored the properties of knowledge networks captured at different geographical levels (Whitington et al., 2009; Galaso & Kovářík, 2020).

One of the major findings shared by these studies, which motivated the first chapter of this disseration, is that spatial distribution along with a structure of interconnections among inventors constitute complex multi-layer structures. Such structures exhibit salient features that play a crucial role in fostering innovation performance. Thus, the third context in which innovation clusters are commonly discussed is the differences in innovation performance of individual inventors who are within and outside of the cluster boundaries. The latter context has crucial policy relevance, as innovation clusters are among the common targets of regional policies aimed at boosting innovations.

In the first chapter, I propose a new cluster identification algorithm that can be used as a tool for analysis of the spatial distribution of innovative activities, and thus provide a useful insight for policy-makers aiming to target both established and emerging innovation clusters. My approach consists of a stepwise restriction of the search space – a set of active inventors – connected via a network of collaboration linkages, along two dimensions: geographical proximity and interconnectedness. My cluster identification algorithm endogenously detects the borders of clusters, illustrating the relationships between the geographical proximity, the interconnectedness of innovators, and the quality of innovations.

The second chapter was inspired by two major aspects of the U.S. patent system that can be observed in publicly available data from the U.S. Patent and Trademark Office – the patent examination procedure and patent reassignments. First, it is notable that, at any given time, there are a large number of pending patent applications, which are idly waiting for the outcome of examination procedure at the USPTO (Mitra & Kahn, 2013) and, thus, are not being fully exploited by technology users or by society as a whole. Second, there is a surprisingly large amount of data about reassignments of intellectual property rights protected by patents, while sale-of-patent transactions were not common and only anecdotal evidences are currently accessible (for example, the IP3 Program originally initiated by Google Inc.).

The latter two aspects of the patent system, and the linkage between a lengthy patent examination and the effectiveness of the market for technology was first studied in Gans, Hsu, and Stern (2008). The key empirical fact they established was that the probability of a settlement between an innovating start-up company and a downstream firm that acquires the property rights for a new technology is significantly higher when the patent application that covers the underlying technology progresses to the final stage of the examination process. It is important to note that Gans, Hsu, and Stern (2008) focus on the hazard rate rather than on the unconditional probability that the commercial agreement will be signed. Therefore, the external validity of this finding is rather limited and applies only to successful examples of the commercialization of patents.

In the second chapter, I contribute to the previous literature by raising a broader question about the links between the length of the patent examination process and commercialization of technologies. I explore whether the earlier result regarding patent examination is also positively related to the unconditional chances of innovating start-ups being able to sell their patents to large firms. My key finding is that shortening the examination time from around 24 to around 12 months for applicants in the USPTO Track One Prioritized Examination program is associated with at least 50% higher probability of reassignment of patents. I suggest that longer pendency time of applications at patent offices may not only lead to a welfare loss due to the deferred commercialization of innovations, but may also create frictions on the market for technology that reduce the overall saleability of patents.

The third chapter, joint with Christian Fons-Rosen and Patrick Gaulé, is about the ageing of the U.S. scientific workforce. It is observed in the literature that the age of scientists at the time when they make major discoveries tend to increase over time. This is not only true for the most prominent scientific achievements, such as a Nobel Prize (Jones, 2009; Jones & Weinberg, 2011), but also for more common career milestones, such as grant awards (Daniels, 2015). This observation has the potential to gain significant policy relevance, if it can be also shown that age influences the quantity and type of knowledge produced by individuals. It is also important for making appropriate policy changes, if there is a need for them, to find the strongest forces driving the upward trend of the typical age among the scientific workforce.

In the third chapter, we build a demographic model of the U.S. academic workforce to shed light on the causes of its ageing. The model leverages novel data on the population of U.S. chemistry faculty members between 1960 and 2010. Combining the flexibility of assumptions in our model and the relatively long time span of the relevant demographic data, we aim to quantify the importance of various factors that potentially contribute to the ageing pattern. We find that changes in the numbers of people hired over time was the major driver in the ageing of our sample. Theis finding has a important implication for science policy. We suggest that policies narrowly focused on increasing the overall size of workforce, may have positive effects only in the short term, whereas the long-term consequences of such policies may counterbalance the early-stage benefits or even generate problematic and longer-lasting conditions, such as an increasing share of the older segment of a scientific population.

Chapter 1

Cluster Identification in Collaboration Networks of Innovators

1.1 Introduction

Ever since Porter's (1990) seminal paper on competitiveness, innovation and growth, innovation clusters, rather than countries or regions, have become increasingly popular as alternative units of analysis in economic geography. Recently, clusters have become key targets of regional policies aimed at boosting innovation.[1] The cluster mapping procedure – delineation of geographical and social boundaries of clusters – is an important technical task demanded by and crucially relevant to policymakers. Hence, there is extensive discussion of cluster mapping in the geography of innovation literature (Delgado, Porter, & Stern, 2015; Ketels, 2017).

As the amounts of micro-level geocoded data on entrepreneurial and innovation activities has grown over time, the focus of cluster mapping techniques have shifted from pre-defined administrative units (states, counties or districts) and well-known case studies (e.g. Silicon Valley, Route 128) to more flexible computable algorithms that can identify clusters spanning across arbitrary areas. Moreover, such algorithms may help to identify emerging clusters in developing regions, without the need to refer to established cases.

In this paper, I propose a new cluster-identification algorithm that can be used as a tool for analysis of spatial distribution of innovative activities, and thus provide a

[1] Examples include the Smart Specialization Strategy as a part of EU innovation policies, and the Regional Innovation Clusters initiative by the U.S. Small Business Administration

useful insight for policymakers aiming to target both established and emerging innovation clusters. My approach consists of a stepwise restriction of the entire search space – a population of active inventors – connected via a network of collaboration linkages, along two dimensions: geographical proximity and interconnectedness. Such a cluster-identification algorithm endogenously detects the borders of clusters, illustrating the relationship between geographical proximity, interconnectedness of innovators, and quality of innovations.

I use the newly-proposed algorithm to identify clusters from the microgeographic data of U.S. inventors – geolocalized with their residence addresses. Unlike previous studies where parameters of the clustering algorithms are set to arbitrary fixed values, I search over a range of possible values that determine a set of identified clusters. I show that the difference in the quality of innovations produced within and outside the clusters varies a lot with parameters of the clustering algorithm. As a result, it is possible to find a specific range of parameter values, that describe the strength of collaboration ties and geographical proximity of inventors, such that their within-cluster performance exceeds the outside-of-cluster counterpart.

Specifically, clusters with more than 170 members, and a maximum of 9-12 kilometers between the closest members, produce patents that outperform those produced outside them for the following quality measures. First, they receive 13% more forward citations on average. Second, these citations are more dispersed across the technological areas, resulting in 20% higher generality. Third, patents produced within these clusters are 55% more likely to belong in the 95% of most-cited patents than other patents produced outside clusters. Hence, the proposed novel algorithm successfully recovers the innovation clusters from the data.

To illustrate a comparative advantage of my algorithm, I consider counterfactual methods of cluster identification – density-based clustering of geographic locations and identification of large connected components in a collaboration network. Neither of these methods allow me to identify clusters that would exhibit a positive and statistically significant difference in innovation performance that is robust to a choice of performance measures. As a consequence, the newly-proposed algorithm captures features that are not recovered by other commonly-used methods of cluster identification.

This paper is structured as follows. Section 2 provides an overview of studies in the geography of innovation and social network analysis literature that constitute the methodological basis of this paper. In Section 3, I propose a formal definition of a 'cluster' and give a detailed description of a novel methodology – a cluster-identification algorithm. In Sections 4 and 5, I focus on the application of the newly-proposed methodology to the U.S. patent dataset. Section 6 concludes.

1.2 Related literature

The methodological basis of this study borrows from two strands of the literature, the geography of innovation and social network analysis. In the geography of innovation literature there are various ways of referring to technology clusters. One group of studies uses specific, usually well-known examples of particular industries or locations, as in Ter Wal (2013), Giuliani (2007) and Huber (2012). There are also studies that refer to administrative units (states, counties or districts) as separate clusters, as in Verspagen and Schoenmakers (2004) or Nomaler and Verspagen (2016).

However, analysis of clusters based on a pre-defined division of space has a crucial limitation, known in spatial analysis as the "modifiable areal unit problem", or the "ecological fallacy". This limitation arises because "one administrative unit may encompass multiple clusters, while one technology cluster may expand across several administrative lines" (Alcacer & Zhao, 2012, p. 741). In order to tackle this inconsistency, diverse distance-based clustering techniques, including partition clustering and hierarchical clustering, have been developed in the geography of innovation literature to identify cluster borders from the actual spatial distribution of innovation activities. These techniques, however, fail to identify clusters of irregular geographic shapes (Alcacer & Zhao, 2016).

Another group of techniques, including the density-based cluster identification approach in Alcacer and Zhao (2016) and city clustering algorithm in Rozenfeld et al. (2008), overcome the limitations of the previous methods, however, they still tend to rely on arbitrary pre-defined parameters, including the neighborhood radium and contour threshold in Alcacer and Zhao (2016) or the granularity of the spatial grid and cell size in Rozenfeld et al. (2008).

As long as an ultimate outcome of the clustering algorithm is contingent on parameters pre-defined by a researcher, an empirical analysis of the innovation performance of clusters identified might lead to inconsistent implications. Thus, the potential evidence of positive externalities arising from the clustering, if any, becomes less robust. In order to address this issue, I seek to endogenize the choice of parameter values in the proposed clustering algorithm by searching over the range of possible values and empirically tracing the relationship between parameter values and the innovation performance of clusters identified. My conjecture is that a choice of parameters can be considered optimal if the resulting clusters exhibit significantly better innovation performance than innovators in the outside environment.

In contrast to the geography of innovation literature, where clustering has been considered to be a geographical concentration of actors and innovative activities, in network analysis, clustering is considered to be a particular pattern of interconnections among a set of unique nodes, represented, for example, by economic actors. In a similar vein, as innovation geographers have developed various techniques to identify technology clusters from the geographic data, network researchers have developed clustering techniques. Such techniques, known as community detection algorithms, aim to partition complex structures of interpersonal connections into cohesive groups of individuals who are tightly connected to, or share similarities with others in the group, and have loose connections or differ from individuals outside the group.[2]

In the context of innovations, a widely-studied and important type of interpersonal connection is knowledge transfer among innovators. Multiple interpersonal ties, in the form of co-authorships, citations or co-inventions, constitute a complex structure of the knowledge network. Positioning in the knowledge network, instances of collaboration and knowledge exchange among individuals have all been studied as potential determinants of innovation performance in the network literature (Cassi & Plunket, 2012; Ter Wal, 2013).

Studies at the intersection of social network analysis and the geography of innovation have considered the role of collaboration networks and knowledge diffusion in combination with geographic proximity, in determining innovation and economic performance of regions (Lobo & Strumsky, 2008; Strumsky & Thill, 2013; Coffano, Foray, & Pezzoni, 2017; Hazır, LeSage, & Autant-Bernard, 2018). They have also explored the properties of knowledge networks captured at different geographical levels (Whitington et al., 2009; Galaso & Kovářík, 2020). One of the major findings that motivated this paper is that a spatial distribution and a structure of interconnections among inventors form a complex multi-layer structure with salient features that play a crucial role in fostering innovation performance.

In this paper, I therefore combine methodologies employed in previous studies to depict innovation clusters as an intersection of two patterns – geographical concentration

[2]Rani and Mehrotra (2019) provide a review of ongoing developments in this field.

and interconnectedness via collaboration ties among inventors. My approach is in line with the definition proposed by Michael Porter: “A cluster is a form of network that occurs within a geographic location, in which the proximity of firms and institutions ensures certain forms of commonality and increases the frequency and impact of interactions.” (Porter, 1998, p. 226)

Furthermore, my approach is in line with the definition of a ‘regional innovation system’ in Asheim and Gertler (2005, p. 299), or ‘innovation ecosystem’ in general. The latter concept emerged in the geography of innovation literature in the early 2000s as an extension of the ‘cluster’ notion.[3] Rallet and Torre (2017) emphasize one of the main distinctive features of an innovation ecosystem – it is not only the economic actors who are located within a close geographic proximity, but also a complex network of interconnections among them that constitutes an extended environment.

In line with this literature addressing innovation ecosystems, I propose two main characteristics of a cluster. First, economic actors within a cluster are interconnected; thus, their interactions might constitute a localized part of a large-scale network. Second, a cluster is characterized by the geographical proximity of its members, which fosters frequent interactions among them. In this paper I refer to both interconnectedness and proximity as essential characteristics of technology clusters and I show how a combination of both characteristics can determine the geographical and social boundaries of clusters.

Firstly, I consider geographical proximity as an initial criterion for identification of clusters. In line with previous studies (Rozenfeld et al., 2008; Catini et al., 2015), I apply a distance-based clustering technique to delimit groups of densely-located innovators, thus considering a purely geographic ‘layer’ of clustering. I refer to such groups as benchmark clusters or ‘geographical components’. Further, I suggest that a set of interpersonal linkages represented by collaboration ties should be considered to be an additional ‘layer’ of the clustering space. Finally, I propose a novel multi-layer clustering algorithm aimed at simultaneously identifying innovation clusters across two dimensions: geographical space and the structure of collaboration ties among innovators.

While showing how a combination of clustering at the geographical and interpersonal levels might shape the boundaries of innovation clusters, I also measure the role of both pure geographic proximity and collaboration in determining the variation observed in innovation performance. I contrast the differences in innovative output produced within and outside the ‘borders’: for a set of connected components in a collaboration network;

[3]Granstrand and Holgersson (2020) review this strand of the literature.

for a set of benchmark clusters – components in a geographical network; and, finally, for a set of technology clusters identified with my clustering algorithm.

Using a panel of innovators I seek to exploit a within-inventor variation in innovation performance and relationship to clusters, controlling for unobserved time-invariant individual characteristics, to measure the degree of association between the quality of innovations produced and relationship to a cluster. I observe a significantly higher quality of innovation output produced within technology clusters than outside them. In contrast, for the purely geographical counterparts, I find little or no significant difference between innovation output produced within and outside them. Thus, interpersonal knowledge flows among geographically proximate innovators are shown to play a reinforcing role in determining the individual performance of innovators. The latter finding is in line with previous studies where it was shown that geographic proximity is a necessary but not a sufficient condition of the success of innovation clusters.

1.3 Methodology

1.3.1 Definition of a cluster

In my analysis of technology clusters, I focus on individual inventors, rather than firms or institutions, as component elements of clusters, and consider a particular type of interconnection between them – collaborations on joint patenting activities[4]. I draw on the network studies discussed in the previous section by referring to the structure of these interconnections as a collaboration network. More formally,

Definition 1. A collaboration network is an undirected graph C represented by an ordered pair $C = (V, E)$ comprising a set V of vertices (individual inventors) together with a set E of edges (collaboration ties), which are two-element subsets of V (co-authors on a given patent).

Therefore, technology clusters are assumed to comprise particular parts of a larger collaboration network, such that cluster members are concentrated at a certain spatial scale and are characterized by geographical proximity. In order to describe a geographical dimension of clusters, consider:

[4]This specification is linked to the microgeographic data that is employed to illustrate the proposed methodology. The latter, however, does not preclude considering a discussion in a different context, such as identification of industrial clusters consisting of firms, or technology clusters consisting of R&D labs interconnected via different types of linkages.

Definition 2. A geographical network is an undirected graph $G = (V, D(x))$ comprising the same set of vertices V from a graph C together with a set $D(x)$ of edges. A pair of elements $(v_i, v_j) \in D(x)$, if a geographical distance[5] between them is less than a threshold of x km.

The two graphs, $C = (V, E)$ and $G = (V, D(x))$, can alternatively be viewed as separate layers of a single multidimensional graph[6] and effectively outline the search space for the clustering algorithm described later in this section.

It is crucial to note that dependence of the set $D(x)$ of proximity relations on a parameter x makes the inputs and subsequent outcome of the cluster identification procedure contingent on the value of x. In Section 3, I verify whether innovations produced within cluster borders are of significantly higher quality than outside them. I show that the latter inference might also be highly susceptible to the choice of the pre-defined parameter. As a result, I suggest that the value of x, the maximum geographical distance separating a node from its nearest neighbor within a cluster, is an essential element of the cluster definition; thus it can and should be specified endogenously so that for a set of identified clusters the implied advantage of being inside a cluster in terms of innovation performance is significantly large and robust to the choice of performance measure.

Based on the defined notions and implied characteristics of clusters, I propose the following formal definition of a cluster:

Definition 3. A technological (innovation) cluster is an induced subnetwork[7] of an entire network $C = (V, E)$, denoted by $C[S(x)]$ that satisfies two conditions:

(i) the vertex set $S(x) \in V$, where $\forall v_i \in S(x)\ \exists v_j \in S(x)$, such that $v_i v_j \in D(x)$;

(ii) there is a path – sequence of edges $i_1 i_2, i_2 i_3, ..., i_{K-1} i_K$ – between any two vertices $(v_i, v_j) \in S(x)$, such that $i_k i_{k+1} \in E$ for each $k \in \{1, ..., K-1\}$, $i_l \in S(x)$ for each $l \in \{1, ..., K\}$ with $i_1 = v_i$ and $i_K = v_j$.

The suggested definition can be thought about from the following perspective: technology clusters comprise groups of closely-located inventors who are also directly or indirectly interconnected with each other. The latter condition might be interpreted as the existence of a common knowledge-sharing environment among members.

[5]Hereinafter I refer to geographical distance as the great-circle distance between two points on a sphere given their longitudes and latitudes calculated using the Haversine formula.

[6]For brevity reasons I use the simpler notation of two separate graphs.

[7]A subset of the vertices and all edges connecting pairs of vertices in that subset.

1.3.2 Cluster identification algorithm

A fundamental goal of cluster identification is to classify the groups of economic agents (or partition a network into groups of vertices), such that they share common properties. According to the proposed definition, vertices within a technology cluster are assumed to share two basic properties – each node has at least one geographical neighbor who is also located within a cluster and each node is connected by at least one path with all other nodes in a cluster via the collaboration ties existing in a cluster.

Consider the following definition:

Definition 4. A component is a maximal subnetwork,[8] such that every pair of nodes in the subnetwork is connected by a path[9] (Jackson, 2008).

Based on the aforementioned Definitions (1-4), I propose a simple algorithm of the cluster identification that takes a set of nodes (innovators with geo-coded locations) as an input and two networks: (*i*) network C^0 of collaboration ties (Figure 1.1 (a)) constructed according to Definition 1 and (*ii*) geographical network G^0 constructed according to Definition 2 that describes the relative allocation of nodes in a geographical space. The following steps thus lead to identification of the technology clusters as per Definition 3:

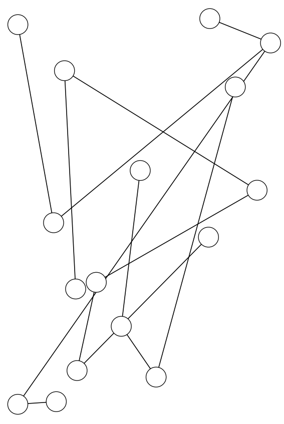
1. In network G^0, find its components (see Definition 4). Consider each component as a purely geographical counterpart of a cluster, hereinafter referred to as a *geographical component*. Each node inside a geo-component has at least one neighbor within a radius of x km, but it might not have any collaboration ties with other nodes (Figure 1.1 (b)).

2. For each geo-component i create an induced subnetwork C_i^1 of the network C^0 constituting all members of a geo-component and collaboration ties among them. For each subnetwork C_i^1 find its connected components (Figure 1.1 (c)).

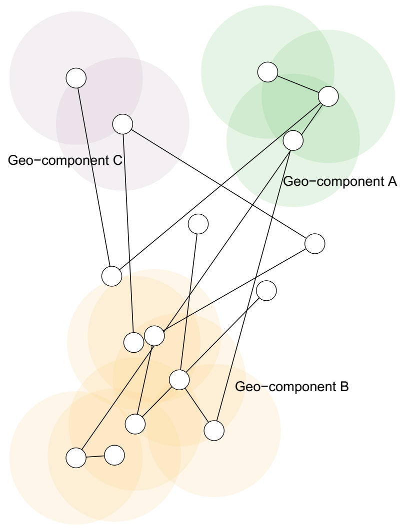
3. If there is more than one connected component per geo-component, for each connected component construct the network G_i^1 according to Definition 2 and find its corresponding geo-components; otherwise consider a candidate from Step 2 as a technology cluster identified at the first iteration.

[8]A subnetwork is maximal if it is not possible to add any other node to the subnetwork without violating the next condition.

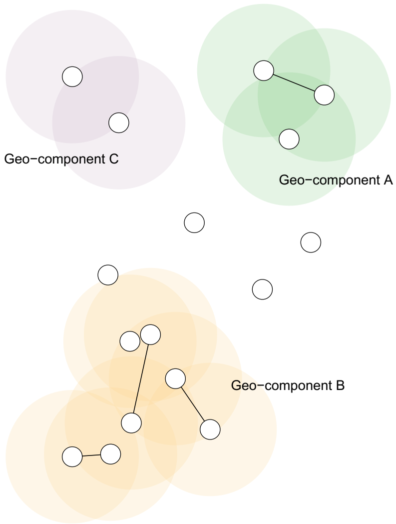
[9]A path is a sequence of intermediating edges between a pair of nodes.



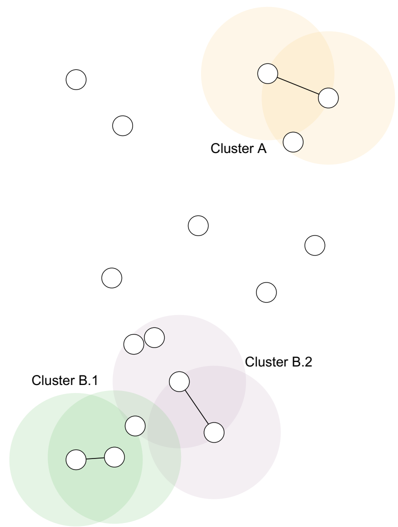
(a) Collaboration network G^0.



