THE UNIVERSITY OF CHICAGO

TECHNOLOGY CLUSTER DYNAMICS AND NETWORK STRUCTURE

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE SOCIAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

KENNETH C. GRIFFIN DEPARTMENT OF ECONOMICS

BY
YULIA ZHESTKOVA

CHICAGO, ILLINOIS
JUNE 2021

To my parents, who went to great lengths to give me the best education possible.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am greatly indebted to my advisors Steve Davis, Ufuk Akcigit and Greg Kaplan for their continuous guidance, persistent help and support not only during my thesis writing and Job Market but throughout my whole graduate school career.

I would like to thank Fernando Alvarez, Mike Golosov, John Grigsby, Doug Hanley, Erik Hurst, Jeremy Pearce, Marta Prato, Robert Shimer, Gustavo de Souza, Chad Syverson, and Liangjie Wu, as well as participants in the Applied Macro Theory, Economic Growth and Macro Group workshops for their feedback and discussion. Your valuable comments and suggestions have truly improved this research.

I was surrounded by incredibly smart colleagues in the Department of Economics and Booth. I have learned a lot from my classmates and I was fortunate to make good friends over the last five years. We shared so many great moments together and I am proud to be a part of class 2016.

I am thankful to my brother Alex, who encouraged me to pursue Economics in High School and got me interested in innovations. None of this would be possible without endless love and unconditional support of my Mom and Dad, Lyudmila and Alexander. Thank you for always believing in me and making sure I have a cabinet full of Russian tea no matter what. The latter was essential fuel for finishing this dissertation.

Most importantly, I am eternally grateful to my wonderful husband John who has been my main support over the last years. His love, patience and encouragement gave me strength not only to make it through the most stressful years of the grad school, but to enjoy them. Thank you for keeping me happy.

ABSTRACT

In many models of endogenous growth, differentiated product lines are produced using improvable technologies. These technology clusters are central to innovation-led growth, but there is a lack of data-driven research on technology cluster dynamics. I introduce a new method of identifying technology clusters in the data based on a match between patent text and Wikipedia articles. A patent belongs to a technology cluster defined by the category of the Wikipedia article that is most similar to the text of the patent. I build a network of relations for technology clusters using links between Wikipedia categories. I apply these new data to the study of technology cluster network structure, entry and evolution since 1980. I show that young and small firms are more likely to discover new technology clusters by undertaking breakthrough innovations. However, it is mostly big and experienced firms that contribute to the growth of existing technology clusters. This growth is a result of reallocation of technology clusters from less productive to more productive firms. While these results hold for any given year, I find that the role of small and young firms as a source of technology cluster entry has declined in the last decades, in line with other measures of firm dynamics and competition documented in the literature. Reallocation of technology clusters among firms has also slowed down which explains the pervasive decline in the growth of technology clusters over time.

CHAPTER 1
TECHNOLOGY CLUSTER DYNAMICS AND NETWORK STRUCTURE

1.1 Introduction

Many theoretical models of innovation-driven growth heavily rely on the concept of a technology cluster – a collection of intellectual property that feeds into a product line of a firm. In these models, firms produce differentiated intermediate goods using improvable technologies and labor. Acemoglu and Cao (2015), Akcigit and Kerr (2018), Acemoglu et al. (2018) and Garcia-Macia et al. (2019) are only several examples of the extensive literature on multi-products firms that rely on innovation input from different technology clusters.

The engine of economic growth in these models is innovations undertaken by firms to improve technologies that are used in the production of their product lines. Minor innovations improve the quality of the firms' existing technology clusters – for instance, a semiconductor producer came up with a method of manufacturing smaller semiconductors, and a telecommunication company found a way to increase the speed of 4G transmission or a dairy farm invented a new pasteurization technique that extends the shelf life of milk. Some firms can draw a major "big-step" innovation which allows them to overtake other product lines because of this newly discovered technology cluster is superior to the existing technologies. For instance, a soda manufacturer came up with a new way of producing better-tasting bottled ice coffee at lower costs and became a big player on this market. Sometimes these major innovations are so radical that they establish new technology clusters that can replace existing technologies like introduction of nanorobotics in semiconductor fabrication triggered a new 5th generation of cellular networks. Sometimes these new technology clusters can even lay a foundation for a new product line that have not even existed before. For instance, invention of high-tech camera sensors lead to appearance of self-driving cars on the roads.

While technology clusters are central to growth in the theoretical innovation literature,

it is rarely the main focus of empirical studies. There is a lack of detailed datasets showing the dynamics of technology clusters, their entry and growth, firm-level and macro-level factors that affect a technology cluster's lifecycle, as well as research on the relations between technology clusters and their evolution over time. In this paper, I introduce a new method of characterizing technology reflected in firms patent applications and classifying this technology in a particular technology cluster. To identify technology clusters in the data, I use the mapping of inventions to business product lines of a firm that can be retrieved from patent data. I focus on innovating firms and treat their patents as a description of their innovation activity. I apply Natural Language Processing techniques to examine patent texts and categorize patents by topics – technology clusters. In order to define the topic of a patent text, I use Wikipedia articles as a "reverse" dictionary. For each patent, I find a Wikipedia article text that is the most similar to the text of the patent. The subcategory of the matched article then defines the technology cluster that this patent belongs to. This procedure allows me to capture a collection of technology clusters that each firm is operating in. At the same time, I also identify breakthrough innovations that establish new technologies – these are the patents that got matched with Wikipedia subcategory that had no patents having been matched to it before.

Using hyperlinks between Wikipedia subcategories, I build a network of technology clusters both on a firm level and on an economy level. While the Wikipedia links are static, arrival of patents to technology clusters determines dynamic evolution of the network. A technology cluster remains inactive until a firm submits a patent related to this cluster. The patenting dynamics in relevant technology clusters drive the evolution of the network both in terms of its size and in terms of its structure. This new methodology allows me to study how and why new technology clusters enter, their life cycle after creation, and the evolution and importance of technological interdependencies since 1980 till 2012.

The main application of the new data product is threefold. First, I study entry of a new technology clusters and in particular, types of innovations and firms that establish new

technologies. I find that patents that lay the foundation for a new technology cluster receive more citations, are based on more recent innovations and fundamental research and have higher market value. Young and small firms are more likely to discover a new technology cluster by undertaking breakthrough innovations. While the size of a firm is determined by its employment, I measure the age of a firm by years since its incorporation as well as by its patenting tenure, i.e. years since its first patent application. Moreover, firms that have just entered innovation market are twice more likely to open a new technology cluster with their first patent.

However, the role of young and small firms as a source of technology cluster entry has declined in the last decades. In 1980, 61% of firms in newly established technology clusters were firms that have just entered innovation market, while in 2010 this share fell to 24%. I also document a decrease in the entry of new technology clusters after 2000, in line with other measures of firm dynamics and competition documented in the literature. The broad fall in business dynamism dates back to the early 1980s as documented by Davis et al. (2007), Davis et al. (2012), Hathaway and Litan (2014), Decker et al. (2016b), Pugsley and Şahin (2019) among many others. Innovating STEM industries used be to an exception from this secular decline in new firm formation and job reallocation until 2000. After 2000, high-tech sector exhibits decline in new business formation and slowdown in entrepreneurship (see Haltiwanger et al. (2014), Decker et al. (2016b) and Decker et al. (2016a)).

While the role of incumbents in establishing new technology clusters is increasing over time, I also look into the contribution of firm's expertise in relevant technologies to laying the foundation for a new technology cluster. Using the global and firm-level network of technology clusters, I find that these incumbent pioneers tend to have prior experience in innovating in technologies related to the new technology cluster. Incumbents from the satellite industries that share a complementary knowledge base with a newly discovered technology cluster are 73% more likely to become pioneers in this new cluster. Conditional on firm's patenting tenure and size, relevant innovation experience in related industries

is conducive to undergoing breakthrough innovations that found a new technology cluster, which suggests the importance of within-firm knowledge spillovers for expanding the network of technologies.

Second, I study the post-entry dynamics of technology clusters. While young and small firms are important for discovering new technology clusters, it is mostly big and experienced firms that contribute to the growth of existing technology clusters. Most of technology cluster growth is accounted by net entry of firms in a cluster and in particular, by reallocation between continuing firms who switch to a technology cluster that they have not patented before. In fact, about 2/3 of firms innovating in a particular technology cluster are newcomers to this cluster. The reallocation forces are quite strong: a 10 percentage point increase in the entry rate of firms to a technology cluster improves the cluster's growth rate by 50%. I show that growth of a particular technology cluster is principally driven by innovation activity moving from less-productive firms to higher-productive firms measured by quality and quantity of their invention portfolio. A change in the leading cluster incumbent (i.e. a firm with the highest share of accumulated patents in the technology cluster) is associated with a 22 percentage point increase in the cluster's growth rate.

While there is a decline in the entry rate of new technology clusters in the recent decades, I also document a persistent slow-down in their average growth rate conditional on age. This is related not only to a declining business formation in high tech sector but also to a decreasing reallocation rates among incumbent firms. Moreover, the contribution of net entry to technology cluster growth is fading over time. Together with an increase in firm's business concentration that we have observed in the last decades, this indicates that within-firm evolution of a technology cluster is becoming a dominant source of technological growth as reallocation forces between firms are getting weaker.

Third, I study the role of innovation market concentration for technology cluster growth and network formation. I use the dynamic network of technologies to explore how firms concentration relates to the position of a technology cluster in the global network. I measure

a technology cluster's innovation concentration as Herfindahl-Hirschman Index of patent shares of each firm that is active in this technology cluster. I find that technology clusters with less concentration of innovating firms tend to both grow faster in size and become more central as a result of getting connected to more satellite clusters. This result holds both across and within technology clusters and implies that besides reallocation to the most productive incumbent, competitive efforts of firms are also important for technological development.

Interestingly, the negative relation between a technology cluster's patenting concentration and its centrality growth breaks after 2000. The decline in firm dynamism relates to the decline in firm competition and through the latter can affect overall slow down in technology cluster dynamics. But the fact that this mechanism is not as pronounced after 2000 suggests that the nature of R&D may have changed in the recent decades. Moreover, as the within-firm component is becoming more important for technological growth over time, it implies that breakthrough innovations now require more specialized "in-house" production. The increase in concentration can be a feature of the new equilibrium in the economy where innovations are becoming more complex and fixed costs of research and development are high. Indeed, I show that after controlling for the changing costs of innovation, higher concentration of firms in a technology cluster is actually associated with higher degree growth (i.e. number of connections that a technology cluster has). Together with the fact that incumbent firms are more likely to discover a new technology cluster if they have patenting experience in its satellite clusters implies that conglomeration spillovers are crucial for innovation production and we should expect their role to become even bigger.

The rest of the paper is organized in the following way. A substantive literature review is deferred to Appendix C. However, each section of the empirical analysis has a brief discussion of the related literature and puts my analysis in the context of previous research. The following section provides a detailed discussion of the methodology and data. Section 1.3 explores the time series patterns of the changing life cycle of technology clusters and summarises the main time trends. I then turn to the cross-sectional analysis of technology

entry and growth. Section 1.4 looks into the entry of new technology clusters. Section 1.5 focuses on the importance of relevant experience in related industries for discovering a new technology cluster. Section 1.6 talks about technology cluster growth and the importance of reallocation between firms. Section 1.7 looks into the role of market concentration for dynamics of technology clusters. Section 1.8 concludes.

1.2 Methodology

The central methodological contribution of this paper is a construction of the match between patents and Wikipedia articles using state-of-the-art methods of Natural Language Processing. This method allows to identify technology clusters of innovating firms in the data and study their dynamics and network structure from 1980 till 2012. The main idea is to use Wikipedia as a "reverse" dictionary to identify the topic of a patent. This computational exercise can be summarized in the following way: we want to machine-read patent text, compare it with the texts of all potentially relevant Wikipedia articles and pick the one which text is most similar to the text of the patent. Then the topic, i.e. the category, of this matched article defines the technology cluster that the patent of interest belongs to. Using the links between Wikipedia categories, I also build a network of relations for technology clusters. This section elaborates on the computational part of this exercise, describes the input datasets and the final data product. Subsection 1.2.1 focuses on the description of USPTO patent database and Wikipedia data, Subsection 1.2.2 introduces text data analysis techniques and describes the measure of text similarity with its application to merging patents to articles. Subsection 1.2.3 discusses main properties of the resulting merge. Subsection 1.2.4 describes the construction of technology cluster network. I discuss potential alternative ways to define technology clusters in the data and talk about the benefits of my approach in Appendix D.

1.2.1 Data

Data on patent applications, including patent titles, year of application, class, firm that the patent was assigned to (assignee) and citations given and received by patents were retrieved from U.S. Patent and Trademark Office (USPTO) PatentView database.[1] USPTO also publishes the full text of patent abstracts. I specify a patent's date to be the year of application, as this most closely reflects the invention's discovery date. I consider only granted patents with application year from 1974 to 2015. The data set includes a name and address of a company that each patent is assigned to, but firm names often have misspellings and the same firm can have different variants of its name under different patents (e.g. with "Inc." at the end of the name and without). The algorithm proposed by Hall et al. (2001) helps to identify unique companies by cleaning and standardizing companies names.[2]

Complete and most up-to-date Wikipedia data in dump files are free and available to download.[3] I use November 2019 version of the files. All data related to Wikipedia articles, including some metadata regarding edits timestamps and conversations among articles' moderators and editors is zipped into less than 100GB XLM files. Such a relatively small size makes parsing of text files quite manageable. The size of a Wikipedia article is determined by principles of readability, editing and technical issues (e.g. compatibility with mobile browsers). Most of the articles have from 4,000 to 10,000 words in readable prose.[4] If an article exceeds this upper bound, it is flagged for potential division – when an article is too large it is split into smaller articles or some of its parts get merged with other existing articles. If an article is too small and has not been growing in size for a while, it is merged with one or several existing articles. The decisions on splitting and merging articles require

1. See http://www.patentsview.org (accessed 11/10/2019).

2. I also use cosine similarity between TF-IDF values with 4-grams to identify same companies among assignees. More details on this method is in Subsection 1.2.2 and on my Github page.

3. See https://dumps.wikimedia.org (accessed 12/05/2019).

4. Readable pose is the amount of viewable text in the main sections of the article, not including tables, lists, or footer sections. See https://en.wikipedia.org/wiki/Wikipedia:Article_size for more details on the borders of Wikipedia articles.

editorial consensus of moderators. More on the principals and guidelines that Wikipedia articles are based on can be found in Appendix E.

To create a network of technology clustered projected from network of Wikipedia articles, I use data on links between Wikipedia subcategories and categories.[5] All Wikipedia articles are grouped into subcategories, which have their own page with the description.[6] Subcategories are then grouped into categories, which are grouped into a parent categories and so on. Each subcategory can appear in several categories. The top of the categorization scheme are 42 classes. Figure F1 in Appendix F shows a concrete example of one branch of such classification. Wikipedia has a very rich categorization structure that does not form a strict hierarchy or tree. Subcategories can share a link either because they belong to the same category (or their categories belong to the same parent category) or because they are closely related to each other topic-wise but not in a subset relation, like "Oven" and "Fireplace". This feature makes it easier for a reader to navigate through articles and easily find other related pages. I defer a more detailed discussion of Wikipedia categorization to Appendix F.

1.2.2 Matching Patents with Wikipedia Articles

Wikipedia has around 6 million articles, most of which will be irrelevant to the final goal of this exercise – identifying technology clusters of patents. To make the match computationally manageable, I follow a three-step procedure to select potentially relevant articles. I defer a detailed description of this procedure to Appendix G. Many Wikipedia articles are very granular and finely-defined. This can be a problem since we do not want technology clusters to be firm-specific or too narrow.[7] Otherwise, we simply will not be able to capture any dynamics within a cluster, nor would there be any between-firm reallocation of technology

5. SQL files can be downloaded from https://dumps.wikimedia.org/enwiki/ (accessed 12/21/2019)

6. Editors can suggest a relevant subcategory of an article while moderators revise the decision.

7. For instance, while there is a Wikipedia article on iPhone and Galaxy Note, the two are brand-specific and cannot be definitions of technology clusters.

cluster (or the same of the clusters with non-zero between-firm dynamics would be biased). To avoid this over-specificity, I am using a subcategory that a given article belongs to as the defining technology cluster of the matched patent rather than the article title itself.[8] After selecting only potentially relevant articles for the match with patent data, I end up with 18,271 subcategories that can be matched with about 4 million patents.

The main idea behind the procedure of matching patents to Wikipedia articles is to come up with metrics that will allow us to compare similarity of two texts. In a nutshell, we want to compare the text of each patent with all Wikipedia articles and find the article which text is most similar to the patent's text. I summarize this procedure in broad strokes below and defer the detailed description to Appendix H. First we need to clean the text of the patents and articles in preparation for the match. This includes elimination of the so-called "stopwords" that leaves us with only meaningful words that convey the content of a text. I also lemmatize words to bring them to their most standard form. For both patents and Wikipedia articles, I leverage the fact that the title of the text document is much more informative than any other phrase in the text. I upweight words in the title of the patent's text by a factor of five and put them at the beginning of a patent's abstract. The same is done for Wikipedia articles and their titles.

My approach to matching patents text to Wikipedia articles text is similar to the procedure described in Kelly et al. (2018) and Argente et al. (2019). As previewed above, the match is based on computing pair-wise text similarity between one text and another. This exercise boils down to vectorization of the texts and computing similarity score of the two corresponding vectors. I transform each cleaned patent text and Wikipedia article text into a vector of size N where N is the number of all unique words from all patents and Wikipedia articles combined. Element i of a vector is equal to one if a word token i is present in the text document that this vector corresponds to and equals zero otherwise. After appending

8. For instance, in the case of iPhone and Galaxy Note, the subcategory of these articles that defines the corresponding technology cluster is "Smartphone".

all M patents and Wikipedia articles into one matrix, we end up with a very sparse matrix of dimension $N \times M$.

Even among meaningful words, relative importance of each word in a text for its content varies a lot. Words that frequently appear in a given text document are more informative than those that appear only once. At the same time, words that are frequent in all text documents are not that informative. In order to take this into account, I weight words by their contribution to a document's content using "term-frequency-inverse-document-frequency" (TF-IDC) transformation of word counts:

$$\omega_{ij} = TF_{ij} \times IDF_i \tag{1.1}$$

where ω_{ij} is weight of word i in document (vector) j. The first item, TF_{ij} is frequency of word i in document j. The second item, IDF_i is inverse frequency of word i in all documents:

$$TF_{ij} = \frac{c_{ij}}{\sum_i c_{ij}} \tag{1.2}$$

$$IDF_i = log\left(\frac{M}{\sum_j \mathbb{1}\{i \text{ in } j\}} + 1\right) \tag{1.3}$$

where c_{it} is number of times that word i appears in a document j and $\sum_i c_{ij}$ is the length of the document measured in words. After adjusting the weights of each vector's elements according to the TF-IDF measure, I normalize them to have unit length.

I measure the similarity score of two texts as cosine similarity between the two corresponding vectors. This similarity score lies in the interval between 0 and 1, where 0 indicates no similarity at all and 1 means the two texts are identical. Then I match each patent text to Wikipedia article that has the highest similarity score with this patent conditional on passing a 30% similarity threshold.

1.2.3 Main Properties of the Final Match

The resulting data product is a sample of 1,122,220 patents that were successfully matched to one of the Wikipedia articles. The articles with a patent pair belong to 6,461 unique Wikipedia subcategories that define technology sectors. For instance, patent #6427584 with a title "System and method for processing citrus fruit with enhanced oil recovery" is matched to subcategory "Citrus production" and patent #8360885 "System and method for using a game to interact with television programs" is directed to "Education television" in Wikipedia. The merge between patents and articles is many-to-one: each patent is matched to exactly one article, while an article can have many patents matched to it. The quality of the match is fairly consistent across years and sectors. On average, slightly more than a quarter of new patents submitted every year is matched to one of the Wikipedia subcategory. More on the properties of the patent-Wikipedia match can be found in Appendix I.

The final sample includes 101,843 firms with more than one patent in their accumulated life-time portfolio. Firms with matched patents in my sample are slightly bigger and have higher tenure than firms in the USPTO universe, but this difference is not significantly big to question representativeness of the dataset in hands. Table 1.1 reports some key summary statistics that describe the final sample. On average, an article subcategory (and the corresponding technology cluster) has 173 patents matched to it, but the distribution is skewed to the left. For a median firm, I am able to match 70% of its patents to Wikipedia articles. Such high share of matched patents for an average firm implies that the selection of the observations in the final sample happens on a firm level rather than independently across firms. Thus, if the algorithm matches one patent of a firm, it is likely to match the others too.

1.2.4 Technology Cluster Network Construction

Once we match firms' patents to Wikipedia articles subcategories, we can identify a set of technology clusters that the firm innovates on. But not all technology clusters are com-

Table 1.1: Summary statistics for the final patent-Wiki article matched sample

	Mean	St.Dev.	5%	25%	50%	75%	95%
#patents per cluster	173.3	606.6	1	3	16	104	811
Firm's patent portfolio size	49.8	752.6	2	2	4	9	84
Share of matched patents	.71	.31	.25	.5	.7	1	1
#unique clusters per firm	5.2	27.48	1	1	2	3	13
Cumulative patenting tenure	13.1	83.9	2	4	9	16	33
Technology cluster's lifespan within a firm	3.9	6.0	1	1	1	4	17

Notes: #patents per cluster captures how many patents are matched to an article subcategory that defines a technology cluster. Firm's portfolio size is number of firm's patents accumulated over whole lifespan of the firm. Share of matched patents shows what share of firm's patents is matched to some article's subcategory. #unique clusters per firm is number of unique subcategories that are matched to firm's patents. Cumulative patenting tenure is the difference between the year of the last firm's patent and the first firm's patent. Technology cluster's lifespan within a firm is the difference between the year of firm's exit and entry into the technology cluster defined by the article subcategory.

pletely independent. Dynamics of a technology cluster may be relevant for the evolution of another technology cluster if they are related to each other. Using links between Wikipedia subcategories and categories, I build a network of relations for technology clusters on the economy level and for each individual firm. Subcategories can be connected to each other either because they belong to the same category, their categories belong to the same parent category or because they are closely related to each other topic-wise (more on this in Appendix F). These links are the main skeleton of the edges between technologies that lay the foundation for the technology cluster network.

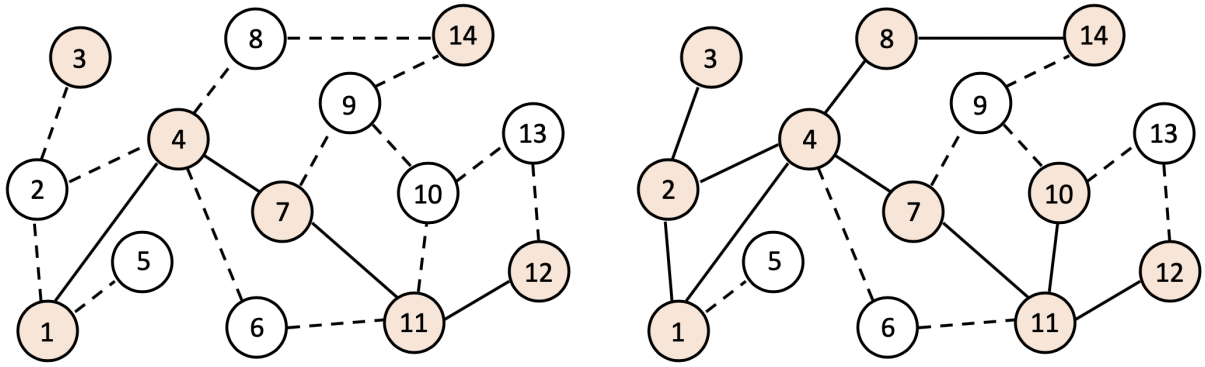
The connections between nodes in this network have different weights. If two Wikipedia subcategories refer to each other firsthand, the weight of this connection is 1. If subcategory A is linked to subcategory B and subcategory B is linked to subcategory C, then A and C share an edge with weight 1/2. If instead of one subcategory B there are two subcategories between A and C (i.e. A refers to X, X refers to Y and Y refers to C), the weight of the edge (A,C) is 1/3.[9] The resulted weighted network is undirected, meaning that all connections are mutual. Following the literature on graphs, I define the degree of technology cluster to

9. Connection by "two handshakes" is the maximum distance allowed in the network. Visual inspection concludes that technology clusters that are connected through more than two other clusters are most of the time not actually related to each other in any economic way.

be the number of connections/neighbours it has. A cluster's degree captures how central and hence important for the structure of the network this technology is. The average number of connections is 7.81, but the distribution is quite skewed with median equal 1.

Wikipedia links allow me to construct only a static network of technology clusters at the time when the Wikipedia data is retrieved. However, even though a set of technology clusters is fixed and the upper limit of their degree is predetermined, the number of *active* technology clusters as well as the number of their *active* neighbors depends on the time-changing patenting activity. The network evolution is determined by the dynamic of patents that are submitted to technology clusters every year. Figure 1.1 shows an artificial network consisting of 14 technology sectors that I will use as an example. Each technology cluster is a node in this network and it is inactive or "muted" until at least one patent belonging to this technology cluster is submitted. Once a firm starts innovating in a cluster, it becomes active.

Figure 1.1: Example of a network consisting of 14 technology clusters



Panel A: Time t Panel B: Time $t+1$

Notes: Orange nodes are technology clusters that are active in a given year because they have received patents. These active nodes determine the size of the technology cluster network. Bold links indicate connections between active clusters. The degree of a technology cluster is number of active neighbors that the cluster has – number of bold links.

Every year, some technology clusters will be active (i.e. the ones that have patents submitted to them) while the others will stay inactive. In the example from Figure 1.1, only 7 technology clusters received patents in year t and as a result, only clusters 1, 3, 4, 7, 11, 12 and 14 are active in this year. The actual technology network is defined by *active*

technology clusters, which implies that it will differ from year to year. In year t, the example network have 7 technology clusters, while in year $t+1$ it has 10 clusters. As a result, a degree of a technology cluster is also time-dependent. In our example, technology cluster 4 can potentially have 5 connections but since in year t only two of its satellites are active, its degree in this year is 2. In year $t+1$, its degree increases to 4 as two of its potential connections enter the economy and become active.