(b) Geographic components in D^0.



(c) Connected components in G^1.



(d) Geographic components in D^1.

(a) Provides an example of an input collaboration network with geocoded positions of vetrices.

(b) Illustrates the first stage of the algorithm and the geo-components obtained. Circles here represent the distance threshold x.

(c) Illustrates the second stage of the algorithm and the connected components obtained within each geo-component. Note that: the connected component in geo-component A already corresponds to the cluster; the geo-component B contains more than one connected component and requires further analysis; the geo-component C does not contain any connections within its borders so will not be considered in the further stages.

(d) Shows the final set of clusters identified. Note that only two out of three connected components in B can be considered to be clusters satisfying the proximity requirement.

Figure 1.1: Cluster identification algorithm.

4. Iterate Steps 2 and 3 until all geo-components of the network G_i^k entirely coincide with the connected components of the subnetwork C_i^k. The latter can be referred to as the technology clusters identified at the k-th iteration of the algorithm (Figure 1.1 (d)).

The final set of clusters identified can be further restricted by a minimum component size threshold measured by the number of cluster members. Hereinafter I refer to the minimum component size threshold $N_{\min}$ and the maximum distance to the closest geographical neighbor x as the pre-defined parameters of the algorithm.

The proposed algorithm of cluster identification is applicable, in general, to the microgeographic data on innovation collaborations that possesses the following features:

1. Economic actors (e.g. firms, researchers or individual innovators) are grouped in collaboration teams (e.g. joint R&D projects, research papers or patents).

2. Each economic actor is associated with a unique and geo-coded location (e.g. registration, residence or working address).

In contrast to previous studies, I aim to endogenize the choice of pre-defined parameters of the cluster-identification algorithm by searching over the range of possible values. This allows me to link different outcomes of the algorithm to the innovation performance observed within the clusters identified, and thus identify an optimal range of the parameter values.

In the following sections, I present an example of how technology clusters can be identified using the proposed algorithm and how the optimal values of pre-defined parameters of the algorithm can be found. The ultimate result of the cluster identification procedure illustrated below is a set of technology clusters. I calculate and analyze various geographic and network characteristics of identified clusters to illustrate the interrelationship between spatial concentration and collaborations among cluster members.

1.4 Data

I focus on co-patenting collaborations among individual inventors as a source of data for cluster identification and utilize patent quality measures to estimate the difference in innovation performance of inventors within and outside the delineated cluster borders. Noteworthily, the patent data combines all the types of information needed for an empirical analysis in the present setup. First, patents constitute an example of collaborations among individual inventors. Second, patent applications contain information about the geographical locations of co-authoring inventors. Last but not least, patents represent a kind of innovation output that can be assessed in terms of its quality and characterizes the innovation performance of individual inventors.

I employ the U.S. Patent Inventor Database retrieved from the Harvard Dataverse (Li et al., 2014), which contains records on more than 4 million patents granted by the United States Patent and Trademark Office between 1975 and 2010. Each patent record in the database comprises disambiguated names[10] of inventors listed in the application, their geographical locations (geo-coded residence addresses at time of filing patent applications), patent grant and application filing dates, and the major technology classes of patents defined according to the U.S. Patent Classification System (USPC). Importantly, each innovator in the dataset is uniquely related to the geo-coded location[11] and geographical distances between all inventors can be calculated in a given application-filing year.

Using these microgeographic data on patenting activities, I aim to identify innovation clusters at the national level, restricting the geographical scope of analysis to the U.S. borders. I consider a subset of the U.S. patents, namely those for which the primary inventor is based in the U.S., which reduces the number of patents in the dataset to 2.3 million.

Although the latter subset of the patent and inventor data would be sufficient for implementation of the cluster-identification algorithm, I merge it with the Examiner Citation Data (Sampat, 2012) retrieved from the Harvard Dataverse, which contains examiner and other backward citations made by the U.S. patents granted between 2001 and 2010, in order to obtain patent quality measures used in the empirical analysis for

[10]See Li et al., (2014) for the description of disambiguation procedure.

[11]This feature of the dataset is essential, though its actual presence is contingent on the accuracy of the preceding disambiguation procedure. In fact, 4% of inventors do not satisfy this condition and are discarded from further analysis.

this study.

Merging the patent and citations datasets imposes the following restrictions on the patent grant and application filing years. First, as an analysis of the patent quality is based on counts and composition of forward citations, and in order to make a comparison of patent quality consistent across time, I consider citations made within a two-year time window starting from the patent grant date of a focal patent. The range of possible patent grant years in the dataset is thus restricted to 2001-2008. Second, as mentioned earlier, unique geographical locations of inventors are available at the time an application is filed and to construct a single geographical network that represents relative locations of inventors, patents filed in the same year should be considered.[12] Therefore, the dataset allows me to construct geographical networks and to identify sets of clusters for each year of application until 2008.

However, as I rely on the citation data that are collected during the time window linked to the patent grant date, and the difference between application filing and patent grant dates is not the same for all patents, ranging from several months to 10 years in the dataset, I explicitly limit the difference allowed between patent grant and application filing dates to at least 1 and at most 5 years.[13]

If, for some patents filed in a year, it is not possible to obtain quality characteristics because of excessive application-to-grant lag I exclude that year. This ensures that the distribution of patents used for the cluster identification are unchanged in any given year. For example, to identify clusters in 2006, patents that were filed in 2006 and were granted between 2007 and 2011 could be considered, but this would require a truncation of the distribution at 2008, since it is the latest possible patent grant year for which the citation data can be collected that respects the two-year time-window restriction. Therefore, the range of application filing years for which I can identify clusters is restricted to 2000-2003.

The final dataset includes 395,033 patent applications filed by 379,358 inventors between 2000 and 2003 for which patents were granted between 2001 and 2008. For each filing year between 2000 and 2003, data on patent applications are processed according to the following procedure:

(1) construct a geographical network based on locations of individual inventors as stated in the corresponding patent applications;

[12] I assume that the residence address details of the inventors are not changed during the year an application is filed.

[13] This condition is satisfied for about 90% of patents in our dataset.

(2) construct a collaboration network of co-patenting ties among inventors;

(3) identify technology clusters, with each cluster comprising a set of individual inventors who belong to it;

(4) consider a cross-section of inventors, where each inventor has a cluster status (within/outside the cluster) and a measure of innovation performance.

As a result, I obtain an unbalanced panel of 551,946 inventor-year pairs with two associated time-variant variables: cluster status and innovation performance equal to the average quality of patent applications filed in a given year by each inventor. The resulting panel data are used to study the relationship between individual innovation performance and presence within a cluster. The strength of such a relationship is used later in this analysis as a criterion for optimizing the pre-defined parameters (i.e. component size and distance thresholds – N_{min} and x, correspondingly) of the cluster-identification algorithm.

1.5 Results

1.5.1 Application of the algorithm

The ultimate outcome of the cluster-identification algorithm (i.e. the geographical borders of clusters and composition of their members) is contingent on the input data (i.e. the exact geographical locations of inventors and structure of collaboration ties among them), as well as the values of pre-defined parameters.

Even though the outcome of the cluster-identification algorithm is not uniquely defined at this stage of analysis, due to the variability of factors described above, I present a 'snapshot' of the cluster mapping for an arbitrarily chosen time period and preset parameter values to illustrate the proposed definition of a 'cluster' and recapitulate its main features.

Figure 1.2 provides an example of clusters depicted on a fragment of a map. Each color corresponds to a separate cluster; each circle depicts an area delimited by the maximum distance threshold around a unique location of inventors. It is clear from the map that a single node is considered to be a part of a given cluster if it falls within a radius of the 'close neighborhood' defined by the distance threshold. It is also apparent that this condition is necessary for being a part of a given cluster, but not sufficient. Therefore, clusters might overlap visually on a geographical map if there are inventors who are

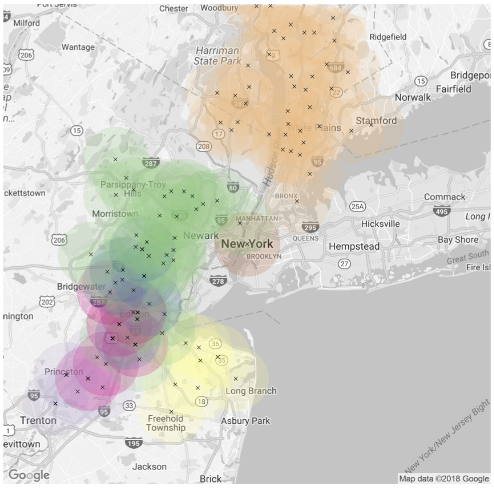


Figure 1.2: Sample map of clusters.

geographically close to each other, but not interconnected via a collaboration network and, hence be identified as members of different clusters.

Figure 1.3 (a) depicts separately the largest cluster on the map with overlaid collaboration linkages. It consists of 80 individual inventors based in the State of New Jersey. Patents produced in this cluster represent innovations in pharmaceutical and biotechnology industries, most of them were assigned to the Schering Corporation[14] and Dendreon Corporation.[15] Figure 1.3 (b) depicts a network of collaboration linkages among individual inventors in this cluster. In this figure, nodes (individual inventors) are placed in a two-dimensional space in such a way[16] that shorter distances between nodes correspond to more collaboration linkages and do not necessarily correspond to closer geographic locations of inventors.

1.5.2 Optimization of the parameter values

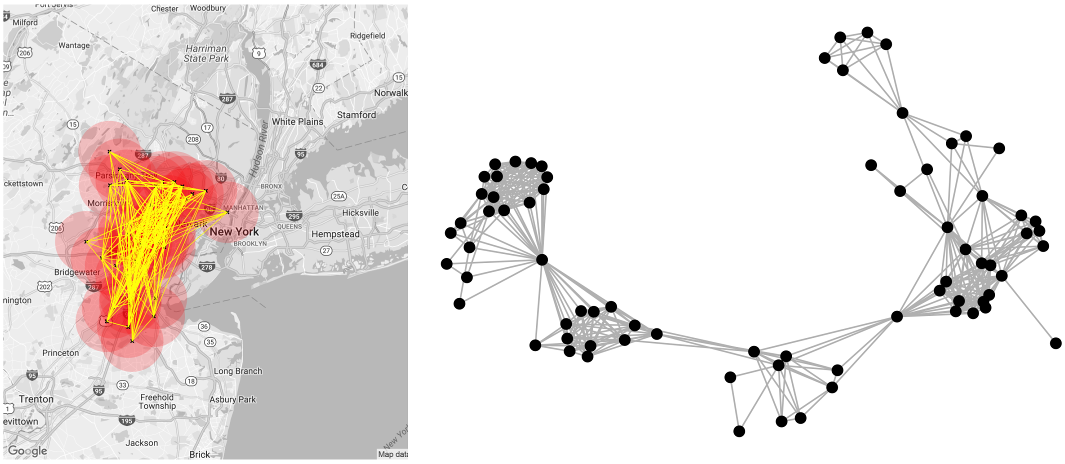
In the empirical analysis I utilize the variability of cluster borders observed across time in order to compare the performance of individual inventors in periods when they fall into and out of the clusters identified. Thus, I aim to measure the degree of association between 'being in a cluster' and delivering a certain level of innovation performance.

Effectively, I compare the difference between innovation performance within and out-

[14] The company's headquarters is in Kenilworth, New Jersey. The company was acquired by Merck & Co. in 2009.

[15] The company had a manufacturing facility in Morris Plains, New Jersey, sold in 2012 to Novartis Pharmaceuticals Corporation.

[16] The Fruchterman-Reingold Layout (Fruchterman & Reingold, 1991) was used for plotting the network of collaboration linkages.



(a) Cluster. **(b)** Network structure.

Figure 1.3: Example of identified cluster.

side the 'borders': for a set of connected components in a collaboration network; for a set of benchmark 'geo-components' that resemble the outcome of conventional distance-based clustering algorithm; and, finally, for a set of technology clusters resulting from the proposed cluster-identification algorithm.

To measure individual innovation performance I calculate the average quality of inventors' patents. I consider three proxies of the patent quality, in line with Acemoglu and Akcigit (2014):

1. Average number of citations received per patent during a two-year window since a patent grant date.

2. Average generality index measuring the average dispersion of forward citations received by patents (during a two-year window) in terms of the technological classes of their citing patents, defined for a patent portfolio with positive citations as: $\frac{1}{N}\sum_{j=1}^{N}\left(1-\sum_{i\in I}s_{ij}^2\right)$ where $i \in I$ denotes a technological class and $s_{ij} \in [0,1]$ denotes the share of citations that patent j receives from patents in technological class I. The index is close to 1 if there is a large number of classes citing a patent and is 0 if all citations come from a single technological class. Let us also assign 0 value to patents with no forward citation.

3. The number of 'tail patents', i.e. patents which, during a two-year window, received a number of citations above the 95$^{\text{th}}$ percentile in the distribution of all patents granted in the same period.

I use each of the three measures to evaluate the degree of association between the presence of an inventor in a connected component, geo-component or cluster on the one side and her innovative output on the other. Table 1.1 summarizes the correlation coefficients, that are positive and statistically significant. Therefore, I suggest that for a given method of grouping inventors the relationship between belonging to the group and delivering a certain level of innovation quality is robust if it has the same sign and significance level across all three measures.

	Measures of innovation performance		
	Citations	Generality	'Tail' patents
Citations per patent	1		
Generality (average)	0.574^{***}	1	
'Tail' patents (number)	0.409^{***}	0.152^{***}	1

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 1.1: Correlations

In the panel of 551,946 inventor-year pairs, I construct three dependent variables that follow the definitions of innovation quality measures introduced above and three binary independent variables that take a value of one if the inventor was a 'member' of one of the identified connected components, geo-components or clusters and zero otherwise. In total, 9 different regressions with inventor fixed effects are estimated according to the following equation:

$$Quality_{it} = \alpha + \beta Member_{it} + \gamma_i + \epsilon_{it} \tag{1}$$

For each independent variable that represents a certain method of grouping inventors (based on their interconnectedness, proximity or both), I obtain three values of the coefficient β to measure the degree of association with one of the three innovation performance measures. By controlling for unobserved heterogeneity γ_i, I capture the within-inventor variation in innovation performance for those inventors who switched between being a 'member' of and being outside the corresponding groups at least once during the time frame covered by the panel. I refer to the output of three alternative regressions for each type of the groups to ascertain the robustness of association between the dependent and independent variables.

Because each method of grouping inventors relies on the minimum component size and the maximum distance thresholds ($N_{\min}$ and x, correspondingly), which affect the composition of groups identified and hence the variation in the right-hand side of the regression equation (1), I reestimate each of the 9 specifications while changing the values of the pre-defined parameters along the grid.[17] For each grouping method I obtain a grid of β estimates that correspond to different combinations of the parameter values. I identify the optimal values such that the β estimates are significantly positive and robust to the choice of an innovation quality measure.

The results of a sensitivity analysis are reported in a concise way using a graphical representation (Figure 1.4). Each of the 9 graphs corresponds to a certain regression specification and plots a grid of possible parameter values. Each dot on the grid depicts properties of the corresponding β estimate: size being proportional to its magnitude[18], color showing the significance level and shape representing a sign of the relationship between the cluster status and innovation performance of individual inventors.

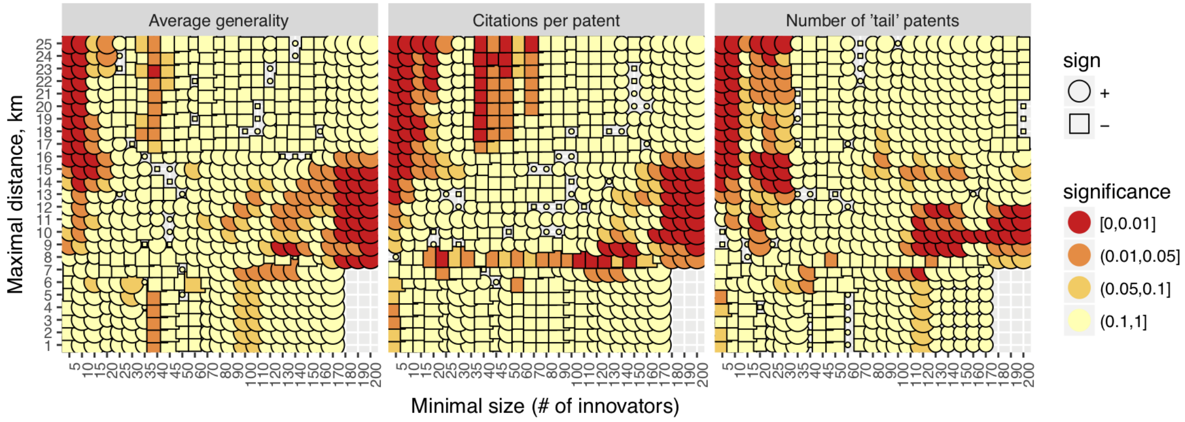
In the following subsection, I identify the optimal values of the parameters (i.e. component size and distance thresholds – $N_{\min}$ and x, correspondingly) used in the clustering algorithm that lead to the positive and significant association, which is robust to the choice of a performance measure. I show below that different methods of partitioning inventors into cohesive groups might exhibit different degrees of sensitivity of the optimal values to the choice of innovation performance measure.

1.5.3 Results of a sensitivity analysis

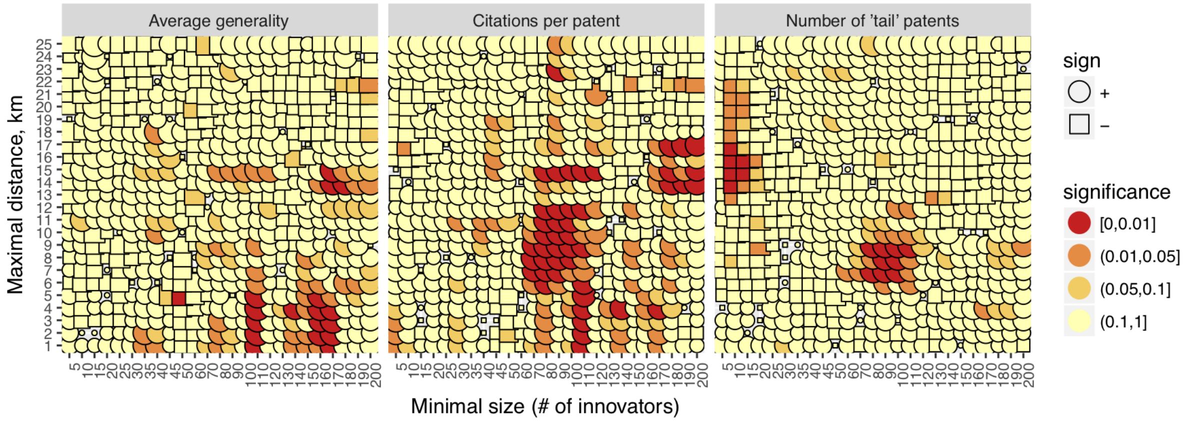
First, consider the network components (Figure 1.4 (c)). Under this approach, inventors in a given year were grouped according to their positions in the collaboration network, disregarding their geographical locations. The vertical axis on the grid of estimates is thus irrelevant and the horizontal axis corresponds to the minimal restricted size of a connected component to which an inventor should belong, in order to turn the value of $Member_{it}$ from zero to one. Note, however, that in the case of the minimum size of connected component being set to one, there would be no variation observed in the right-hand side of the regression equation (1), since, by definition, every node in the

[17] I consider 25 possible values of the maximum distance threshold ranging from 1 to 25 km and 25 possible values of the minimum component size threshold ranging from 5 to 200 inventors.

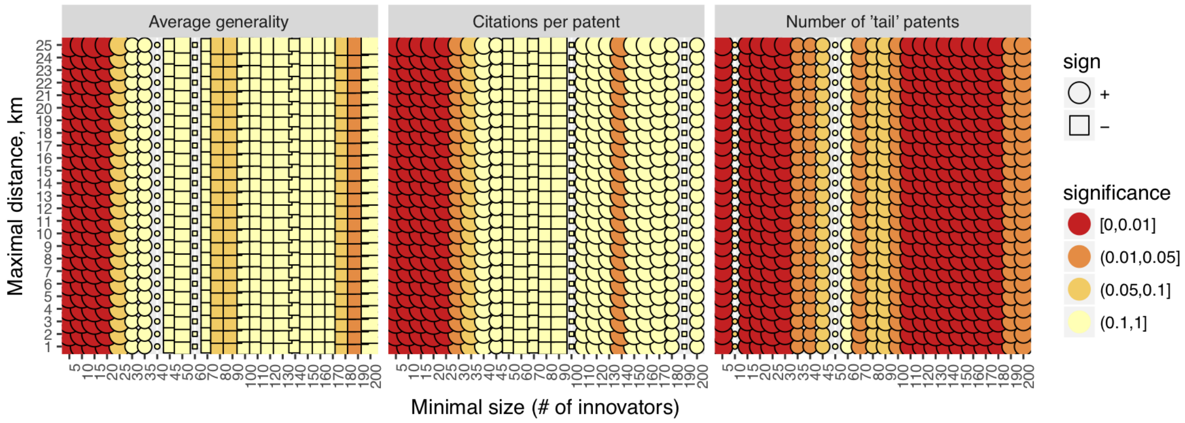
[18] Size of the dot represents a relative (percentile) magnitude of the estimate β as compared to other estimates on the same grid.



(a) Clusters



(b) Geo-components



(c) Network components

Figure 1.4: Results of sensitivity analysis.

network would belong to some connected component of a size equal to or greater than one. Thus for small values of the component size threshold, the within-inventor variation in the $Member_{it}$ can be approximately interpreted as 'switching' from a single-authorship to collaboration or vice versa. The significant and positive βs along the smallest values of the horizontal range commonly observed on all three graphs of panel (c) suggests the presence of a positive relationship between collaborating with other inventors, rather than working alone, and achieving a higher average quality of innovations. It is not, however, enough to consider connectedness to other inventors through collaboration ties as the sole criterion of cluster identification.

For the geographical components (Figure 1.4 (b)), the method of assigning inventors to the densely populated groups according to the distances to their nearest geographical neighbors closely resembles the City Clustering Algorithm in Rozenfeld et al. (2008) and the Network Analytical Approach in Catini et al., (2015). I refer to geo-components as the 'benchmark' clusters and I aim to contrast the proximity-based methodology of cluster identification to the proposed algorithm. In contrast to the grouping method analyzed in panel (c), the proximity-based methodology disregards the structure of interpersonal connections in the collaboration network of inventors. The maximum distance to the nearest neighbor in a geographical component is an important parameter of such a clustering method and the vertical axis of the grid in panel (b) corresponds to its possible values. Increasing the value of this parameter would lead, in the limit case, to the inclusion of all inventors in a single geo-component that would cover the whole area of the country.

Finally, consider the results of sensitivity analysis (Figure 1.4 (a)) for the technology clusters identified with our proposed methodology. The latter combines the clustering criteria of the previous two methods. According to Definition 3, the group of inventors constitute a technology cluster if the distance of each inventor to her nearest geographical neighbor does not exceed the maximum threshold value and if there is at least one sequence of collaborations between any two inventors.

It is apparent from the graphs in panels (a) and (b) that imposing an additional criterion of 'interconnectedness' in the proximity-based clustering algorithm leads to substantial changes in the pattern of association between presence within the cluster borders and delivering a certain level of innovation performance. It is not possible to endogenously determine the optimal values of the distance and size parameters for the proximity-based algorithm as none of the parameters lead to robust evidence of a positive and statistically significant degree of association. At the same time, all three graphs in panel (a) clearly exhibit the following common patterns.

First, as the distance restriction becomes irrelevant (when the maximum threshold is increased above the 15 km value), the positive relationship between 'being present' in a cluster and producing innovations of a higher quality is apparent only for small sizes of the clusters, which mirrors the pattern from the graphs in panel (c) and effectively suggests the dominant role of collaborations as a cluster-identification criterion.

Second, and most importantly, it is easy to identify the range of optimal parameter values that lead to robust evidence of a positive and statistically significant relationship between the variables $Member_{it}$ and $Quality_{it}$, which represent the cluster status and innovation performance of individual inventors. The optimal range includes the values of the distance threshold x between 9 and 12 km and the minimum component size $N_{\min}$ greater than 130 members.

In the following subsection, I characterize the clusters identified in terms of their network structure, geographic locations, and degree of concentration in order to illustrate the interrelationship between spatial concentration and collaborations among cluster members.

1.5.4 Characteristics of identified clusters

The range of optimal parameter values yields a set of 19 distinct clusters. Because a search space (i.e., geographic and collaboration layers of the network) used as an input for the cluster-identification algorithm is not fixed over time (i.e. geographic locations and collaboration linkages among inventors are specified for a given moment of time), the exact geographic and social boundaries of clusters may change over time. In order to characterize the identified clusters in general, not in each moment of time, I focus on the most-recent state of clusters and consider the following measures:

– *Size* defined as the total number of nodes (inventors) in a cluster, which is limited by the minimum component size threshold set at the optimization stage.

– *Diameter* defined as the maximum length of the shortest path (sequence of collaboration ties) between any pair of nodes (inventors) within a cluster. Diameter values can vary in a range from 1 (i.e. each inventor has at least one collaboration tie with any other inventor within a cluster) to $N-1$, where N is a size of cluster (i.e. all inventors are connected in a chain of collaborations).

– *Global clustering coefficient* defined as follows. Consider a triplet (3-set of inventors) where each node is directly (collaboration tie) or indirectly (chain of two collaborations) connected to other two nodes within a triplet, and all nodes in a triplet are members of an identified cluster. The global clustering coefficient is the number of closed triplets (3-sets of inventors having at least one collaboration tie with each other) divided by the number of all triplets (3-sets of inventors with at least two pairs having at least one collaboration tie). A range of possible values is from 0 (no closed triplets) to 1 (all triplets are closed and, thus, each inventor has at least one collaboration tie with any other inventor within a cluster).

– *Spatial spread* measured by the maximal geographical distance between any pair of nodes (inventors) in a cluster. It has a similar interpretation of a diameter, but in a purely spatial context. This measure can vary in a range between 0 (i.e. all inventors being concentrated on a single location) and $(N-1)x$, where x is the maximum distance threshold (i.e., all inventors are located on a straight line and are separated by exactly x km from each other.

Table 1.2 reports summary statistics on five cluster characteristics defined above. The results suggest that clusters identified using the proposed methodology tend to be relatively small; the median size of cluster is only twice the minimum threshold of 130 inventors and the largest cluster is slightly above 500 – the size of a small business.[19] At the same time, the relatively small size of the clusters can be explained by the stringency of conditions that need to be satisfied according to the proposed cluster-identification algorithm: each cluster member has to collaborate on at least one patent with other inventors within a cluster, and the distance to the closest neighbor within a cluster should not exceed a threshold of about 10 km.

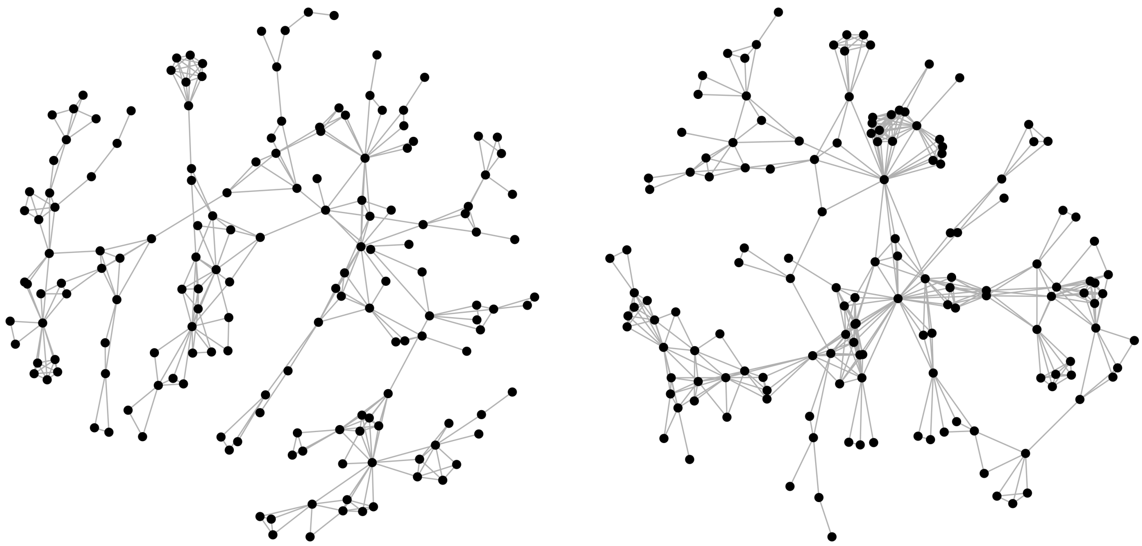
Cluster characteristic	Mean	Median	Min	Max
Size	305	267	135	592
Diameter	16	16	8	32
Global clustering	0.52	0.55	0.35	0.69
Max geo-distance	63.00	60.90	11.25	140.08

Table 1.2: Summary statistics on cluster characteristics

A typical value of the clustering coefficient is relatively large – 0.55, i.e., about half of the collaboration triplets are closed. This means that most inventors within identified clusters tend to form multiple teams, recombining different members around the cluster, rather than repeatedly collaborating with the same coauthors. At the same time, a

[19] According to the U.S. Small Business Administration:http://www.sba.gov/size

relatively large average diameter of clusters can be explained by the existence of several collaboration hubs within clusters, so that teams are typically formed by the members of such hubs (increasing the global clustering coefficient), but only a few inventors form collaborations among the hubs, which makes inventors from different hubs within a cluster relatively distant from each other in terms of collaboration ties (increasing the diameter of a cluster). Figure 1.5 illustrates such a structure of collaboration ties observed in some of the clusters identified.



(a) Cluster in Albany (NY) dominated by General Electric Company

(b) Cluster in Appleton (WI) dominated by Kimberly Clark Worldwide Inc.

Figure 1.5: Examples of collaboration networks within clusters.

Even though clusters are characterized by a relatively large distance between teams in terms of collaboration ties, the geographical spread of inventors within clusters tends to be rather small. The most-distant cluster members are, on average, located 63 km from each other. This can be explained by the maximum distance threshold x imposed by the cluster-identification algorithm and the typical radius of urban agglomerations.

Therefore, a typical cluster is concentrated within an area characterized by a sufficiently high density of population. A higher spatial concentration of inventors within such areas presumably leads to a higher frequency of social interactions and increases the chances for new collaborations. However, it is evident from the characteristics of the clusters identified that connected components in collaboration networks – communities of inventors, where each member is linked with at least one other member of the same community via a collaboration tie – do not necessarily span the entire area of a high spatial concentration of inventors. The latter fact may explain the relatively small sizes of the clusters identified. Large values of a global clustering coefficient and relatively large distances in social networks of clusters show that collaboration ties are likely to be concentrated within small groups that are linked with each other via only few a collaboration ties (Figure 3b).

Therefore, a crucial role in the establishment of a connected network structure within areas of high spatial concentration of inventors, and thus in the formation of innovation clusters, might be played by a small number of so-called 'gatekeepers'[20]. Further extensions of this study may include analysis of the role played by 'gatekeepers' in connecting collaboration hubs into innovation clusters.

In this study, the difference in innovation performance observed within clusters and outside the cluster borders was used as a criterion for optimization of the cluster-identification parameters. Future research based on the results obtained in this study may explore a differential contribution by internal collaborations (among cluster members) and external collaborations (with at least one inventor from outside the cluster borders) to the innovation performance of clusters.

Last but not least, another potential extension of the current study may investigate how the geographical proximity of cluster members and the network structure of collaboration ties among them is related to the technological specialization of clusters.

1.6 Conclusions

It has been widely observed in empirical studies that frequent interactions between innovation firms, organizations or individual inventors are more likely to occur within a relatively short space. The latter naturally leads to a strengthening of agglomeration forces and geographic clustering of innovation activities.

The latter tendency has recently attracted great interest to clusters, among both economic researchers and policymakers. Because of this popularity, the notion of a 'cluster' has been discussed and reinterpreted many times in the literature, though there is still no consensus among researchers on its clear-cut definition, or a commonly-accepted

[20] The term 'gatekeeper', in the context of innovations and collaboration networks, refers to an economic actor (individual or firm) which links a localized knowledge network to outside sources. Graf (2011) discusses the role of 'gatekeepers' in regional innovation systems.

cluster-identification procedure.

This paper builds on the previous work and proposes a novel method of cluster identification. I suggest that boundaries of innovation clusters should be sought simultaneously across two dimensions – geographical space and the structure of collaboration ties among innovators. I propose a novel multi-layer clustering algorithm and exploit it to identify technology clusters in the network of collaboration ties among innovators with geo-coded locations.

Bearing in mind that the ultimate outcome of the cluster-identification algorithm is contingent on the input data, as well as the values of pre-defined parameters, I seek to endogenize the choice of parameter values in the proposed clustering algorithm by searching over the range of possible values and empirically tracing the relationship between the choice of parameter values and the innovation performance observed within identified cluster borders.

I search for the optimal values of the parameters used in the clustering algorithm that lead to the positive and significant association between 'being in a cluster' and delivering a certain level of innovation performance, which is robust to the choice of a performance measure. I find that the sign, magnitude and significance of such association is highly vulnerable to the choice of pre-defined parameter values, as well as the clustering method.

The sensitivity analysis shows that the cluster-identification algorithm proposed in this study performs better than existing methods in identifying successful clusters with a significantly better innovation performance in the terms of several patent measures, and within a consistent range of pre-defined parameters.

Chapter 2

The Impact of Prioritized Examination on Commercialization of Patents

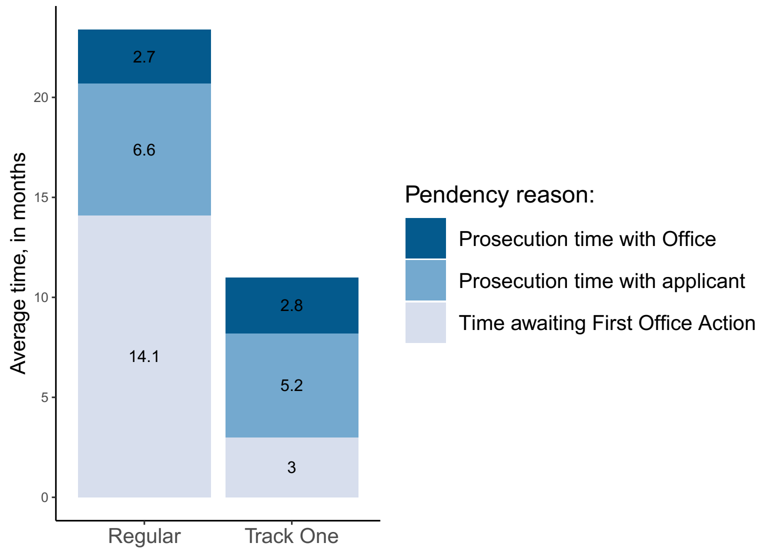
2.1 Introduction

Patents play an important role in facilitating the transfer of knowledge and enabling commercialization of innovative ideas via the market for technology. There are many well-known cases in which patent sales have been accompanied by multi-billion dollar deals between technology giants, such as the $12.5 billion purchase of Motorola Mobility and its 20,000 patents by Google Inc. in 2011. Anecdotal evidence has shown that even small startups and individual inventors can turn their patents into salable commodities in rather short periods. On April 27, 2015, Google Inc. announced its Patent Purchase Promotion, offering to buy patented technology for the price set by the patent owner. The time during which sale offers could be submitted was limited to several weeks. This short-term experimental marketplace was just the beginning of a series of such events[1] supported by many other large companies across different industries. This has proven how quickly patents owned by small startups and individual inventors may be converted into cash flow. The crucial question that remains is how quickly innovative ideas can be patented.

In practice, an inventor cannot immediately obtain patent protection for her invention, since the patent system requires time to process applications before the patent office grants formal property rights to the invention claimed by the applicant. The total time to

[1] IP3 Program (https://www.ast.com/ip3/).

patent, known as the pendency time, is not strictly determined. It may vary substantially from one case to another and its length depends on many different factors that have been examined in the literature (Harhoff & Wagner, 2009, Mejer & Potterie, 2011, Liegsalz & Wagner, 2013, Tong *et al.*, 2018). Currently, the total pendency time averages 2-4 years across the largest patent offices[2]. Although the length of pendency is largely determined by the time of certain actions taken by the patent office and the applicant, more than half of the total pendency may be spent waiting in a queue of unprocessed applications (Figure 2.1).



Source: Data Visualization Center of the USPTO.

Notes: This figure compares the average total pendency time and its determinants for applications under regular and prioritized examinations. Numbers correspond to the most recent statistics as of January 2019.

Figure 2.1: Total Pendency Time

As the demand for patents increases, the patent backlog also increases, due to the numbers of applications waiting for examination by patent offices (Mitra & Kahn, 2013). This trend has attracted increasing public concern over the last decade and is commonly referred to as "global patent warming" (Mejer & Potterie, 2011). One of the negative externalities arising from this trend is a large volume of idled inventions that may be associated with substantial social costs of the delayed benefits from a technological change.

[2]The range corresponds to the total pendency time averages at the USPTO, EPO, JPO, and SIPO.

According to London Economics, the estimated[3] overall harm to the global economy caused by an additional year of pendency for all current applications at the three largest patent offices, the European Patent Office (EPO), the Japan Patent Office (JPO) and the United States Patent and Trademark Office (USPTO), is as large as $9 billion per annum (London Economics, 2010).

While some applicants may intentionally postpone the outcome of the examination process (Henkel & Jell, 2010, Zahringer, Kolympiris & Kalaitzandonakes, 2016) and applicants' preferences related to the length of pendency time are, in general, ambiguous (Rassenfosse & Zaby, 2015), patent offices are generally willing to redistribute their limited examination capacity in favor of welfare generating inventions and to examine these patent applications as a matter of priority. Therefore, so-called 'accelerated examination' is commonly used by patent offices to promote 'green' technologies[4], cancer research, e.g. the "Patents 4 Patients" pilot program at the USPTO, and earthquake disaster recovery support-related innovations (at the JPO) by issuing patent grant decisions for these types of inventions on an accelerated basis.

The privilege of faster examination may be granted only for certain groups of patent applications selected by patent offices, as in the cases above. Such policies, however, do not directly address the issue of patent backlogs, and many other inventions with potentially high value that do not belong to the selected groups may remain idled in pendency. To cope with the patent backlog more generally, as a part of the America Invents Act (AIA) enacted in September 2011, the USPTO introduced Track One Prioritized Examination – an option for applicants to obtain priority in the list of pending applications and, thus, to expedite the examination process, for an extra fee. This intervention was partly motivated by the need to promote innovations produced primarily by small firms in technology sectors with short product life cycles[5].

Wider access to faster examination at the USPTO may have brought many potential benefits to technology markets, including private benefits for start-up firms seeking to obtain a competitive advantage in the R&D race on markets with short product life cycles or those seeking financing in sectors with scarce funding sources (Fischer & Ringler, 2014). The Track One option has also been widely promoted by U.S. patent attorneys encourag-

[3]The cost of lost innovation estimated from the average patent values in the PATVAL survey (Gambardella *et al.*, 2006) and the assumption that the value of a patent is proportionally spread over its lifetime.

[4]See Lu (2013) for an overview of existing policy practices to expedite examination of 'green' patents in different countries.

[5]https://www.uspto.gov/

ing applicants to take advantage of the faster examination at the USPTO. Nevertheless, to our knowledge, the introduction of the USPTO Track One Prioritized Examination and its consequences on the market for technology have not yet been studied in the literature.

In this paper, we raise two empirical questions about the prioritized examination of patent applications. First, we summarize the main statistics related to participation of applicants in the USPTO Track One program and ask whether the program's target group – start-up firms – was effectively reached during the first year of the program. Second, we ask whether participation in the program brought pecuniary benefits to start-ups. More specifically, we verify whether prioritizing an innovation in a queue of pending applications increases its saleability in the market for technology. Hence, we shed some light on frictions in the market for technology potentially created by the pendency time of patent applications.

Our empirical strategy aims to address three main challenges related to the questions posed above: (1) how to measure the saleability of both pending and granted patents; (2) how to avoid confounding due to the fact that, under the limited time coverage of the data, prioritized applications are observed for a longer period of time after a patent is granted than regular applications (3) how to disentangle the effect of participation in prioritized examination on the saleability of a patent, which may be confounded by other observable and unobservable characteristics of the patent.

First, the extensive coverage of the dataset released by the Office of the Chief Economist of the USPTO, which contains patent assignments recorded by the USPTO (Marco *et al.*, 2015), allows us to track the history of the reassignment of property rights for regular patent applications, and for those that have undergone the prioritized examination. We closely follow the refinement procedure employed in Serrano (2010)[6] to select reassignment records that most likely correspond to the sales transactions from the start-up firms to larger corporate entities. The most important feature of the USPTO Patent Assignment Dataset is that it contains virtually all records of both pending and granted patent sales, as it is required that patent sale transactions are filed with the USPTO and, thus publicly recorded, to be legally binding (Dykeman & Kopko, 2004; Serrano, 2011).

Second, in measuring the saleability of patents, we take into account the fact that the limited time coverage of the Patent Assignment Dataset restricts the length of the forward-looking time window starting from the application filing date during which we can track the reassignment history of a pending or granted patent and, thus, conclude whether

[6]Described in detail in Serrano (2008) – the working paper version of the referenced study.

or not it was sold by the start-up firm. We also take into account the empirical observation first documented in Gans, Hsu and Stern (2008) and recently confirmed in Gaessler (2016) that the distribution of the timing of patent commercialization agreements (licensing and reassignment) peaks immediately after the patent allowance. Therefore, granted patents which underwent the prioritized examination would be observed longer during the post-allowance period than granted patents that underwent a regular, longer examination process. In view of this, we first verify whether the limited time window causes underestimation of the saleability of patents in the regular, longer system.

Third, to eliminate potential bias driven by omitted characteristics of patents correlated both with the probability of application for prioritized examination and the probability of commercial reassignment, we compare patent applications filed before and after the Track One inception date with high predicted propensity for prioritization implied from their observable characteristics. Thus, if there are any confounding patent characteristics strongly correlated with the probability of prioritization, they would be equally distributed among the two groups of applications filed before and after the program start date. This conjecture relies on the assumption that there was no evidence of manipulation of the filing date in anticipation of the announced Track One program. As we find strong evidence of strategic filing immediately after the program start date, we eliminate its impact on the main results by excluding applications filed closely around this date. We also eliminate the effects of other factors that may have affected all applications, including those less likely to be prioritized, by estimating a standard difference-in-differences model. Finally, we control for differential non-linear time trends of applications with high and low propensity for prioritization to eliminate the effect of potential pre-program changes in saleability of patents that persisted after the Track One inception date.

We find that shortening the examination time for participants in the Track One program is associated with at least a 1.56 percentage point increase in the probability of commercial reassignment, which is in fact 50% of the average reassignment rate of patent applications filed by the start-ups. This suggests that the USPTO Track One Prioritized Examination, in addition to its obvious advantage of providing an earlier disposition of patent applications, may have resulted in private pecuniary benefits to applicants who opted for prioritization, and that a large overall benefit to the market for technology remains unrealized due to low participation in the program.

This study contributes to the literature on the commercialization of innovations via the market for technology in two respects. First, in the data on transfers of formal property rights across firm boundaries, we find evidence for the sales of patents, in line with Serrano (2008, 2011), Galasso *et al.* (2013) and Gaessler (2016). Second, we find that, other things being equal, patent applications that undergo prioritized examination are significantly more likely to be commercialized via the market for technology. This finding suggests that longer pendency time of applications at patent offices may not only lead to a welfare loss due to the deferred commercialization of innovations (Gans, Hsu & Stern, 2008), but may also create frictions on the market for technology that reduce the overall saleability of granted and pending patents. We thus contribute to evidence from other studies (Galasso *et al.*, 2013, Harhoff & Stoll, 2015, Hegde & Luo, 2018) of frictions in the market for technology, and analyze different sources of those frictions.