Some technology clusters have more patents than others. The *stock* of patents that is submitted to a technology cluster *by* a certain point in time is the cluster's size, which also evolves over time. To put some context into this discussion, Table B1 in Appendix B reports the largest technology clusters by patenting and technology clusters with the highest degree – number of all possible connections in the static network. Note that these two lists are very different – technology clusters with highest degree are those that are more general and thus related to many other technologies, but they are not necessarily the largest in terms of size.

The relationship between technology clusters can have upstream or downstream nature. It can be defined by technology clusters substitutability and complementarity. The latter in its way can characterize the links between technologies either based on the product space or innovation knowledge space. Regardless of the nature of their relationship, we say that connected technology clusters are more relevant for each other's growth and development than those technologies that are not connected. This is the nature of the network I am constructing and using in the following analysis. Later on, when we turn to the discussion of the importance of relevant industry experience, I am going to identify and focus on a particular type of inter-cluster links with a specific economic meaning.

1.3 The Changing Life Cycle of Technology Clusters

In this section, I will take a first step in studying the life cycle of technology clusters by exploring how technology cluster entry, exit and growth rate have changed over time. In particular, I am looking at the expansion of the global network of technologies from 1980

to 2012 and explore growth of an average technology cluster's size and degree. I capture the main time trends of technology cluster entry and exit, compare them with the entry rate of innovating firms and look at the predictive power of size and age for a technology cluster survivorship. This macro view on the technology clusters evolution will help us to understand the key features of the data and serve as an introduction to the cross-sectional micro-level study of technology cluster dynamics and network structure in the following parts of the paper.

Since entry and exit are one of the central concepts in this paper, I start with introducing the definitions of entering, active and exiting technology clusters. An obvious way to define a year of technology entry would be the year of the fist patent submitted to this technology cluster. However, there are many cases in the data when a firm would submit a patent related to a new technology cluster but there is no patenting activity in this cluster in the following several years. If the next patent appears more than 5 years after this "first try" and only this second invention is followed by an actual innovation growth in the technology cluster, we should consider this second patent as an actual breakthrough invention in the technology cluster rather than the very first attempt that did not trigger any further innovation activity.

For instance, in February 1981, NASA filed a patent application "Digital interface for bi-directional communication between a computer and a peripheral device" that technically established a new technology cluster – *USB*. This patent described a new device for data transmission from a compute to a peripheral file storage. Even though it laid a ground for the USB technology, it took several years for this technology cluster to start growing, after M-Systems proposed an actual prototype of a portable USB flash drive in their patent "Architecture for a USB-based Flash Disk" in 1999. IBM jumped on the bandwagon in the same year and became the biggest producer of USB flash drives (under ThumbDrive trademark) in US. In this case, we would consider 1999 as the year of USB technology cluster entrance, not 1981.[10]

10. USB remained an active technology sector till the end of the data sample and have over 200 different

Formally, I define the year of a technology cluster's *first entry* as 1) the year when the first patent related to this technology is submitted *and* 2) there is positive patenting in this technology cluster in the following 5 years. The year of a technology cluster's *exit* is the year of its last patent's submission after which there are no patents related to this technology cluster in more than 5 years. Once a technology cluster exits it becomes inactive again but this state is not absorbing and the technology can re-enter again. The year of a technology cluster's *re-entry* is the application year of the first patent related to this technology, with no patents in the last five years and positive patenting in this technology cluster in the following 5 years.[11] Throughout the whole paper, the concept of *entry* includes both *first entry* and *re-entry*. Naturally, a *continuing* technology cluster has patents submitted to it in the last 5 years and in the next 5 years (i.e. time between entry and exit). A technology cluster is *active* in a given year if we observe positive patenting in this cluster in the last 5 years.[12]

Through the paper, I am focusing on years from 1980 to 2012, even though the final data product of patents matched to Wikipedia articles runs from 1974 to 2015. Truncation of the first six and the last three years of the sample is necessary to avoid the bunching of entry and exit caused by sample beginning and end.[13] As discussed before, the technology network in a given year consists only of the technology clusters that are active in this year. This global network of technology clusters changes a lot over time as technologies enter and exit. Figure 1.2 captures the state of the network in 1980 and 2010. The size of each node is the share of patents submitted to the corresponding technology cluster in a given year, with the top-20 technology clusters labeled.

This first glance at the data implies that technology network is becoming more dense

assignees that has contributed to its growth.

11. For instance, there is a patent in a technology cluster submitted in year t and the last patents in this cluster were no later than $t-6$. If we also observe positive patenting in this technology cluster from t to $t+5$ then year t is the year of the technology's re-entry.

12. Note that exiting technology clusters are also consider active.

13. For instance, we would see abnormal entering of technologies in the first years of observations because the pretrend is not observable. By the same token, there would be a mechanically high exit rate in the last years due to the end of the sample.

Figure 1.2: Global network of technology clusters in 1980 and 2010.

Notes: The size of each node is the relative size of the corresponding technology cluster measured as share of patents in this cluster submitted in 1980 or in 2010 relative to all patents submitted in 1980 or in 2010. Brightness of a node color indicates the degree of the node. The largest technology clusters are labelled. The network captures only active technology clusters (entering, continuing or exiting).

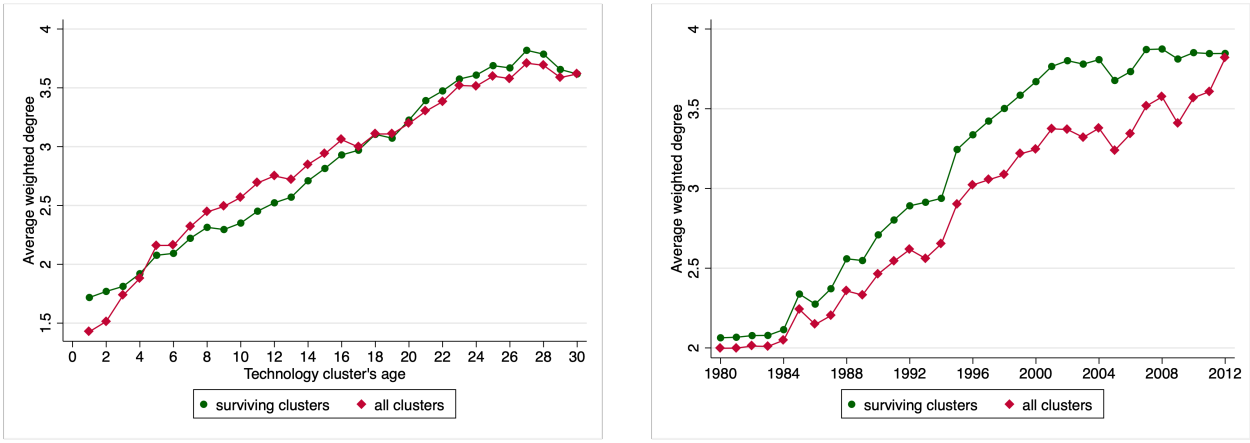
both in terms of numbers of connections and in terms of number of active clusters. We can also notice that the size of nodes is smaller in 2010 than it is in 1980. Indeed, Figure A1 in Appendix A shows that the concentration of patents in technology clusters is decreasing over time as share of patents belonging to an average technology is decreasing. Figure A2 in Appendix A shows the evolution of size (number of active technology clusters) and order (number of all unique connections between technologies) of global technology cluster network over time. There is a steady growth both in number of active technology clusters and in number of links between the active clusters up until the Great Recession, but after 2008 both of these indicators drop precipitously.

The growth in the number of total edges in the network can be a mechanical consequence of an increase in total number of nodes. However, an average technology cluster is getting more connections both over time and with age as depicted in Figure 1.3. The positive trend in an average degree of a technology cluster holds both for the whole sample and for the survivors subsamples. The red line uses the whole sample of technologies, while the green line takes care of a potential composition bias and focuses only on the sample of technology clusters that have been active for at least 30 years.[14] However, there is a difference between having 5 connections in a network of size 10 and 5 connections in a network of size 100. Since the size of the network is also growing over time, Figure A4 in Appendix A takes care of this concern and shows the dynamics of an average degree centrality – degree of a node divided by the network size. The size-adjusted degree is also increasing with a technology cluster's age and over time, implying that the technology network is becoming more connected not only as a result of an increase in number of active technology clusters but also because the existing technologies are getting more connected to each other creating groups of clusters.

While a cumulative number of technology clusters is going up over the years, annual entry of new technologies does not show such a linear positive trend. Figure 1.4 tracks

14. A median technology cluster lives for 28 years considering observations from 1974 to 2015. I define technologies that are active for more than 30 years as long-run survivors.

Figure 1.3: Evolution of an average degree of a technology cluster over its age and over time.



Notes: Survivors are technology clusters that stay active for more than 30 years.

the dynamics of new technology clusters entering economy every year since 1980 till 2012. Before 2000, the number of new entrants is increasing, but after 2000, the trend reverses to a sharp decline. We can also capture this change in new technology formation by looking at the dynamics of the entry *rate* of new technology clusters rather than the absolute values. This trend is not driven by any particular sector of technology clusters – entry rate of new technologies is decreasing in all sectors (Figure A3 in Appendix A).

Figure 1.4: Technology cluster entry over time

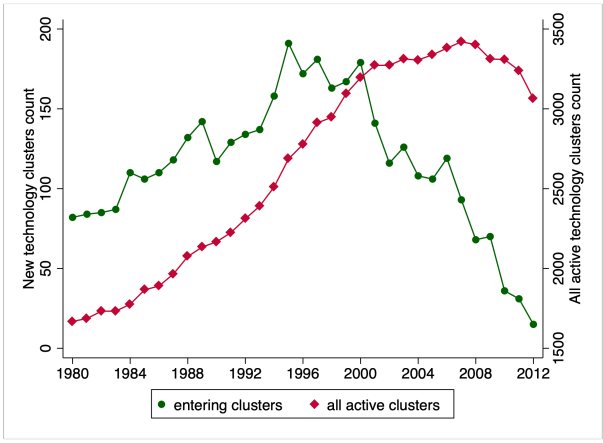
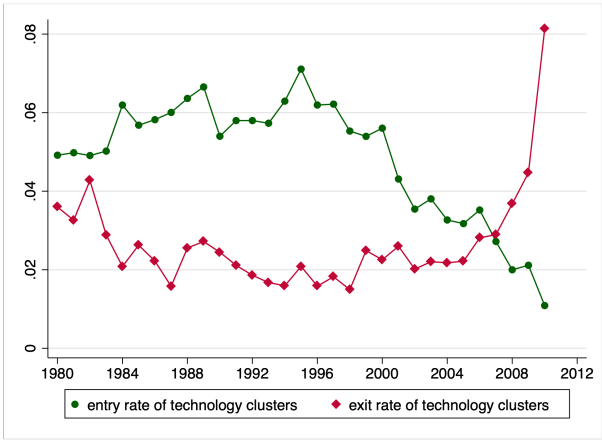


Figure 1.5: Technology cluster entry and exit rates over time

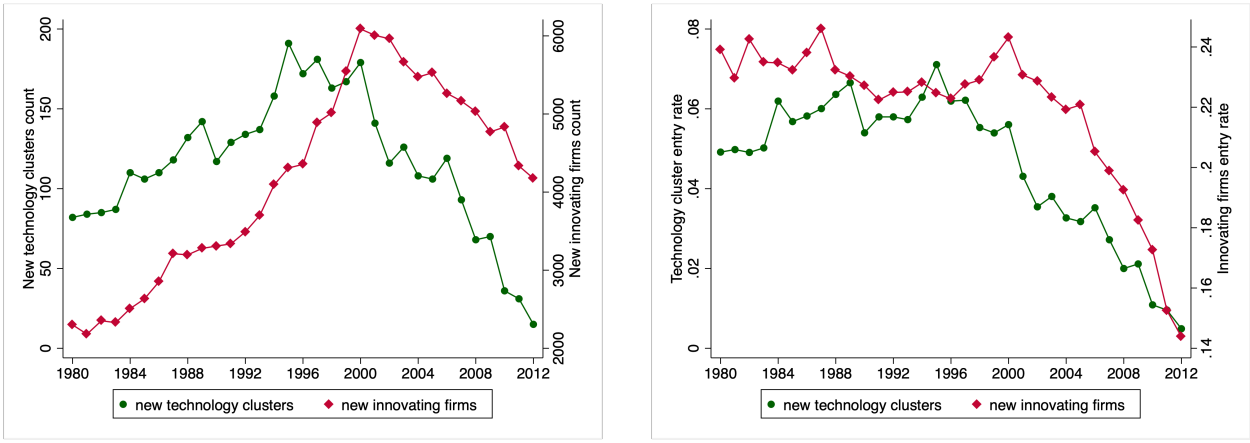


Notes: Active technology clusters are entering, continuing or exiting clusters in a given year. Entry/exit rates are computed as number of entering/exiting technologies divided by number of active technologies in this year.

This changing trend in technology cluster entry is in line with other measures of firm

dynamics documented in the literature. While a secular decline in business formation and reallocation dates back to the early 1980s, high-tech sector used to be an important exception from this trend.[15]. But after 2000, innovating STEM industries started to exhibit the same alarming patterns of a declining new business formation and entrepreneurship.[16] I replicate this result by looking at the number of new innovating firms over time. Figure 1.6 shows that the number of firms entering innovation market is decreasing after 2000, both in absolute values and in rate. This figure implies that new business formation in high-tech sector and new technology formation are strongly correlated. I am going to elaborate on this relationship and explore its micro foundation in Section 1.4.

Figure 1.6: Technology cluster and innovating firm formation over time



Panel A: Counts Panel B: Rates

Notes: Left figure shows the evolution of number of new technology clusters (left vertical axis) and number of new innovating firms (right vertical axis) over time. New innovating firms are firms with the first patent application submitted in a given year in the sample of USPTO assignees with more than one patent application. Right figure shows the entry rate of new technology clusters and new innovating firms. For the former, it is the ratio of new technology clusters over all active technology clusters. For the firms entry rate, it is the ratio of firms with the first patent application relative to all firms with a patent application in a given year.

Perhaps even if there are less new technology clusters getting discovered every year, the existing ones has showed a faster growth in the recent decades. Unfortunately, that is not

15. See Davis et al. (2007), Davis et al. (2012), Hathaway and Litan (2014), Decker et al. (2016b), Pugsley and Şahin (2019) for a more detailed discussion on the decline of business dynamism in US

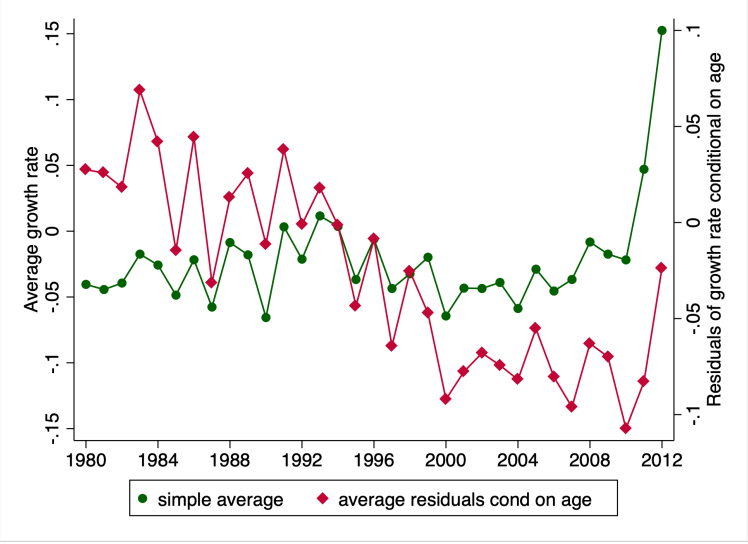
16. (Haltiwanger et al., 2014), (Decker et al., 2016b) and (Decker et al., 2016a) discusses why decline in high-tech sector dynamism is especially concerning.

the case. Figure 1.7 plots average growth rate of technology clusters over time, measured by the growth rate of the share of annual patents related to this technology:

$$growth\_rate_{ct} = \left(\frac{patents_{ct}}{patents_{t}}\right) / \left(\frac{patents_{ct-1}}{patents_{t-1}}\right) - 1$$

where c is a sector and t is a year. Unconditional on a technology cluster's age, average growth rate is basically constant throughout the sample, with some acceleration after 2010. However, if we take out the variation that can be explained by technology cluster's age, we see that on average, technology clusters are growing at a decreasing rate over time. This implies that a decline in the technology entry rate may not only affect the reallocation of growth from young technology clusters to mature ones, but can also slow innovation growth in aggregate. The decrease in technology entry and the slow-down in average growth rate of technology clusters may have the same fundamental source but may also have quite different forces standing behind them. The following sections of the paper are trying to solve this puzzle.

Figure 1.7: Growth of technology clusters over time



Notes: Green line (left vertical axis) plots average growth rate of a technology clusters among all the clusters that are active in a given year. Red line (right vertical axis) plots the residuals of the same measure after regressing growth rate on a technology cluster's age.

The literature on firm dynamics consider firms' age and size as two of the main predictors of firms' growth and survivorship. I now turn to the discussion of technology survival rate and explore whether age and size have the same predictive factor of survivorship for technology clusters as they do for firms. Figure 1.8 shows that survival rate of a technology cluster is increasing with its size.[17] This result resonates with the empirical facts on firm's survival rate (for instance, see the canonical paper by Dunne et al. (1989)).[18] Figure 1.9 plots the survival rate of a technology cluster depending on its age. Interestingly, the survival rate is mostly flat throughout a cluster's age and is decreasing for old technologies. This implies that technology clusters that have been around for a while are more likely to have obsolete innovations and thus are more likely to be replaced by new superior technologies. Note that the patterns of firm survival rate are quite different in this respect. Firm's age is one of the main predictor of its survivorship.[19]

To conclude, I summarize the main empirical facts that has been discussed in this section:

1. The technology cluster network is growing both in terms of number of nodes and number of unique edges until Great Recession when the trend reverses.

2. The average degree of a technology cluster is increasing both with its age and over time.

3. The number of new entering sectors increases till 2000 and drops precipitously afterwards.

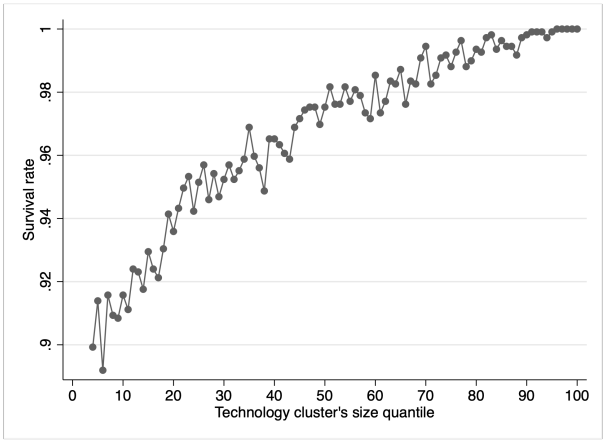
4. The time trends of technology cluster entry strongly correlate with the trends of inno-

17. Size is measured as number of accumulated patents in a given technology cluster.

18. It also helps us understand the nature of my method to characterize and classify firms' innovation activity into clusters. If the boundaries of a technology cluster were determined by its size rather than by similarity of innovation in the same group then at least starting from a certain size, the survival rate of a cluster should be decreasing. Technology clusters that are too big would have to split into several new clusters to maintain the predetermined boundaries. The fact that we do not see such patterns implies that the suggested identification and the resulting definition of technology clusters is not just nominal and size-dependent.

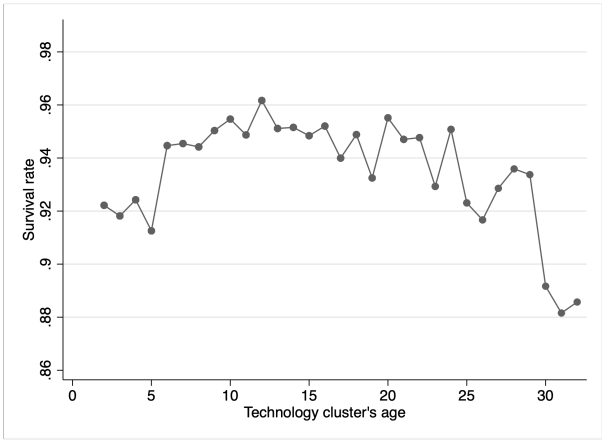
19. Evans (1987) and Hopenhayn (1992) are among the key pioneer papers on this.

Figure 1.8: Survival rate of technology cluster by size quantiles



Notes: Size is defined as number of accumulated patents in a given technology cluster

Figure 1.9: Survival rate of technology cluster by age



Notes: Age is defined as the difference between the year of technology exit and the year of its entry.

vating firm entry.

5. Conditional on age, the average growth of a technology cluster is slowing down over time.

6. The survival rate of a technology cluster is

   (a) increasing with a technology cluster's size;

   (b) fairly constant throughout a technology cluster's age and starts decreasing with age after a cluster turns 20 years old.

I now turn to studying cross-sectional patterns of technology cluster entry and move to a micro-level analysis. The following section explores what type of innovations lay a foundation for a new technology cluster and what kind of firms are the pioneers who establish new technology clusters by undertaking these innovations.

1.4 Entry of New Technology Clusters

This section focuses on studying how technology clusters appear and how they enter the economy. First, I explore what kind of innovations lay a foundation for a new technology

cluster. Then I turn to a firm-level analysis which goal is to ascertain the firm characteristics that are conducive to discovering a new technology cluster. I show that the patents that give a start to a new technology cluster receive more citations from other patents, they are based on more fundamental research, they follow the step of more recent innovations and have higher market value. In other words, it is indeed radical breakthroughs that establish new technologies. Using data on various characteristics of patenting firms, I draw a picture of a typical firm that is a pioneer in a technology cluster. I also look at what stage of a firm's life it has more chances to undergo a breakthrough innovation that will lay a ground for a whole new technology cluster. I show that it is small, young firms who are more likely to establish in new technology cluster as a result of their innovation activity. Moreover, firms that have just entered innovation market are twice more likely to bring a new technology cluster with them. However, the role of young and small startups in discovering new technology clusters is decreasing over time.

1.4.1 Ideas that Establish New Technology Clusters

In theoretical models of endogenous growth, an appearance of a new technology cluster is either a result of expanding varieties as in Romer (1990) and Grossman and Helpman (1991) or goes hand-in-hand with the size of innovation step as in Acemoglu and Cao (2015) and Akcigit and Kerr (2018). For the former, innovation characteristics are unrelated to a discover of a new variety and hence, a new technology cluster. On the other hand, size of innovation step that firms undertake in Akcigit and Kerr (2018) is the key factor. Firms can draw a "small-step" innovation that is only incremental and designed to improve the quality of existing technology clusters and their product lines. A new technology is discovered when some firm draw a "big-step" innovation – a radical breakthrough.

The role of new technology cluster can go beyond being an addition to a current set of technologies used in the production of existing product lines in the economy. Due to technological superiority, new technologies can make one or several of the existing technology

clusters obsolete – the creative destruction effect that canonical Schumpeterian models are based on. In this section, I empirically validate one of the key assumption of the endogenous growth models with heterogeneous step size and study the relationship between breakthrough ideas and new technology sectors using the new dataset. As a by-product, I also discuss several proxies for innovation step in the data based on different measures of patent quality, originality, its scientific and economic value.

In the heart of this exercise is a comparison between various patent characteristics of pioneer patents in a new technology cluster and other new patents in the same year of application. The level of observation is patent-year but the exercise has a cross-sectional nature. The main regressor is an indicator that is equal to one if patent p submitted in year t has established a new technology cluster[20] and is equal to zero if patent p belongs to a continuing technology cluster:

$$Characteristic\_of\_Patent_{pt} = \beta \times \mathbb{1}\{Entry\}_t + \gamma_t + \varepsilon_{pt} \tag{1.4}$$

Table 1.2 summarizes the results of this exercise. All dependent variables are standardizes to have zero mean and unit standard deviation. Summary statistics of pre-normalized variables is reported in Table B2 in Appendix B. I consider six patent characteristics that can capture the quality, scientific contribution, economic value and innovativeness of a patent. The first one is a common proxy of a patent quality that is often used in innovation literature – received citations. The number of other patents that cite a given patent up to five years from its submission captures the value of this patent's contribution for the follow-up innovations and thus reflects its R&D value.[21] Column 1 shows the result of regression (1.4) with received citations count as a dependent variable. We see that pioneer patents in a

20. As mentioned above, the patent that generates technology entry should be followed by at least one other patent within 5 years of this first patent.

21. Five-year truncation is needed to avoid bias in citations for the most recent patents. The idea is that patents from 1980s have more years too accumulate citations than patents from 2010s. Focusing on a five-year horizon helps to overcome this issue.

technology cluster receive .2 more citations in standard deviations in comparison to patents in continuing clusters submitted in the same year.

The second patent characteristic that I use to capture the innovativeness of a patent is how current the contribution of a patent is. In particular, I look at whether it is based on more recent innovations and thus relates to an actively growing segment of the technology frontier. A proxy for that is average difference between the year of application for patent p and the year of applications for the patents that p cites. If this gap is small, then the inventions that serve as a knowledge base for the patent are more recent and thus the patent can be considered to be more up-to-date. Column 2 reports a negative coefficient on the citation age suggesting its negative correlation with pioneer patent indicator. That means the innovations that establish new technology clusters cite more recent patents, and thus are based on more current inventions and belong to dynamically evolving spheres of innovation progress.

With the third measure, I aim to capture how fundamental the knowledge underlying a patent is. The number of scientific articles cited by a patent helps to distinguish basic invention from inventions of a more applied nature. The reason why we want to compare pioneer patents by this characteristic is to test whether breakthrough innovations contain more fundamental knowledge in comparison to follow-up innovations that can be more applied. As a proxy for patent basicness, I use data on the number scientific articles that a patent cites provided by Marx (2019). Column 3 shows a positive association between number of academic publications that a patent has and its likelihood of being a pioneer in a technology cluster. Hence, the patents that establish new technologies encompass more fundamental knowledge and rely on academic research more heavily. This result speaks to the literature on innovation growth that distinguish basic and applied research, arguing that basic innovations generate stronger spillover effect that spreads across many industries (see (Gersbach et al., 2013) and (Akcigit et al., 2013) among others).

However, scientific and economic value of a patent are not perfectly correlated as demon-

strated in Kogan et al. (2017). The three patent characteristics that I have introduced above can be good proxies for a patent's scientific contribution, but they may not capture a patent's private value for a firm. I use stock market's opinion on the value of a patent as a proxy for its economic value as opposed to research contribution. Kogan et al. (2017) provides the measure of patent private value based on stock market movements in prices immediately after the patent grant. They look at the changes in stock prices of a company that gets a patent grant during a three-day announcement window. This short-run price volatility (with a couple of distributional assumptions to account for noisy stock movements unrelated to the patent) allows to identify the economic value of a patent. A detailed explanation of the methodology and index construction can be found in Kogan et al. (2017). The final index of a patent value if measured in millions of US dollars. Column 4 reports a positive coefficient on the market value of a patent, implying that the private market value of the patents that "open" a technology cluster is significantly higher compared to other patents that appeared in the same year. Hence, pioneer patents in technology clusters do not only have a superior scientific contribution but also have a higher economic value.

Another indicator of a patent quality and innovativeness is patent originality as defined in Hall et al. (2001).

$$Originality_p = 1 - \sum_{j}^{n} s_{pj}^2 \tag{1.5}$$

where s_{pj} is share of citations that patent p gives to patents from class j and n is number of possible classes. If a patent is a "small-step" innovations then it is considered as a minor improvement of existing technology clusters in the endogenous growth models. That means, this patent is a follow-up invention that should have strong ties with a particular innovation branch. Citations allow us to trace this connection. The originality score (5) is high when patent p cites patents from many different classes, which implies that this patent is based on a variety of previous inventions rather than being a direct follow-up from one particular innovation branch. Positive association between patent originality and its indicator of a pioneer in a technology cluster (Column 5) implies that cluster-founding patents indeed

have a broader range of inventions that they are building on and thus are likely to be a result of a "big-step" R&D draw.

The final patent characteristic I am considering is related to the originality measure proposed by Hall et al. (2001). I look at the number of broadly defined sectors that a given patent is classified to. All the patent applications submitted to USPTO are assigned to one (or several) out of nine possible sectors.[22] When a patent is attributed to more than one general sector, it implies that it's innovation contribution is quite broad and hard to pin down or not well-defined. From the negative coefficient in Column 6, we see that the patents that open new technology clusters have a well-defined specific contribution that can be clearly attributed to one big sector.

Table 1.2: Pioneering patents in new technology clusters

	#Cit	Cit age	Acad cit	Mark val	Original	#Ind
	(1)	(2)	(3)	(4)	(5)	(6)
$\mathbb{1}\{\text{Entry}\}$	0.190***	-0.0313***	0.0384***	0.0692***	0.0496**	-0.0838***
	(0.0609)	(0.0100)	(0.0109)	(0.0188)	(0.0190)	(0.0159)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,123,274	1,009,199	1,123,274	418,149	910,124	1,123,051
R-squared	0.053	0.076	0.011	0.029	0.201	0.003

Notes: All dependent variables are normalized to have zero mean and unit standard deviation. Summary statistics for pre-normalized dependent variables is reported in Table B2. Dependent variable is an indicator of whether a patent is a pioneer in a technology cluster or not. Mean of $\mathbb{1}\{\text{Entry}\}$ is .012 with standard deviation of .11. Standard errors are clustered on the FE level. *** p<0.01, ** p<0.05, * p<0.1

Figure A5 in Appendix A is a visual representation of these qualitative results. It plots the evolution of these six patent characteristics over a technology cluster's age after taking out the variation that can be attributed to time fixed effects. The main message of the findings is that technology clusters are indeed established by breakthrough ideas that have both high scientific value for innovation community and high economic value for a firm. Thus,

22. The nine sectors are Human Necessities, Performing Operations and Transportation, Chemistry and Metallurgy, Textiles and Paper, Fixed Construction, Mechanical Engineering and Lightening plus related sections, Physics, Electricity, General Tagging on New Technological Developments.

the models with heterogenous innovations are a better fit for exploring entry dynamics of technology sectors since the variety expansion framework does not capture the importance of patent characteristics for establishing a new technology cluster at all.

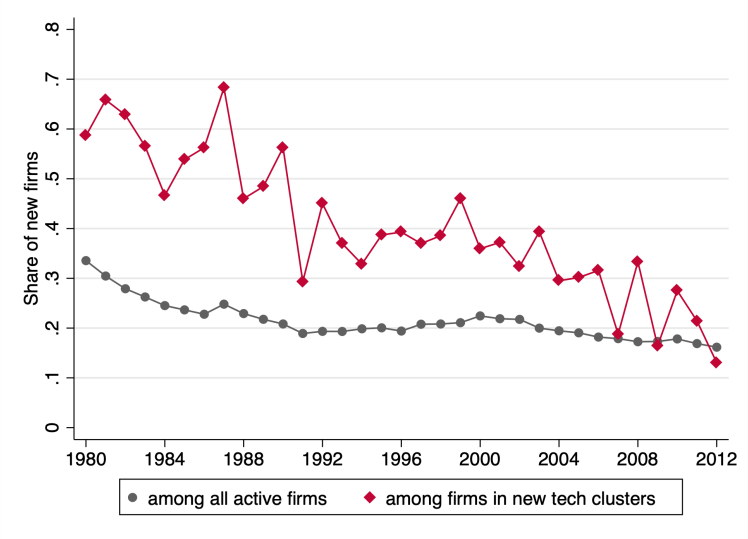
1.4.2 Firms that Establish New Technology Clusters

Literature on innovation-driven growth does not suggest one unambiguous answer on whether it is new or incumbent firms that introduce new technologies to the market. Early Schumpeterian models suggest that all innovations are undertaken only by new firms because of the Arrow replacement effect (see (Arrow, 1962)). A canonical example is Aghion and Howitt (1990) where each firm is allowed to have only one product and all innovations are coming from new-comers that replace incumbents by improving the quality of a product and capturing the whole monopoly profit. On the other hand, Klette and Kortum (2004) assumes that firms with more product lines are more likely to overtake other product lines, which implies that industry incumbents have an advantage over new firms.

In all these early Schumpeterian papers, the set of varieties is fixed and thus all the "creation" is destructive – each new technology replaces an old one. More recent models of endogenous innovation growth avoid this trap of absorbing creative destruction by giving both incumbents and new entrants a chance to discover new technologies that can lead to new product lines. In Acemoglu and Cao (2015) and Akcigit and Kerr (2018), newly formed firms are engaged in more radical innovations while continuing firms improve the quality of existing product lines. My findings speak to the structure of these endogenous growth models. I show that young firms are more likely to come up with radical innovations that lay a foundation to a new technology cluster. In fact, first-time patenting firms are even more likely to establish a new technology cluster as they enter the innovation market. I show that these pioneers are smaller in size measured by employments and patent portfolio but nevertheless, they have large innovation human capital that allows them to work on "big-step" inventions.

Note that if new entrants and incumbent firms equally likely to discover new technologies, then the share of entrants in new technology clusters would be exactly equal to the share of entrants in the economy overall.[23] Figure 1.10 plots the evolution of these two shares over time. Until recently, the share of new firms among the companies that have discovered new technology clusters was much bigger than the share of new firms among all active firms in the economy. This implies that newcomers are more present in frontier technologies and thus are more likely to undertake radical innovations that lay a foundation for a new technology clusters. However, the two shares are converging over time and intersect around 2006, suggesting that the role of new firms in establishing new technology clusters is not as crucial in the recent years as it used to be.

Figure 1.10: Distribution of new firms among all innovating firms in USPTO data and in new technology clusters



Now I will turn to a more granular analysis of firm characteristics that are conducive to a discovery of a new technology cluster. What is the image of a typical firm that pushes the technological frontier further? The particular characteristics of a firm I am going to consider is employment, age, innovation human capital, and internal (firm-level) network

23. Note that this is different from testing a model assumption that new firms and existing first have the same odds of coming up with a new technology that leads to a new product line. The data used in this exercise reflect equilibrium allocation, meaning we are observing the "results" rather than the 'assumptions".

of technology clusters that a firm has. I will also zoom in on the within-firm variation to understand at what stage in life a firm has more chances to discover a new technology cluster.

The data on firm employment and age comes from the Orbis global database, provided by Bureau van Dijk. It is a large database that includes firm-level statistics on financial statements and production activity of both private and public companies. The advantage of this database is that is has a pervasive coverage of firms from different industries and of different sizes and thus has been used in many research papers.[24] I match firms from Orbis to the firms from my sample of patents by company names using a similar TF-IDC procedure described earlier with a more conservative similarity score threshold[25] and a different n-gram structure.[26] This allows me to identify around 30% of patenting firms from my initial sample of USPTO assignees with patents matched to Wikipedia articles in Orbis. The resulting panel has 29,533 firms with observations from 1980 to 2012. The main regression specification is as follows:

$$\mathbb{1}\{Entry\}_{it} = \beta_1 Patents_{it} + \beta_2 Emp_{it} + \beta_3 IHC_{it} + \beta_4 Age_{it} + \beta_5 First_{it} + \\ + \beta_6 Network_{it} + \gamma_t + \varphi_i + \varepsilon_{it} \tag{1.6}$$

where the dependent variable is an indicator of whether firm i discovers a new technology cluster in year t or not. $Patents_{it}$ is number of patents that firm i has accumulated by year t: a proxy for patenting experience of firm i in innovation market. Emp_{it} is total employment of a firm in a given year and IHC_{it} is innovation human capital measured as number of inventors working for firm i in year t.[27] Age_{it} is firm age from the year of its incorporation,

24. See Kalemli-Ozcan et al. (2015) for a comprehensive description of the database, its features, representativeness, initial source of the raw data, advantages and potential flaws.

25. The lower boundary for the match is 0.8 similarity score.

26. The difference is the definition of a "word token" in the patent-article match and in the company names match. If for the former I use lemmatized words and word combinations, in this case, I am using 4-grams: a sequence of 4 characters.

27. This information is retrieved using all patents of a firm submitted in year t. Innovation human capital

$First_{it}$ is an indicator of whether year t is the first year when firm i applies for a patent and enters innovation market. $Network_{it}$ is the average degree of a technology cluster in a firm's technology network consisting of all technology clusters that firm i has ever patented by time t. Note that a technology sector degree in a firm-specific network are not necessarily equal to its degree in a global network. If a technology cluster has 5 related technologies in the global network but firm i has not ever innovated in any of them, the local firm-specific degree of this technology is 0. Average degree of firm's technology clusters captures how dense firm's network is and thus how specialized firm i is. If this indicator is equal to zero, none of the technology clusters that firm i innovates in are connected to each other and thus the firm's innovation activity is quite dispersed.

The results of regression (1.6) are reported in Table 1.3. All specifications are linear. Accumulated patents, employment, inventor's count and firm's age are normalized to have zero mean and unit standard deviation. Columns from (1) to (4) explore between firm variation while columns (5) and (6) focus on within firm variation.

Strong negative coefficients on employment and accumulated patents suggest that small firms are more likely to establish new technology clusters. Quite often, these firms are complete newcomers to the innovation market and discover new technologies in the first year of their patenting record. At the same time, a positive coefficient on inventors count implies that innovative human capital is indeed important for the production of radical innovations that lay a foundation for a new technology cluster. Existing technology clusters of the firms that come up with new technologies have higher average degree, conditional on age, size and experience. This implies that pioneer firms are in general more specialized in a certain innovation field and their R&D activity is concentrated in one area. While the sign reverses if we are looking at within firm variation, this is not surprising given a negative relationship between firm age and likelihood of a discovering new technologies. Since firms are more likely to establish new technology sectors earlier in their life when they do not have much

equals to number of unique inventors stated in these patents.

Table 1.3: Pioneering firms in new technology clusters

	Technology entry indicator					
	(1)	(2)	(3)	(4)	(5)	(6)
Employment	-0.046***	-0.031***	-0.030***	-0.033***	0.003	-0.008
	(0.014)	(0.010)	(0.010)	(0.009)	(0.017)	(0.018)
Inventors count	0.108***	0.156***	0.158***	0.001**	-0.017	0.031
	(0.029)	(0.048)	(0.049)	(0.042)	(0.023)	(0.028)
Accum patents		-0.073**	-0.069**	-0.126***	-0.249**	-0.302***
		(0.033)	(0.033)	(0.035)	(0.114)	(0.094)
$\mathbb{1}$\{First patent\}			0.387***	0.631**	0.546**	0.640***
			(0.099)	(0.249)	(0.215)	(0.215)
Av firm degree				0.185***	-0.236**	-0.556***
				(0.047)	(0.109)	(0.126)
Age					-1.20***	
					(0.226)	
Year FE	Yes	Yes	Yes	Yes	No	Yes
Firm FE	No	No	No	No	Yes	Yes
Observations	299,875	299,875	299,875	283,429	279,396	283,429
R-squared	0.003	0.003	0.004	0.004	0.039	0.039

Notes: All specifications are linear. Accumulated patents, employment, inventor's count and firm's age are normalized. Dependent variable is measured out of 100%. Mean of the dependent variable is .65%. Standard errors are clustered on FE level. *** p<0.01, ** p<0.05, * p<0.1

experience, their internal technology network will be quite sparse.

Overall, we can conclude that a typical firm that discovers a new technology cluster is young and small. It is very likely to be a complete newcomer to the innovation market but nevertheless has a large team of inventors working on breakthrough inventions. Firms have better chances to discover new technologies early in their life. We can think about these pioneer firms as startups or research incubators with extensive innovation human capital and sharp focus on a specific industry.

1.5 Incumbent Firms in New Technology Clusters

The preceding analysis shed some light on what innovations lay a foundation for a new technology cluster and what kind of firms are the pioneers who discover new technologies leading to new products. The fact that that young firms come up with radical innovations while mature firms are more likely to undergo incremental innovations speaks to the implications of many endogenous growth models. Even though young small firms have better chances to be pioneers, this average pattern that does not rule out the possibility of mature firms discovering a technology cluster. Moreover, from Figure 1.10 we see that the role of startups in pushing the technological frontier is decreasing over time.[28]

While startups do not have any observable industry experience, mature firms have their accumulated knowledge base: a collection of technology clusters they have been working on. As a result, we can observe the past-dependency of a firm's choices to innovate in its own existing technology clusters or to switch to other clusters. If the latter is the case, firms can choose whether to work in a technology cluster that is related to their accumulated knowledge base or turn to something quite different from their core experience.

There are pros and cons of both of these strategies. Building on your previous experience is less risky but there might be less chances to come up with radical innovations around the fields that you have been working on before. Moving to uncharted waters can be more risky and require higher fixed costs of innovations, but there may be more chances to stumble upon a breakthrough invention. This section provides an empirical evidence on whether companies with a relevant experience in related technology clusters have better chances to discover a new technology. As mentioned before, I am focusing on mature firms in this section and thus the following analysis will be conditional on firm's age and overall patenting tenure.

28. For instance, touch screen technology was invented in the beginning of the 70s by a recently established company *Elographics*, founded by the inventor of the touch sensor himself ((Bellis, 2018)). However, a more recent technology of a flexible foldable display was introduced in 2013 by Samsung – a giant industry incumbent ((Blagdon, 2013)).

1.5.1 Role of Related Experience in Discovering New Technologies

In order to capture firms' experience in related technology clusters, I am going to use firm-specific technology network. While the global technology network is based on patents from all firms that are active in the economy at a given point in time, firm-specific network captures only technology clusters that a particular firm has ever innovated on and thus is based on this firm's accumulated patents. We can think of a firm's network as a map of the firm's cumulative innovation activity. The presence of an edge between technology clusters in a firm-specific network is determined the same way as in the global network – technologies are connected if the corresponding Wikipedia subcategories are linked.

When a firm discovers a new technology cluster to its network by submitting a patent to it for the first time, this new technology enters a firm's network and becomes a node there. If this firm has patents in technology clusters that are related to this newly discovered technology that the firm has discovered, then the new technology cluster will get connected to them in firm's network. In this case, a new technology cluster will have positive degree (i.e. positive number of connections) in the firm's network. The higher degree of this new technology sector is, the more satellites among firm's technology clusters it has and the more in line with firms existing knowledge base it is. Hence my measure of related firm's experience is a degree of a new technology cluster in a firm's network once the firm has discovers it.

Note that we can only observe a degree of a new technology cluster in a network of the firms that have actually discovered it. If a firm has not patented in this technology cluster and as a result has not added this technology to their network, we do not observe its degree and thus cannot measure related industry experience for this "non-entering" firm. It does not mean that this firm have no technology clusters that would connect with the new technology if it were added to the firm's network. In fact, we can impute the measure of related industry experience for these non-entering firms by manually adding a new technology cluster to these firms' network and computing its firm-specific degree.

The first step in this exercise is to identify firms that have actually discovered a new technology cluster in the data (i.e. submitted the first patent ever to this technology cluster) and compute the new technology's degree in this pioneer firm's network. That will give us the measure of related industry experience for the firms that open new technology clusters straight from the observed data. Then for each of these "pioneer firm - new technology" pairs, we need to construct the measure of related industry experience for their peer firms that have not discovered this new technology. In order to do so, I manually add the new technology cluster to the networks of these non-pioneer firms (as if they have discovered this technology) and compute its firm-specific degree the same way I did with the actual pioneer firms. This imputation allows me to capture the related industry experience for the firms with zero entry outcome.

Execution of this procedure is quite computationally costly since for every "pioneer firm - new technology" pair we need to construct the degree of this new technology cluster in every other firm in the sample. Thus, I restrict the range of non-entering firms that I am comparing the pioneer firm to. For each 'pioneer firm - new technology" pair, there are five non-entering firms that are most similar to the entering firm in this pair. In particular, I match pioneers to non-pioneers by firm's age, size and main broad sector of innovation activity. The latter means that I require the pioneer and the potential non-pioneer peer to have the majority of their cumulative patents in the same broadly defined sector.[29] Then I compute the squared percentage difference between the size of a pioneer firm and all other active non-pioneers, as well as the difference between firm's patenting tenure (i.e. years since first patent). The final match indicator is a square-root of the sum of these size and tenure gap. Five firms with the smallest indicator are then assigned to be the non-pioneer comparison firms for a given "pioneer firm - new technology" pair.

The level of observation in this exercise is a firm-technology pair. The main idea is

29. The definition of a general sector of operation comes from CPC classification and has 9 sectors. A firm's main sector is determined by allocation of its patents: a sector with the highest share of firm's cumulative patents is the main sector of a firm.

to see whether experience in related technology clusters is conducive to discovering new technologies or the pioneer firms are complete new-comers to this innovation area and have never patented in any similar clusters. If the latter is the case, incumbent pioneer firms would basically be not much different from startups in a sense that their relevant knowledge base is zero.

I start with a simple mean comparison of pioneer firms and their matched non-pioneer peers. An average degree of a new technology cluster in the network of *pioneer* firms that have actually discovered this technology is 1.31, while an average degree of this technology cluster in the network of *non-pioneer* firms is 2.13, nearly 60% bigger. The gap between the means is statistically different from zero. This implies that pioneer firms have actually *less* experience in related fields when they discover a new technology cluster. While startups are "global" newcomers, this result implies that mature firms that open new technology cluster are "local" newcomers.

Obviously, a simple mean comparison is a fairly weak evidence given the fact that the event of a technology cluster's discovery highly correlates with firm's age and other characteristics. In the previous section, we saw that young firms are more likely to establish a new technology cluster, especially if it is their first year of patenting. Mechanically, these firms will not have any related industry experience because they just do not have any observable experience at all. We need to explicitly control for this selection when comparing related industry experience of pioneer and non-pioneer firms. The main regression specification I am using for that is the following:

$$\mathbb{1}\{Entry\}_{ict} = \gamma New\_Degree_{ict} + \beta \mathbf{X}_{it} + \gamma_t + \varphi_c + \varepsilon_{ict} \tag{1.7}$$

where $\mathbb{1}\{Entry\}_{ict}$ equals one if company i discovers technology cluster c in year t and zero otherwise. New_Degree_{ict} is our proxy for related industry experience described above – a degree of newly discovered technology cluster c in firm's i network measured in the year of the

discover t. $\mathbf{X}_{it}$ is a vector of firm-level controls such as number of patents submitted by time t, employment, innovation human capital, average firm-specific degree of firm's technology clusters, patenting tenure (years since first patent application) and an indicator of whether t is the year of firm's first patent. I am also exploring different levels of variation by including year and technology cluster fixed effects.

The results of this exercise are reported in Table 1.4. The coefficient on new technology cluster's degree is negative and significant in all different versions of the regression (1.7). The specifications with both year and sector fixed effects in Columns 3 and 5 show that one more connection to a new technology cluster *decreases* the chances of a firm to actually discover this technology by 5.2-5.8 percentage points (or 31%-35%). This implies that firms that have *not* operated in related technology clusters before are actually more likely to establish a new technology cluster. This result is quite surprising as it suggests that firms with relevant experience are actually less likely to exploit that experience. Industry newcomers seem to be more likely to undertake radical innovations that will give a start to a new technology cluster, regardless of whether they are "global" or "local" newcomers.

However, a big caveat that needs to be made here is the lack of proper economic interpretation of the links between technologies so far. There are many possible reasons why two technology clusters can be connected as these links are built based on the connections between the related Wikipedia subcategories. In the following section, I elaborate on the nature of the links between technology clusters, select the ones that connect technology clusters with complementary knowledge base and as a result, refine the definition of related industry experience.

1.5.2 Refining Economic Meaning of the Links between Technologies

Remember that the links between technology clusters are based on the links between subcategories in Wikipedia categorization network. The main unambiguous property of these links that has been utilized so far is that connected technology clusters are more related to

Table 1.4: Role of previous experience in connected technology clusters

	Technology entry indicator				
	(1)	(2)	(3)	(4)	(5)
New tech degree	-0.0413***	-0.0345***	-0.0583***	-0.0159***	-0.0524***
	(0.0146)	(0.0084)	(0.0123)	(0.0054)	(0.0131)
Average firm degree	-0.216***	-0.583***	-0.699***	-0.450***	-0.621***
	(0.0261)	(0.0326)	(0.0246)	(0.0315)	(0.0326)
$\mathbb{1}\{$First patent$\}$		-0.0951***	-0.0442***	-0.0553**	-0.0230
		(0.0066)	(0.0057)	(0.0246)	(0.0364)
Accumulated patents		0.236***	0.202***	0.108***	0.109***
		(0.0244)	(0.0269)	(0.0173)	(0.0231)
Inventors count		0.203***	0.201***	0.165***	0.172***
		(0.0307)	(0.0304)	(0.0253)	(0.0289)
Patenting experience		0.0379***	0.0307***	0.0307***	0.0280***
		(0.0039)	(0.0039)	(0.0033)	(0.0069)
Age				0.0605***	0.0439***
				(0.0053)	(0.0070)
Employment				0.0130	-0.0002
				(0.0095)	(0.0135)
Year FE	Yes	Yes	Yes	Yes	Yes
Technology FE	No	No	Yes	No	Yes
Observations	30,822	30,820	30,820	8,049	8,049
R-squared	0.220	0.610	0.750	0.484	0.697

Notes: Accumulated patents, employment, inventors count, age and patenting experience are normalized. Mean of the dependent variable is .167. Standard errors are clustered on year level. *** p<0.01, ** p<0.05, * p<0.1

each other than those that do not have a bipartite connection. However, economic interpretation of this connection can be quite heterogeneous and require more careful investigation. This implies that the results of the previous section should be interpreted differently since they are based on the technology network that contains links of several different economic nature. In this section, I am going to refine the definition of "relevant industry experience" by focusing on technology clusters that are linked due to their knowledge and technology base complementarities. Once I have a technology network with more homogeneous con-

nections that have a desired economic interpretation, I am going to repeat the methodology from the previous section to address the question of the role of related industry experience in discovering new technologies.

Visual examination of the links between technology clusters shows that in general, there are four different economic interpretations of a connection. First, technologies are connected if one of technology cluster is a narrowly-defined "subsector" of a more general technology cluster. For instance, cottage cheese and cheese are connected because the former is a particular type of cheese. While some companies can have a broader business activity and submit patents related to different types of cheese and cheese production in general (for instance, Land O'Lakes Inc.), others can focus their innovations on a particular type of cheeses (for instance, Good Culture LLC specializes on cottage cheeses) and refine production process specific to this cheese.

Another interpretation of a connection between two technology clusters is based on complementarity and substitutability in product and knowledge space. Complementarity (substitutability) in a product space simply means that the two products that these technologies are feeding into are used together in the production of some final good (i.e. can be used interchangeably) or are consumed together (i.e. can replace each other in a consumption basket). To fix this idea, let's consider a more formal definition of product space complements. Suppose there are N different technology clusters and each of them have their unique innovation output x_1, x_2,..., x_N. This innovation output is used in the production of a range of final goods, including consumption utility. The production function of final good q^i can be described as $q^i = Q^i(x_1, x_2, ...x_N, \mathcal{V})$, where $\mathcal{V}$ is a vector of inputs that are not related to innovations. If a cross-derivative of Q^i with respect to x_j and x_k is positive, $Q^i_{jk} > 0$, then technology clusters k and j are product-complements (or as sometimes called, q-complements). If $Q^i_{jk} < 0$, then they are product-substitutes.

On the other hand, technology clusters can be linked because they are complementary in the knowledge space. Suppose each technology cluster i has a technology-specific knowledge

input associated with it, h_i, that is a primary factor of production for the technology's output x_i. Technology clusters can also rely of knowledge input from other technologies, h_{-i}. Production function of innovation input in technology cluster i is then $x^i = F^i(h_i, h_{-i}, \mathcal{W})$, where $\mathcal{W}$ is a vector of inputs unrelated to knowledge, i.e. effort. Most of the other technology clusters' knowledge bases are irrelevant for innovation input in technology i: $F^i_j = 0$, where j is a technology cluster that is irrelevant for i. But if technology i and j share a similar knowledge base, their inputs will be complementary to each other: $F^i_j > 0$. In this case, we say that technology clusters i and j are knowledge-complements.

Some technology clusters have induced complementarity in a product space but are irrelevant in knowledge space. For instance, camera lenses and camera flashes are product-complements as they are used in the production of cameras together. But they are based on quite different innovation inputs and the knowledge base that is needed to produce camera lens is different from the knowledge base for camera flashes. On the other hand, speakers and headphones are product-substitutes but are complements in the knowledge space. Firms that are producing speakers are often also producing headphones because it is very easy to enter a technology cluster that is based on similar knowledge that you have already implemented in your current innovation activity.