2.2 Institutional Context and Hypotheses

2.2.1 The America Invents Act and Track One Prioritized Examination Program

At the USPTO, patent examination, the process that precedes the issuance of a patent, typically takes about two years. A final disposition – allowance or final rejection – for a patent application is reached, on average, within 24 months of the filing date. About two-thirds of that time is spent awaiting the first office action – the start of communication between the patent office and the applicant on the merits of the application (Figure 2.1). Thus, a major part of pendency time at the USPTO is spent waiting in a line of other filed applications awaiting examination.

While earlier approval of a patent application may hasten commercialization of innovation via the market for technology (Gans, Hsu & Stern, 2008; Gaessler, 2016), it is not common for all applicants to seek a shorter examination time. Some inventors, for example, may need more time to secure investment or generate revenue needed to convert the invention into a marketable product, and therefore they may tolerate or even prefer a deferred patent grant. Moreover, owners of pending patent applications may be able to realize up to three-forths of the returns that would be generated under full patent protection by strategically creating uncertainty for competitors about the patentability of an innovation and the risk of infringement of a future patent (Harhoff, Rudyk & Stoll, 2016).

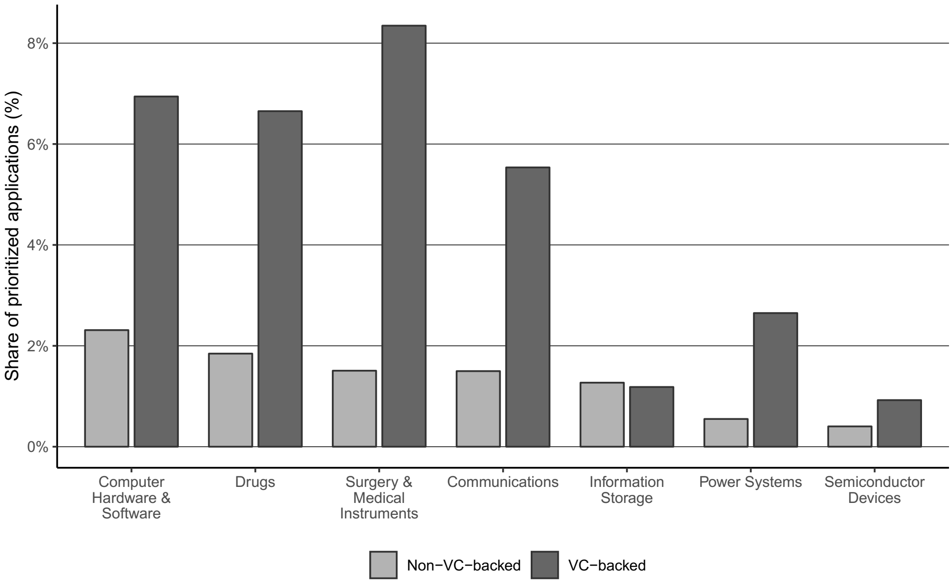
To allow inventors to shorten waiting times and release the potential value of innovations trapped in a backlog of pending applications, the USPTO introduced the Track One Prioritized Examination program as a part of the America Invents Act (AIA) enacted in September 2011. Under this program, up to 10,000 nonprovisional utility patent applications filed each year, beginning on September 26, 2011 can obtain prioritized status in the examination process. This option is offered for a fee that ranges from $1,000 for a micro-entity to $4,000 for a large entity, which compares to the minimal overall cost of all stages from an application filing to an issued patent[7] ranging from $715 to $2,860 depending on the applicant's status.

In exchange for the prioritization fee, applicants in the Track One Prioritized Examination are effectively allowed to obtain a final disposition for their applications in about half the time of the regular examination (Figure 2.1). Since the usual 20-year term of a patent starts from the filing date of a application, earlier final disposition and a patent grant imply a longer enforcement time, during which a patent can be enforced by its owner against potential infringers. Moreover, in some cases, a pending application may severely delay product market entry when legal protection of the technology is crucial for a producer to secure itself against unauthorized infringement. In many other cases, such an aggressive R&D race between large corporations or participation in limited-time marketplaces for innovations, such as the Patent Purchase Promotion[8], faster examination of patent applications can create a substantial competitive advantage for innovators. Shorter pendency time can also reduce transaction costs faced by start-up innovators commercializing their technologies via licensing contracts (Gans, Hsu & Stern, 2008). Further, start-up firms seeking external funding with a lack of tangible assets to secure the loan may benefit from earlier issue of a patent, since the patent can be immediately provided to a lender as an alternative form of collateral (Fischer & Ringler, 2014).

Introduction of the Track One program was initially motivated by the USPTO as a promotion mechanism for innovations produced by small firms in technology sectors with short lifecycles and high speed R&D races[9]. Since the program's inception, the advantages of faster examination at the USPTO have been widely promoted by patent attorneys around the U.S. (Whitt, 2015; O'Brien, 2017). Surprisingly low demand, however, is evidenced by the fact that the limit of 10,000 prioritization requests per fiscal year has never been achieved (Merchant, 2015). The overall participation rate in Track

[7]A sum of filing, search, examination and issue fees.
[8]IP3 Program (https://www.ast.com/ip3/)
[9]http://www.uspto.gov/

One Prioritized Examination during its first year averaged only 1.2% of all eligible utility patent applications filed at the USPTO. We find that the participation rate was notably higher among VC-backed start-ups, averaging 4.8% and peaking in the following technology sectors: Computer Hardware & Software, Communications, Surgery & Medical Instruments and Drugs, with the shares of prioritized applications ranging between 6-8% (Figure 2.2). To see how the group of innovators targeted by the USPTO and characterized by more active participation was affected by the Track One program, we focus on the applications initially owned by the VC-backed start-up firms.



Notes: This figure compares the shares of applications in our sample that were prioritized, across technology sectors defined in Hall, Jaffe & Trajtenberg (2001) and between two groups of applicants: VC-backed start-ups and all others.

Figure 2.2: Track One Participation Rates

2.2.2 Empirical hypothesis

In our empirical analysis, we study the implications of faster examination for the commercialization of innovations made by VC-backed start-up firms via the market for technology. We ask how the probability of commercial reassignment of a granted or pending patent from a VC-backed start-up to a large corporation is affected by the length of pendency time at the patent office.

There are several potential ways prioritized status of a patent application can affect its saleability on the market for technology. First, in sectors characterized by a short product life cycle and incremental innovations, new technologies developed by innovative start-ups may quickly become obsolete and lose the interest of potential buyers – practicing firms willing to acquire patents for production or strategic use. Thus, other things being equal, faster examination would naturally increase the probability of commercial reassignment of a given patent. Second, when the marketplace for technology is limited in time[8], the possibility to expedite examination of a pending application may become a deciding factor in a competition among sellers. Last but not least, the innovator may convey an informative signal about the intrinsic value of a pending or granted patent to its potential buyers by filing a prioritized application (Harhoff & Stoll, 2015).

Thus, we believe that introduction of the prioritized examination track at the USPTO may have affected the market for technology by reducing frictions between buyers – start-up innovators – and sellers – large corporations – and by increasing the saleability of patents undergoing prioritized examination. We exploit a variation in the length of patent examination generated by the introduction of the USPTO Track One Prioritized Examination to test our hypothesis of the existence of a difference in the saleability of granted and pending patents that undergo regular versus prioritized examination.

2.3 Data

We consider patent applications filed at the USPTO within one year before and after the effective inception date of the Track One program, that is between September 26, 2010, and September 26, 2012. These include 803,621 applications. We merge several data sources to find the necessary details about the application characteristics, their prosecution, and their reassignment history. First, we use the USPTO Patent Examination Research Dataset (PatEx), including technology class, number of inventors, small entity status of the applicant and a detailed transactions history between the applicant and the patent office, including the filing date, notice of allowance date, and the date of granting the prioritized status. Second, we use the USPTO Patent Assignment Dataset to track the reassignment history of each patent application in a sample, and identify its initial owner and the first assignee involved in a commercial transaction with the initial owner. Third, we retrieve the names of VC-backed firms and the dates of their funding rounds from the VentureXpert database, and match them with the names of the initial owners of the patent applications. Finally, we obtain the subset of 15,458 applications filed within one year before and after the program inception date and initially owned by the VC-backed start-ups; that is, VC-backed firms that had their first round of funding no later than five years before September 26, 2010 – the beginning of the time window. Additionally, we use the OECD Triadic Patent Families database and the PATSTAT data to obtain the application characteristics that are not available in the PatEx database, including patent family size and the triadic status of the patent.

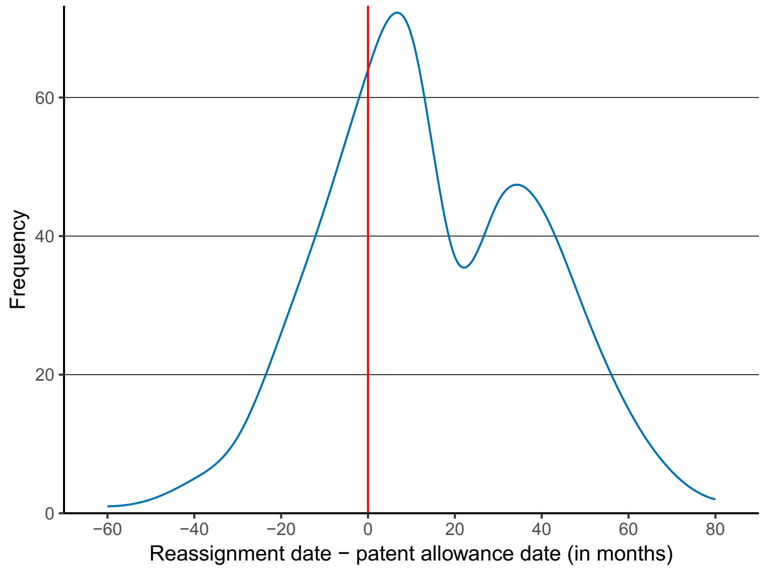
To construct the outcome variable – saleability of patent applications – measured by the probability of commercial reassignment, we use the Patent Assignment Dataset[10] released by the Office of the Chief Economist of the USPTO. This dataset contains patent assignments – transactions in patents executed by an interested party prior to or after a patent is granted and recorded by the USPTO (Marco *et al.*, 2015). Though, in general, the disclosure of patent assignments to the USPTO is not mandatory, to be legally binding, it is necessary for patent sale transactions to be filed with the USPTO and publicly recorded. The latter condition implies a no or negligibly small selection issue in subsequent empirical analysis (Dykeman & Kopko, 2004; Serrano, 2011). Most of the recorded transactions, however, are not associated with a genuine transfer of property rights across firm boundaries and, thus, with the sale of a patent (Serrano, 2008, 2010; Galasso et al., 2013; Gaessler, 2016). There are several types of records, including individual inventor assignments, security agreement assignments, name change records, patent assignments associated with mergers and acquisitions, and transactions between subsidiaries and patent companies, that are usually disregarded in analyses of commercial reassignments. We closely follow the refinement procedure employed in Serrano (2010)[11] to filter out non-relevant transactions and to select reassignment records that most likely correspond to the sales transactions between the VC-backed start-up firms and the large corporations. We construct a dichotomous variable which is equal to one if a given patent application is reassigned from its initial owner – a VC-backed start-up firm – to another corporate entity. A more detailed description of the data construction and refinement procedures is presented in Appendices C, D, and E.

When identifying whether a given patent application has been reassigned or not, we take into account the fact that the time coverage of the Patent Assignment Dataset is limited by its most recent update, and applications with earlier filing dates in a sample are observed in the Assignment Dataset for a longer time than applications with more recent

[10]http://www.uspto.gov/economics

[11]Described in detail in Serrano (2008) – the working paper version of the referenced study.

filing dates. We thus consider a fixed five-year forward-looking time window starting from the application filing date, during which we track the reassignment history of all applications in a sample.



Notes: We plot the distribution of the reassignment lag (the difference between the reassignment and patent allowance dates) to confirm that the probability of commercialization (conditional on patent allowance) peaks right after the date of patent allowance, which was previously documented in Gans, Hsu and Stern (2008) and Gaessler (2016).

Figure 2.3: Distribution of Difference Between Reassignment and Patent Allowance Dates

We also take into account the fact that the probability of commercial reassignment is not evenly distributed along a patent's life, and peaks right after its allowance date (Figure 2.3; Gans, Hsu & Stern, 2008; Gaessler, 2016). Therefore, given a limited five-year forward-looking time window, patents that undergo prioritized examination and which are, on average, allowed within twelve months from the filing date, would have a longer exposure time to a higher probability of reassignment than the non-prioritized patents that are allowed within twenty-four months on average. To verify whether a five-year time window causes underestimation of the reassignment rates of regular versus prioritized patents, we calculate reassignment rates of patents filed in 2006 and granted within twelve months of the filing date, considering a restricted five-year time window and a counterfactual time window of a maximum of ten years. The former reassignment rate is 36% lower than the latter due to the truncated distribution of timing of reassignments. It turns out, however, that this discrepancy is not much different for patents filed in the same year and granted within twelve months of the filing date. For the latter patents, reassignment rates within the five-year time window is 38% lower than the time window of a maximum of ten years. We thus conclude that the restricted time window does not cause any implicit differences in the reassignment rates of regular and prioritized applications.

2.4 Empirical results

2.4.1 Treated vs untreated applications

Only applications filed on or after September 26, 2011 – the effective date of the policy change – were eligible for a prioritized examination request. However, since the requests were initiated by the applicants, not all applications filed after the policy change were actually treated. In fact, fewer than 5% of all eligible applications[12] had prioritized status (Table 2.1). Thus, it is not possible to directly utilize a discontinuity in the length of examination time of all applications filed around the policy change to estimate the effect of shortened examination time on commercialization.

Another way of assessing the degree of association between shortened examination time and the saleability of patent applications – a simple comparison of means (Tables 2.1 and 2.6) – shows that eligible applications (filed after the program inception date) which underwent the prioritized examination had a 1.56 p.p. higher reassignment rate, thus, were on average 50% more frequently commercialized by VC-backed start-ups than applications that underwent the regular examination. It is, however, also evident from the summary statistics (Table 2.1) that the two groups of applications – prioritized and non-prioritized – differ in terms of their observable characteristics. If a clustering of some of these characteristics in either group affected both the probability of commercial reassignment and treatment status, they cannot be directly compared in terms of their average saleability.

[12]We assume that only the filing date of an application is a relevant eligibility criterium. Even though other eligibility criteria were set by the USPTO, such as the maximum total number of claims and the maximum number of independent claims, those additional criteria were satisfied by more than 80% of all applications in our sample.

	Means (SD)				
	Whole sample	Period		Track One status	
		before Sep 26	after Sep 26	untreated	treated
	(1)	(2)	(3)	(4)	(5)
Outcome variables					
Reassignment rate	0.031 (0.175)	0.032 (0.175)	0.031 (0.174)	0.031 (0.172)	0.048 (0.215)
Track One rate	–	–	0.047 (0.213)	–	–
Applicant characteristics					
Small entity status	0.44 (0.50)	0.44 (0.50)	0.44 (0.50)	0.43 (0.50)	0.53 (0.50)
Age of the firm	1.78 (2.31)	1.81 (2.26)	1.75 (2.34)	1.78 (2.34)	1.22 (2.39)
Application characteristics					
Top-tier law firm	0.21 (0.41)	0.21 (0.41)	0.21 (0.41)	0.21 (0.41)	0.31 (0.46)
Triadic status	0.12 (0.32)	0.13 (0.33)	0.12 (0.32)	0.12 (0.32)	0.09 (0.29)
Patent family size	4.81 (8.39)	5.06 (8.37)	4.59 (8.40)	4.45 (8.22)	7.39 (11.05)
Number of inventors	2.92 (1.86)	2.92 (1.93)	2.93 (1.79)	2.92 (1.77)	3.22 (2.06)
Allowance lag (months)	26.4 (12.05)	28.53 (12.9)	24.59 (10.96)	25.13 (10.66)	14.57 (11.53)
Allowance rate	0.68 (0.47)	0.68 (0.47)	0.69 (0.46)	0.68 (0.47)	0.75 (0.43)

Notes: This table reports the summary statistics (means and standard deviations – in parentheses) of the outcome variable in the difference-in-differences model and predictor variables used for predicting the treatment status, separately for untreated applications filed before the Track One program start date, and treated and untreated applications filed after the program start date.

Table 2.1: Summary statistics

In fact, as the results show (Table 2.3), there are patent application and firm characteristics that are correlated both with the outcome variable (column 1) and the treatment status (column 2) of patent applications in our sample. In particular, applications prepared and prosecuted by top-tier patent attorneys are more likely to be examined under the prioritized Track One procedure and are also more likely to be commercialized within five years of the filing date. It is also evident that the number of applications in the patent family and the number of inventors listed in the application increase both the probability of prioritized examination and of reassignment.

Finally, we show that characteristics of patent applications predict both, probability of reassignment and prioritized examination by plotting fitted values of the linear probability models presented in columns 1 and 2 of Table 2.3. It is clear from Figure 2.4 that there is a significant positive correlation between probability of reassignment and propensity for prioritization, both predicted using the same set of application and firm characteristics. Thus, due to these characteristics the effect of prioritized examination on the probability of commercial reassignment is confounded and the estimate in a simple comparison of

	Reassignment ($mean = 0.031$)
Intercept	0.0309***
	(0.0020)
TrackOne	0.0181**
	(0.0091)
Observations	8,151
F-statistic	3.916

Notes: Simple comparison of means of the outcome variable computed for the treated and untreated patent applications filed after the Track One program start date suggests a higher reassignment probability of prioritized applications as compared to the applications that underwent the regular examination process.

Standard errors are in parentheses; $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

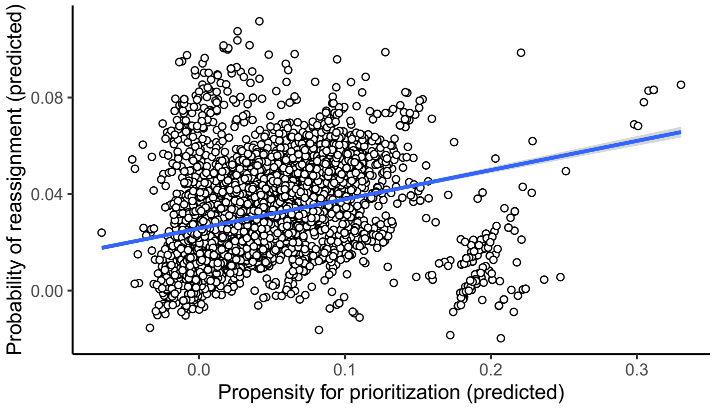
Table 2.2: Difference in means: treated vs. untreated

means presented earlier in Table 2.6 is biased.

To address the issue of bias caused by non-random assignment of applications into the treated (prioritized) and untreated (non-prioritized) group, we compare applications with high predicted propensity for treatment filed before and after the policy change. This allows us to match applications with similar observable characteristics that predict both the probability of commercial reassignment and the propensity for treatment (Table 2.3). Thus, if we observe a difference in the probability of commercial reassignment between applications with high predicted propensity for prioritization filed before the introduction of Track One Program and similar applications filed after the policy change, this difference cannot be attributed to observable application and firm characteristics. We also seek to eliminate the effect of other factors that may have affected all applications, including those that were less likely to be prioritized, by estimating a standard difference-in-differences model.

2.4.2 Difference-in-differences framework

In this setup, we do not distinguish patent applications based on their actual treatment status. Instead, we define several levels of treatment intensity inferred from the observable application and firm characteristics. Using the difference-in-differences framework, we compare the difference in the probability of commercial reassignment before and after



Notes: This figure plots the fitted values of regressions presented in Table 2.3 – predicted probability of commercial reassignment and predicted propensity for prioritization of applications filed after the policy change.

Figure 2.4: Predicted Probability of Reassignment vs Predicted Propensity for Prioritization

the policy change for applications with high and low propensity for prioritization.

Under the null hypothesis of no effect of prioritized examination, the difference in reassignment probability between high and low propensity groups would be the same around the date of the policy change. Alternatively, if the difference in probability of reassignment between high and low groups increased after the implementation of Track One Prioritized Examination, this may suggest the presence of an effect of a prioritized examination on the probability of commercial reassignment.

Propensity for prioritization

To predict the propensity for prioritization of applications filed before and after the policy change, we start with the linear probability model of the form:

$$TrackOne = Z\beta + \varepsilon \tag{1}$$

where $TrackOne$ is a binary outcome variable that takes a value of one if an application was examined on the prioritized track, and Z denotes the matrix of observable applicant and application characteristics listed in Table 2.1. Based on the subset of 8,285 patent applications filed after the policy change, we estimate equation (1). Coefficient estimates, their standard errors, and significance levels are reported in column 2 of Table 2.3.

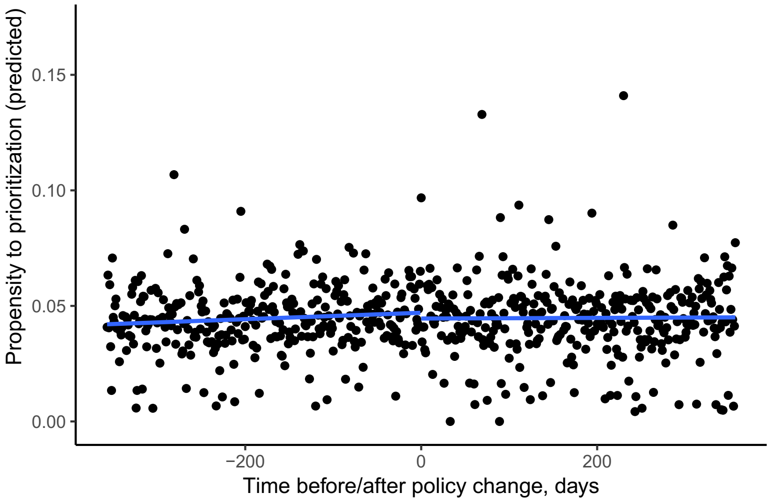
	Linear probability model	
	Reassignment (*mean* = 0.031)	Track One status (*mean* = 0.047)
Applicant characteristics		
Small entity status	−0.0079**	0.0128**
	(0.0038)	(0.0062)
Age of the firm	−0.004***	−0.0011
	(0.0010)	(0.0016)
Small × Age of the firm	0.0044***	−0.0024
	(0.0013)	(0.0021)
Application characteristics		
Top-tier law firm	0.0151***	0.0217***
	(0.0035)	(0.0058)
Triadic status	−0.0072	−0.0254***
	(0.0046)	(0.0077)
Patent family size	0.0003*	0.0019***
	(0.0002)	(0.0003)
Number of inventors	0.0026***	0.0034**
	(0.0008)	(0.0013)
Technological sector dummies	Yes	Yes
Observations	15,458	8,285
R^2	0.008	0.042

Notes: This table reports the estimated coefficients (LPM) of the variables predicting commercial reassignment and treatment – prioritized status of patent applications. Standard errors are in parentheses; $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 2.3: Reassignment probability and propensity for prioritization

Most characteristics are significant predictors of prioritization status. Notably, if the application is prosecuted by a top-tier patent attorney, its probability of prioritization increases by 2.17 p.p. Filing an application at multiple jurisdictions in different countries, thus increasing its family size and broadening the market scope (Harhoff et al., 2003), increases the probability of prioritization. However, applications that are part of a triadic patent family undergo prioritized examination less frequently. Applications authored by larger teams of inventors are more likely to be prioritized. Small entity status increases the probability of prioritization by 1.28 p.p. We also find that prioritization of applications is requested more frequently by smaller and younger firms, though the magnitude of the relationship is not statistically different from zero.

To find similar applications in terms of their propensity for prioritization filed before and after the policy change date, we estimate model (1) with a large set of controls and their interaction terms refined using the Lasso method[13]. We use a random subset from our sample of applications filed after the policy change to train the model. Then, we use the remaining applications to validate the model[14].



Notes: This figure plots the predicted propensity for prioritization over time. Each dot corresponds to the average propensity score of applications filed within a one-week interval.

Figure 2.5: Predicted Propensity for Prioritization

We use the Lasso model to make an out-of-sample prediction of the prioritization status of all 15,458 applications filed within one year around September 26, 2011, and initially owned by the VC-backed start-ups. Outliers in the top and bottom 1% of the overall distribution of predicted values were excluded, resulting in 15,150 observations used in the subsequent analysis (Figure 2.5).

Based on the predicted values of propensity for prioritization, we assign applications into two groups with different levels of treatment intensity. Applications with predicted propensity for prioritization above the optimal cut point are assigned to the high-level group, while the remaining applications are assigned to the low-level group[15].

[13]We chose the Lasso (least absolute shrinkage and selection operator) method to select the most significant controls and improve the prediction power of the model.
[14]More details about the estimation process can be found in Appendix 2.A.4.
[15]More details about the optimal cut point can be found in Appendix 2.A.4.

Assumptions

Our empirical strategy of performing a difference-in-differences comparison of applications with similarly high or low predicted propensity for commercialization relies on several assumptions. We state them and discuss their validity in this section.

First, we assume that applications filed before the policy change with relatively higher predicted propensity for prioritization would have been exposed to the policy change to a larger extent if it had been implemented earlier. The validity of this assumption relies on the choice of predictors used to predict the propensity for prioritization. We test for this assumption by running a simple OLS of actual treatment status (prioritized vs. non-prioritized) of 8,135 applications filed after the policy change on their predicted propensity for prioritization (Table 2.4). In fact, 10.8% – the highest participation rate – corresponds to applications with high predicted propensity; significantly larger than in the case of applications with low predicted propensity.

	Track One ($mean = 0.047$)
$Intercept$	0.0230***
	(0.0027)
$High$	0.0846***
	(0.0051)
Observations	8,135
F-statistic	272.9

Notes: We regress the treatment variable – prioritization dummy – on a dummy for the high treatment intensity level. The results reported in this table show that actual prioritization rate is significantly higher in the group of applications with high (above the optimal cut point) as opposed to low predicted propensity for prioritization. Standard errors are in parentheses; $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 2.4: Participation rates across treatment intensity groups

Second, we implicitly assume that observed characteristics used to predict the propensity for prioritization to some extent confound the effect of a prioritized examination on reassignment probability. Thus, we expect significant discrepancies between the reassignment rates at different levels of the treatment intensity, if our assumption is valid. We test for the presence of the latter discrepancies by running a simple OLS of the reassignment probability on a $High$ group dummy along with an intercept corresponding to the base

	Reassignment ($mean = 0.031$)
$Intercept$	0.0285***
	(0.0017)
$High$	0.0112***
	(0.0032)
Observations	15,150
F-statistic	12.46

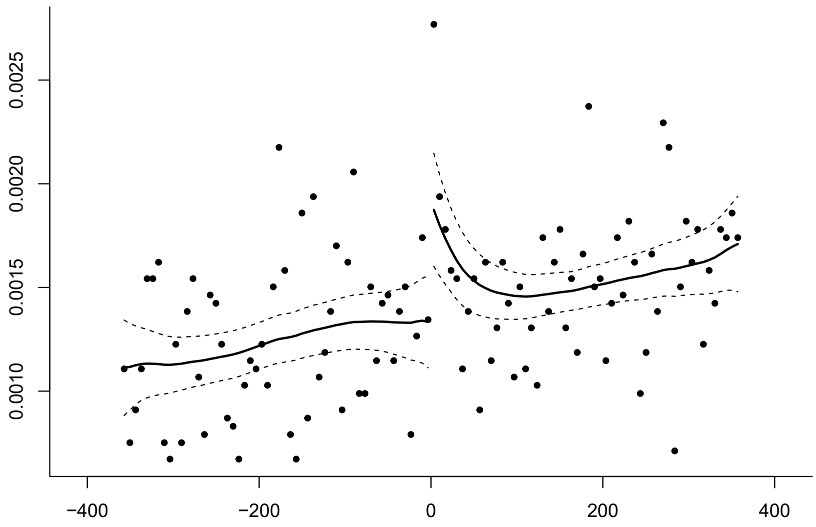
Notes: We regress the outcome variable – reassignment probability – on a set of dummies for the treatment intensity levels. The results reported in this table show that, in the absence of treatment, higher predicted propensity for prioritization is associated with higher reassignment probability. Thus, the effect of prioritization on the outcome variable may be confounded by other characteristics of patent applications that are associated both with the treatment and the outcome variables.
Standard errors are in parentheses; $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 2.5: Reassignment rates across treatment intensity groups

group, using all 15,150 observations in the sample (Table 2.5). The average reassignment rate in the high-level group is in fact significantly higher than in the base group.

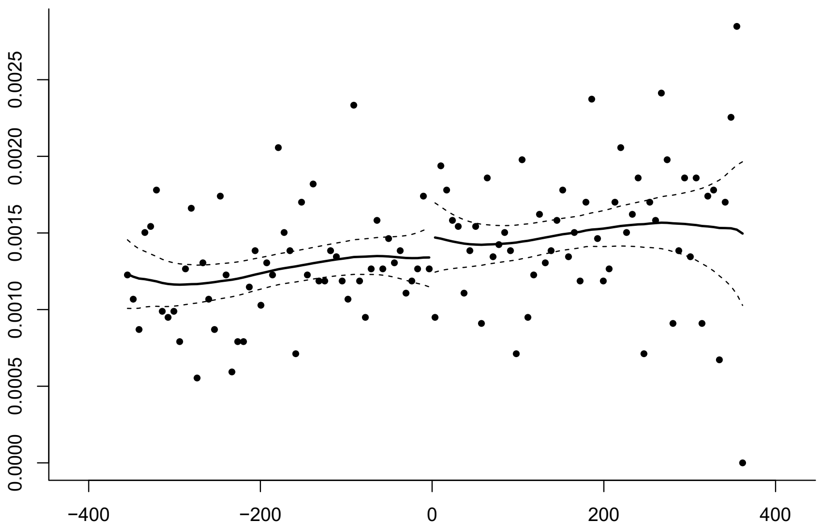
Third, we assume that confounding patent characteristics strongly correlated with the probability of prioritization are equally distributed among applications around the program start date, making applications that were filed before the inception date comparable counterfactuals of those that were filed after. To explore evidence of manipulation – strategical postponement – of the filing date by the applicants in anticipation of the announced Track One program, we test for discontinuity in a distribution of filing dates of applications with a high propensity for prioritization at the program start date. The "manipulation test" (McCrary, 2008) clearly rejects the null hypothesis of no discontinuity at a 99% significance level (Figure 2.6). To avoid the potential impact of this filing pattern on our results, we exclude from our sample applications filed closely around the program start date. The result of the test confirms the absence of any statistically significant evidence of manipulation after excluding applications filed within one week around the program start date (Figure 2.7), which results in a sample of 14,625 applications used in the difference-in-differences analysis.

Last but not least, we assume that changes in the reassignment probability in high and low groups over time before the policy change followed parallel paths and were on the same increasing, decreasing or constant trajectories at the time of the policy change.



Notes: This figure plots the distribution of filings with high propensity for prioritization around the Track One program start date. "Manipulation test" rejects the null of no discontinuity at the program start date.

Figure 2.6: Distribution of filings in a high treatment intensity group



Notes: Excluding applications filed within one week around the Track One program start date mitigates the issue of manipulation of the filing date in anticipation of a new policy. A repeated "manipulation test" does not reject the null of no discontinuity at the program start date.

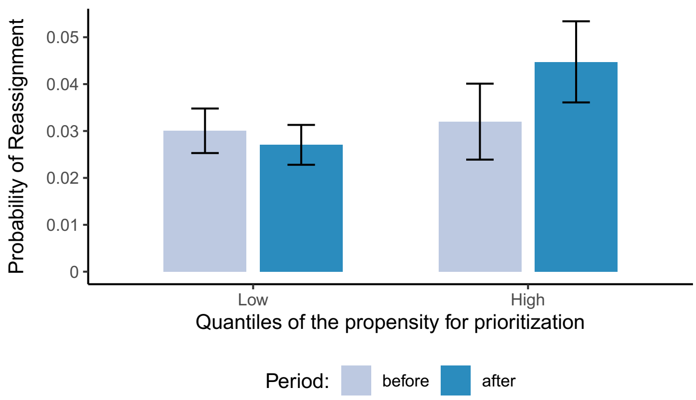
Figure 2.7: Excluding one week around the program start date

Results

We start with a comparison of the means of an outcome variable – reassignment probability – across four cells constructed on the interaction of two treatment intensity groups – high and low – and two time intervals relative to the policy change date – before and after. The results (Figure 2.8) suggest that the high treatment intensity group exhibiting about the average level of reassignment probability (within one year) before the policy change, experienced a significant increase in the outcome (within one year) after the policy change, whereas the low treatment intensity group seems unaffected by the introduction of Track One Prioritized Examination. We formally test for the latter finding by estimating a difference-in-differences model:

$$\begin{aligned} Reassign = \beta_0 + \beta_1 After + \beta_2 High + \\ + \beta_3 After \times High + \varepsilon \end{aligned} \tag{2}$$

where $Reassign$ is a binary outcome variable that takes a value of one if an application was reassigned from its first assignee (VC-backed startup) to a corporate entity within five years of its filing date, the $After$ dummy takes a value of one if the filing date is on or after September 26, 2011, and the $High$ dummy corresponds to the high treatment intensity group defined above. The coefficient of the interaction term, β_3, is a difference-in-differences estimator and constitutes the focus of interest in our analysis.



Notes: This figure plots the average probability of reassignment in each treatment intensity group. The results suggest that a higher predicted propensity for prioritization is associated with a larger positive difference in the reassignment probability before and after the policy implementation.

Figure 2.8: Comparison before and after

The results (Table 2.6) suggest that the introduction of the Track One Prioritized Examination by the USPTO on September 26, 2011, did not change the saleability of patent applications with low propensity for prioritization. At the same time, compared to the change in the probability of commercial reassignment of patent applications with low propensity for prioritization before and after the policy change, patent applications that were more likely to be prioritized were also, by 1.57 p.p. (50% of the mean reassignment rate), more frequently reassigned once the Track One Prioritized Examination was introduced.

	Reassignment ($mean = 0.031$)
$After$	-0.003
	(0.0034)
$High$	0.0020
	(0.0048)
$After \times High$	0.0157^{**}
	(0.0065)
Observations	14,625
F-statistic	5.567

Notes: The baseline results of our standard difference-in-differences model are reported in this table. We compare applications with high propensity for prioritization with the base group – low-propensity applications – before and after the Track One program start date. A significantly higher reassignment rate after the policy change is observed only among applications with high propensity for prioritization. Standard errors are in parentheses; $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 2.6: Difference-in-differences (baseline results)

To visualize the assumption of parallel paths implicitly made in the model (1), we allow for differential changes over time in pre- and post-treatment periods for both treatment intensity groups. Thus, we consider a linear model with time fixed effects and interactions between the $High$ group dummy with time period dummies:

$$Reassign = \sum_t \beta_{H,t} T_t \times High + \sum_t \beta_t T_t + \varepsilon \tag{3}$$

where the $High$ dummy corresponds to the high-level treatment intensity group and T_t are time period dummies corresponding to four 90-day lags preceding the policy change indexed as $-4, -3, -2, -1$ and four 90-day leads following the policy change indexed as $1, 2, 3, 4$.

Estimates of the coefficients and their significance levels are reported in Table 2.7. In Figure 2.9, we plot the values of the coefficients $\beta_{H,t}$ which correspond to the departures of the outcomes of high-propensity group from the base – low-propensity group in each period. Coefficients $\beta_{H,t}$ in the pre-treatment periods were not significantly different from zero, which confirms the validity of our assumption about the parallel trends posed in the previous section. At the same time, coefficients $\beta_{H,t}$ in the post-treatment periods show that the probability of commercial reassignment of the high-propensity group with the largest share of applications which were actually prioritized visibly increased and remained significantly higher than the probability of commercial reassignment of the low-propensity group, which was unlikely to be exposed to the effect of the Track One Prioritized Examination.

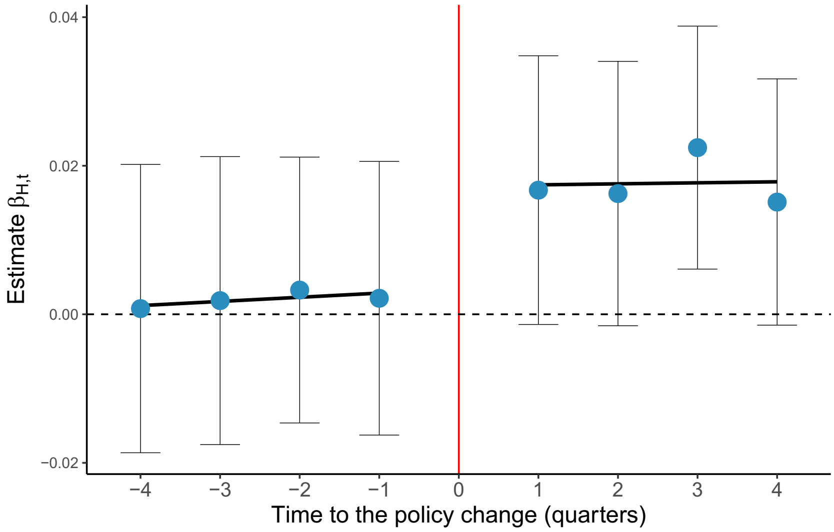
	Reassignment ($mean = 0.031$)		
$High_{-4}$	0.0008	$High_1$	0.0167*
	(0.0099)		(0.0092)
$High_{-3}$	0.0018	$High_2$	0.0163*
	(0.0099)		(0.0091)
$High_{-2}$	0.0033	$High_3$	0.0224***
	(0.0091)		(0.0083)
$High_{-1}$	0.0022	$High_4$	0.0151*
	(0.0094)		(0.0085)
Observations	14,625		
F-statistic	31.25		

Notes: This table reports the estimation results of the extended version of the baseline difference-in-differences model that allows for a flexible specification of the time trends.
Standard errors are in parentheses; $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 2.7: Difference-in-differences (flexible model)

Finally, our extended difference-in-differences model suggests that applications which underwent prioritized examination may have about 1.56 p.p. (the difference between the averages of the $\beta_{H,t}$ coefficients in the pre- and post-treatment periods) higher reassignment rates, that is 50% of the average reassignment rate of applications filed by the

VC-backed start-up firms.



Notes: This figure depicts the statistical significance of the upward trend in the high-propensity group after the Track One program start date. Error bars correspond to the 95% CI.

Figure 2.9: Difference-in-differences (flexible model)

2.4.3 Discussion of results

Our empirical results suggest that saleability of pending and granted patents is strongly associated with the length of pendency time at the patent office. Thus, a growing patent backlog and consequently longer time periods before the First Office Action, which already accounts for half of the total pendency time at the USPTO, may not only delay commercialization of innovations, but it may also create frictions on the market for technology and deteriorate potential economic gains from innovation.

This leads to several important implications of the Track One Prioritized Examination Program introduced by the USPTO. Firstly, the introduction of the fast examination track affected the market for technology by facilitating commercial agreements between start-up innovators who opted for prioritization of their patent applications and large corporations.

There are several possible mechanisms that can drive the relationship between the prioritized status of patent application and its attractiveness to investors. First, investors may perceive an intention of inventors to move their patent applications forward in line of pending applications as a signal of greater economic value. Second, investors may particularly value the longer patent enforcement period implied by an earlier patent grant. Third, an earlier litigation procedure against patent infringements can be anticipated by large corporations buying prioritized patents. Fourth, potential buyers can pursue a competitive advantage through earlier entry to the product market with a technology patented in a shorter time. Last but not least, companies may anticipate rapid obsolescence of certain technologies and, thus, will be interested in obtaining patents for such technologies in a shorter time frame.

Finally, we conjecture that weak demand for prioritization, potentially caused by a vague awareness of inventors about the Track One Program, implies that the large potential benefits of the program to the market for technology, that have not been anticipated by the USPTO, remain unrealized.

2.5 Conclusion

Previous research has shown that timely granting of patents plays a crucial role in the commercialization of innovations via the market for technology (Gans, Hsu & Stern, 2008). Even though in some cases innovators may choose to strategically postpone the outcome of a patent office examination, it is widely claimed that longer application pendency has a detrimental impact on the social value of innovations. Therefore, patent offices around the world have been implementing various policy programs targeted at accelerating the examination process of innovations with the highest social value.

Starting September 26, 2011, the USPTO offered its applicants the option to choose a faster examination track – Track One. As evidenced from the data, this option allows applicants to obtain a final disposition in half the usual time. Several empirical observations made in this study concern the participation activity of applicants in the Track One program. First, we find that the group of applicants – small start-ups – which were primarily targeted by the USPTO indeed participated much more actively than others. Second, we find that dissemination of information about the benefits of the Track One program by patent attorneys may have had a persuasive impact on innovators' decisions to apply for the prioritized examination. Third, despite the overall low demand for the prioritized examination documented by the USPTO, we find evidence of bunching of strategically postponed filings right after the program start date, suggesting possible anticipation of some benefits made available by participation in the program.

Using the difference-in-differences approach, we compare the average saleability of granted and pending patents, which we assign into two groups according to their predicted propensity for prioritization before and after the program start date. We find that shorted examination time or a decision to apply for prioritized status have a positive impact on the probability of commercialization of a patent via the market for technology. We suggest that this empirical finding may have important policy implications for patent offices willing to minimize the social costs of pending innovations and to reduce the frictions on the market for technology. Our findings may also be relevant in the context of innovation management within start-up firms seeking formal protection for their intellectual property rights and commercialization of ideas via the market for technology.

2.A Appendix

2.A.1 Data Refinement: Applications of VC-backed start-ups

To construct a sample of patent applications initially owned by VC-backed start-ups we implement the following refinement procedure:

1. **Create a subset of applications filed within one year around the program start date:** We started with a set of 803,621 applications filed between September 26, 2010 and September 26, 2012 retrieved from the USPTO Patent Examination Research Dataset (PatEx). 578,963 of them had at least one assignment record in the USPTO Patent Assignment Dataset.

2. **Identify the first assignee of a patent application:** for 572,986 applications that were assigned by inventor(s) at least once it was possible to identify a unique assignee (employer) in 541,088 cases; in the other 6,732 cases it was possible to identify a unique non-academic employer (some inventors may assign their patent applications to multiple employers – academic and business entities). Out of the 5,977 applications that were never assigned by the inventor(s), in 2,961 cases there was just one assignment of the assignor's interest, and assignors of such applications were treated as the first assignees. In total, for 28,182 applications (fewer than 5%) with at least one assignment record, it was not possible to identify a unique first assignee.

3. **Match the names of assignees with the names of VC-backed firms:** before identifying VC-backed firms among the first assignees of patent applications, we unified both names of assignees and known VC-backed firms by simplifying and deduplicating them. Specifically, both groups of names were cleaned of special characters and numbers, names containing such strings as "also known as" and "formerly known as" were split into separate names of the same company. All resulting names were deduplicated using the Rosette API[16] and assigned unique identifiers that allows to match the names of first assignees with the list of names of VC-backed firms and the dates of their funding rounds retrieved from the VentureXpert database.

4. **Create a subset of patent applications owned by the VC-backed start-ups:** out of 550,781 applications for which first assignees could be identified, 45,190 applications had first assignees with names that matched those of VC-backed firms. 15,458 were owned by the VC-backed firms that had their first round of funding no later than five years before September 26, 2010 and, thus, were considered in our analysis as applications owned by the VC-backed start-ups.

2.A.2 Data Refinement: Commercial reassignments

To construct an outcome variable – probability of commercial reassignment – we implement the following refinement procedure:

1. **Assignments from VC-backed start-ups:** Out of 57,349 assignment records associated with a set of patent applications defined above, 2,789 transactions originated from VC-backed first assignees.