The type of a connection that allows to measure relevant industry experience is knowledge-base complementarities.[30] In the following analysis, I will consider only the links between technology clusters that are dictated by knowledge-complementarity to define relevant industry experience. In particular, I will focus on a special type of these links that connect a new technology cluster with the technology that was creatively destructed by it. Consider a pager and a mobile phone. These technology clusters are knowledge-complements – inventors of mobile phones took pagers technology as a base, improved it and made pagers obsolete. I am going to isolate the cases when a new technology cluster significantly replaces

30. Continuing with the previous example, even though camera lenses and camera flashes are connected in the network, a current producer of camera lens has no related experience in the creation of camera flashes.

an old knowledge-complementary technology. The ultimate goal of this exercise is to understand whether the companies that has discovered a new technology cluster have a record of innovating in the replaced technology cluster too.

In order to identify technology clusters that are connected to each other due to knowledge complementarity *and* have a creative destruction nature of their relationship (i.e. one technology has replaced the other one in the innovation market), I start with identifying the following patterns in the data. Among all the technology clusters that are connected in the global network, I choose the connections where one technology (old technology) has entered the economy before the other technology (new technology) and also exits before this new technology in the pair. I also require the exit of the old technology to be less than 5 years after the new technology cluster has entered. By looking for this pattern in the data, I identify technology clusters that have entered after its linked neighbor technology and this neighbor cluster exited the economy soon after this event. Then I visually inspect the resulting dataset and make sure that the final sample has only links of knowledge complementarity nature.[31]

This procedure gives me about nine hundred pairs of technology clusters where one of the technologies have plausibly replaced each other by creative destruction.[32] Each new technology cluster in these pairs is associated with a firm (or sometimes several firms) that have discovered it (i.e. submitted first patents to this technology cluster). If these companies have also innovated in the old technology cluster from this pair then they are not complete newcomers to this innovation fields and have some related experience and knowledge base. For the companies that have not discovered a new technology, we can still measure whether they have related experience in the replaced technology cluster just by checking whether the company has ever submitted a patent associated with this old technology. As in the

31. For instance, digital signatures and public-key cryptography, fighter aircraft and attach aircraft, cottage cheese and cheddar cheese.

32. For instance, photography film (replaced) and digital photography (new), X-ray (replaced) and CT scan (new), text messaging (replaced) and online chat (new).

previous section, I will focus on five non-pioneer companies as a comparison group for each pioneer company that are most similar to it by size, patenting tenure and main broad sector of innovations.

The regression specification is similar to (1.7):

$$\mathbb{1}\{Entry\}_{ict} = \gamma \mathbb{1}\{Related\_technology\}_{ict} + \beta_1 \mathbf{F}_{it} + \beta_2 \mathbf{O}_{it} + \alpha_i + \varphi_c + \gamma_t + \varepsilon_{ict} \quad (1.8)$$

but now the main regressor is a binary indicator $\mathbb{1}\{Related\_technology\}_{ict}$ that equals one if company i has innovated in the old technology cluster related to the new technology c and zero otherwise. $\mathbf{F}_{it}$ is a vector of firm controls that includes accumulated number of patents, inventors count, patenting experience of a firm, average degree of a firm's technology cluster and an indicator of whether year t is the first year of patenting for a firm. $\mathbf{O}_{it}$ is a vector of firm controls that use Orbis data such as firm's employment and age since establishment. Since Orbis data is available only for a third of firms from the sample, including these controls will significantly reduce the number of observations. If previously when we were using the whole sample of technology clusters it would not create a big issue, in this exercise the subsample of observations becomes critically small. To maintain a relatively large sample, the main specification I am referring to does not have Orbis controls.

The results are summarized in Table 1.5. Strong positive coefficients on *Related technology exposure* in the first three columns suggest that previous experience in technology clusters with knowledge complementarities is actually conducive to discovering new, potentially superior technologies. In fact, the magnitude of this relation is quite large even after controlling for technology cluster and year fixed effects – firms with past experience in related technologies increase their chances to discover a new technology cluster by 12 percentage points or by 73%. This result suggests that incumbents do not stop innovating in their fields several years after their entry but, quite on the contrary, are more likely to advance their technology fields than outsiders and newcomers from different innovation area.

Table 1.5: Importance of relevant innovation experience in complementary technologies

	Technology entry indicator				
	(1)	(2)	(3)	(4)	(5)
Related tech exposure	0.182***	0.113**	0.122**	0.0567	0.0724
	(0.0392)	(0.0456)	(0.0599)	(0.0523)	(0.0799)
Technology degree	-0.00019***	-0.00038	0.00267	-0.00019	0.0643
	(0.00007)	(0.00031)	(0.00326)	(0.00057)	(0.0467)
Average firm degree		0.0306*	0.106***	0.0175	0.148***
		(0.0167)	(0.0219)	(0.0205)	(0.0393)
$\mathbb{1}$\{First patent\}		0.329***	0.433***	0.368	0.505*
		(0.0475)	(0.0556)	(0.254)	(0.271)
Accumulated patents		-0.0725***	-0.0742**	-0.0666**	-0.0722*
		(0.0226)	(0.0300)	(0.0268)	(0.0367)
Patenting experience		0.0371***	0.0507***	0.0789***	0.0973***
		(0.0116)	(0.0142)	(0.00580)	(0.0171)
Inventors count		0.0721***	0.0692***	0.0743***	0.0556***
		(0.0159)	(0.0182)	(0.0162)	(0.0200)
Employment				0.00384	-0.0116
				(0.0167)	(0.0218)
Age				0.00568	0.0228
				(0.0107)	(0.0167)
Year FE	Yes	Yes	Yes	Yes	Yes
Technology FE	No	No	Yes	No	Yes
Observations	4,788	4,200	4,200	1,273	1,273
R-squared	0.005	0.041	0.073	0.075	0.204

Notes: Accumulated patents, employment, inventors count, age and patenting experience are normalized. Mean of the dependent variable is .167. Standard errors are clustered on year level. *** p<0.01, ** p<0.05, * p<0.1

An important reminder here is that these results do not contradict or overrule our previous findings about the crucial point of young small firms for discovering new technology clusters. They just emphasize that when mature firms do come up with radical innovations that establish new technologies, they are likely to have some expertise in the technology clusters that share a similar knowledge base. After all, perhaps startups that open new technology

clusters also have some related experience in the corresponding innovation field, but we just do not observe it in the patent data.

1.6 Growth of Technology Clusters

The analysis in the previous section showed that young and small firms are more likely to discover a new technology cluster. However, creation of new breakthrough technologies and production of follow-up research that generates actual innovation growth in technology clusters are two different phases of a technology cluster life cycle. Forces that stand behind formation of new technologies can be different from the factors that determine these technology clusters evolution.

In this section, I study the growth dynamics of technology clusters after their entry. I decompose technology cluster growth into between-firm, within-firm and net entry components and highlight the important role of between firm reallocation. This exercise shows that reallocation – firms refocusing their innovation efforts from one pursuit to another – is the main engine driving technology growth. In particular, reallocation between incumbent firms that switch their innovation activity from one technology cluster to another is more conducive to technology growth than reallocation between new firms entering the innovation market and exiting firms.

I elaborate on the role of reallocation for technology growth by exploring how a change in a technology cluster's leading firm relates to patenting activity in this cluster. I show that technology clusters are getting reallocated to more productive firms over time and this reallocation is associated with an increase in growth rate by 22 percentage points. Alarmingly, reallocation of technology clusters is slowing down over time, which can explain a decline in growth rates that we have seen in Figure 1.7.

I also explore cross-sectional patterns of technology cluster growth. Following the previous analysis on firm characteristics that are conducive to discovering a new technology cluster, I study what firm-specific and technology-specific factors are crucial for technol-

ogy cluster growth in a controlled environment. The results suggest that although young and small firms are dominant among the founders of new technology clusters, it is mature incumbent firms that generate technology cluster growth after entry.

My results relate to the literature on the role of reallocation for firm dynamics, industry development and economic growth in general. Starting from Lentz and Mortensen (2005), many models of innovation-driven growth emphasize the importance of across-firm reallocation. Their model generates quite an extreme result when all growth comes from reallocation of resources to firms that come up with superior products from firms that lose the market. Among more recent studies, Garcia-Macia et al. (2019) show that most growth is coming from the incumbent firms and its primary source is improvement of existing product lines, not introduction of new product varieties. Moreover, they show that the contribution of new firms and creative destruction declined from 1983-1993 to 2003-2013. Finally, Argente et al. (2018) study the cyclicality of product creation and destruction using Nielson scatter data from 2007 to 2013 that focus on consumer goods. They show that most of the goods reallocation is happening between continuing firms. This part of the paper can be seen as a continuing research on product lines reallocation that goes beyond consumer goods, looks at a longer time horizon, and as a result is able to address a bigger variety of the questions related to the importance of reallocation for innovation and industry dynamics.

1.6.1 Decomposition of Technology Cluster Growth

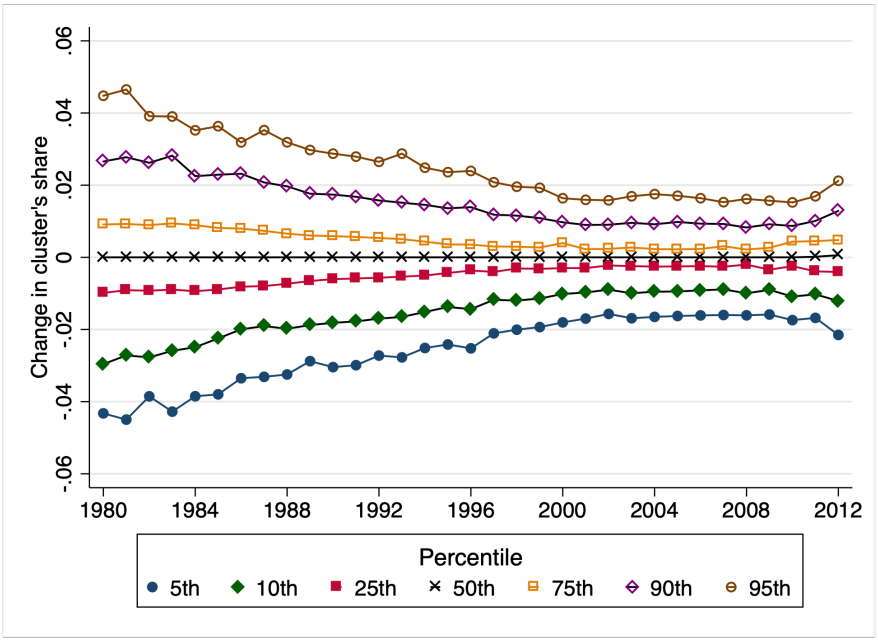
In this section, I decompose annual changes in a technology cluster size into three main components: between firm reallocation, within-firm growth and net entry contribution. Technology cluster size is measured as a share of patents that belong to this technology cluster relative to all patents in the economy submitted in a given year.[33]

For the following decomposition, I define growth of a technology cluster as an annual

33. I use relative measure as a size indicator rather than absolute number of patents in a technology sector to account for the changing dynamics of aggregate patents in USPTO data. While overall number of patents is increasing over time, I am interested in the distribution of this growth across different technology clusters.

change in its patent share. Figure 1.11 captures the evolution of this change in technology cluster's share over time for technologies in different growth percentiles. While before 2000 the growth of technology clusters was quite dispersed, it is converging to the median for the technologies in all percentiles. A potential explanation for this convergence is a decline in entry rate of new firms due to higher barriers of entry and the resulting decrease in selection. If there are less new firms entering the economy, there is less competitive pressure on the industry incumbents to boost their productivity. At the same time, higher entry barrier shifts the cutoff for a marginal firm (i.e. firm that is indifferent between entering and paying higher costs of entry and not entering). Now this marginal firm should be more productive than the marginal firm at the time when entry is easy – otherwise, expected profits would not cover the fixed costs of entry.

Figure 1.11: Annual changes in technology cluster size by percentiles



The share of a technology cluster c in year t can be rewritten as a composition of firms' size and a technology cluster's within-firm share:

$$x_{ct} = \frac{pat_{ct}}{pat_t} = \sum_i \frac{pat_{ict}}{pat_t} = \sum_i \frac{pat_{it}}{pat_t} \frac{pat_{ict}}{pat_{it}} = \sum_i s_{it} z_{ict} \tag{1.9}$$

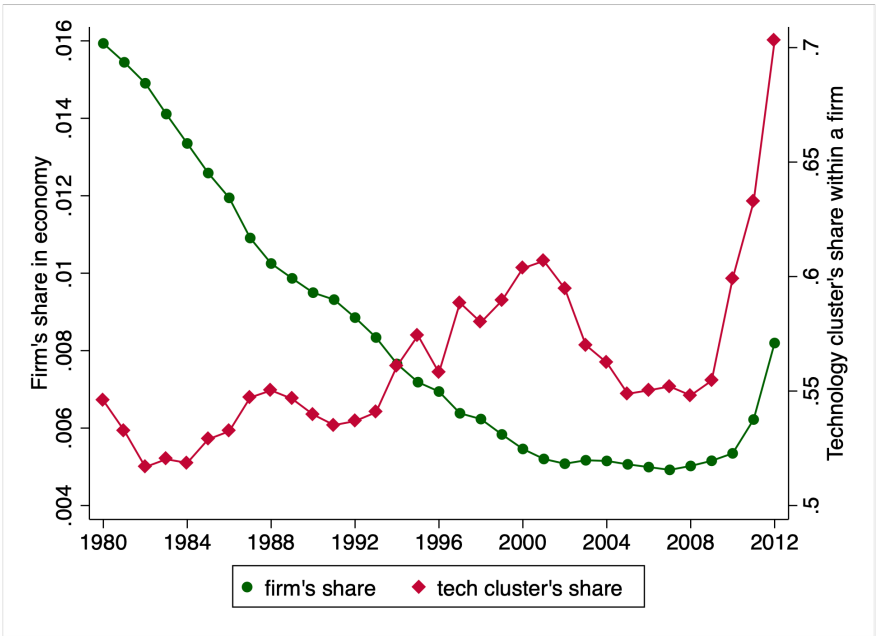
where pat_{ict} is number of patents in technology cluster c submitted by firm i in year t, pat_{it} is number of patents submitted by firm i in all technologies and pat_{ct} is overall number of patents in technology cluster c that are submitted in year t. s_{it} is then a share of a firm in the innovation market and z_{ict} is a share of technology cluster c within firm i – the relative concentration of firm i in technology cluster c. This representation of technology growth resembles the productivity decomposition from Olley and Pakes (1992) who define aggregate productivity at time t as a size-weighted average of firm's productivity. While I am adding another dimension to this exercise by focusing on technology cluster growth instead of aggregate growth, but the nature of it is very similar.

Before the decomposition of changes in x_{ct}, let's look at how each of the items from (9) evolves over time. Figure 1.12 plots the time trends of average firm's share in the economy (s_{it}) and average technology cluster's share within a firm (z_{ict}). Average firm's share was decreasing before 2000 but it stabilized after 2000 on a constant level until the Great Recession when it started going up. As we saw from Figure 1.6 before, there has been a persistent decline in the entry of new innovating firms since 2000. The increase in firm concentration after the Great Recession can be a consequence of this preceding decline in new entry.[34] At the same time, we can see that average share of a technology cluster within a firm was steadily increasing until 2000 when the trend changed in the opposite direction. However, since the beginning of the Great Recession an average technology share within a firm has resumed its growth at even higher pace. This implies that lately firms are becoming more specialized in a few specific technology clusters rather expanding their activity to many different lines of innovation.

I now proceed to the decomposition of technology cluster growth by within-firm and reallocation components. The first step is to split annual changes in technology cluster's size

34. The literature on decline in dynamics and increase in market concentration is very large at this point. Just a few examples are Decker et al. (2014b), CEA (2016), Diez et al. (2019), Covarrubias et al. (2020). I discuss market concentration and its implications for technology cluster growth in Section 1.7.

Figure 1.12: Evolution of the average firm's share in the economy and average technology cluster share within a firm



Notes: Green line, firm's share, is the average share of patents that belongs to one firm relative to all patents submitted in a given year (s_{it}) measured in percentage points for simplicity of labeling. Red line, technology cluster's share, is the average number of patents in technology cluster c relative to all patents that a firm submits in a given year (z_{ict}).

by the contribution of net entry and continuing firms:

$$\Delta x_{ct} = x_{ct} - x_{ct-1} = \Delta x_{ct}^{NE} + \Delta x_{ct}^{CNT} = \Delta x_{ct}^{NE} + x_{ct}^{within} + \Delta x_{ct}^{between} \tag{1.10}$$

The net entry component is the difference between technology cluster growth coming from newcomers to the technology and exiting firms:

$$\Delta x_{ct}^{NE} = \sum_{new} s_{it} z_{ict} - \sum_{exit} s_{it-1} z_{ict-1} \tag{1.11}$$

Note that the definition of entry and exit in this decomposition is tied to a firm, not a technology cluster. It refers to *firms* entering or exiting a technology cluster, not a technology cluster entering or exiting the economy. A firm enters a technology cluster when it submits a patent to this technology for the first time and has a record of positive patenting activity

in this cluster in the following five years. It can be a brand new firm or an incumbent from different technology clusters. By the same token, exiting firms are those that stop patenting in a given technology cluster but they can still remain active and submit patents related to other technologies.

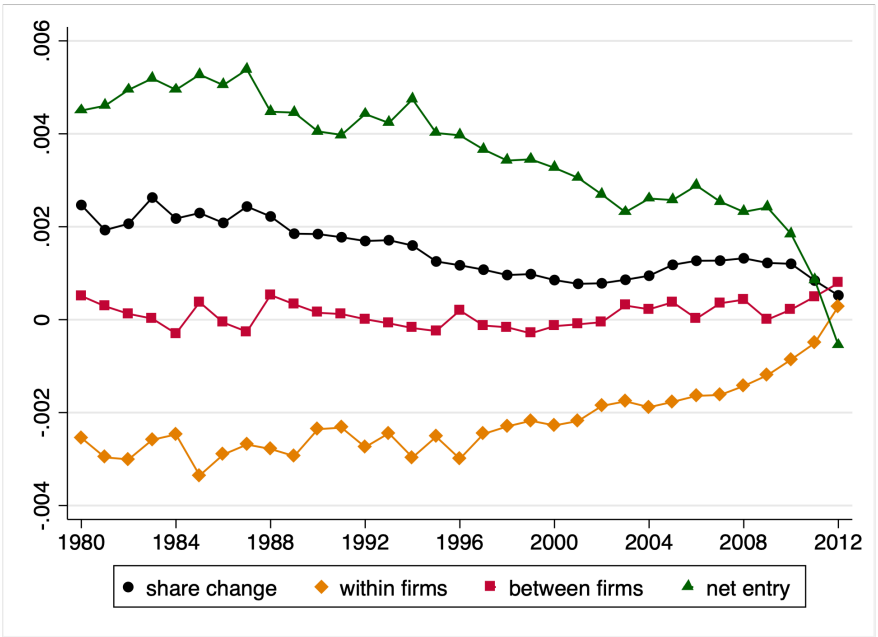
The definition of entering firm, exiting and continuing firms echoes the definition of entering, exiting and continuing technology clusters. A continuing firm is a firm that has previously entered a technology cluster and have submitted at least one patent related to this technology cluster in the last five years. Hence, a firm can be a new entrant in one technology cluster, an exiting firm in some other cluster and a continuing firm in a different one. The contribution of continuing firms comes from within-firm and between-firm components:

$$\Delta x_{ct}^{CNT} = \Delta x_{ct}^{within} + \Delta x_{ct}^{between} = \left[\sum_{CNT} s_{it-1}(z_{ict} - z_{ict-1})\right] + \left[\sum_{CNT} (s_{it} - s_{it-1}) z_{ict}\right] \tag{1.12}$$

The within-firms component focuses on the changes in the share of a technology cluster within a firm, holding the share of a firm in the economy constant. Between-firms component looks at the changes in a firm's market share, keeping the share of a technology cluster within a given firm fixed.

The results of the decomposition are reported in Figure 1.13. Focus first on the relative importance of each component for technology cluster growth at any given point in time before the Great Recession. We see that most of the growth is coming from net entry – reallocation of technology clusters between entering and exiting firms. For instance, Blackberry used to be a leader in the market of smartphones until the late 2000s. When Apple introduced iPhone in 2007 with its superior technology, Blackberry failed to compete with Apple and had to leave the market of mobile electronics eventually. Apple was not a brand new firm – it has been a big producer of personal computers for a long time. Nor was Blackberry exiting the market completely. While the company left the mobile electronics industry, it has successfully switched its specialization to enterprise software.

Figure 1.13: Decomposition of technology cluster growth



Notes: Share change is the average change in a share of patents that belong to a technology cluster between year t and year $t-1$ (Δx_{ct} from (10)). Within firms component is the average change in technology cluster's c share within a firm, holding the firm's overall patent share fixed (Δx_{ct}^{within} from (12)). Between firms component is the average change in a firm's patent share in the economy, holding the share of a cluster c fixed ($\Delta x_{ct}^{between}$ from (12)). Net entry is the average difference between technology cluster growth coming from newcomers to the cluster and firms that are exiting the cluster (Δx_{ct}^{NE} from (11)).

As mentioned above, the definition of entry is firm-specific and embeds four different types of entry: new firms entering new technology clusters, new firms entering continuing technologies, continuing firms entering new technologies and continuing firms entering continuing technologies. Net entry component in Figure 1.13 aggregates entry of all the four groups. Without disaggregation of this overall net entry trend, we cannot see what type of reallocation is more growth-inducing: reallocation between startups and closing firms or reallocation between continuing firms that are exploring the technologies they have not been worked on before. Figure 1.14 looks separately at the four types of entry while Figure 1.15 shows the same disaggregation for exit. Most of the growth through reallocation is coming from continuing and new firms entering continuing technology clusters on the entry side and continuing and exiting firms leaving continuing technology clusters on the exit side. Thus, most of the growth of technology clusters is actually coming from the reallocation of clusters

between continuing firms rather than between new firms and closing firms.

In fact, we can find many specific cases in the data confirming this result. For instance, in July 1989, Detroit Institute of Children submitted a patent "Augmentive communications system and method" which main contribution is an augmentive communications system and method for enabling handicapped patients to generate sentences and to control external devices. This patent laid a foundation to a technology cluster *augmentative and alternative communication*. Several months later, New England Center Hospital together with MIT submitted a closely related patent "Method for selecting communication devices for non-speaking patients" in the same technology cluster. Since then, there has been an active patenting activity in this cluster with new patents coming every other year from over a dozen of different assignees. These patents were mostly submitted by newcomers to this technology cluster, such as The University of Southern California and Research Foundation of State University of New York, which became the biggest contributor to this technology cluster starting from the middles of 1990s. Around the same year, the initial pioneer in this technology, Detroit Institute of Children, submitted their last patent related to augmentative and alternative communication, while New England Hospital exited this technology cluster even earlier.

Figure 1.14: Entry component decomposition

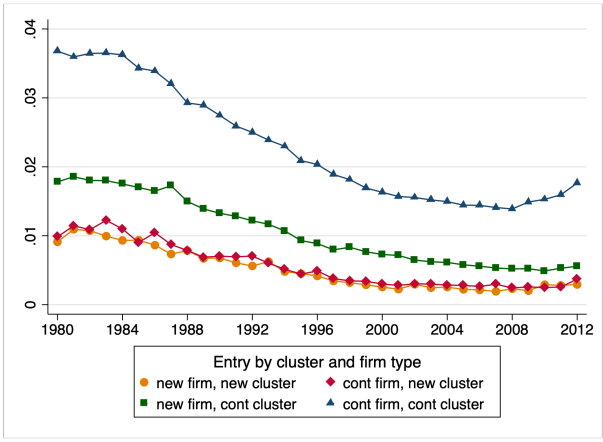
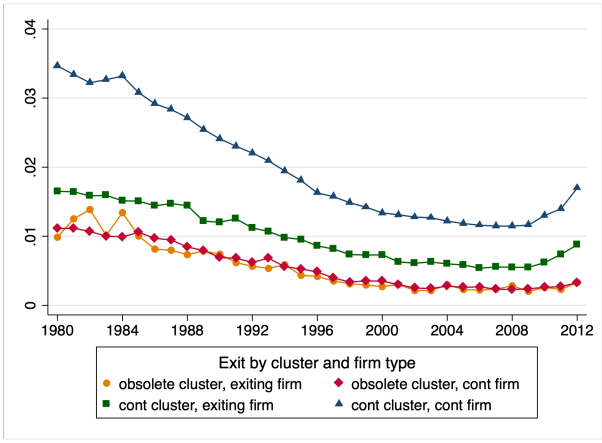


Figure 1.15: Exit component decomposition



Now focus on the time trend of the decomposition from Figure 1.13. We see that net entry component is decreasing over time and within firms component is going up. The former resonates with the decline in business dynamism that has accelerated in the high-tech industry since 2000. There are not only fewer new firms and new technology clusters entering the economy, but there is also a significant decrease in the entry of continuing firms to new technology clusters. While the reallocation of technologies is going down over time, firm's are becoming more specialized in their business activity. We can see it both by an increase in within firms component of the decomposition and by an increase in the average specialization of firms proxied by share of patents related to a particular technology cluster for a given firm, z_{ict} (Figure 1.12).

The analysis in this section has shown that reallocation of technology clusters between firms is the crucial source of technology cluster growth. Although this result is well-established in the theoretical literature on innovation growth, it has not got much validation from the empirical research. In particular, we saw that reallocation between continuing firms plays an essential role for technology growth. At the same time, reallocation rate are slowing down over time and especially in the recent decade. Beyond its absolute decline, the contribution of reallocation to technology cluster growth has been steadily decreasing too.

1.6.2 Reallocation of Technology Clusters and Growth

The decomposition of the annual changes in technology cluster size showed that reallocation of technology clusters between firms is important for technology growth. In this section, I elaborate on the role of reallocation for development of technology clusters and estimate the magnitude of the association between reallocation and technology growth.

I start with an event study that investigated patenting activity in a technology cluster around the time of reallocation from one technology leader to another one. For this exercise, I define reallocation of a technology cluster as a change in a firm that owns most of the patents that has ever been submitted to this technology cluster (as of a certain point in

time). For instance, if in year $t-1$ firm A has the biggest share of all patents in a technology cluster c among all other firms then firm A is a leader in this c at time $t-1$. Suppose in year t firm B has submitted so many patent applications to technology cluster c that its share of accumulated patents in c is now bigger than firm A's. Then firm B is now a new leader in cluster c and time t is the year of the reallocation event.[35]

The main event-study specification is as follows:

$$Patents_{ct} = \sum_{\tau=-5}^{5} \beta_\tau \mathbb{1}\{EventYear_{ct} = \tau\} + \gamma_t + \varphi_c + \varepsilon_{ct} \tag{1.13}$$

where the dependent variable, $Patents_{ct}$ is annual number of patents submitted to a technology cluster c in year t. Event year is defined as the year of reallocation of a technology cluster c from one leading incumbent to another as defined above. I am focusing on a 10-year horizon: for the event at time t, I am looking at the annual patenting in a cluster from $t-5$ till $t+5$. γ_t and φ_c are time and technology cluster fixed effects correspondingly.

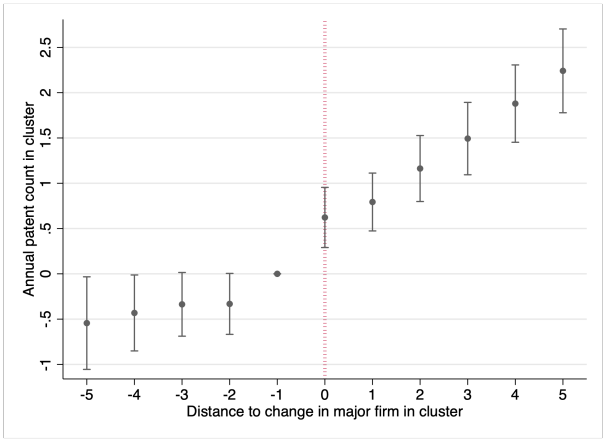
Figure 1.16 reports the results and plots the coefficients on the binary indicator of the distance to the reallocation event β_τ for $\tau = [-5, 5]$. Panel A shows the results with time fixed effects only while Panel B is based on the specification with both time and technology cluster fixed effects. We see that innovation activity in a technology cluster goes up after a change in the leading firm in this cluster. Note that this is not a one-time discontinuous jump, but rather a persistent increase in annual patenting observable during all the five years after the reallocation event that we are following a technology cluster.

Technology cluster dynamics and firm dynamic are closely related. Since startups are the main driving force standing behind the creation of new technology clusters, a decline in new business formation in high-tech sectors inevitably slows down entry of new technology clusters as we have seen previously. On the other hand, mature firms are essential for the growth of technology after entry. In this case, firm dynamism of incumbents should be a

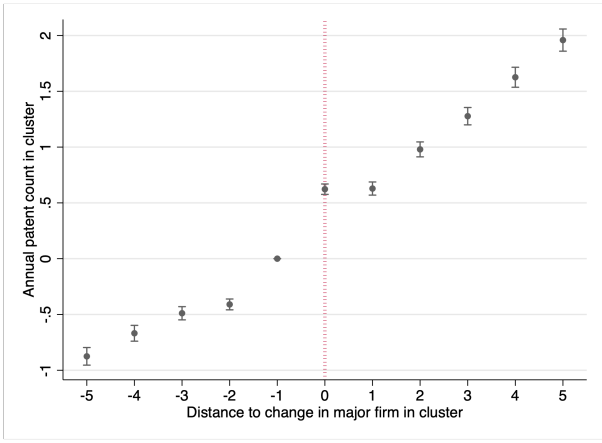
35. Note that technology clusters can get reallocated more than once in their lifetime. If this is the case, such technologies appear in the reallocation event study more than once.

Figure 1.16: Results of the event study according to specification (1.13).

Panel A: With time fixed effects only



Panel B: With time and technology cluster fixed effects



Notes: Reallocation event is defined as a change in the technology cluster's leading incumbent: a firm with the highest share of accumulated patent applications in a given technology cluster. The dependent variable is number of annual submitted patent applications that belong to a technology cluster c in year t. The omitted category is year before the reallocation even, $t-1$.

determining factor for technology cluster evolution. In the following exercise, I am going to explore how different measures of firm dynamic relates to technology cluster growth. The main regression specification for this is as follows:

$$Technology\_growth_{ct} = \beta Firm\_Dynamism_{ct} + \gamma Controls_{ct} + \gamma_t + \varphi_c + \varepsilon_{ct} \qquad (1.14)$$

where $Technology\_growth_{ct}$ is growth rate of technology cluster size, measured as share of patents submitted to cluster c in year t relative to all patents submitted in year t.[36] $Firm\_Dynamism_{ct}$ is one of the three measures of firm dynamism in this technology cluster. The first one is the same measure of reallocation as I used in the previous event study. It is an indicator of whether a leading firm in a technology cluster c has changed in year

36. In the rest of the paper, I am going to measure growth in annual rates rather than annual changes as in the decomposition exercise. While it is much easier to decompose *changes* in technology cluster size into several additive terms, growth *rate* is more intuitive for the interpretation of the regression coefficients. The results with annual difference as a dependent variable are available upon request and are qualitatively similar to the results of the main version of the analysis.

t. The second measure is a vector consisting of number of entering and exiting firms in technology cluster c in year t. The third measure of firm dynamism is net entry and between firms components from the growth decomposition (10). $Controls_{ct}$ is vector of technology cluster controls such as number of all active firms in a cluster, cluster age and cluster size. Regressions with only time fixed effects look at the across technology clusters variation and focus on exploring what kind of technologies grow faster. When we including technology cluster fixed effects, we are zooming on within a cluster variation which is a better fit for understanding what can augment a given technology growth.

The results of regression (1.14) are reported in Table 1.6. The first two columns focus on the first measure of firm dynamism in a technology cluster, columns (3) and (4) consider number of entering and exiting firms and the last two columns look at the relationship between reallocation components of the decomposition (9) and technology cluster growth rate. First and foremost, note that all the measures of firm dynamism in a technology cluster has a strong positive relation with the cluster's growth rate. In particular, a change in a leading firm in a technology cluster is associated with an increase in its growth rate by 22 percentage points. Not only firm entry rate to a technology cluster is growth-inducing but also firm exit rate.

Note that here I do not draw a distinction between entry of startups and entry of continuing firms to a technology cluster. Analysis in Appendix J distinguishes these two types of newcomers. I show that conditional on aggregate entry rate, more new startups in a technology cluster is actually associated with its slower growth. Section J also explores what kind of firms are conducive to technology growth by looking at the role of different characteristics of a typical firm in a technology cluster for its post-entry evolution. In particular, technology clusters with bigger and more mature incumbents grow faster than their peers.[37]

A change in a leading incumbent of a technology cluster can be a necessary but not

37. If we focus on within-cluster variation, it is also true that a technology cluster grows faster when bigger and older firms submit their patents to it.

Table 1.6: Firms reallocation and technology cluster growth

	Technology growth rate					
	(1)	(2)	(3)	(4)	(5)	(6)
$\mathbb{1}$\{Leader change\}	0.255***	0.221***				
	(0.025)	(0.027)				
#Entering firms			0.098***	0.122***		
			(0.006)	(0.008)		
#Exiting firms			0.010***	0.028***		
			(0.003)	(0.006)		
Net entry					18.91***	18.70***
					(0.635)	(1.499)
Between firms					13.17***	13.07***
					(0.331)	(1.095)
#All active firms	0.011***	0.011***	-0.021***	-0.030***	0.002	0.004*
	(0.002)	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)
Technology age	0.011***		0.007***		0.012***	
	(0.002)		(0.002)		(0.002)	
Technology size	-273.8***	-451.8***	-223.8***	-443.1***	-236.1***	-315.8***
	(59.29)	(125.1)	(57.74)	(120.0)	(51.61)	(82.96)
Observations	91,236	90,785	91,236	90,785	90,342	89,878
R-squared	0.314	0.352	0.330	0.377	0.448	0.479
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Technology FE	No	Yes	No	Yes	No	Yes

Notes: $\mathbb{1}$\{Leader change\} equals one if a firm with the highest share of accumulated patents in a technology cluster changes in year t. # Entering firms is number of firms that have entered a technology cluster in year t. # Exiting firms in number of firms that have exited a technology cluster in year t. Net entry and between firms are the corresponding components of the growth decomposition in (10) measured in percentage points. Mean of the dependent variable is 0.062. Standard errors are clustered on year level. *** p<0.01, ** p<0.05, * p<0.1

a sufficient condition for improvement of the cluster's evolution. There is evidence that technology clusters are getting reallocated to firms that are more productive in comparison to their previous incumbent leaders. On average, a new leading firm in a technology cluster has 17.6% bigger lifetime patent portfolio than an old leading incumbent of this cluster. When it comes to the difference between the quality of an average patent of a new dominant

firm and an old one, the former has on average 20.2% more citations per patent than the latter. If we compute a time-specific difference in quality of a new leader in a technology cluster and its old leader, we still see that new for any given year a new dominant firm is more productive than the old one (see Figure 1.17 and Figure 1.18). However, the average gap between the new and the old leader in a technology cluster is decreasing over time. If a new dominant firm is similar to the previous leader in a technology cluster then we should expect a smaller innovation boost resulting from this change. Hence, besides the decrease in reallocation rates that are slowing down technology cluster growth, it can also be suppressed due to the declining effect of reallocation on growth.

Figure 1.17: Average difference in lifetime patent count between a new leading firm in a technology cluster and an old leading firm.

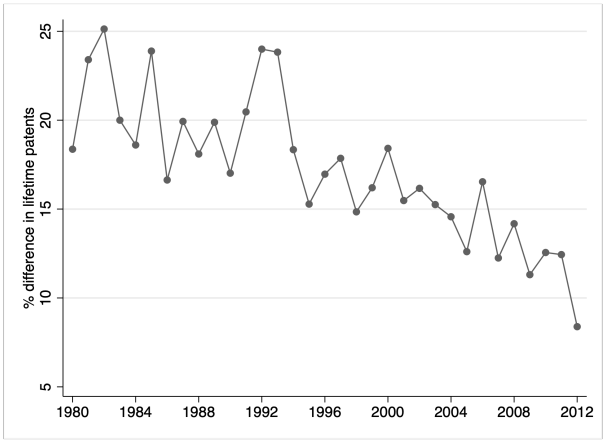
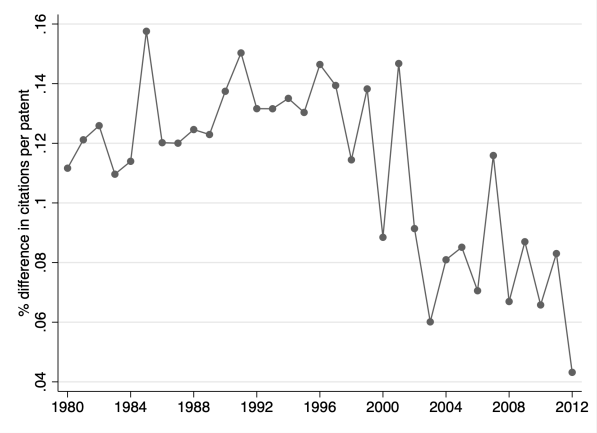


Figure 1.18: Average difference in citations per patent between a new leading firm in a technology cluster and an old leading firm.



Notes: Percentage difference is counted in percents of the old leading firm's value: $\frac{\text{new firm's value} - \text{old firm's value}}{\text{old firm's value}}$

1.7 Innovation Market Structure and Technology Cluster Network

In this section, I continue the research on the role of firm dynamism for technology cluster evolution but shift the focus to market structure of technology clusters. In particular, I am going to explore the importance of firms' concentration for the evolution of technology

clusters over time. If reallocation among firms is an important driver of technology cluster growth, we would expect that technology clusters with lower firm concentration and higher competition would grow faster. In this section, I test this hypothesis by regressing the growth of a technology cluster on the index of firms concentration in this cluster. I show that innovation market concentration plays an important role in making a technology cluster well-connected. Higher market competition and more diverse R&D efforts from different firms are conducive to a technology cluster's mobility from the periphery of the technology network to a more central position. Interestingly enough, this holds only for the observations before 2000 and has not been the case in the more recent decades. In what follows, I am going to discuss the potential forces standing behind this change.

The literature on the role of market concentration for innovation-driven growth does not give one unambiguous answer on whether competition is growth-inducing or not. In one of the first models of product variety from Romer (1990), competition decreases incentives to innovate and as a result suppresses growth. In the step-by-step innovation model where firms first have to catch-up with the industry leaders and only then can challenge them in a monopolistic competition, the implications are the opposite. Aghion et al. (2001) shows that escape the competition effect dominates Schumpeterian effect, making the relationship between competition and innovation growth positive. Later research by Aghion et al. (2005) reconcile these two opposite arguments and show that the relationship between competition and growth has rather an inverted-U shape.

1.7.1 Technology Cluster Growth and Competition

I start the analysis with exploring the relationship between innovation market concentration and growth of a technology cluster. Understanding this relationship is not only important per se, but also allows to look at the changing dynamics of technology cluster from a different angle. In the recent decades, we have seen a pervasive increase in market concentration across different industries, as shown in Autor et al. (2017) and Grullon et al.

(2019). Increasing concentration can affect the dynamics of technology clusters, which makes the analysis of this relationship valuable from the forecasting perspective as well. I use a patent-based Herfindahl-Hirschman index ss a measure of technology cluster concentration. Step one is to construct the distribution of citation-weighted patents across firms for each technology cluster in a given year. Figure **??** in Appendix A plots this distribution for some selected industries in 2000. The horizontal axis is number of citation-weighted patents submitted to this cluster by a firm in 2000. The vertical axis is share of firms who have submitted given number of citation-weighted patents. As before, I consider a firm to be active in a technology cluster if it has submitted patents to this cluster in the last five years. Thus, I take into account all the cluster-related patents of a firm submitted within the last five years. I summarize this technology- and year- specific distribution in Herfindahl-Hirschman index computed as:

$$HH_{ct} = \sum_{i} \left(\frac{pat_{ict}}{pat_{ct}} \right)^2 \tag{1.15}$$

where pat_{ict} is number of citation-weighted patents related to technology cluster c that firm i submitted from year $t-4$ to t. pat_{ct} is total number of citation-weighted patents related to technology cluster c submitted from year $t-4$ to t. HH index takes values from 0 to 1, where higher value is associated with higher concentration of firms in a technology cluster. The main regression specification is as follows:

$$Size\_gr_{ct} = \beta_1 HH_{ct} + \beta_2 Size_{ct-1} + \beta_3 Age_{ct} + \gamma_t + \varphi_c + \varepsilon_{ct} \tag{1.16}$$

where $Size\_gr_{ct}$ is the annual growth rate of the size of technology cluster c between t and $t+1$ measured in shares of patents submitted to this cluster. HH_{ct} is the firm concentration index of a technology cluster as defined in (15). I control on the initial size of a cluster by including lag of technology cluster size $Size_{ct-1}$ and on technology cluster age Age_{ct}.

Table 1.7 reports the results. Columns (1) and (2) use citation-weighted HH index while columns (3) and (4) use the unweighted version (i.e. each patent is counted with weight 1).

A consistently negative coefficient on the HH index in all specifications implies that higher concentration of firms in a technology cluster is detrimental for its growth. This holds regardless of whether we look at the within year variation or zoom in on the within year-sector variation. Note that the fact that we observe this negative relationship between growth rate and concentration controlling on year and technology cluster fixed effects implies that this result is not driven by selection and changing composition of active technology clusters.

Table 1.7: Market concentration and technology cluster growth

	Technology cluster's size growth rate			
	(1)	(2)	(3)	(4)
HH, weighted	-1.1301***	-1.0345***		
	(0.0462)	(0.0561)		
HH, unweighted			-1.3721***	-1.3110***
			(0.0397)	(0.0484)
Technology cluster size	-129.5***	-387.2***	-123.9***	-386.1***
	(18.51)	(77.05)	(18.52)	(76.67)
Technology cluster age	-0.0038***		-0.0045***	
	(0.0009)		(0.0009)	
Year FE	Yes	Yes	Yes	Yes
Technology FE	No	Yes	No	Yes
Observations	95,152	94,686	95,152	94,686
R-squared	0.327	0.358	0.331	0.361

Notes: Dependent variable mean is .054. Standard errors are clustered on the fixed effect level. *** p<0.01, ** p<0.05, * p<0.1

1.7.2 Technology Cluster Degree and Competition

Size of a technology cluster is not the only measure of the cluster's importance for the economy. Technology clusters that have many neighbors in the global network can transmit shocks to a big range of other technologies. Degree of a technology cluster captures the role of the cluster in the global technology network. As Table B1 in Appendix B implies, the most connected technology clusters are not necessarily the largest ones. As more and more

of their potential neighboring technologies are becoming active, some technology clusters are getting more central over time. Such dynamics imply that these technologies have bigger inter-cluster spillovers and can generate knowledge spin-offs. Other technologies remain isolated and keep standing alone with a zero degree throughout their whole lifetime.

In this section, I discuss the role of innovation market competition in determining the network spillovers that a technology cluster can generate. I measure the scale of these spillovers by the growth rate of a technology's degree or, in other words, growth rate of the number of a technology's links in the network. The goal of this exercise is to understand whether a single incumbent can bring a technology cluster from the periphery of the global technology network to the center or such evolution requires efforts of multiple firms. If the latter is the case, then as in the previous section, we should expect a negative relationship between firms concentration in a technology cluster and its degree growth rate.

The main regression specification is the same as in (1.14), but now the dependent variable is annual growth rate of technology cluster's degree:

$$Degree\_gr_{ct} = \beta_1 HH_{ct} + \beta_2 Size_{ct-1} + \beta_3 Age_{ct} + \gamma_t + \varphi_c + \varepsilon_{ct} \tag{1.17}$$

The results are reported in Table 1.8. Columns (1) and (2) use citation-weighted HH index and columns (3) and (4) use unweighted HH index. The negative sign of the coefficient on the HH index implies that technology clusters with less firm concentration have better chances of increasing their centrality in the network and generating spillovers. However, this relationship is to a large extent driven by the observations before 2000. If we separately consider a before-2000 and after-2000 subsamples, we see that there is no significant relationship between innovation market concentration and degree growth rate after 2000 (Table 1.9).[38]

38. For the analysis of the relationship between market concentration and technology cluster growth, splitting the sample into before-2000 and after-2000 parts does not change the significance of the negative coefficient. However, the magnitude is much smaller for the after-2000 subsample. The results are available in Table K1 in Appendix B.

Table 1.8: Market competition and network degree of a technology cluster

	Technology cluster's degree growth rate			
	(1)	(2)	(3)	(4)
HH, weighted	-0.0123***	-0.0095***		
	(0.0020)	(0.0021)		
HH, unweighted			-0.0142***	-0.0130***
			(0.0022)	(0.0023)
Technology cluster size	-0.553	-2.695***	-0.495	-2.769***
	(0.531)	(0.946)	(0.545)	(0.966)
Technology cluster age	-0.00028***		-0.00028***	
	(0.00008)		(0.00008)	
Year FE	Yes	Yes	Yes	Yes
Technology FE	No	Yes	No	Yes
Observations	118,403	118,290	118,403	118,290
R-squared	0.007	0.024	0.007	0.024

Notes: Dependent variable mean is 0.014. Standard errors are clustered on the fixed effect level. *** p<0.01, ** p<0.05, * p<0.1

Table 1.9: Change of relationship between market competition and technology degree after 2000

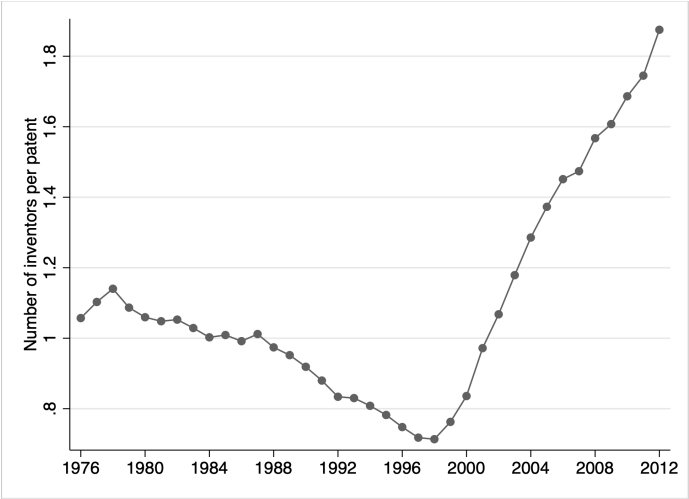
	Technology cluster's degree growth rate			
	(1)	(2)	(3)	(4)
HH, weighted	-0.0170***		-0.0011	
	(0.0025)		(0.0038)	
HH, unweighted		-0.0011***		-0.0009
		(0.0024)		(0.0057)
Technology cluster size	-1.927	-1.989	-3.053	-3.046
	(1.738)	(1.728)	(1.862)	(1.887)
Sample	pre-2000	pre-2000	post-2000	post-2000
Year FE	Yes	Yes	Yes	Yes
Technology FE	Yes	Yes	Yes	Yes
Observations	69,760	69,760	48,381	48,381
R-squared	0.035	0.035	0.030	0.030

Notes: Dependent variable mean is 0.014. Standard errors are clustered on the fixed effect level. *** p<0.01, ** p<0.05, * p<0.1

The analysis in this section establishes another connection between technology cluster growth, decreasing market competition and reallocation among firms. This decline in reallocation and competition can imply a change in the optimal resource allocation and market structure due to a shift in the production process. If the complexity of innovations is going up and it is harder to come up with new inventions (see (Bloom et al., 2020)), the fixed costs of research and development should go up, especially for newcomers. This in its turn should put an upwards pressure on the barriers of entry and, despite the concerns of anti-competitiveness effect on welfare, make more concentration optimal (see (Davis and Murphy, 2000) for a relevant case study and theoretical interpretation).

In the rest of this section, I elaborate on the plausibility of the increasing costs of innovation hypothesis. One of the key inputs in innovation production is human capital. If innovations are becoming more costly over time and in particular, require more human capital, this should be reflected on inventors employment. We would expect to see firms hiring more inventors to produce the same amount of innovation output. One way we can test this hypothesis is to look at how many inventors are needed to produce a unit of innovation and how this indicator is changing over time. As a measure of human capital for innovation, I use number of inventors stated in a patent filing. In order to account for patent quality and its innovation contribution, I divide the size of the inventors team by number of citations received by a patent (truncated at a 5-year horizon as before). This is my indicator of human capital costs per unit of effective innovation input. Figure 1.19 shows its evolution over time for an average patent. Interestingly, we see that this indicator is decreasing before 2000 but then the trend turns in the opposite direction – after 2000, more and more inventors are needed to create a patent of the same quality. This picture by itself would be interesting to study to pin down the cause of the drastic change in the trend. Appendix K discusses the possibility of innovation costs being an omitted variable in regressions (1.16) and (1.17) and show how the results of this and the previous section change if we include this regressor.

Figure 1.19: Average size of inventor's team per citation-weighted patent



1.8 Conclusion

This paper is an empirical study of technology cluster dynamic and network. One of the key contributions of this paper is methodological. I introduce a new method of identifying technology clusters in the data based on a match between patent text and Wikipedia articles. A patent belongs to a technology cluster defined by the category of the Wikipedia article that is most similar to the text of the patent. Hyperlinks between the articles and the granular structure of Wikipedia categorization scheme allows me to build a network of relations for technology clusters on the economy level and on a firm level.

I apply these new data to the study of technology cluster network structure, entry and evolution from 1980 till 2012. I find that young and small firms are more likely to discover a new technology cluster by undertaking breakthrough innovations. While startups are the main driving force behind technology cluster creation, a pervasive decline in business formation in high-tech sector has suppressed the entry of new technologies after 2000. At the same time, the share of technology clusters discovered by mature incumbent firms has been growing in the recent decades. These incumbent pioneers tend to have innovation expertise in technologies related to the new technology cluster, suggesting the importance of within-firm knowledge spillovers. Experienced incumbents are the main contributors to technology

cluster growth after its entry. This growth is a result of reallocation of technology clusters from less productive to more productive firms. However, reallocation forces are slowing down over time which puts a downward pressure on technology cluster growth. The decline in growth is also related to increasing concentration of firms in innovation market.

This paper shows one potential application of text analysis to economic research on innovations. I encourage researchers to use the opportunities that various text data resources, and in particular, patent data and Wikipedia data can offer. For example, the global network of technologies can help us to understand how a cluster-specific shock can transmit in the economy. Wikipedia articles make it possible to connect patents to particular events in history and study how demand shocks affect the direction of innovations. Wikipedia data give an opportunity to identify process innovations from product innovations. Pairing Wikipedia-based network with citation-based network can help us to build chains of inventions and study the path from an initial fundamental breakthrough to its first application in a mass-produced good. Finally, Wikipedia articles on inventors and firms can shed more light on people and organizations who are standing behind breakthrough innovations. Now, when the tools and techniques of machine learning are well-developed and computational power of our computers makes their implementation easy, it is a fantastic time to innovate on our study of innovations.

REFERENCES

Acemoglu, D., P. Aghion, and F. Zilibotti (2003). Vertical integration and distance to frontier. *Journal of the European Economic Association 1*(2-3), 630–638.

Acemoglu, D., P. Aghion, and F. Zilibotti (2006). Distance to frontier, selection, and economic growth. *Journal of the European Economic association 4*(1), 37–74.

Acemoglu, D., U. Akcigit, H. Alp, N. Bloom, and W. Kerr (2018). Innovation, reallocation, and growth. *American Economic Review 108*(11), 3450–91.

Acemoglu, D. and D. Cao (2015). Innovation by entrants and incumbents. *Journal of Economic Theory 157*, 255–294.

Aghion, P., N. Bloom, R. Blundell, R. Griffith, and P. Howitt (2005). Competition and innovation: An inverted-u relationship. *The quarterly journal of economics 120*(2), 701–728.

Aghion, P., R. Blundell, R. Griffith, P. Howitt, and S. Prantl (2009). The effects of entry on incumbent innovation and productivity. *The Review of Economics and Statistics 91*(1), 20–32.

Aghion, P., C. Harris, P. Howitt, and J. Vickers (2001). Competition, imitation and growth with step-by-step innovation. *The Review of Economic Studies 68*(3), 467–492.