2. **Assignments of assignor's interest:** 1,136 of 2,789 records were of the "assignment of assignor's interest" type identified based on the conveyance text, and those records were considered to be commercial reassignment candidates. 291, however, were excluded as their execution dates were prior to the filing dates of reassigned applications. Another 81 records were excluded because their assignors' and assignees' names were identified as the same names. A further 26 records where applications were assigned to entities that later reassigned the same applications back to the first VC-backed assignees were also excluded and, finally, 18 records

[16] https://developer.rosette.com/features-and-functions

that duplicated earlier records with the same assignor-assignee-application combination were excluded from the sample. 720 assignment records associated with 712 patent applications remained.

3. **Probability of commercial reassignment:** in our analysis, we treat a patent application as commercially reassigned if the execution date of commercial reassignment is within the five years after the filing date. The length of the time window is dictated by availability of the data on assignments (Figure 9) recorded up to 2016 in the original Patent Assignment Dataset, and additional data on assignments recorded in 2017 that were retrieved via the API interface of the web-based Patent Assignment Database[17]. The length of the forward-looking time window is, thus, set to five years based on the maximum possible value for the most recent applications in our sample. Out of 712 commercially reassigned applications in our sample, 606 were reassigned within five years of the filing date, thus implying a 3.1% average commercialization rate.

2.A.3 Data Construction: Applicant and application characteristics

1. **Track One status:** indicator variable constructed from the Transaction History Data, part of the PatEx dataset, that takes a value of one if there is a specific type of transaction coded as "Mail Track 1 Request Granted" in the transaction history of the patent application, available only on and after September 26, 2011, when the USPTO began to accept requests for prioritized examination.

2. **Small entity status:** indicator variable retrieved from the PatEx dataset that takes a value of one if the applicant is either an individual inventor, a collaboration of individual inventors, a nonprofit organization, or a company with fewer than 500 employees[18].

3. **Age of the firm:** a numerical variable indicating the difference in years between September 26, 2010 (start of the sample time frame) and the date of first funding round of the VC-backed firm – first assignee of the patent application – constructed

[17] https://assignment.uspto.gov/,https://assignment-api.uspto.gov/documentation-patent/
[18] https://www.uspto.gov/web/offices/pac/mpep/s2550.html

from the VentureXpert data. Firms aged less than or equal to two years are considered ‘young’.

4. **Top-tier patent attorney:** indicator variable constructed from the information on the attorneys and patent agents who have been granted power of attorney with regard to the corresponding subject applications, that takes a value of one if the name of a patent agent appears in the list of 123 “Best Law Firms for Patent Law" compiled by the U.S. News & World Report[19].

5. **Triadic status:** indicator variable retrieved from the OECD Triadic Patent Families database, February 2015, that takes a value of one if the patent application is a part of a patent family formed by patents filed at the European Patent Office (EPO), the Japan Patent Office (JPO) or the United States Patent and Trademark Office (USPTO).

6. **Patent family size:** numerical variable retrieved from the PATSTAT database indicating the number of patents that cover exactly the same technical content as a focal patent application.

7. **Number of inventors:** numerical variable constructed from the PatEx dataset indicating the number of individual inventors listed in a patent application.

8. **Technological sector:** categorical variable indicating 36 two-digit technological sectors aggregated from 457 classes of the U.S. Patent Classification (USPC) System based on the mapping constructed by Hall, Jaffe & Trajtenberg (2001).

2.A.4 Lasso estimation: Predicted propensity for prioritization

To construct the dummy variable used in the difference-in-differences analysis, which distinguishes patent applications with high and low predicted propensity for treatment (prioritization), we implement the following procedure:

1. We construct an extended set of controls including applicant and application characteristics described in Appendix 2.A.3 plus: 1) a set of dummies for all possible combinations of values of the young firm and small entity indicators; 2) a set of dummies for all possible combinations of values of the top-tier patent attorney indicator and each of two indicators – for a young firm and a small entity.

[19] https://www.usnews.com/rankings

2. We take a random subsample constituting 80% of applications filed after the policy change, for which the actual treatment status – prioritized or non-prioritized – is observed. We use the latter applications to train the Lasso model and the remaining 20% of applications are stored for cross-validation purposes.

3. Using the random subsample of applications and the extended set of controls, and technology sector dummies, we run the Lasso algorithm to select the most significant predictors of the prioritized status of patent applications.

4. We make an out-of-sample prediction of the propensities for treatment (prioritization) for the remaining set of applications and use the distribution of predicted values to choose the cut point (threshold value), such that applications can be divided into high and low treatment intensity groups.

5. For each possible value of the cut point (threshold value), we contrast the actual treatment status (prioritized or non-prioritized) with the predicted treatment intensity (high or low) and calculate the corresponding values of true positive and false positive rates. We find the optimal cut point, such that the true positive rate is maximized and the false positive rate is minimized.

6. We use the optimal cut point obtained in the previous step to split all 15,150 applications in our sample filed before and after the policy implementation into two groups – with high and low predicted propensity for prioritization, and construct the corresponding dummy variable used in the difference-in-differences analysis.

Chapter 3

Why has Science become an Old Man's Game? (co-authored with Christian Fons-Rosen and Patrick Gaulé)

3.1 Introduction

Scientific knowledge is increasingly produced by older people. John Goodenough was 97 when he received the 2019 Nobel Prize for chemistry. He has since published ten papers in highly respected peer-reviewed journals. Beyond this (arguably extreme) example, the average age of National Institutes of Health grant recipients has increased from 39 to 51 between 1981 and 2008 (Daniels, 2015). Among U.S. chemistry faculty members, the mean age has increased from 37 in 1960 to 53 in 2015. Moreover, the age at which Nobel-prize-winning discoveries are made has risen steadily over the course of the 20th century (Jones, 2009; Jones & Weinberg, 2011).

The ageing of the scientific workforce may not necessarily impact the rate and direction of scientific progress. Scientific discoveries can be made by scientists in middle age – Wilhelm Röntgen discovered X-rays at the age of 50 – or indeed later. To the extent that age has an impact on cognitive function, this effect could be small and counterbalanced by experience. Some studies find only a small effect of age on scientific productivity (Diamond, 1984, Turner & Mairesses, 2003).

If, however, age influences the quantity and type of knowledge produced by individuals to any degree, then understanding the reasons for general ageing of the scientific workforce is important to shape appropriate policy responses. One influential explanation of the ageing of the scientific workforce is the 'burden of knowledge' hypothesis (Jones, 2009). As the stock of human knowledge accumulates over time, new entrants need to spend more time in training in order to reach the knowledge frontier. This leads to a secular increase in the age at which scientists begin their careers and make keys discoveries, empirical patterns that are observed among Nobel Prize winners (Jones, 2009, 2010). From this perspective, policymakers should seek to improve the quality of training and preserve incentives to start scientific careers, but may want to refrain from giving large grants to young scientists (Jones, 2011). An alternative explanation for the ageing of the scientific workforce is declines in the retirement rate of older scientists, a process facilitated by the elimination of mandatory retirement in U.S. universities (Blau & Weinberg, 2017). If declines in retirement rates drive the ageing of the scientific workforce, policymakers might – depending on the productivity of older scientists – reconsider end-of-career incentives and policies.

It may seem that rising ages at entry and retirement dynamics are the only factors impacting the age composition of the scientific workforce, but this is not so. Entry age and exit dynamics clearly matter in the age composition of the scientific workforce, but a potentially important third factor is the number of people hired over time. Consider what would happen if the U.S. government made a large investment in science – a new Apollo or Manhattan project, perhaps. As the total demand for scientific labor goes up, the extra positions would be disproportionately filled by younger people graduating from universities (as opposed to older individuals moving from non-research to research jobs). As a consequence, the scientific workforce would immediately become younger. The effect of such a hiring spree on the age composition of the scientific workforce would not be limited to the short run: as the dis-proportionally large cohort of new entrant aged, so would the scientific workforce as a whole.

In this paper, we build a demographic model of the U.S. academic workforce, to shed light on the causes of its ageing. The model leverages novel data on the population of U.S. chemistry faculty members between 1960 and 2010. Having set up the model to mimic observed empirical patterns in the data, we can then use it to quantify the importance of various channels – changes in entry age; retirement dynamics; and the number of new hires – to the ageing of the workforce. For instance, we can ask what would happen to the age composition of faculty members if entry ages had remained at their initial (1960) level, but retirement dynamics and hiring patterns still evolved as they did.

We find that changes in the numbers of people hired over time is the major driver in the ageing of our sample. In the data, the mean age of chemistry faculty members rises from 37 in 1960 to 53 in 2010. While the age at which individuals become faculty members has indeed increased (as predicted by the 'burden of knowledge' hypothesis), this only accounts for about 20% of the increase in mean faculty age. Further, changes in retirement dynamics have no sizeable effects in our sample. By contrast, variation in the numbers of people hired over time appear to be the major factor in the ageing over time. In the 1960s, the number of new hires exceeded retirements by a factor of four. If faculty hiring had proceeded at the same pace in subsequent decades (instead of declining a level roughly in line with retirements), the mean age of chemistry faculty members would have stabilized around 40, instead of rising to 53 as it did.

Our results have a number of implications for science policy. First, the age composition of the academic workforce need not simply reflect fundamental trends in the nature of knowledge production or of overall societal ageing. Instead, it may (and perhaps should) be seen as the result of past and present policy choices, particularly in terms of the number of people hired. Second, hiring more new faculty could generate disproportionate returns. Because new hires tend to be younger, they generate a sort of 'demographic dividend' in that their productivity tends to be higher than the average faculty, at least in the beginning. Given that currently less than 10% percent of graduate students become faculty members, and that faculty positions are generally considered attractive (Ganguli, Gaule & Vuletic, 2022), there is no shortage of talented young scholars interested in taking up faculty positions. Third, steps could be taken to mitigate the bias against young scholars in grant allocation. This could include, for instance, putting a greater weight on a proposal itself relative to the track record of the principal investigator.

This paper contributes to the literature on the causes of the ageing of the scientific workforce (Jones, 2009; Jone, 2010; Blau & Weinberg, 2017), highlighting a new cause (a slowdown in hiring) as quantitatively most important factor across the post War War II era.

The rest of the paper is organized as follows. Section 2 provides institutional context on the changing age structure of the U.S. scientific workforce, U.S. Science Funding since World War II, and chemistry as a scientific discipline. Section 3 describes the data, Section 4 presents the methodology of the simulation and section 5 presents the results. Finally, section 6 concludes.

3.2 Institutional context

In this section, we provide three pieces of institutional context that are relevant to understanding the causes and consequences of ageing of the scientific workforce. First, we provide some descriptive statistics on the ageing of the U.S. scientific workforce from 1960 to 2010. Second, we describe how funding for basic research has evolved in the U.S. from World War II to the present, as this will be relevant for understanding how hiring patterns have changed over time. Third, since our analysis will focus on chemistry, we provide some background information on how knowledge production is organized in this discipline.

3.2.1 Age structure of the U.S. scientific workforce

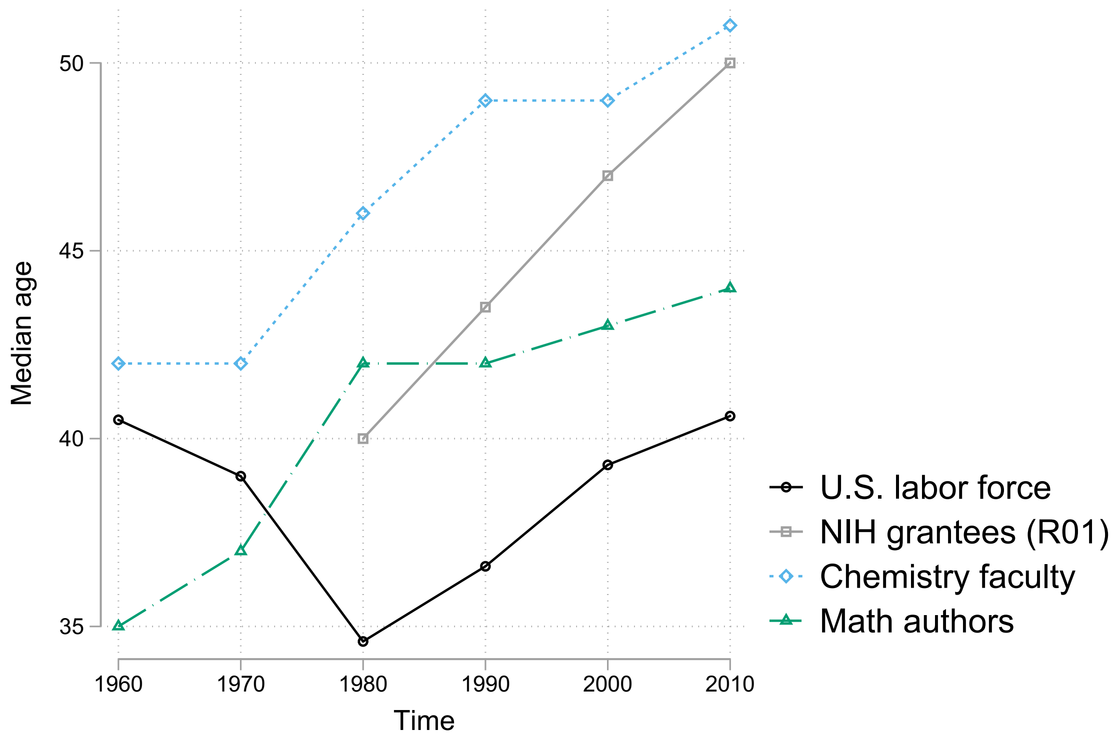
The U.S. scientific workforce is ageing. Many observers have noted changes in the age composition of National Institutes of Health grantees (see, e.g., Daniels, 2015). From 1980 to 2010, for instance, the median age of NIH R01[1] grant recipients increased from 40 to 50 years of age. Among chemistry faculty members, the median age increased from 41 years in 1960 to 51 in 2010 (Figure 3.1).[2] Even in mathematics, traditionally seen as the preserve of the young, the median age of authors increased from 35 in 1960 to 44 in 2010.

There appears to be something distinct and specific about the scientific workforce compared to the labor force as a whole. To put the ageing of the U.S. scientific workforce into context, it is useful to consider the ageing of the U.S. labor force as a whole. It is true that the U.S. labor force in general grew older from 1980 to 2010, with the median age rising from 35 to slightly above 40. However, across a longer time period, the trends are less clear: the U.S. workforce had a median age slightly above 40 as early as 1960. Moreover, even within 1980-2010, the rate of ageing of the U.S. workforce (with the median age rising by less than 1.5 years per decade) is clearly lower than for the NIH grantees, for instance, where the median age rose by more than 3 years per decade.

[1]Research Project Grants (R01) is the original and historically oldest grant mechanism used by the NIH. Grants are meant to support a specified project to be performed by a principal investigator in an area representing the investigator's specific interest and competencies, based on the mission of the NIH. NIH R01 grants constitute the bulk of NIH external grant giving.

[2]This figure is based on our dataset of U.S. chemistry faculty members, which we describe in the next section.

Figure 3.1: Age structure of the U.S. scientific and labor workforce



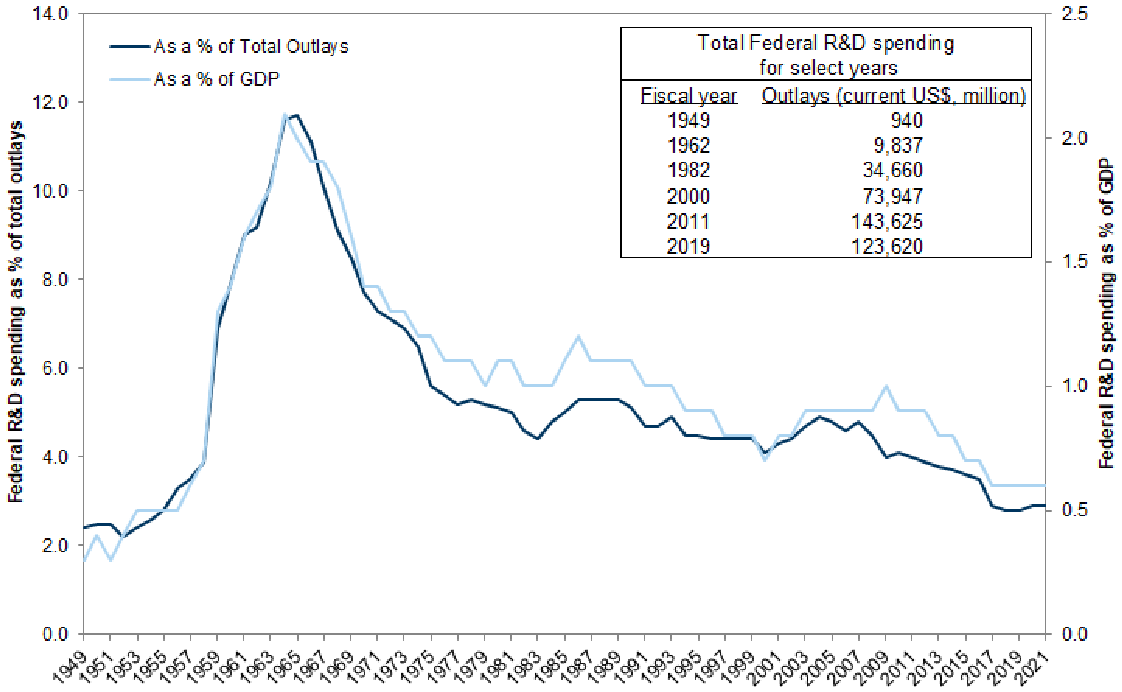
3.2.2 U.S. science funding after World War II

World War II was a powerful demonstration of the practical utility of science and innovation. The development of the radar, for instance, gave the United Kingdom a key advantage in the Battle of Britain (a series of aerial battles between German and British air forces over the skies of Britain). Most ominously, World War II ended shortly after deployment of the atomic bomb at Hiroshima and Nagasaki. The atomic bomb was developed under the Manhattan Project, a gargantuan research project mobilizing most of the best physicists residing in America.

After World War II, a consensus emerged that basic research was important both for national prosperity and in the ongoing geopolitical rivalry with the USSR. The argument was best encapsulated in Vannebar Bush's influential report to the U.S. report 'Science, The Endless Frontier' (Bush, 1945). In 1957, the USSR launched the first satellite into space, sparking fears that the U.S.A was falling behind in technology. This 'Sputnik moment' provided further impetus for investment in U.S. science and education.

After 1945, the U.S. federal government began to support basic research on a systematic basis, and public R&D investments became very substantial. In 1950, a brand new federal agency, the National Science Foundation, was funded. Further, the National Institutes of Health were instituted, building on prewar institutions, bolstered by a growing mandate and budgets. Federal funding for R&D grew tenfold from 1949 to 1962. As the share of GDP, federal funding for R&D grew from 0.5% in 1949 to more than 2% in 1964, before declining subsequently (Figure 3.2).

Figure 3.2: Federal Expenditure on R&D



Notes: This Figure is reproduced from Goldman Sachs Research (2020)

The influx of federal funding resulted in a bonanza for U.S. universities. In an anecdote recounted by Paula Stephan (Stephan, 2018), federal grant agencies were sending representatives to universities to encourage faculty members to apply for funding. In the late 1960s and early 1970s, U.S. universities also benefited from a surge in demand for education from the 'baby boom' generation entering college. In response to increased demand for both research and teaching, U.S. universities hired large numbers of new faculty members, a pattern we clearly see in our data, and will discuss later.

Federal R&D funding peaked in the mid 1960s before decreasing substantially in the late 1960s and 1970s. As a percentage of GDP, federal R&D spending fell from above

2% in 1964 to around 1% in 1980. Thereafter, there was a further slow decline over time, with federal R&D spending reaching 0.6% in 2019. The decline in federal R&D was briefly interrupted by the doubling of the NIH budget between 1998 and 2003, but this increase was not permanent. Moreover, the rapid increase in NIH spending and ensuing deceleration created substantial adjustment problems in the market for research (Freeman & Von Reenen, 2009).

3.2.3 Chemistry

Chemistry is the scientific study of the properties and behavior of matter. While chemistry is a physical science, large parts of it relate to living organisms, so that chemistry is also closely related to life sciences. Apart from some smaller sub-disciplines such as theoretical chemistry, chemistry is largely a lab-based science. Besides being a physical space with instruments and research materials, the lab is also an organizational structure through which a faculty member (principal investigator) obtains funding for the lab, directs research projects, and appears as a coauthor on all publications. While faculty members in elite institutions are normally supported by a relatively large number of graduate students, postdocs and technical staff working in their lab, support staff varies to some extent across time and institutions.

Traditionally, faculty members/lab directors appear as the last author of scientific publications. The graduate student or postdoc who has done most of the day-to-day work on the research project typically receives first authorship. Authors in the middle of the authorship list have normally made relatively minor contributions to the projects, though this clearly varies across papers.

Research by chemistry faculty members is supported through a mix of federal, state, and industry sources. Depending on their speciality, chemistry faculty members may apply to the NIH, the NSF, or to other federal agencies. Given that research in chemistry often has practical applications, industry funding through R&D contracts is common.

3.3 Data

To investigate the causes and consequences of the ageing of the scientific workforce, we assembled an original data set combining multiple sources. The core and most original component is the longitudinal database of academic scientists derived from the ACS directory (described below). We complement this longitudinal database with information on publications.

The ACS Directory. Our main data source is the 'ACS Directory of Graduate Research' (hereafter: ACS Directory). The ACS directory was a biennial publication of the American Chemical Society than ran from 1953 to 2015, when it was discontinued. The publication aimed to provide prospective graduate students with information on U.S. chemistry departments offering PhD degrees. Initially published as a book, the ACS directory was also later disseminated via an electronic version (first on CD-ROM, and then on a dedicated website). The publication included lists of faculty members by name, year of birth, gender, educational history, and current affiliation, among other information (Figure 3.8). The information on the year of birth is particularly interesting and valuable, as it is otherwise hard to find on a systematic basis.