Aghion, P. and P. Howitt (1990). A model of growth through creative destruction.

Akcigit, U., D. Hanley, and N. Serrano-Velarde (2013). Back to basics: Basic research spillovers, innovation policy and growth.

Akcigit, U. and W. R. Kerr (2018). Growth through heterogeneous innovations. *Journal of Political Economy 126*(4), 1374–1443.

Angrist, J. D. and J.-S. Pischke (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Argente, D., S. Baslandze, S. Moreira, and D. Hanley (2019). Patents to products: Innovation, product creation, and firm growth.

Argente, D., M. Lee, and S. Moreira (2018). Innovation and product reallocation in the great recession. *Journal of Monetary Economics 93*, 1–20.

Arrow, K. J. (1962). The economic implications of learning by doing. *Review of Economic Studies 29*, 155–173.

Atkeson, A. and A. T. Burstein (2010). Innovation, firm dynamics, and international trade. *Journal of political economy 118*(3), 433–484.

Autor, D., D. Dorn, L. F. Katz, C. Patterson, and J. Van Reenen (2017). Concentrating on the fall of the labor share. *American Economic Review 107*(5), 180–85.

Azar, J., I. Marinescu, and M. Steinbaum (2020). Labor market concentration. *Journal of Human Resources*, 1218–9914R1.

Balsmeier, B., M. Assaf, T. Chesebro, G. Fierro, K. Johnson, S. Johnson, G.-C. Li, S. Lück, D. O'Reagan, B. Yeh, et al. (2018). Machine learning and natural language processing on the patent corpus: Data, tools, and new measures. *Journal of Economics & Management Strategy 27*(3), 535–553.

Bartelsman, E. J. and M. Doms (2000). Understanding productivity: Lessons from longitudinal microdata. *Journal of Economic literature 38*(3), 569–594.

Bellis, M. (2018, 7). The inventor of touch screen technology. *ThoughtCo.*. Retrieved on 06/30/2020.

Bernard, A. B., S. J. Redding, and P. K. Schott (2010). Multiple-product firms and product switching. *American Economic Review 100*(1), 70–97.

Blagdon, J. (2013, 1). Samsung shows off curved phone prototype using flexible display. *The Verge*. Retrieved on 06/30/2020.

Bloom, N., C. I. Jones, J. Van Reenen, and M. Webb (2020). Are ideas getting harder to find? *American Economic Review 110*(4), 1104–44.

CEA (2016). Benefits of competition and indicators of market power. *Council of Economic Advisors*.

Covarrubias, M., G. Gutiérrez, and T. Philippon (2020). From good to bad concentration? us industries over the past 30 years. *NBER Macroeconomics Annual 34*(1), 1–46.

CPC (2017). Guide to the cooperative patent classification. *www.cpcinfo.org*.

Davis, S. J., R. J. Faberman, and J. Haltiwanger (2012). Labor market flows in the cross section and over time. *Journal of Monetary Economics 59*(1), 1–18.

Davis, S. J., R. J. Faberman, J. Haltiwanger, R. Jarmin, and J. Miranda (2010). Business volatility, job destruction, and unemployment. *American Economic Journal: Macroeconomics 2*(2), 259–87.

Davis, S. J., J. Haltiwanger, R. Jarmin, J. Miranda, C. Foote, and E. Nagypal (2007). Volatility and dispersion in business growth rates: Publicly traded versus privately held firms. *NBER macroeconomics annual 21*, 107–179.

Davis, S. J. and K. M. Murphy (2000). A competitive perspective on internet explorer. *American Economic Review 90*(2), 184–187.

De Loecker, J., J. Eeckhout, and G. Unger (2020). The rise of market power and the macroeconomic implications. *The Quarterly Journal of Economics 135*(2), 561–644.

Decker, R., J. Haltiwanger, R. Jarmin, and J. Miranda (2014a). The role of entrepreneurship in US job creation and economic dynamism. *Journal of Economic Perspectives 28*(3), 3–24.

Decker, R., J. Haltiwanger, R. Jarmin, and J. Miranda (2014b). The secular decline in business dynamism in the US. *Unpublished draft, University of Maryland 3*.

Decker, R., J. Haltiwanger, R. Jarmin, and J. Miranda (2016a). Where has all the skewness gone? The decline in high-growth (young) firms in the US. *European Economic Review 86*, 4–23.

Decker, R. A., J. Haltiwanger, R. S. Jarmin, and J. Miranda (2016b). Declining business dynamism: Implications for productivity. *Brookings Institution, Hutchins Center Working Paper*.

Diez, M. F. J., J. Fan, and C. Villegas-Sánchez (2019). *Global declining competition*. International Monetary Fund.

Dunne, T., M. J. Roberts, and L. Samuelson (1989). The growth and failure of us manufacturing plants. *The Quarterly Journal of Economics 104*(4), 671–698.

Evans, D. S. (1987). The relationship between firm growth, size, and age: Estimates for 100 manufacturing industries. *The journal of industrial economics*, 567–581.

Foster, L., J. C. Haltiwanger, and C. J. Krizan (2001). Aggregate productivity growth: lessons from microeconomic evidence. pp. 303–372.

Garcia-Macia, D., C.-T. Hsieh, and P. J. Klenow (2019). How destructive is innovation? *Econometrica 87*(5), 1507–1541.

Gersbach, H., M. T. Schneider, and O. Schneller (2013). Basic research, openness, and convergence. *Journal of Economic Growth 18*(1), 33–68.

Goldschlag, N., T. J. Lybbert, and N. J. Zolas (2019). Tracking the technological composition of industries with algorithmic patent concordances. *Economics of Innovation and New Technology*, 1–21.

Gomez, R. (2019). Geonamescache. *MIT license*.

Grossman, G. M. and E. Helpman (1991). *Innovation and growth in the global economy*. MIT press.

Grullon, G., Y. Larkin, and R. Michaely (2019). Are us industries becoming more concentrated? *Review of Finance 23*(4), 697–743.

Hall, B. H., A. B. Jaffe, and M. Trajtenberg (2001). The nber patent citation data file: Lessons, insights and methodological tools.

Haltiwanger, J. (2011). Firm dynamics and productivity growth. *European Investment Bank Papers 16*(1), 116–136.

Haltiwanger, J., R. Decker, and R. Jarmin (2015). Top ten signs of declining business dynamism and entrepreneurship in the US.

Haltiwanger, J., I. Hathaway, and J. Miranda (2014). Declining business dynamism in the us high-technology sector. *Available at SSRN 2397310*.

Hashmi, A. R. (2013). Competition and innovation: The inverted-u relationship revisited. *Review of Economics and Statistics 95*(5), 1653–1668.

Hathaway, I. and R. E. Litan (2014). What's driving the decline in the firm formation rate? a partial explanation. *The Brookings Institution*.

Hopenhayn, H. A. (1992). Entry, exit, and firm dynamics in long run equilibrium. *Econometrica: Journal of the Econometric Society*, 1127–1150.

Hyatt, H. R. and J. R. Spletzer (2013). The recent decline in employment dynamics. *IZA Journal of Labor Economics 2*(1), 5.

Kalemli-Ozcan, S., B. Sorensen, C. Villegas-Sanchez, V. Volosovych, and S. Yesiltas (2015). How to construct nationally representative firm level data from the orbis global database: New facts and aggregate implications.

Kelly, B., D. Papanikolaou, A. Seru, and M. Taddy (2018). Measuring technological innovation over the long run.

Klette, T. J. and S. Kortum (2004). Innovating firms and aggregate innovation. *Journal of political economy 112*(5), 986–1018.

Koehrsen, W. (2018). Wikipedia data science: Working with the world's largest encyclopedia. *Towards Data Science*.

Kogan, L., D. Papanikolaou, A. Seru, and N. Stoffman (2017). Technological innovation, resource allocation, and growth. *The Quarterly Journal of Economics 132*(2), 665–712.

Lentz, R. and D. T. Mortensen (2005). Productivity growth and worker reallocation. *International Economic Review 46*(3), 731–749.

Luttmer, E. G. (2007). Selection, growth, and the size distribution of firms. *The Quarterly Journal of Economics 122*(3), 1103–1144.

Luttmer, E. G. (2011). On the mechanics of firm growth. *The Review of Economic Studies 78*(3), 1042–1068.

Marx, M. (2019). Patent citations to science. *Ann Arbor, MI: Inter-university Consortium for Political and Social Research*.

Olley, G. S. and A. Pakes (1992). The dynamics of productivity in the telecommunications equipment industry.

Pugsley, B. W. and A. Şahin (2019). Grown-up business cycles. *The Review of Financial Studies 32*(3), 1102–1147.

Romer, P. M. (1990). Endogenous technological change. *Journal of political Economy 98*(5, Part 2), S71–S102.

APPENDIX A

ADDITIONAL FIGURES

Figure A1: Evolution of network: average size of active technology clusters, measured in absolute terms (number of patents submitted to a technology cluster annually) and in relative terms (share of patents submitted to a technology cluster relative to all patents in a given year)

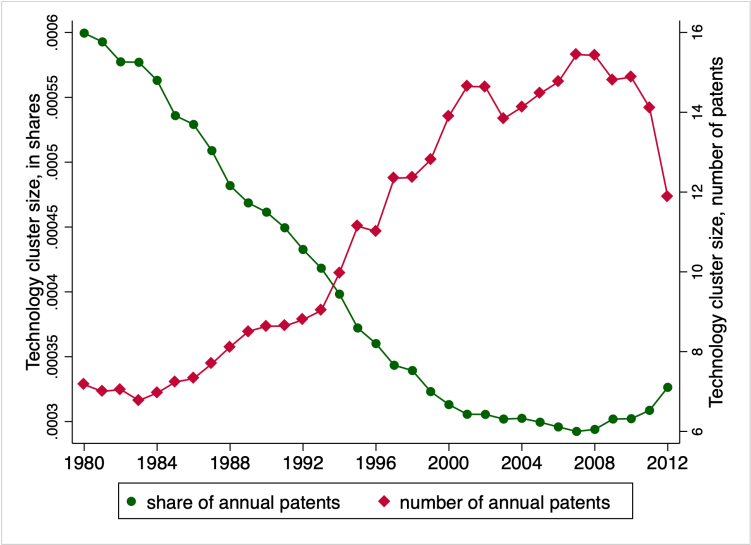


Figure A2: Evolution of network: active technology clusters and links between them

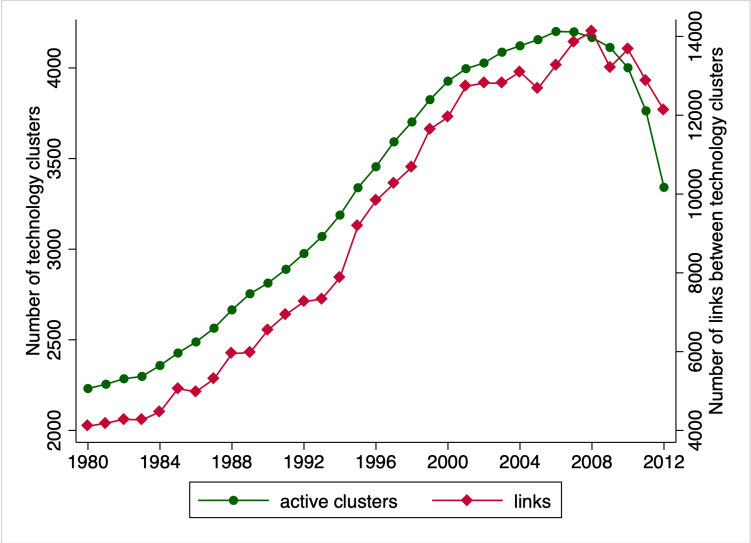
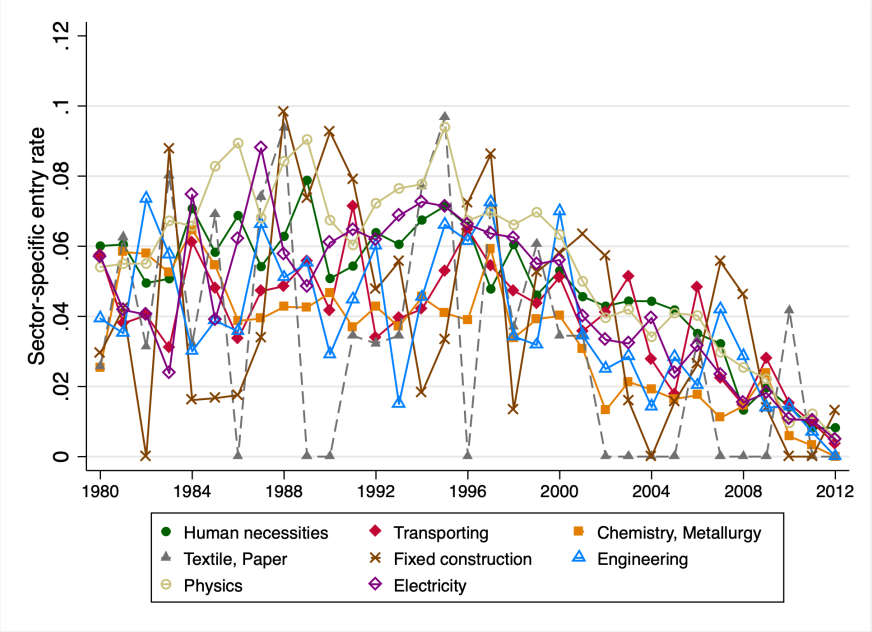
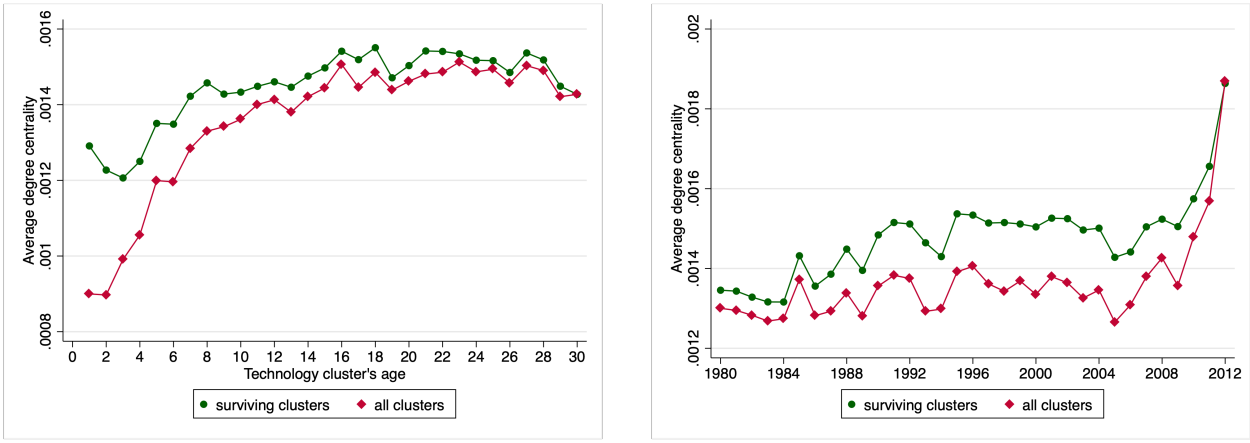


Figure A3: Entry rate of technology clusters by innovation sector



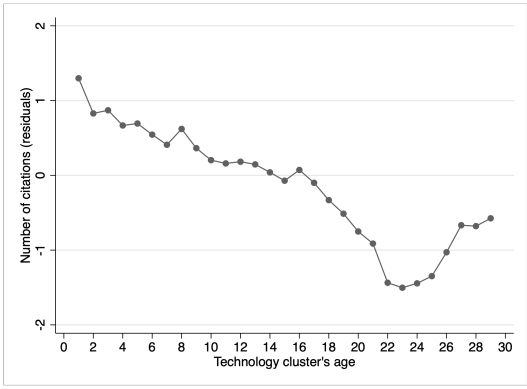
Notes: Entry rate is a ratio of number of new technology clusters to number of all active technology clusters. Technology clusters are allocated to an innovation sector where the majority of technology cluster's lifetime patents belong to.

Figure A4: Evolution of average degree centrality of a technology cluster over its age and over time

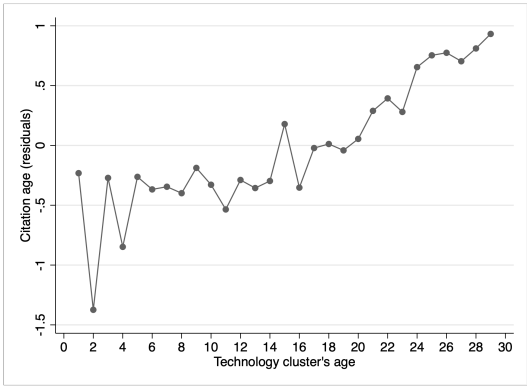


Notes: Survivors are technology clusters that stay active for more than 30 years.

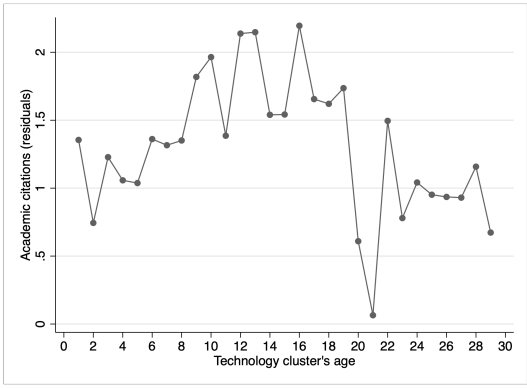
Figure A5: Evolution of different patent characteristics over technology cluster's age



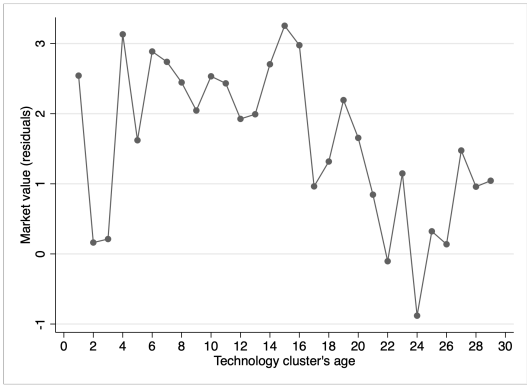
(a) Number of received citations



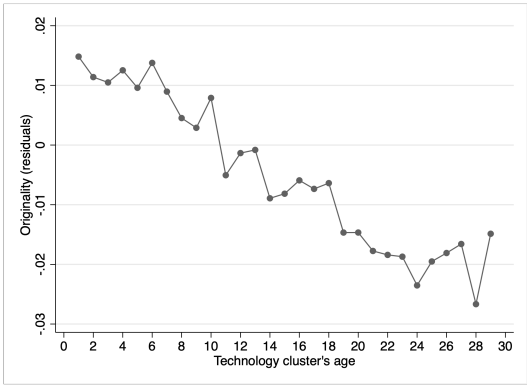
(b) Citation age



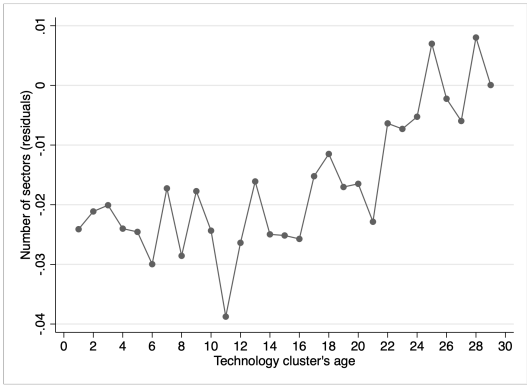
(c) Number of academic citations given



(d) Market value



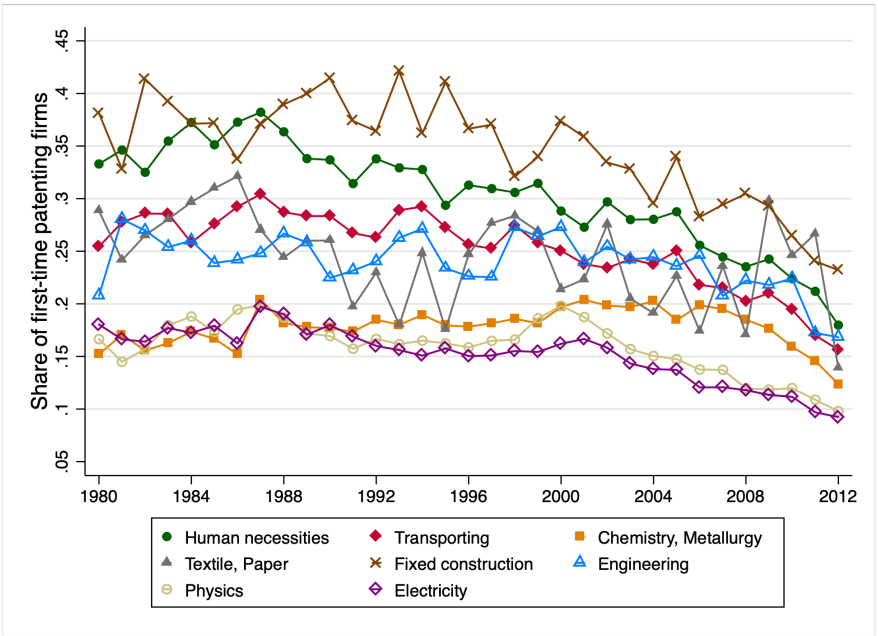
(e) Originality



(f) Number of sectors

Notes: For every cluster and every age, the average across patents belonging to this technology cluster is computed. Then these averaged indices are regressed on year fixed effects. Horizontal axes plots residual values of the corresponding dependent variable averaged by cluster's age.

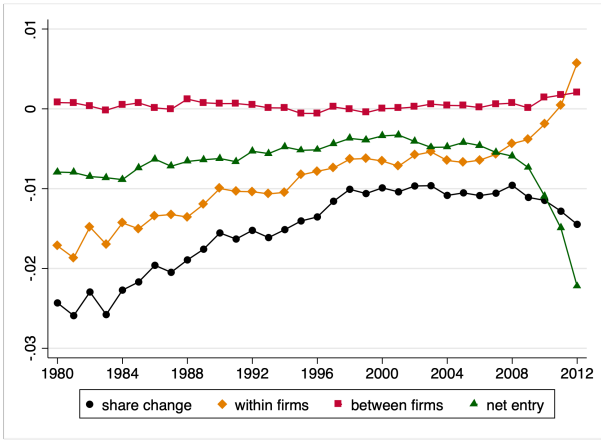
Figure A6: Share of new firms among all active firms in different innovation sectors



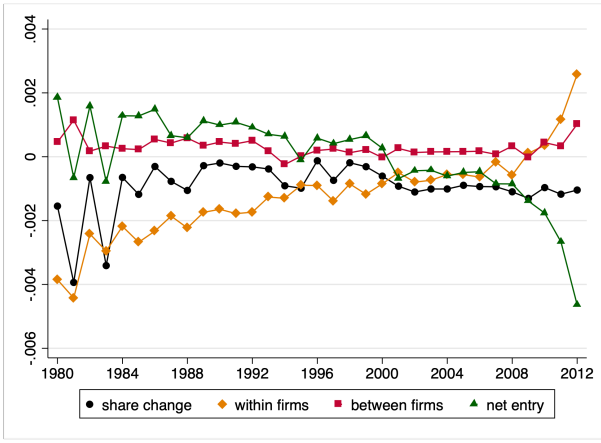
Notes: New firms are defined as firms with the first lifetime patent. Technology clusters are allocated to an innovation sector where the majority of technology cluster's lifetime patents belong to.

Figure A7: Distribution of exiting firms in the economy and in exiting technology clusters

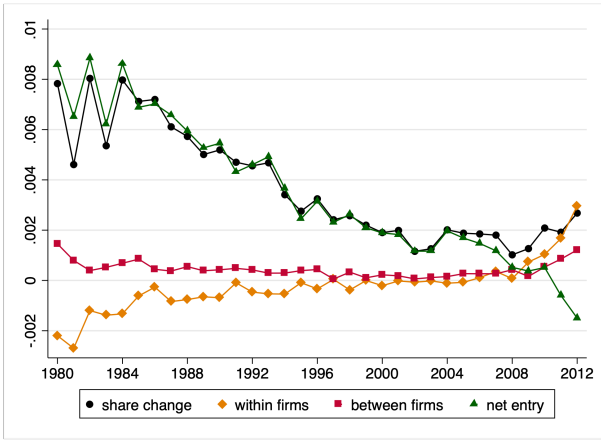
Figure A8: Decomposition of technology cluster growth



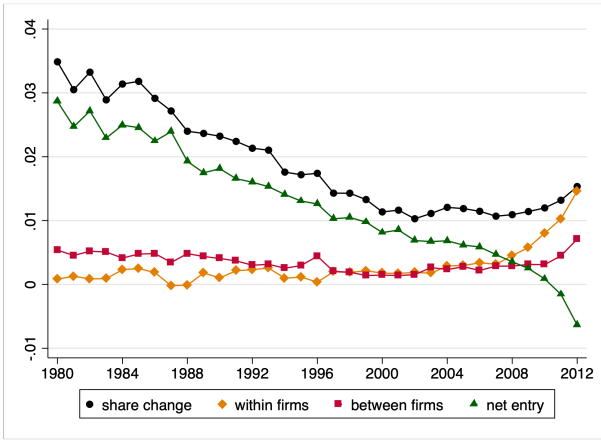
(a) 1st quartile



(b) 2nd quartile



(c) 3rd quartile



(d) 4th quartile

Figure A9: Standard deviation of technology cluster growth by year

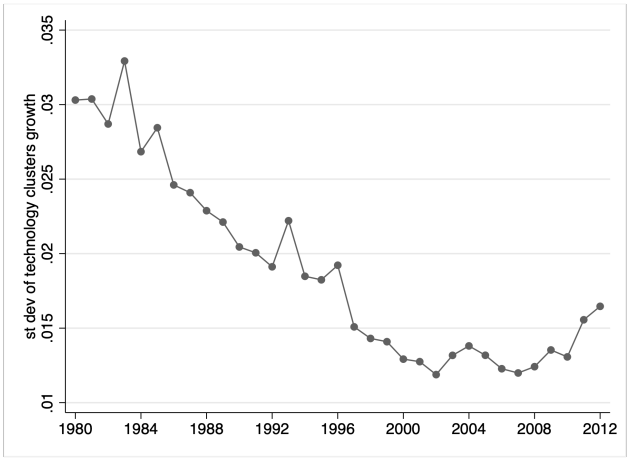
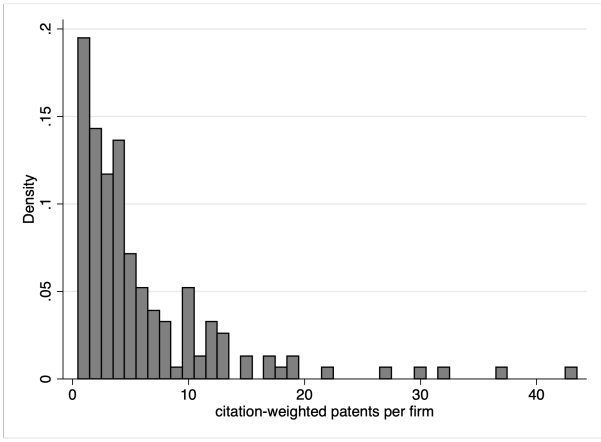
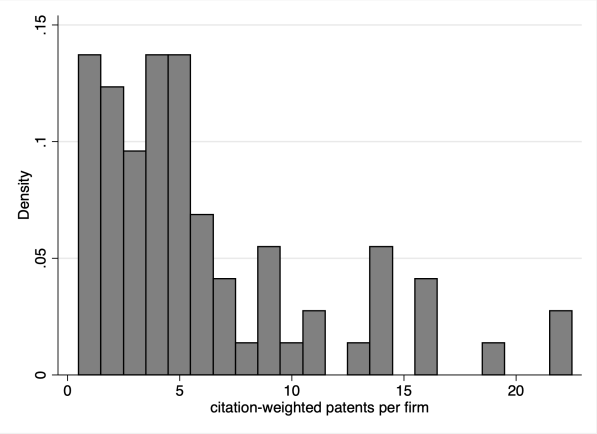


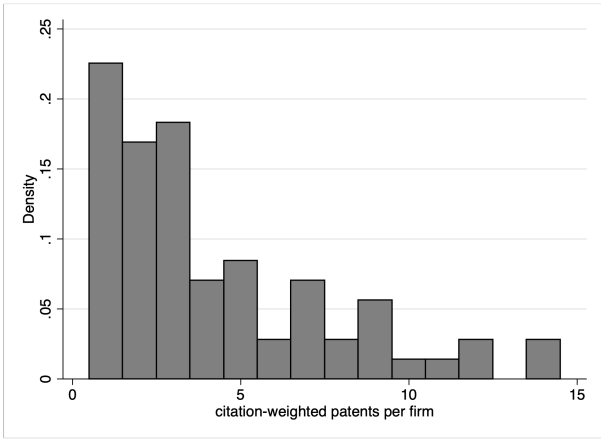
Figure A10: Examples of citation-weighted distribution of firm's patents in different technology clusters in year 2000



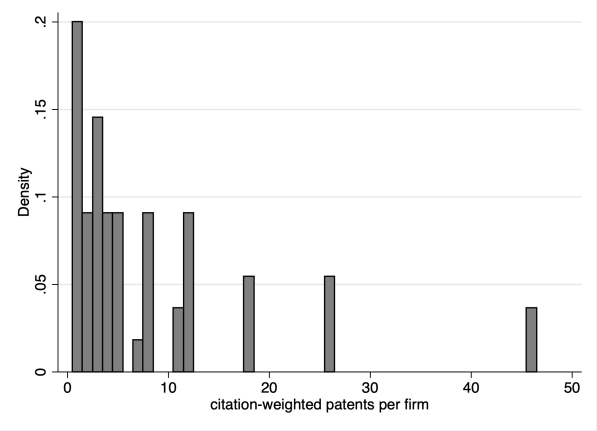
(a) Electrical cable



(b) Coated paper



(c) Packaging machinery



(d) Vitamin

APPENDIX B

ADDITIONAL TABLES

Table B1: Top technology clusters by patenting and degree

Largest	Share	Most connected	Degrees
Optical fiber	1.64%	Human activities	151
Internal combustion engine	1.38%	Engineering	133.2
Battery	1.03%	Management	109.2
Laser medicine	0.94%	Chemistry	105.5
Digital circuits	0.77%	IT Management	104.5
Semiconductor fabrication	0.74%	Applied Physics	101.7
Network synthesis filters	0.71%	Design	101
Electrical network	0.66%	Data management	84.7
Printing	0.61%	Chemical elements	84.17
Cleaning	0.55%	Computer security	73.5

Notes: Left part reports the largest technology clusters measured by number of patents submitted to them. The right part reports the most connected technology clusters measured by cluster's degree. This statistics is not time-specific and uses all patents in the sample.

Table B2: Summary statistics of patent characteristics used in Table 1.2

	Mean	St Dev	5%	25%	50%	75%	95%
Number of citations	4.36	9.26	0	0	2	5	17
Average citation age	7.39	6.43	2	4.25	6.6	9.67	15.5
Number of academic citations	4.24	20.15	0	0	0	2	17
Dollar value of a patent (mln)	14.67	36.13	0.08	2.01	6.20	13.11	56.06
Originality	.46	0.29	0	.24	.50	.69	.82
Number of industries	1.23	0.47	1	1	1	1	2

Table B3: Market competition and technology growth before 2000 vs. after 2000

	Technology cluster size growth rate			
	(1)	(2)	(3)	(4)
HH, weighted	-1.1401***		-0.5250***	
	(0.0449)		(0.0723)	
HH, unweighted		-1.3438***		-0.8635***
		(0.0429)		(0.1002)
Technology cluster size	-468.0***	-465.0***	-1,079**	-1,079**
	(31.81)	(31.99)	(411.7)	(411.7)
Sample	pre-2000	pre-2000	post-2000	post-2000
Year FE	Yes	Yes	Yes	Yes
Technology FE	Yes	Yes	Yes	Yes
Observations	53,283	53,283	41,019	41,019
R-squared	0.169	0.175	0.455	0.456

Notes: Dependent variable mean is .054. Standard errors are clustered on the fixed effect level. *** p<0.01, ** p<0.05, * p<0.1

APPENDIX C

LITERATURE REVIEW

The literature on growth through innovations do not have an unambiguous answer on whether it is new or incumbent firms that introduce new technologies to the market. Early Schumpeterian models predict that all innovations are undertaken only by new firms because of the Arrow replacement effect (see (Arrow, 1962)). A canonical example is Aghion and Howitt (1990) where a firm is allowed to have only one product and all innovations are coming from new firms that replace incumbents by improving the quality of a product and capturing the whole monopoly profit. Klette and Kortum (2004) give a start to Schumpeterian models that allow firms to have multiple product lines but the set of possible varieties is fixed. Luttmer (2011) extends the Klette and Kortum model of endogenous growth model with firms producing a set of differentiated commodities to capture size distribution of firms in US economy and connect it with organization capital rather than heterogeneity in productivities per se.

The main problem with the pioneer Schumpeterian models where growth is generated only by new entry is that their main implication is not supported by the data even in earlier studies. Foster et al. (2001) shows that only a quarter of productivity growth in U.S. is driven by new entry while the rest is contributed to continuing plants. More recent models of endogenous innovation growth are built to capture the contribution of incumbent firms to growth. In Acemoglu and Cao (2015), new entrants are engaged in more radical innovations while continuing firms improve the quality of existing product lines. Even though their model is Schumpeterian by nature, the argument that three quarters of productivity growth is coming from incumbent firms holds in their calibration. My findings also resonate with this conclusion. I show that it is new young firms that are more likely to come up with radical innovations that lay a foundation to a new technology cluster. But when it comes to growth, the contribution of continuing firms is the main force behind technology cluster growth. Akcigit and Kerr (2018) add another level of heterogeneity in endogenous growth models

and look at the innovation strategies of firms of different size. According to their growth decomposition, more than half of the innovation-driven growth is coming from incumbent firms capturing product lines of other firms or creating new product the ones.

This result implies that reallocation of resources and products across firms is an important force standing behind growth of firms, industries and economy as a whole. My paper speaks to this brunch of literature by studying reallocation of technology clusters across different kinds of firms, looking at how these patterns are changing over time and what potential causes and consequences of this changing dynamic are. Like Akcigit and Kerr (2018), many models of innovation-driven growth emphasize the importance of across-firm reallocation starting from Lentz and Mortensen (2005) who explain the role of worker reallocation for productivity growth in an innovation model. All growth in their framework comes from the reallocation of resources to firms that come up with superior products from firms that lose market.

Among more recent research, Garcia-Macia et al. (2019) build a model where innovation can come in three different ways: classic creative destruction mechanism, improvements of existing product lines or introduction of new varieties. Using U.S. LBD data they show that most of the growth is coming from the incumbent firms and its primary source is improvement of existing product lines, not introduction of new varieties. Creative destruction also does not have such a strong effect on growth as predicted by early Schumpeterian models. Moreover, they show that the contribution of new firms and creative destruction declined from 1983-1993 to 2003-2013. My paper discovers the same trend using patent data from 1980 to 2012. I show that even though reallocation between old and new firms has never been the main driver of growth, relative contribution of newcomers to the innovation market has been steadily decreasing in the last decades.

Acemoglu et al. (2018) elaborate on the importance of reallocation from low-capacity to high-capacity firms and present a model with multi-product firms of heterogeneous quality where reallocation is the main driver of innovation growth. They show that large fraction of

small firms are not growth-oriented and thus most of the reallocation happen between large incumbent firms. At the same time, their model is far from being 'one-sided' – the firms can exogenously switch their types from high-capacity to low-capacity, which can lead to large mature firms ceasing to innovate and falling behind eventually. On the contrary, the selection forces implies that there should be more and more high-type firms in the economy over time.

One of the papers that is most relevant for this research is Argente et al. (2018) who explicitly focus on the reallocation of products across firms (UPCs). They study the cyclicality of product creation and destruction using Nielson scatter data from 2007 to 2013 that focuses on consumer goods and show that most of the reallocation is done by continuing firms. My paper can be seen as a continuing research on product lines reallocation and goes beyond consumer goods, looks at the long time horizon, and as a result is able to address a bigger variety of the questions related to innovation and industry dynamics.

The evidence of the importance of reallocation for growth can be found outside of the literature on innovation-driven growth. Bartelsman and Doms (2000) summarize findings from the early research on the role of entry and resource reallocation for productivity growth based on longitudinal microdata. Even in these early micro studies it was clear that reallocation of resources explain a large share of productivity growth, which make the Schumpeterian idea of creative destruction even more reassuring at that time. Haltiwanger (2011) discuss cross-country comparison of allocative efficiency and its relationship to growth. Bernard et al. (2010) is one of the earliest papers on product switching using quinquennial US Manufacturing Censuses from 1987 to 1997. They define "product" as one of approximately 1,500 five-digit Standard Industrial Classification (SIC) categories and show that half of the firms in their sample change their product portfolio every five years.

The results of this paper are closely connected to the literature on the decrease in firm dynamism. Decker et al. (2016b) review the existing literature documenting the decline in business dynamism in US and discuss the potential explanations through the lens of canonical

models of firm dynamics. I show that young and small firms are crucial for introduction of new technology clusters and as a result, the slow-down in firm formation echoes in the decrease of technology cluster entry rate. Davis et al. (2007) connect the decline in entry and exit rates with the decline in volatility and growth dispersion among private firms in US. Since volatility is decreasing with firms age, a shift towards older business would lead to less aggregate volatility. I document a decrease in growth dispersion both across technology clusters and within.

The decrease in firm dynamics is quite pervasive across different industries. Decker et al. (2014b) find most of the decline in the dynamism is within a sector, not due to changes in the composition of sectors in the economy. They also emphasize that information and high-tech business sectors showed rise in startup activity before 2000 while the dynamism Mom and Pop startups in retail and service sectors showed a decline in dynamism for a long time. However, after 2000 high tech businesses also showed slower entry and decline in high-growth businesses. Haltiwanger et al. (2014) specifically focuses on high-tech startups and show that they are not an exception from a pervasive decline in business dynamism after 2000. Hathaway and Litan (2014) takes a first step to provide an explanation for the decline in firm formation rate and decline in business dynamism. They emphasize the role of slowing population growth and higher rates of business consolidation. Pugsley and Şahin (2019) connect the decline in startup activity to increase in import competition and changes in the demographic structure of the population. The argument that a decrease in trade costs can lead to a decrease in entry of new firms and products was earlier raised by Atkeson and Burstein (2010) in a general equilibrium model where increase in trade discourages so-called "product innovations" – creation of new firms with its new firm-specific product.

The juries are still out on the implication of the decline in firms dynamism for the economy. Davis et al. (2010) relate a decline in business variation to job destruction rate and discuss the implications of these processes for the changes in the unemployment flows. At the same time, Hyatt and Spletzer (2013) attribute a fairly modest explanatory power of decline

in employment dynamics to net job creation and changing composition of businesses. Decker et al. (2014a) show that it is not just the rate of business formation that has plummeted in the last decades but also the share of employment accounted by young firms have been decreasing for the last three decades. As a result that should slow down the reallocation of resources among firms. Decker et al. (2016a) connect a decline in skewness and across-firm dispersion of growth rate to a decline in the role of young firms in the economy and their decreasing role in generating growth.

My paper also relates to the literature on the role of market concentration for growth. I show that higher market competition in a technology cluster associates with faster growth in technology cluster size and its position in the global technology. However, the relationship between technology degree growth and firms concentration has changed after 2000. The literature on the role of market competition for innovation-driven growth is very rich with different arguments about the causal channels between competition and innovations. Non-Schumpeterian AK-style models cannot say anything on this due to perfect competition assumption. In one of the first models of product variety from Romer (1990), competition decreases incentives to innovate and as a result suppresses growth. That is based on the assumption that all innovations are made by firm-outsiders who either become monopolists if they succeed or get nothing if they fail.

In the step-by-step innovation model where firms first have to catch-up with the industry leaders and only then can challenge them in a monopolistic competition, the implications are the opposite. Aghion et al. (2001) show that escape the competition effect dominates Schumpeterian effect, making the relationship between competition and innovation growth positive. Aghion et al. (2005) reconcile these two opposite arguments in the model where innovation incentives depend on incremental profits from innovation rather than post-innovation profits. On the one hand, competition discourages laggard firms from innovations since the chances of entering the market are low. On the other hand, it boosts innovations among neck-and-neck firms who can only escape the competition if they win the innovation race. The inverted-U

relationship between competition and innovations comes from the endogenously determined composition of industries. When competition is low, neck-and-neck firms are predominant in most industries but when it is high, a large share of industries would have laggard firms performing most of innovations.

Several studies show that distance to technological frontier is another factor that should be taken into consideration when talking about the effect of competition of innovations. Aghion et al. (2009) show that higher competition encourages innovations in frontier firms while discouraging them in the firms that are far from the frontier. While Aghion et al. (2005) calibrate their model using U.K. data, Hashmi (2013) finds a mildly negative relationship between growth and competition in US data after instrumenting out the possible engodeneity of competition (using trade-weighted average of industry exchange rate). He explains these results by the difference in composition of neck-and-neck industries in the U.K. and U.S. data – the exact shape and steepness of this relationship can depend on the average technological gap between firms or, in other words, share of firms that are more neck-and-neck. Acemoglu et al. (2003) make a different argument on how competition can affect innovations. They show that if we take into account time endowment of managers and span of control, managerial overload resulting from excessive vertical integration would suppress innovations making high concentration suboptimal.

To a large extent, understanding the forces standing behind the evolution of market concentration are crucial for studying the relationship between concentration and growth and why it can be changing over time. Otherwise, there is a risk f omitting a potential control variable – the factor that affects both market concentration and technology cluster growth in size and network centrality. According to Luttmer (2007), selection can be one of these factors. Lack of selection due to either increase in the costs of imitation or entry cost would lead to a decrease in firm concentration. While I discuss the importance of reallocation for growth in the main body of the paper, selection of firms into the economy has an impact on growth as well. However, the scale of the effect that selection has on the

innovations and growth can also be heterogeneous across countries and years. Acemoglu et al. (2006) show that in the world where both innovation and adoption from the frontier are possible, selection of the right entrepreneurs is becoming more important as a country moves towards the frontier. Such countries are more likely to innovate than adopt and this requires more talent and highly capable human capital for that. Luck of selection in such economic environment would mean a decrease in innovation growth.

The most relevant literature on potential consequences of the increase in market concentration besides a slow-down in growth is fairly recent but expands rapidly. Covarrubias et al. (2020) discuss possible causes of decrease in competition, labor share and investment share in US and tries to understand the consequences of these shifts depending on their sources. If this increase in concentration is caused by increasing elasticity of substitution or technological changes leading to increasing returns to scale and intangible capital deepening then the concentration should not be a worrisome phenomenon. On the other hand, if the barriers of competition have risen then concentration is not optimal. They conclude that the increase in competition that was documented before 2000 can be attributed to "good" factors but after 2000 market concentration was inefficiently high and its continued increase was driven by higher barriers to entry.

On a similar note, Grullon et al. (2019) show that increase in market concentration does not lead to an increase in operational efficiency, questioning the positive returns to scale argument for many M&A deals in the recent decades. Even though the return on asset is getting higher, it is mainly due to higher profit rather than to higher efficiency. While higher market concentration is associated with higher markups, De Loecker et al. (2020) show that the increase in markup is mainly driven by the firms in the upper tail of the distribution, while the median remains unchanged. The reallocation from low-markup to high-markup firms is within industries and can explain the declining labor share and decrease in labor market dynamism. Azar et al. (2020) connect increase in product market concentration with increase in labor market concentration and recent changes in monopsony power.

Finally, from the methodological point of view this paper relates to the literature that uses natural language processing in research on innovations using patent data. Balsmeier et al. (2018) show how natural language processing can be applied to the patent data to disambiguate inventors, companies and create new indicators of innovation activity. Kelly et al. (2018) uses natural language processing to build similarity metric for pairs of patents to identify the most important patents. Their approach of computing the similarity score between pairs of patents is very similar to my measure of similarity between patents and articles. Argente et al. (2019) use the same approach of similarity score comparison to match patents to about a thousand of Nielson consumer products by their description. They point out that focusing on patents in research on products can show us only a tip of the iceberg since many product innovations are not patented. But for patenting firms, patent-based metrics of innovations are correlated with product-based measures of innovation quality and quantity.

APPENDIX D

ALTERNATIVE WAYS TO DEFINE TECHNOLOGY CLUSTERS

In this section, I discuss other potential approaches of defining technology clusters in the data. I consider several possible measures and elaborate on the advantages of using my approach over the other alternatives. I start with summarizing the features of my identification of technology clusters that make it a good fit for speaking to the models of endogenous growth and for dynamic analysis.

Property 1: Capture and summarize firms' innovation activity

Property 2: Narrow enough to encompass multi-product firms operating in multiple technology clusters

Property 3: General enough to avoid firm-specific technologies and study long-run dynamics

Property 4: Broad sample with many industries presented

Property 1 simply requires that a technology cluster measure reflects the nature of the concept used in models of innovation-driven growth. It should group innovations of similar type that feed into a certain product line. While in many models the map between a technology type and product line type is one-to-one or even one-to-many, this should not necessarily be true. One product line can use several different technologies in its production while the same technology cluster can be applied to multiple different product lines. Thus, the assumption that identification of product lines in the data also retrieves technology sectors due to one-to-one mapping is with loss of generality.

Property 2 is important for capturing the framework of multi-product firms that can have several product lines and innovate on multiple technologies simultaneously, enter new technology clusters or exit their current ones without completely leaving the economy. While

appropriate granularity of the measure is crucial for capturing entry and exit of technology clusters, a certain level of generality is essential for the analysis of technologies reallocation between firms. Property 3 implies that iPhone or Tide Pods cannot be technology clusters, but smartphone and detergent pods can. By the same token, technologies cannot be tied to one very specific product – otherwise we will not see any dynamics in a technology cluster besides creation and destruction. Finally, property 4 requires a fairly broad representation from different industries for the sake of external validity of results.

The measure that I introduce in this paper checks all the four boxes. By design, it captures innovation activity of patenting firms since I am using a detailed description of their inventions from patent texts to infer the corresponding technology cluster they are working in. It is narrow enough that a median firm have multiple technology clusters at a given point in time and thus speaks to the models of multi-product firms. Due to sufficient granularity, we can also capture entry and exit of technology clusters that affects the dynamics of the product lines where they are used as an input. At the same time, the proposed technology clusters measure is general enough to explore long-run evolution of technologies and allow several firms to operate in the same technology cluster simultaneously. Finally, it is fairly broad to have a good distribution of firms across different industries and fields of innovation, from food and household suppliers to bio-pharma and software developers.

One of the possible alternatives for technology cluster proxy is the North American Industry Classification System. NAICS is a standard used by many government agencies to classify business establishments by type of their economic activity. It does capture firm's business activity and permits to work with a representative sample of firms from various industries. However, it allows us only to approximate technology clusters and relies on the "one technology – one product line" assumption. Even if we are willing to accept this assumption, NAICS codes are too general to identify product lines in the way that will allow us to explore their dynamics. The most narrow classification, NAICS-6 digit code, have only 1057 unique industries and almost all companies have only one unique code that describes

the main industry of a firm's operation with rare changes between codes throughout the lifespan of a firm. For instance, Amazon has NAICS code 454110 that hasn't changed for decades – Electronic Shopping and Mail-Order House. It does describe the main business activity of the company in general, but it does not capture the whole range of its products such as its streaming service Amazon Prime Video, virtual assistant Amazon Alexa or no checkout store Amazon Go. Such lack of granularity makes NAICS a bad choice for the analysis of technology clusters dynamics and reallocation.

Another alternative is Universal Product Code (UPC) from Nielson Retail Scanner Database. This approach also leans on the assumption of one-to-one mapping between technology clusters and product lines, but in comparison to NAICS codes, UPC does capture actual product lines with about 1100 unique product categories in the data. However, the sample includes only consumer goods that can be purchased in grocery and convenience stores. As a result, such a choice of business lines categorization restricts our sample of firms and products to a particular selection. In addition, this data set starts only in 2006 which makes studying long-run trends of sectors dynamism challenging.

Since I am focusing on innovating firms and relying on patents as a source of firms' business description, patent classification used by USPTO would be a natural candidate. The two patent classifications employed by USPTO are their traditional USPC (United States Patent Classification) system and a fairly new system in developed in collaboration with European Patent Office CPC (Cooperative Patent Classification). In 2013, USPTO switched from using its traditional classification USPC to CPC and stopped maintaining USPC records for patent application submitted after 2013. While the crosswalk between USPC and CPC classifications allows to transform one classification into another, this concordance is only probabilistic. This lack of bijection is due to different granularity of old and new classification. While this paper's main working sample stops in 2012 (and full processed data stops in 2015), my approach is easily extrapolated to further years without any frictions as new data.

Both of USPC and CPC have the same underlying goal and application. The main objective of these patent systems is to classify structural features and contribution of patents over previous inventions (i.e. patent claims), not necessarily the field of application and technology overall (see (CPC, 2017)). This is done to facilitate search of patent claims and help inventors with proper citations of preceding patents. This objective for patent classification implies that a firm's patent class summarize technical contribution of a firm, but not its activity. An easy way to see it is look at the mapping between patent classification and industry classification and realize that it is far from being one-to-one. Goldschlag et al. (2019) matches patent classes to NAICS industry codes by comparing a patent's title and abstract with industry code description to find the most relevant industry codes for this patent. A median USPTO patent class has 6 different NAICS-6 digit codes that can be associated with it, while as discussed above, NAICS codes are more general than product lines.

While this feature may actually be desirable as it definitely does not rely on a "one technology - one product" assumption, the fact that a median patent class can be matched to two 1-digit NAICS sectors implies a very loose mapping between patent classes and product lines. If we look at the probabilistic distribution of these potential industries, it is not the case that most of the classes are matched to one main industry with high probability – the mean probability of the match between USPTO patent class and NAICS-6 industry code is 0.15 with median of .06 (among matches with positive probabilities).[39] Visual examination of this match and patent content implies that a patent assigned to "Electrical resistor" patent class can refer to Computer and Electronic Products, or Construction of Buildings, or Fabricated Metal Products, or Electrical Equipment. A patent from USPTO class "Lamp" is more likely to refer to Paper Manufacturing than to Electrical Equipment.

Though patent classifications are carefully constructed to classify technical innovation

39. USPC has 438 classes. For 3 digit NAICS, a median USPTO class is matched to 4 industries, for 1 digit NAICS – to 2. Median weights of these concordances are .09 for NAICS-3 and 0.39 for NAICS-1.

contribution of patents, they are too coarse to uncover potential application of inventions and product lines that are related to them. In addition, these classification systems are either too narrow and too general depending on the level of the hierarchy that we are considering. CPC classification divides all patents in 639 groups which is a fairly broad grouping. But the next level is already too fine – 134,725 subgroups. As a result, a median subgroup has only 12 patents in it, while most of the firms in USPTO sample have as many patents over their lifespan as the number of unique subgroups on their record.

APPENDIX E

PRINCIPLES AND GUIDELINE OF WIKIPEDIA ARTICLES

Wikipedia is operated by the non-profit Wikimedia foundation, but it is a self-governing project run by the community which can be joined by everyone for free. The policies and guidelines are also developed by the community members as a result of a preceding discussion. Policies are more formal and less flexible rules that summarize the standards that editors should follow. Guidelines are a set of best practices that editors are expect to follow but occasional exceptions are possible. Whether a policy or guideline is an accurate description of best practice is determined by the community through consensus. The decisions regarding articles editing, articles diffusion or merging, article structure and categorization are also made through consensus between the editors of the article in question. Consensus does not mean unanimity and does not result from a vote. Consensus is built through editing and discussion on the associated talk pages and ideally arrives with an absence of objections by settling for a wide agreement. In rare cases when the editors of an article/category cannot reach a consensus, third-parties are getting involved in the discussion, e.g. other editors, administrators or the community as a whole. In the lack of consensus, the principle of "status quo" is applied and the previous version is retained. Articles for which much of the factual accuracy is actively disputed have a corresponding warning at the top.