Building a dataset of U.S. chemistry faculty members based on the ACS directory. We procured the 1961, 1971, 1981, 1991, 2001 and 2011 editions of the ACS Directory. For the 1991, 2001 and 2011 editions, we used electronic versions.[3] For the earlier versions (1961, 1971 and 1981), we digitalized the respective copies (for a total of more than 3 thousand pages) using optical character recognition (OCR) software and freelance assistants to correct OCR mistakes. The resulting data yields six snapshots of the distribution of faculty members in U.S. chemistry departments between 1961 and 2011. We then generated a longitudinal database linking individual faculty members across the different editions of the book, using names, birth years, and educational histories to create linkages. This longitudinal dataset also enabled us to obtain a proxy for entry and exit from the profession, through the year of first listing in the directory and the year of last listing, respectively.

3.4 Methodology

We build a demographic model simulating the evolution of the U.S. academic workforce in chemistry to shed light on the causes of its ageing. Having set up the model to mimic observed empirical patterns in the data, we can then use it to quantify the importance of various channels – changes in entry age; retirement dynamics; and numbers of new hires – to the ageing of the workforce.

[3] A version of that database covering 1993 to 2009 has been used in Gaule (2014), Gaule & Piacentini (2018), Catalini, Fons-Rosen & Gaule (2020) and Ganguli, Gaule & Vuletic (2022).

The simulation starts with the sample of 3,464 scientists who are active as of 1961 in the data. Six editions of the directories are used with an interval of ten years between consequential snapshots, producing a range of years between 1961 and 2011.

Assumptions. We assume that someone who appeared in a given snapshot s for the first time actually entered the dataset in $\underline{s} = s - 6$. Due to the left truncation, we cannot make any assumptions about the actual time of entrance for people who were active as of the initial snapshot $s = 1961$, that is, $\underline{s} = 1955$. Therefore, after excluding 1955, our range of entry years is: $\underline{s} \in 1965 + 10k_{k=0}^{4}$. Similarly, we assume that someone who was active in a given snapshot s for the last time actually exited the dataset in $\overline{s} = s + 4$. Due to the right truncation, we cannot make this assumption in the last snapshot $s = 2011$, that is, $\overline{s} = 2015$. Therefore, we end up with the range: $\overline{s} \in 1965 + 10k_{k=0}^{4}$.[4]

We refer to the middle-decade points throughout the simulation as the simulated periods t. In other words, whenever we discuss time periods, we are referring to these middle-decade points.

Transitions between simulated periods. The dataset, starting from the initial data we have available for 1961, is subject to hirings and exits over time that lead to changes in the *number of people*. Furthermore, we identify three channels leading to changes in the *mean age*: (1) Exits; (2) Hirings; (3) Age composition of hirings. Below we sequentially describe these three channels.

Exits. For each person who is active in the current simulated period t, we assign an updated status – "active" or "exited" – for the next period $t + 10$ based on age-specific and time-specific exit probabilities e_{at} calculated from the actual data, where a is the age decade (i.e., people in their 20s, 30s, etc.) and t is time. This implies that we account for the fact that older scientists are more likely to exit than younger ones, and that these likelihoods also evolve over time.

For every middle-decade point t and 10-year-wide age group a, we calculate probabilities $e_{at} = N_{at}^{e}/N_{at}^{a}$. The numerator, denoted as N_{at}^{e}, is the number of people in age group a who are active at time t but are no longer active at time $t + 10$. The denominator, denoted as N_{at}^{a}, is the total number of currently active people in the age group a at time t.[5]

In our simulations, suppose that a group of individuals (defined by an age range a

[4]In other words, we assume that the hiring of people who appear for the first time in snapshot s and the exit of people who were active for the last time a decade earlier in snapshot $s - 10$, actually happen at the middle-decade point.

[5]For clarify, the subindex e stands for "exits" and the subindex a stands for "active".

at time t) has a likelihood of exit defined by e_{at}, say, equal to 20%. In this case, we randomly choose one fifth of these individuals and assign them an "exited" status in the next time period.

Hirings. In each simulated period t, we simulate hiring $\hat{N_t}^h$ people, which we refer to as the "extensive" margin of hiring, meaning that we do not account for scientist age.[6] It is determined by $\hat{N_t}^h = \hat{N_t}^e \cdot h_t$, where $\hat{N_t}^e$ is the number of simulated exits including all age groups, and h_t is an expansion rate, i.e. the number of simulated hirings for each simulated exit in the same period. We define this expansion rate as $h_t = N_t^h / N_t^e$, that is, calculating the ratio from actual data on hires and exits in a given time period (N_t^h and N_t^e).

The exits channel directly affects both the intensive and extensive margins, because each exiting individual is assigned an age that we track. But the hirings channel only affects the extensive margin, because no age is assigned to entrants yet. To allocate an age to these hires, we need to add a third channel that will be described below. In other words, we distinguish between the "extensive" (number of people) and "intensive" (age composition) margins of *hiring* represented by two separate channels, while the *exit* channel is not divided into two parts.

Age composition of hires. In each simulated period t, birth years of newly hired people ($\hat{N_t}^h$) are assigned so that the resulting age composition of simulated hirings reconstructs the actual age composition of newly hired people in the data.[7] All parameters calculated from actual data (i.e., N_t^h and N_t^e, e_{at}, and all $P_t^{x\%}$) are time-specific data moments.

Extensive vs intensive margins. It follows from these three definitions that exits and hirings affect *not only* the total number of scientists *but also* the age composition of active scientists in each simulated period t. We justify the asymmetry by which the exit channel does not incorporate an intensive margin channel as follows.

First, we show that the results of a simulation relying on time-specific data moments will be equivalent, independently of whether the exit activity is collapsed into a single channel or instead is decomposed into two different channels – the number and the age composition of exits.

Consider an alternative definition of exits in which we explicitly account for the two channels: in each simulated period t, we first impose a fixed number of total exits N_t^e and

[6]The subindex h stands for "hiring" and the hat notation stands for "simulated".

[7]Specifically, we assign $(P_t^{min} + P_t^{10\%})/2$ to 10% of randomly chosen simulated hires, $(P_t^{10\%} + P_t^{50\%})/2$ to 40%, $(P_t^{50\%} + P_t^{75\%})/2$ to 25%, $(P_t^{75\%} + P_t^{90\%})/2$ to 15% and $(P_t^{90\%} + P_t^{95\%})/2$ to 10% of hires, where $P_t^{x\%}$ is a corresponding percentile of the actual age composition of hires at the time t.

then divide them into age groups according to age-specific shares of the exits – E_{at}. We then randomly assign an "exited" status to a number of people from a given age group according to $\hat{N}^e_{at} = N^e_t \cdot E_{at}$. In this way, we explicitly distinguish between the "extensive" and "intensive" margins of exits.

While this alternative share of exits E_{at} only relies on one age-specific component, N^e_{at}, in our preferred share of exits e_{at} we additionally incorporate a second age-specific component, N^a_{at}. Thus, without a loss of generality, we can replicate the actual data by relying on *either* a broader set of data moments in our initial definition *or* a narrower set of data moments with the alternative definition.

Second, it follows from defining simulated exits as $e_{at} = N^e_{at}/N^a_{at}$ that we simulate the number of exits for the age range a in the following way: $\hat{N}^e_{at} = N^a_{at} \cdot e_{at}$. Summing up across all age ranges, the total number of exits is therefore $\hat{N}^e_t = \sum_a \hat{N}^e_{at} = \sum_a N^a_{at} e_{at}$. Because hirings are a function of exits, $\hat{N_t}^h = \hat{N_t}^e \cdot h_t$, we can rewrite this expression as $\hat{N_t}^h = h_t(\sum_a N^a_{at} e_{at})$. Thus, our approach allows us to account for the age-specific stock of active scientists in a given period t, N^a_{at}, not only in the exits, but also in the hirings channel.

Counterfactual scenarios. Starting with the initial sample of scientists described above, we generate transitions to the next periods through the three channels described above by: (1) adding new people to the dataset; (2) randomly dropping people conditional on each given age group; (3) randomly assigning age to each newly added hire based on statistics from actual data.

For the three channels, we differentiate between two modes: (1) "ON" – time-specific data moments (N^h_t and N^e_t, r_{at}, and $P^{x\%}_t$) are used in each time period; (2) "OFF" – only the initial 1965 period data moments (N^h_{1965} and N^e_{1965}, r_{a1965}, and $P^{x\%}_{1965}$) are used for all periods.

As a clarification, under our baseline scenario "ALL ON", all three channels are simulated in the "ON" mode. The simulated data moments ($h_t = N^h_t/N^e_t$, r_{at}, and $P^{x\%}_t$) for a given time period are generated based on statistics of actual data in that same time period. Instead, under the baseline scenario "ALL OFF" only actual data from 1965 is used to generate the simulated data moments of all time periods.

We evaluate alternative counterfactual scenarios by switching on only one channel at a time. For example, suppose that the time-varying expansion rate is "ON", while the exits and age distribution of newly hired are "OFF". In this scenario, actual data is used for the parameter $h_t = N^h_t/N^e_t$ across all decades. Instead, data from 1965 (r_{a1965}, $P^{x\%}_{1965}$)

is used for the exits and moments of the age distribution of new hires across all decades.

The other two counterfactual scenarios switch on either the exits or the age distribution of the new hires. That is, we use actual data from all decades only for the parameter r_{at} or $P_t^{x\%}$, respectively, while the remaining channels are in the "OFF" mode and only data from 1965 is used.

3.5 Results

3.5.1 Baseline scenarios

In each simulated period t, we calculate a mean age of all scientists active in the current period and plot the dynamics of it over time and across different counterfactual scenarios.

Scenario	Mean 2005 Age
Actual	51.4
ALL ON	51.7
ALL OFF	39.1
Age of hired ON	41.7
Exits ON	39.1
Expansion rate ON	47.3
Mean 1965 Age	38.5

Notes: This table reports the mean age of active scientists as of the first (1965) and the last (2005) periods in each simulated scenario.

Table 3.1: Summary of simulations results

"ALL ON" – time-varying dynamics for all three channels is switched on

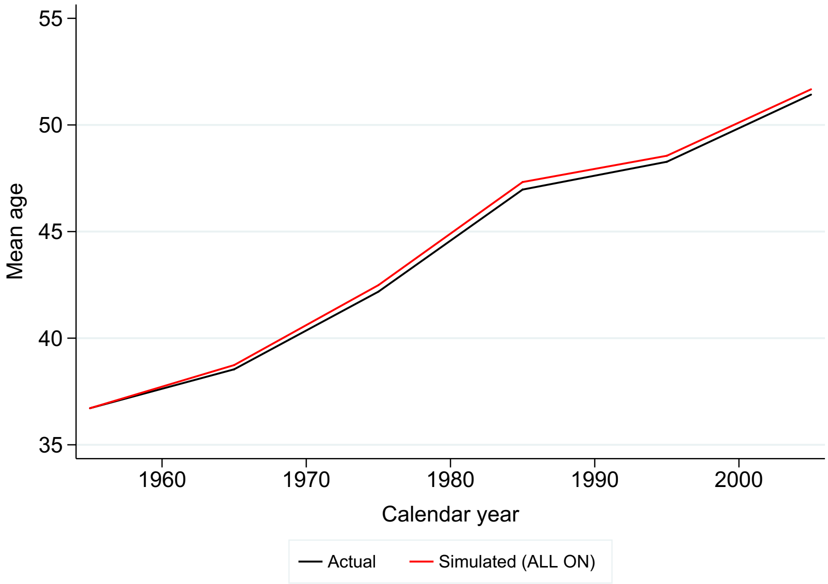
Under this scenario, we switch on all three channels: hirings, exits, and age composition of hired scientists. This means that we will use the time-varying data moments (1965-2005). This baseline scenario demonstrates a good match in terms of replicating both the evolution (Figure 3.3) and the absolute increase (Table 3.1) of the mean age in the actual data.

Therefore, the increasing mean age of scientists between 1955 and 2005 can be explained by the joint dynamics of time-varying expansion rate, exit probabilities, and age distribution of new hires.

“ALL OFF” – time-varying dynamics for all three channels is switched off

This scenario switches off all three time-varying channels, leading to a flat trajectory of the mean age trend (Figure 3.4). We do not observe any change in the absolute levels of the mean age between 1965 and 2005 (Table 3.1).

Figure 3.3: Baseline scenario “ALL ON”



Notes: This Figure shows that the trend in the mean age of scientists simulated with the use of all time-varying moments calculated from the actual data closely matches the actual trend observed over the last 50 years.

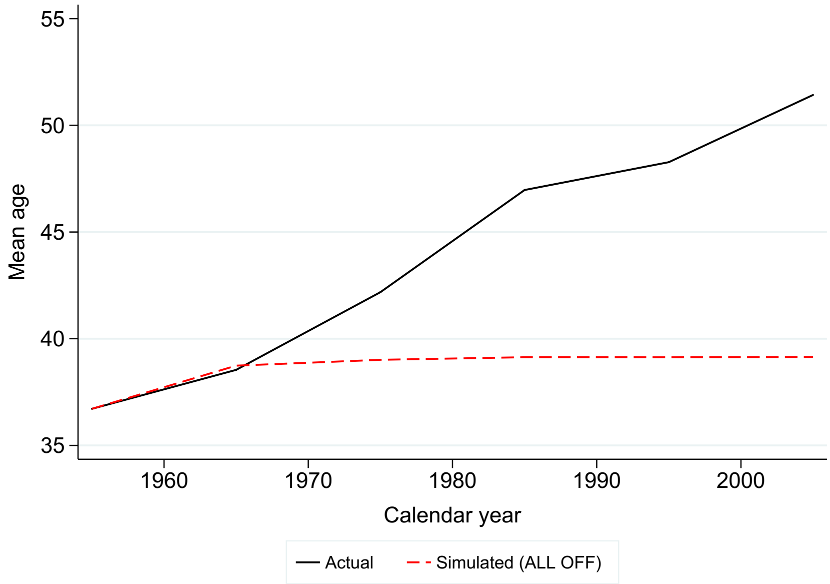
Consider the total gap in the mean age in 2005 – the last period simulated – between the “ALL ON” and “ALL OFF” counterfactual scenarios (51 − 39 = 12 years of age). We decompose this 12-year absolute increase into individual contributions via the three channels described previously. In particular, we will deviate from the “ALL OFF” scenario by switching on only one particular channel in each counterfactual scenario.

3.5.2 Counterfactual scenarios: switching on only one channel

Case 1: Switching on the time-varying age distribution of hired scientists

Unlike the “ALL OFF” case, this scenario allows the age of hired scientists to be matched to actual data moments ($P_t^{x\%}$) in each simulated period. The remaining two channels remain switched off.

Figure 3.4: Baseline scenario "ALL OFF"



Notes: This figure shows that, in the counterfactual scenario assuming no changes over the last 50 years in the age distribution of new hires, the probability of exits, and the expansion rates, the mean age of scientists would have stayed close to that of the 1960s.

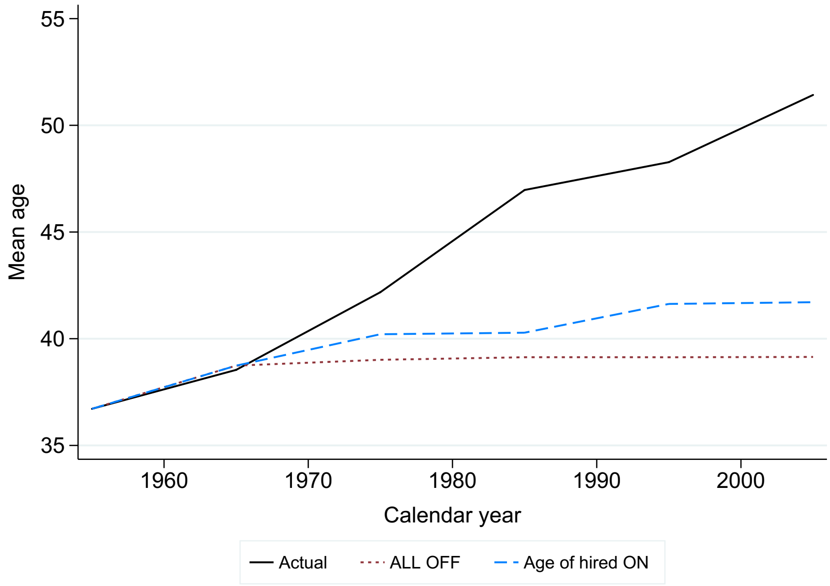
In this counterfactual scenario, there is a clear upward trend in the mean age over time (Figure 3.5). The absolute increase in mean age from 38.5 in 1965 to 41.7 in 2005 is consistent with the notion that entry age is constantly rising as claimed in the 'burden of knowledge' hypothesis (Jones, 2009). Comparing the 2005 levels (Table 1) obtained in this scenario to the "ALL OFF" case, we can explain $(41.7 - 39.1)/(51.7 - 39.1) = 20.6\%$ of the 12-year increase.

Case 2: Switching on the time-varying age-specific exit rates

Unlike the "ALL OFF" case, this scenario uses actual data to obtain the rate of age-specific exits (e_{at}) in each time period.

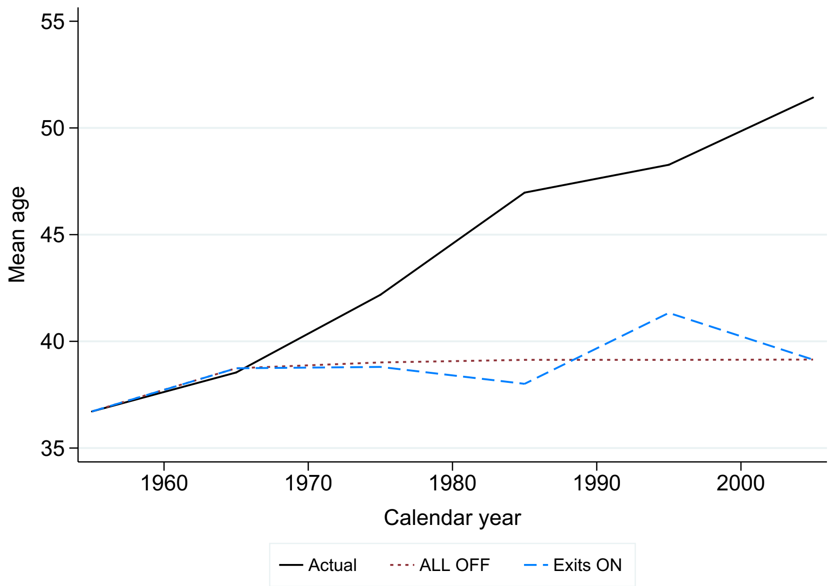
There is no clear trend change in the mean age over time (Figure 3.6). This can be explained by the fact that age-specific exit rates e_{at} did not exhibit a consistent upward or downward pattern between 1965 and 2005 (Figure 3.10). Instead, increases in exit rates in one period are followed by decreases in the following period, thereby driving the volatility of the simulated mean-age trend. Moreover, when we compare the 2005 levels (Table 1), we cannot explain any portion of the total 12-year increase by the dynamics

Figure 3.5: Time-varying age distribution of hired is "ON"



Notes: This Figure shows how the changes observed during the last 50 years in the age distribution of new hires (the difference between the blue and red dashed lines) contributed to the increasing mean age of scientists (black line).

Figure 3.6: Time-varying exits are "ON"



Notes: This figure shows how the changes observed during the last 50 years in the probability of exits (the difference between the blue and red dashed lines) contributed to the increasing mean age of scientists (black line).

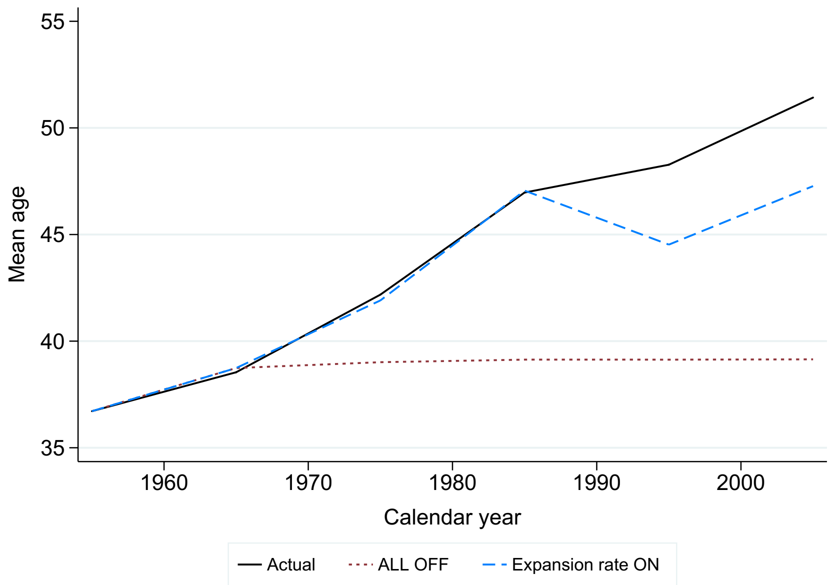
of exits channel isolated from other two channels, since the mean age remains unaltered at a value of 39.1.

Case 3: Switching on the time-varying expansion rates

Unlike the "ALL OFF" case, this scenario uses actual data to account for time-varying expansion rates $h_t = N_t^h / N_t^e$. In the previous two counterfactual scenarios, the exit rate used for all time periods was 4.05. This number was chosen because, in the actual data, we observe 4.05 hirings in snapshot $s = 1971$ for each exit in the previous snapshot $s = 1961$.

As this channel is now switched on, we allow the number of new hires in each period to be determined as: $\hat{N_t}^h = h_t \cdot \hat{N_t}^e$, where $\hat{N_t}^e$ is the number of simulated exits and h_t is an expansion rate that changes every period based on actual data.

Figure 3.7: Time-varying expansion rate is "ON"



Notes: This Figure shows how the changes observed during the last 50 years in the expansion rate (the difference between the blue and red dashed lines) contributed to the increasing mean age of scientists (solid black line).

Looking at actual data on expansion rates across all periods, it turns out that using a fixed value of 4.05 based on 1965 data is a very high expansion rate. This initial level is more than two times the average of 1.76 across the entire interval 1965-2005. The expansion rate remained lower in all subsequent periods (Figure 3.11). Furthermore, the expansion rate never surpassed a value of 2 in any period after 1965.

This means that the major expansion in the scientific workforce observed in 1965 was not persistent over time. Due to the left-truncation of the data, we cannot identify whether this expansion began earlier or not. However, we suggest that it is this unprecedented magnitude and its abrupt end in 1975 that led to the pronounced upward trend in the mean age observed in actual data between 1955 and 2005.

In fact, the simulated trend of the mean age in this counterfactual scenario closely follows the actual trend between 1955 and 1985 (Figure 3.7), and the unexplained gap after 1985 is due to the volatility in the expansion rate observed in the actual data. That is, this counterfactual scenario relying on time-varying expansion rates can explain most of the mean-age increase over time.

Overall, comparing the levels reached in 2005 (Table 1), we conclude that the dynamics of the expansion rate observed in actual data (i.e., a sharp drop in 1975 followed by constantly low expansion rates thereafter) helps to explain 66% of the total 12-year increase in the mean age. Therefore, the hiring channel is clearly the main driver of the ageing pattern.

From a policy perspective, it is relevant that the academic expansion rates that we observed early in our dataset are unparalleled in their magnitude when compared to more recent expansions of the academic workforce. This staggering stock of academic scientists has led to a gradual increase in the age of academics. Combined with the fact that older scientists tend to produce less impactful research, the results in this paper suggest that scientific productivity can be substantially increased via considerable increases in the scientific workforce.

3.6 Discussion

The U.S. scientific workforce has aged considerably over the past 60 years. This phenomenon has been noted by many observers, with the rising age of NIH grant grantees receiving particular attention (Kaiser, 2008, Daniels, 2015). Yet the causes of the ageing of the scientific workforce remain imperfectly understood.

Previous literature on the causes of ageing in academia has focused on the rising age at entry into science (Jones, 2009), as well as the tendency of scientists to retire later (Blau & Weinberg, 2017), our work highlights a third distinct reason the scientific workforce may be be ageing: compositional changes arising from a slowdown in hiring over time. In a simulation based on detailed data on U.S. chemistry faculty members between 1960 and 2010, changes in hiring over time appear to drive most of the change in the age composition of scientists. In the 1960s – a period in which universities were expanding significantly – new faculty hires outnumbered retirements by a factor of four. Because new hires tend to be relatively younger, this led to a large influx of young people in academic profession. However, as hiring slowed down in subsequent decades, the cohorts of 1960s entrants (and to a lesser extent, 1970s entrants) became disproportionally large, resulting in an ageing scientific workforce.

One important consequence of ageing being largely driven by changes in hiring over time is that the ageing of the scientific workforce is not then the inevitable result of an ageing society or of fundamental forces in the production of knowledge. Rather, much of it could be the reflection of a specific set of historical circumstances and policy choices. In the U.S. post WWII, there was a massive expansion of the university sector, followed by a period of no growth in faculty numbers.

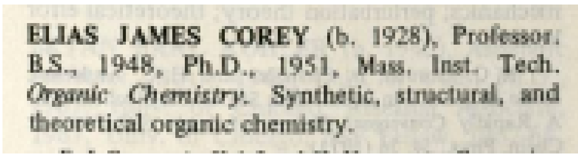
Just as countries that go through a demographic transition experience a demographic dividend, countries that expand their scientific sector may experience a demographic dividend of a kind with an overall younger and more productive scientific workforce in the transition phase. It is quite surprising that systematically increasing the hiring of faculty members is not often discussed in policy circles, despite its obvious potential to increase scientific productivity – both directly, by increasing the number of researchers, and indirectly through rejuvenation of the academic workforce.

We conclude by noting two directions for future research. This paper, and indeed much of the related literature, has focused on the U.S. scientific workforce. However, much less is known regarding other countries. Of particular interest is the case of China, which has made large investments in universities over the last two decade. Research on changes in the age of the scientific workforce outside the U.S. and in China in particular would be welcome. Another area that merits attention is the age dynamics in industrial R&D, where employment relationships are rather different from those in academia. To what extent has the industrial R&D workforce aged and is that ageing also influenced by hiring sprees? Research along either of theses lines of inquiry could further elucidate whether changes in the age composition of the scientific workforce are driven more by fundamental forces or by particular historic circumstances.

3.A Appendix

Figures 3.8 and 3.9 below depict examples of the data – entries from the ACS directory. Among other information, each directory entry lists the name, birth year, and education history of the person. We can also deduce in which department the person was working at the time, from the department under which the listing appears. In this sample entry, the person listed is Elias Corey, a distinguished American chemist who spent most of his career at Harvard and won the Nobel Prize in Chemistry in 1990. Whereas biographical information about famous scientists such as Corey is available from many sources, the strength of the directory is that it includes all faculty members employed in the reporting departments irrespective of their status in their profession.

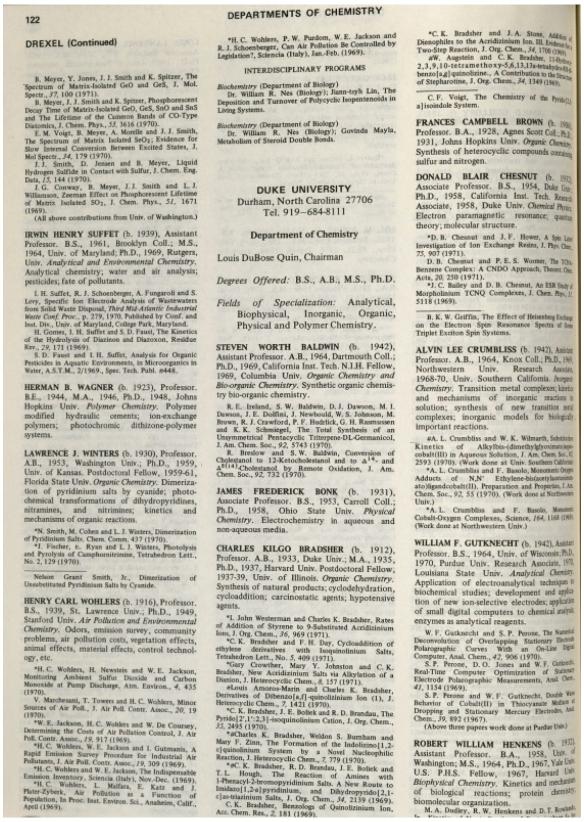
Figure 3.8: A sample entry from the ACS directory



ELIAS JAMES COREY (b. 1928), Professor. B.S., 1948, Ph.D., 1951, Mass. Inst. Tech. *Organic Chemistry.* Synthetic, structural, and theoretical organic chemistry.

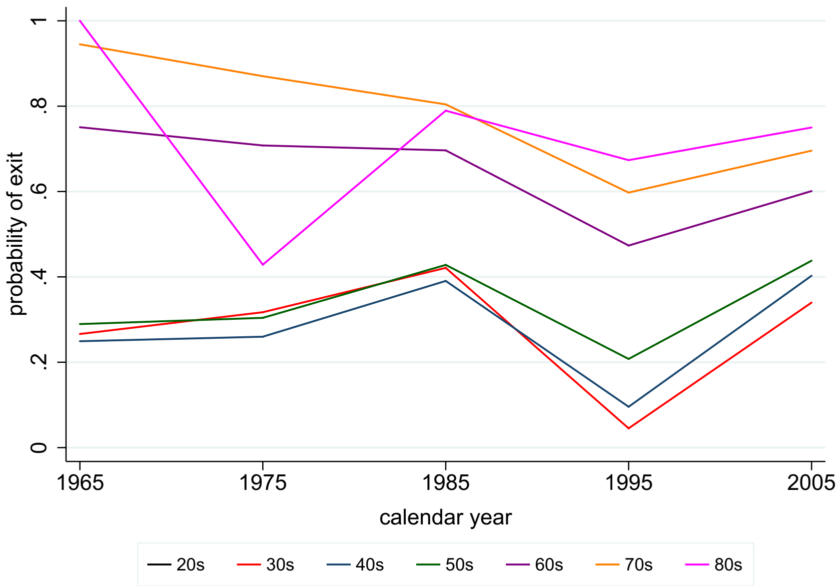
Notes: This Figure displays a sample entry from the 1971 ACS directory – see Figure 3.9 for a full page from the directory.

Figure 3.9: A sample page from the ACS directory



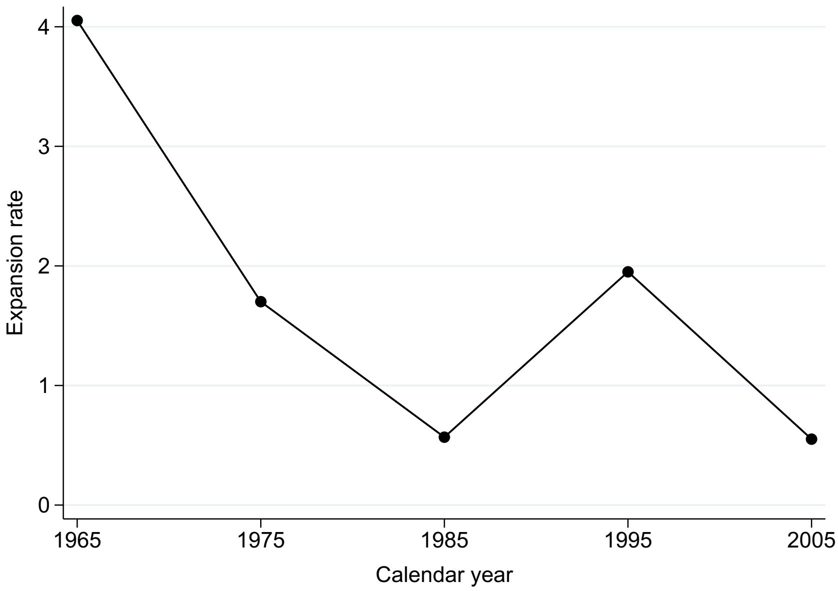
Notes: The Figure displays a sample page from the 1971 ACS directory. In the directory, departments list their faculty members in alphabetical order.

Figure 3.10: Age-specific probabilities of exits across time



Notes: This Figure shows how the probabilities of exits in each age group evolved over time. There is a clear trend only in groups at the ages of 60s and 70s.

Figure 3.11: Evolution of the expansion rate over time



Notes: This Figure shows how the ratio of the number of hirings to the number of exits evolved over time. There is a clear downwards trend.

Bibliography

Acemoglu, D., Akcigit, U., & Celik, M. A. (2014). Young, Restless and Creative: Openness to Disruption and Creative Innovations. *SSRN Electronic Journal.*

Alcacer, J., & Zhao, M. (2012). Local R&D Strategies and Multilocation Firms: The Role of Internal Linkages. *Management Science, 58*(4), 734-753.

Alcacer, J., & Zhao, M. (2016). Zooming In: A Practical Manual for Identifying Geographic Clusters. *Strategic Management Journal, 37*(1), 10-21.

American Chemical Society Directory of Graduate Research (1961 and subsequent editions).

Asheim, B., & Gertler, M. S. (2005). The Geography of Innovation: Regional Innovation Systems. In J. Fagerberg, D. C. Mowery, & R. R. Nelson (Eds.), *The Oxford Handbook of Innovation* (pp. 291-317). Oxford University Press.

Blau, D., Weinberg, B. (2017) Why the US science and engineering workforce is aging rapidly. *Proceedings of the National Academy of Sciences, 114*, 3879–3884.

Bush, V. (1945). *Science, the endless frontier. A report to the President*. Washington: U.S. Office of Scientific Research and Development

Cassi, L. & Plunket, A. (2012). Research Collaboration in Co-Inventor Networks: Combining Closure, Bridging and Proximities (MPRA Paper No. 39481).

Catalini, C., Fons-Rosen, C., & Gaulé, P. (2020). How Do Travel Costs Shape Collaboration? *Management Science, 66*(8), 3340–3360.

Catini, R., Karamshuk, D., Penner, O., & Riccaboni, M. (2015). Identifying Geographic Clusters: A Network Analytic Approach. *Research Policy, 44*(9), 1749-1762.

Coffano, M., Foray, D., & Pezzoni, M. (2017). Does inventor centrality foster regional innovation? The case of the Swiss medical devices sector. *Regional Studies, 51*(8), 1206-1218.

Daniels, R. J. (2015). A generation at risk: Young investigators and the future of the biomedical workforce. *Proceedings of the National Academy of Sciences, 112*(2), 313–318.

Delgado, M., Porter, M. E., & Stern, S. (2015). Defining clusters of related industries. *Journal of Economic Geography, 16*(1), 1-38.

Diamond, A. M. (1984). An economic model of the life-cycle research productivity of scientists. *Scientometrics, 6*(3), 189–196.

Dykeman, D. & Kopko, D. Recording Patent License Agreements in the USPTO. *Intellectual Property Today, 11*(8), 18-19.

Fischer, T., & Ringler, P. (2014). What patents are used as collateral? – An empirical analysis of patent reassignment data. *Journal of Business Venturing,29*(5), 633-650.

Freeman, R., & Van Reenen, J. (2009). What If Congress Doubled R&D Spending on the Physical Sciences? *Innovation Policy and the Economy, 9*, 1–38.

Fruchterman, T.M.J. & Reingold, E.M. (1991). Graph Drawing by Force-directed Placement. *Software - Practice and Experience, 21*(11), 1129-1164.

Gaessler, F. (2016). The Timing of Patent Transfers in Europe. *Enforcing and Trading Patents*, 105-145.

Galaso, P., & Kovářík, J. (2020). Collaboration networks, geography and innovation: Local and national embeddedness. *Papers in Regional Science, 100*(2), 349-377.

Galasso, A., Schankerman, M., & Serrano, C. J. (2013). Trading and enforcing patent rights. *The RAND Journal of Economics,44*(2), 275-312.

Ganguli, I., Gaulé, P., & Čugalj, D. V. (2022). Chasing the academic dream: Biased beliefs and scientific labor markets. *Journal of Economic Behavior & Organization, 202*, 17–33.

Gans, J., Hsu, D., & Stern, S. (2008). The Impact of Uncertain Intellectual Property Rights on the Market for Ideas: Evidence from Patent Grant Delays. *Management Science, 54*(5), 982-997.

Gaulé, P. (2014). Who comes back and when? Return migration decisions of academic scientists. *Economics Letters, 124*(3), 461–464.

Gaule, P., & Piacentini, M. (2018). An advisor like me? Advisor gender and post-graduate careers in science. *Research Policy, 47*(4), 805–813.

Giuliani, E. (2007). The Selective Nature of Knowledge Networks in Clusters: Evidence from the Wine Industry. *Journal of Economic Geography, 7*(2), 139-168.

Graf, H. (2011). Gatekeepers in regional networks of innovators. Cambridge *Journal of Economics, 35*(1), 173-198.

Granstrand, O., & Holgersson, M. (2020). Innovation ecosystems: A conceptual review and a new definition. *Technovation, 90–91*, 102098.

Hall, B. & R. Ziedonis (2001). The Patent Paradox Revisited: an Empirical Study of Patenting in the US Semiconductor Industry, 1979-1995. *RAND Journal of Economics,32*(1), 101-128.

Harhoff, D., & Wagner, S. (2009). The Duration of Patent Examination at the European Patent Office. *Management Science,55*(12), 1969-1984.

Harhoff, D., Rudyk, I., & Stoll, S. (2016). *Deferred Patent Examination.* Paper presented at CESifo Area Conference on Applied Microeconomics, Munich.

Hazır, C. S., LeSage, J., & Autant-Bernard, C. (2018). The role of R&D collaboration networks on regional knowledge creation: Evidence from information and communication technologies. *Papers in Regional Science, 97*(3), 549-567.

Hegde, D., & Luo, H. (2018). Patent Publication and the Market for Ideas. *Management Science,64*(2), 652-672.

Henkel, J., & Jell, F. (2010). Patent Pending – Why Faster Isn't Always Better. *SSRN Electronic Journal.*

Huber, F. (2012). On the Role of Interrelationship of Spatial, Social and Cognitive Proximity: Personal Knowledge Relationships of R&D Workers in the Cambridge Information Technology Cluster. *Regional Studies, 46*(9), 1169-1182.

Jackson, M. O. (2008). *Social and economic networks.* Princeton, NJ: Princeton University Press.

Jones, B. 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.

Jones, B. F. (2010). Age and Great Invention. *Review of Economics and Statistics, 92*(1), 1–14.

Jones, B. F. (2011). As Science Evolves, How Can Science Policy? *Innovation Policy and the Economy, 11*, 103–131.

Jones, B. F., & Weinberg, B. A. (2011). Age dynamics in scientific creativity. *Proceedings of the National Academy of Sciences, 108*(47), 18910–18914.

Kaiser, J. (2008). Zerhouni's Parting Message: Make Room for Young Scientists. *Science, 322*(5903), 834–835.

Ketels (2017), *Cluster Mapping as a Tool for Development*, Institute for Strategy and Competitiveness, Harvard Business School.

Li, G., Lai, R., D'Amour, A., Doolin, D. M., Sun, Y., Torvik, V. I., . . . Fleming, L. (2014). Disambiguation and Co-authorship Networks of the U.S. Patent Inventor Database (1975–2010). *Research Policy, 43*(6), 941-955.

Liegsalz, J., & Wagner, S. (2013). Patent examination at the State Intellectual Property Office in China. *Research Policy,42*(2), 552-563.

Lobo, J., & Strumsky, D. (2008). Metropolitan patenting, inventor agglomeration and social networks: A tale of two effects. *Journal of Urban Economics, 63*, 871-884.

London Economics (2010). *Economic Study on Patent Backlogs and a System of Mutual Recognition.* Newport: Intellectual Property Office. Concept House.

Lu, B. (2013). Expedited patent examination for green inventions: Developing countries' policy choices. *Energy Policy,61*, 1529-1538.

Marco, A. C., Graham, S. J., & Apple, K. (2015). The USPTO Patent Assignment Dataset: Descriptions and Analysis. *SSRN Electronic Journal.*

Mccrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics,142*(2), 698-714.

Mejer, M., & Van Pottelsberghe De La Potterie, B. (2011). Patent backlogs at USPTO and EPO: Systemic failure vs deliberate delays. *World Patent Information,33*(2), 122-127.

Merchant, M. A. (2015). *The America Invents Act (AIA) Handbook: A guide to the patent law reform of 2011.* Law Journal Press.

Mitra-Kahn, B. (2013). *Patent backlogs, inventories and pendency: An international framework.* Newport: Intellectual Property Office.

Nomaler, 'O., & Verspagen, B. (2016). River deep, mountain high: Of long run knowledge trajectories within and between innovation clusters. *Journal of Economic Geography 16*(6), 1259-1278.

O'Brien, M. (2017, February 02). Expediting Your Patent Application. Retrieved from: https://www.obrienpatents.com/expediting-patent-application/

Porter, M. (1990). *The competitive advantage of nations.* London: Macmillan.

Porter, M. (1998). *On Competition.* Harvard Business School Press.

Rallet, A., Torre, A. (2017). Geography of Innovation, Proximity and Beyond. In H. Bathelt, P. Cohendet, S. Henn, & L. Simon (Eds.), *The Elgar Companion to Innovation and Knowledge Creation* (pp. 421-439). Edward Elgar Publishing.

Rani, S., & Mehrotra, M. (2019). Community Detection in Social Networks: Literature Review. *Journal of Information & Knowledge Management, 18*(02), 1950019.

Rassenfosse, G. D., & Zaby, A. K. (2015). The Economics of Patent Backlog. *SSRN Electronic Journal.*

Rozenfeld, H. D., Rybski, D., Andrade, J. S., Batty, M., Stanley, H. E., & Makse, H. A. (2008). Laws of population growth. *Proceedings of the National Academy of Sciences, 105*(48), 18702-18707.

Sampat, B. (2012). *Examiner Citation Data* [Data file]. Retrieved from https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/18735

Serrano, C. (2008). The Dynamics of the Transfer and Renewal of Patents. https://doi.org10.3386/w13938

Serrano, C. J. (2010). The dynamics of the transfer and renewal of patents. *The RAND Journal of Economics,41*(4), 686-708.

Serrano, C. J. (2011). Estimating the Gains from Trade in the Market for Innovation: Evidence from the Transfer of Patents. *SSRN Electronic Journal.*

Stephan (2018) Verbal comment at NBER Summer Institute Science of Science meeting.

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.

Ter Wal, A. (2013). Cluster Emergence and Network Evolution: a Longitudinal Analysis of the Inventor Network in Sophia-Antipolis. *Regional Studies, 47*(5), 651-668.

Tong, T. W., Zhang, K., He, Z., & Zhang, Y. (2018). What determines the duration of patent examination in China? An outcome-specific duration analysis of invention patent applications at SIPO. *Research Policy,47*(3), 583-591.

Turner, L., Mairesse, J. (2003). *Individual productivity differences in scientific research: An econometric study of publishing of French physicists*. Paper presented at the Zvi Griliches Memorial Conference Paris, France.

Verspagen, B., & Schoenmakers, W. (2004). The spatial dimension of patenting by multinational firms in Europe. *Journal of Economic Geography, 4*(1), 23-42.

Whitt, M. R., Q.C. (2015, April 30). Expedited Examination Options for Patent Applications Filed in the United States. Retrieved from: https://www.bennettjones.com/en/Publications-Section/Updates/Expedited-Examination-Options-for-Patent-Applications-Filed-in-the-United-States

Whittington, K. B., Owen-Smith, J., & Powell, W. W. (2009). Networks, Propinquity, and Innovation in Knowledge-Intensive Industries. *Administrative Science Quarterly, 54*(1), 90-122.

Zahringer, K., Kolympiris, C., & Kalaitzandonakes, N. (2016). Time to patent at the USPTO: the case of emerging entrepreneurial firms. *The Journal of Technology Transfer, 43*(4), 923-952.