Wikipedia does not have hard-and-fast rules. This is even postulated as one of the five fundamental principles of Wikipedia. It has policies and guidelines, but their content and interpretation can change over time. It is mostly based on the spirit of the encyclopedia with as unbiased information as possible and thus some exceptions are possible. Discussion and communication between editors is held at community discussion page and noticeboards. In general, encyclopedic style and formal tone are the guiding principles when it comes to the language of the article and information presentation. At the same time, material should be presented in the most widely understandable manner and if it is possible to do without scarifying the content. This especially applies to potentially technical topics. That

being said, the articles are not necessarily targeted to the general reader. Articles that are approachable only by knowledgeable or even expert reader are also welcome and usually have the corresponding label. Ideally, some share of the article should be understandable by the general reader while other parts would require more background knowledge on the subject.

Wikipedia has about 6 million articles, and 92% of them have more then 200 symbols, which is approximately 33 words. There are three principles that determine the optimal size of an article: 1) Reader issues, such as attention span, readability, organization, information saturation and content contribution; 2) Editors issues, such as consensus in the forum discussion between editors; 3) Technical issues, such as limitation of mobile browsers. When article is too large, it is split into smaller articles or part of it is moved to a new article or merged with an existing article. If it is too small, the article is merged with one or several existing articles. As mentioned above, these decisions require editorial consensus of moderators. If an article is too long and some of its sections are spun off into their own articles, there should still be a summary of these new relevant articles in the original article. For instance, article "Radio" includes such sections as "Radar", "Data communication", "Two-way radio", etc. that have their own Wikipedia pages while the parent article has only an introductory part from these pages.

There are no fixed criteria about the size of an article. Maximum allowed article size is 2048kB including all files, images and references but readability is the key principle here. A page of about 30kB-50kB of readable prose, which roughly corresponds to 4,000-10,000 words, takes about 30-40 minutes to read at an average speed and thus 4,000-10,000 words are the recommended boundaries of the articles. From the other angle, mobile version of Wikipedia requires sections to be not too big to navigate. As a result, mobile browsers and internet speed also make this rough limit optimal for easy accessibility from different devices. The guidelines for editors and moderators when deciding whether to split an article or merge it are the following. An article with less than 1kB of readable prose (less than 200 words) and no growth in the last month is recommended for merging with a related article. An

article with more than 50kB of readable prose (over 10,000 words) is flagged for a potential split.

APPENDIX F

CATEGORIZATION SYSTEM OF WIKIPEDIA ARTICLES

The main idea behind the category system in Wikipedia is to provide navigational links to Wikipedia pages through categories. Categories and subcategories of an article are stated in the bottom of the page. The key guiding principle in categorization of articles is focusing on the defining characteristics of the article's subject, not the type of an article.For instance, an article about infrared waves should not be attributed to "Physics Education". The pages that are assigned to a certain category have a set of defining characteristics that make them related to each other for a reader. The principle of modularity is essential when categorizing articles – each page should be assigned to the most specific subcategory that fits the page but not directly to categories where this subcategory belongs. Every article should belong to at least one subcategory and if there is no suitable one among the existing subcategories or categories, a new one should be created. At the same time, each subcategory belongs to a category which in its turn belong to another parent category and so on. At the top, there are 42 main big topics of classification. Concrete examples of one of the branches of Wikipedia categorization system is demonstrated in Figure F1 and Figure F2.

Figure F1: Example of a classification branch in Wikipedia

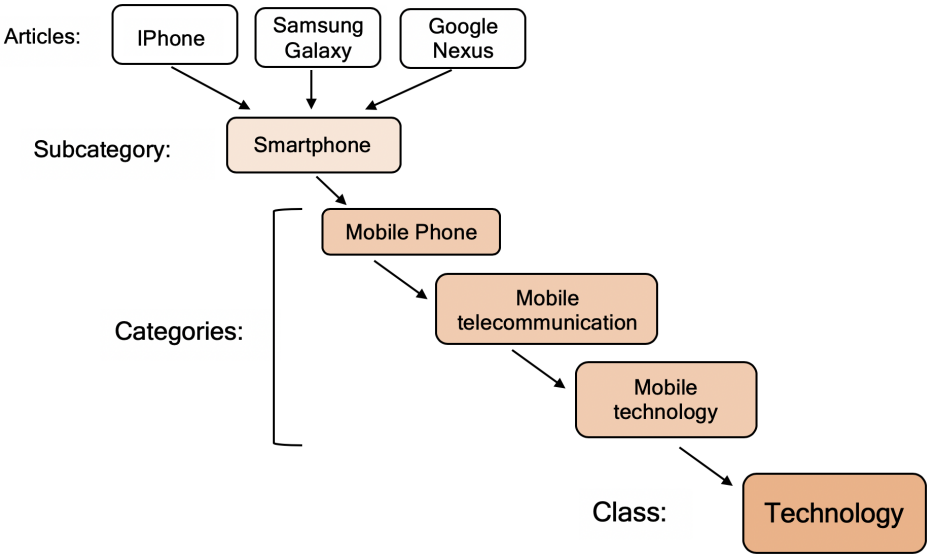
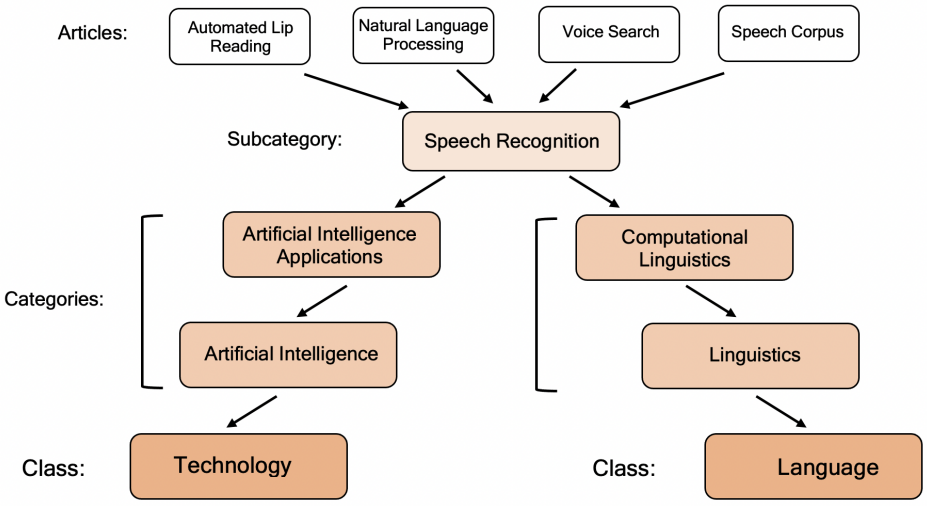


Figure F2: Example of a classification branch in Wikipedia



Each category has a page with a description of the category and a list of pages and subcategories belonging to it (the same is true for the subcategories). The description of the category defines its proper content. Ideally, the category description makes a direct statement about the criteria by which pages should be selected for inclusion in the category. It also contains links to other subcategories and categories that are related to it. Overall, the choice of a category that an article belongs to should be clear from the verifiable content of the article. As with writing, editing and naming article, it is crucial to maintain neutral point of view when categorizing articles. It is expected that one article is a member of many subcategories and one subcategory often belongs to more than one category. For instance, "Espresso Machines" is a subcategory that belongs to both "Coffee preparation" category and "Cooking appliances" category. Categories do no form a strict hierarchy or tree – each article can appear in more than one subcategory and each subcategory can appear in more than one category and so on. That means there are multiple categorization schemes existing simultaneously. Categories are often linked if they are closely related but are not in a subset relation, like 'Oven' and 'Fireplace'. There is no hard limit on the size of categories or subcategories, but large categories eventually are broken down into smaller, more specific subcategories.

Names of categories should be specific, neutral and inclusive. No articles and no categories can have the same name. To avoid confusion, sometimes a short distinguishing information is included in the name after parentheses. For instance, "Social network" would lead to the article about "Social network" as a theoretical concept used in social and behavioral sciences but it can also refer to such articles as "Social network (sociolinguistics)" or "Social networks (journal)" or "The Social Network (movie)". Sometimes there are more than one appropriate title for an article or a category. In this case, the editors choose the best title by consensus. In general, the main guiding principles for choosing the title are recognizability, naturalness, precision, conciseness and consistency. When it comes to formatting, sentence case, singular form, nouns and avoidance of definite and indefinite articles are recommended unless a proper name requires the opposite.

Table F1: Size of categories and subcategories in Wikipedia classification

	Mean	St Dev	5%	25%	50%	75%	95%
#pages: categories and subcategories	47.01	3630	1	3	7	18	90
#pages: subcategories	47.46	3622	1	2	5	13	75
#pages: categories	46.30	3225	1	2	6	18	100
#subcategories: categories	5.46	348.8	1	1	2	5	16
#pages: selected subcategories	34.20	52.58	2	7	17	40	128

Notes: #pages and #subcategories are number of pages and number of subcategories accordingly. Subcategories do not have any other levels below them in the categorization structure besides pages (article). Categories have at least one subcategory belonging to it. Selected subcategories are the subcategories that are used in this paper to define technology sectors.

References

The information in this section is based on several Wikipedia sources:

- Wikipedia: Policies and guidelines. Available at:
https://en.wikipedia.org/wiki/Wikipedia:Policies_and_guidelines

- Wikipedia: Consensus. Available at:

https://en.wikipedia.org/wiki/Wikipedia:Consensus

- Wikipedia: Article titles. Available at:

https://en.wikipedia.org/wiki/Wikipedia:Article_titles

- Wikipedia: Article size. Available at:

https://en.wikipedia.org/wiki/Wikipedia:Article_size

- Wikipedia: Topic categories. Available at:

https://en.wikipedia.org/wiki/Wikipedia:Categorization#Topic_categories

- Wikipedia: Category names. Available at:

https://en.wikipedia.org/wiki/Wikipedia:Category_names

- Commons: Categories. Available at:

https://commons.wikimedia.org/wiki/Commons:Categories

- Wikipedia: FAQ Categorization. Available at:

https://en.wikipedia.org/wiki/Wikipedia:FAQ/Categorization

APPENDIX G

IDENTIFYING RELEVANT ARTICLES IN WIKIPEDIA DATA

The first step in working with Wikipedia data is to parse the XML files to extract only article titles and texts.[40] Wikipedia has around 6 million articles, most of which will be not relevant for the final goal of this exercise – identifying technology sectors of patents. To make the match computationally manageable, I follow a three-step procedure to select potentially relevant articles. First, the vast majority of the articles are tagged according to the template that has been used for creating an infobox – a fixed format table the top right-hand corner of an article that summarizes unifying features of the article (for example, Figure G1 shows an infobox for a certain dog breed). The templates of infoboxes differ depending on the category of an article. All the types of infoboxes are tagged and standardized which makes it easy to identify and delete large groups of irrelevant articles, such as movies, books, people, dogs, etc. The next step is to eliminate all articles related to geographic locations. Using MIT dictionary of geolocations[41], we can directly find articles on countries, cities, states, etc. by their titles or categories they belong to and delete them from the potential sample.

After these two steps, we are still left with many Wikipedia articles that are not relevant for the final data product in mind. The final step in the procedure of identifying potentially useful articles is based on patent classes. The main idea is to construct an extensive dictionary that is descriptive of different patent classes and find articles that have any title, subcategory or category intersection with any word token from this dictionary. The first source of descriptive class dictionary is all meaningful words in the description of the patent class provided by USPTO.[42] The second source of word tokens in the class-specific dictionaries are patent titles. For each patent class, I split the titles of all patents belonging to

40. My parsing code is based on Koehrsen (2018) articles that propose a method of fast parsing of Wikipedia articles.

41. Created by Gomez (2019) and available at https://pypi.org/project/geonamescache/

42. Meaningful words are all words excluding stopwords that search engines ignore: articles, propositions, etc.

Figure G1: Example of an infobox in Wikipedia

this class by words and two-word combinations, stem and lemmatize them (i.e. put every words in a standard form) and choose 1000 most frequent combinations of nouns plus verbs, nouns plus adjectives and alone standing nouns. These tokens from patent titles are also added to a class-specific "dictionary of tokens". As a result, each of the 438 USPTO patent classes has a descriptive dictionary of about fifteen hundred word tokens. I combine all these class-specific dictionaries together to form one big collection of tokens that relate to patents. Then I select all Wikipedia articles that have any non-empty intersection in their title, subcategories or categories with at least one of the tokens from this big dictionary.

Remember that the goal of this computational exercise is to use Wikipedia articles to identify technology sectors of firms. Many Wikipedia articles are very granular and finely-defined. This can be a problem since we do not want technology sectors to be firm-specific or too narrow – we simply will not be able to capture any dynamics within a sector. For instance, while there is a Wikipedia article on iPhone and Galaxy Note, the technology sector we want to attribute these articles should not be brand-specific. Otherwise, there will be no between firm reallocation of sectors or the sample of sectors that will have such dynamic is

biased. To avoid this over-specificity, I am using the subcategory that a given article belongs to as the corresponding "technology sector" rather than the title of the article. In the case of iPhone and Galaxy Note, the subcategory of these articles that defines the associated technology sector is "Smartphone". After the selection procedure described above, I ended up with 18,271 subcategories that have potentially relevant articles that can be matched with 4 million patents.

APPENDIX H

USING NLP TO MATCH A PATENT TO THE MOST SIMILAR WIKI-ARTICLE

The main idea behind the procedure of matching patents to Wikipedia articles is to come up with metrics that will allow us to compare similarity of two texts. In a nutshell, we want to compare the text of each patent with all Wikipedia articles and find the article which text is most similar to the patent's text. The ultimate goal of this method is to match 4 million USPTO patents to 18,271 potential Wikipedia articles. Note that these numbers are upper borders both on the patent and on the Wikipedia side. Some patents may not be similar enough to any of Wikipedia articles in which case the algorithm should not force a match. On the other hand, even after elimination of irrelevant Wikipedia pages, some articles in the final sample may not be related to any invention in the end. Thus, ideally we want to match as many patents as possible while preventing the algorithm from excessive greediness and allowing no matches on both sides.

First and foremost, we need to clean the text of the patents and articles in preparation for the match. I start with deleting all the words that do not have any actual meaning: the so-called "stopwords" – articles, prepositions, pronouns, determiners and conjunctions. That leaves us with only meaningful words that convey the content of a text. The next step is lemmatization – the process of standardizing all words to their dictionary form. The goal is to remove inflectional endings of the words, which will allow our matching algorithm to recognize different forms of verbs and nouns as the same word. For instance, "producing" and "produces" will be transformed to produce after lemmatization. "Components" will be turned into "component".[43]

For both patents and Wikipedia articles, we want to leverage the fact that the title of the text document is much more informative than any other phrase in the text. In order

43. I am using NLTK Python module that relies on WordNet lexical database available at wordnet.princeton.edu.

to reflect the higher importance of the title words, I upweight words in the title of a patent text by a factor of five and put them at the beginning of a the patent's abstract. The same is done for Wikipedia articles and their titles.[44]

My approach to matching patents text to Wikipedia articles text is similar to the procedure described in Kelly et al. (2018) and Argente et al. (2019). As previewed above, the match is based on computing pair-wise text similarity between one text and another. This exercise boils down to vectorization of the texts and computing similarity score of the two corresponding vectors. The first step is to transform each patent text and article text into a very long vector of the same dimension N. This dimension N is the number of all unique words based on the vocabulary of all text documents we are working with. To construct the vocabulary of size N, I merge all patents and articles text together and select all unique word tokens in this combined text. Each patent text and each Wikipedia article text is a vector of size N. Each element of such vector corresponds to a unique word token from the vocabulary. Element i of a vector is equal to one if a word token i is present in the text document that this vector corresponds to and equals 0 otherwise. After we transform each patent text and article text into a vector and combine them as columns of one matrix, we end up with a very sparse matrix of dimension $N \times M$, where N is number of unique words in our vocabulary and M is number of text documents (i.e. patents plus articles).

Even among meaningful words, relative importance of each word in a text for its content varies a lot. Words that frequently appear in a given text document are more informative than those that appear only once. At the same time, words that are frequent in all text documents are not as informative as we may think just looking at the frequency within a particular text document (for instance, we are very likely to encounter "device" or "method" in any text of a patent). In order to take this into account, I weight words by their contribu-

44. Most of the patents abstracts are less than 6000 characters in length. For easiness of comparison, I truncate Wikipedia article to first 10000 characters. Since the most important information is always summarized at the top of an article, such truncation also allows us to focus on the most informative word tokens that characterize the article.

tion to a document's content using "term-frequency-inverse-document-frequency" (TF-IDC) transformation of word counts:

$$\omega_{ij} = TF_{ij} \times IDF_i \tag{H1}$$

where ω_{ij} is weight of word i in document (vector) j. The first item, TF_{ij} is frequency of word i in document j. The second item, IDF_i is inverse frequency of word i in all documents:

$$TF_{ij} = \frac{c_{ij}}{\sum_i c_{ij}} \tag{H2}$$

$$IDF_i = log\left(\frac{M}{\sum_j \mathbb{1}\{i \text{ in } j\}} + 1\right) \tag{H3}$$

where c_{it} is number of times that word i appears in a document j and $\sum_i c_{ij}$ is the length of the document measured in words. Note that words that appear frequently in a text have high TF as well as words that appear rarely in many different texts. Thus, high $TFIDF$ for a given word means that it is relatively important in conveying content of a document.

After adjusting the weights of each vector's elements according to the TFIDF measure, I normalize them to have unit length. Finally, we are ready to compute the similarity score between each possible pair of a patent text and Wikipedia article text. I measure the similarity score as cosine similarity between the two corresponding vectors. Cosine similarity is equal to the cosine of an angle between two non-zero vectors of an inner product space. We can also think about it as an inner product of these two vectors after normalizing their length to one. This similarity score lies in the interval between 0 and 1, where 0 indicates no similarity at all and 1 means the two texts are identical.

I match each patent text to Wikipedia article that has the highest similarity score with this patent conditional on passing a 30% similarity threshold. Since the similarity score is just a cosine similarity, particular quantitative levels do not have any economic meaning. In other words, 30% similarity score does not mean that 30% of word tokens are the same. Note that it is enough to have at least one word in common to have a positive similarity

score. Due to the richness of the text data, it is inevitable that each patent will have at least one article with a positive similarity score. As a result, even positive but low similarity scores suggest that this match is not informative. The threshold score is determined by visually studying pairs of different quality and manually examining matches using various lower bounds on similarity.

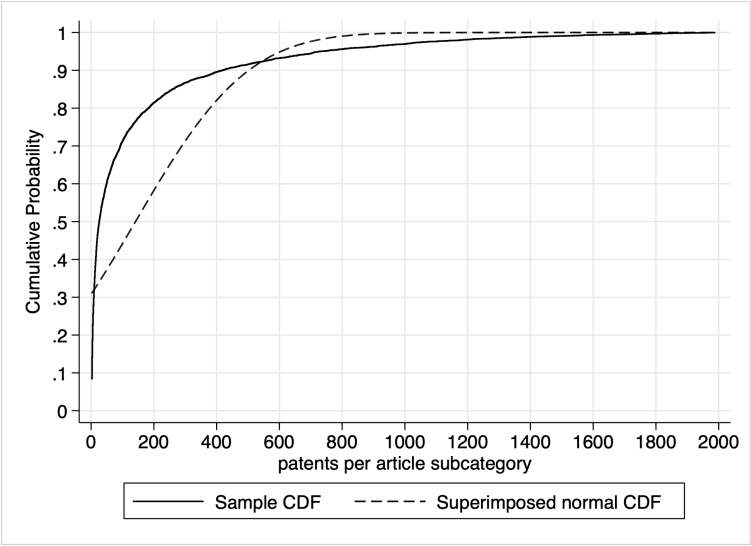
APPENDIX I

PROPERTIES OF THE PATENT-ARTICLE MATCH

The imposed lower border on the similarity score helps controlling the quality of the match between patent text and Wikipedia article text but at the same time implies that some patents remain unmatched – there are no Wikipedia articles that are similar enough to the text of the patent and thus it is impossible to identify a technology sector to which this patent belongs. The resulting data product is a sample of 1,122,220 patents that were successfully matched to one of the Wikipedia articles. The articles with a patent pair belong to 6,461 unique Wikipedia subcategories that define technology sectors. For instance, patent #6427584 with a title "System and method for processing citrus fruit with enhanced oil recovery" is matched to subcategory "Citrus production", patent #8360885 "System and method for using a game to interact with television programs" is directed to "Education television" in Wikipedia, patent #6829283 "Electro-absorption vertical cavity surface emitting laser modulator and/or detector" is matched to "Laser medicine" subcategory and patent #4550733 "Electric dental analgesia apparatus and methodology" goes to "Dental software" subcategory.

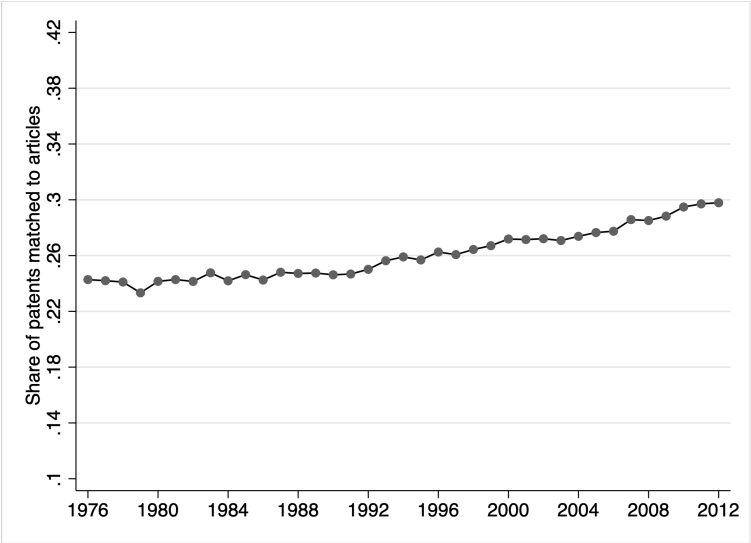
The merge between patents and articles is many-to-one: each patent is matched to exactly one article, while an article can have many patents matched to it. Figure I1 below shows the CDF of patents per article subcategory. The quality of the match is fairly consistent across years. On average, slightly more than a quarter of new patents submitted every year is matched to one of the Wikipedia subcategory (see Figure I2 below). The sectors composition of the matched patents is slightly dominated by Textiles, Paper and Physics but the over-representativeness of these sectors is not considerable and should not create any biases (Figure I3 below). In order to allocate technology clusters into sectors, I use data from Cooperative Patent Classification (CPC). According to CPC, each patent belongs to at least one of nine large sectors. A technology cluster is allocated too a sector where most of its life-time accumulated patents belong to.

Figure I1: CDF of patents per article subcategory in a resulting match.



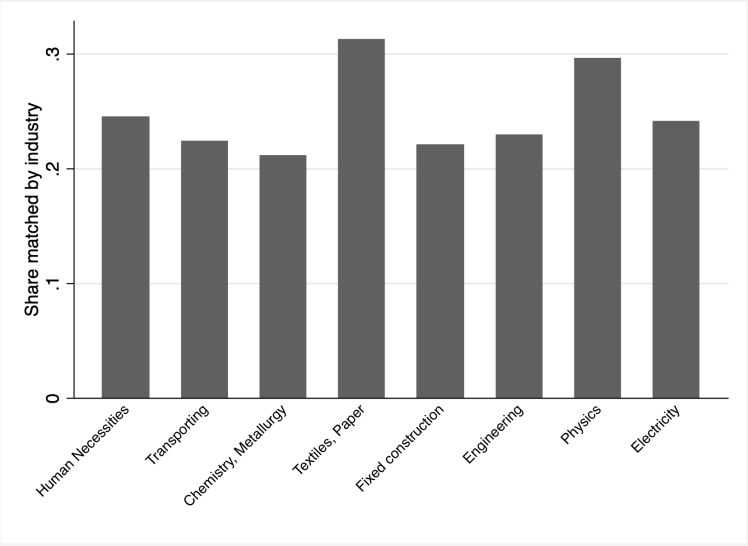
Notes: Superimposed normal CDF has the same mean and standard deviation as sample CDF.

Figure I2: Share of patents matched to an article subcategory by year of patent application



The sample includes 101,843 firms (i.e. unique patent assignees who own the patent) that have more than one patent in their accumulated life-time portfolio. More than half of the assignees in have only one patent. Following the literature, I drop firms with only one patent because these firms are not considered as innovating ones or there is a severe misspelling in the name of this assignee that makes a researcher identify it as a separate independent company. Figures I4 and I5 below compare the distribution of firm's size and tenure in the

Figure I3: Share of patents matched to an article subcategory in every sector



full matched sample that I work with later on and in the whole USPTO dataset. Firms with matched patents in my sample are somewhat bigger and have higher tenure than firms in the USPTO universe, but this difference is not significantly big to question representativeness of the dataset in use.

Table 1.1 reports some key summary statistics that describe the final sample. On average, an article subcategory and its corresponding technology cluster has 173 patents matched to it, but the distribution is skewed to the left. For a median firm, I am able to match 70% of its patents to Wikipedia articles. Such high share of matched patents for an average firm implies that the selection of the observations in the final sample happens on a firm level rather than independently across firms.[45] If the algorithm matches one patent of a firm, it is likely to match the others too. The latter is a desirable property of the dataset – it is better to have a sample of firms but capture most of their innovation activities than to have most of the firms with selected patents that we can observe.

45. In fact, most of the firms with only one patent were not in the sample even before their elimination by this attribute.

Figure I4: Matched sample representativeness: portfolio size

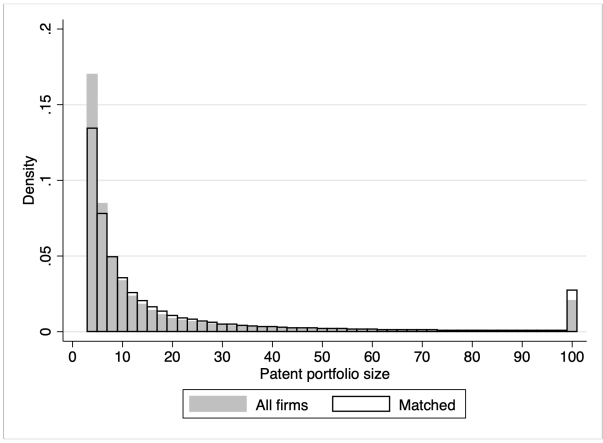
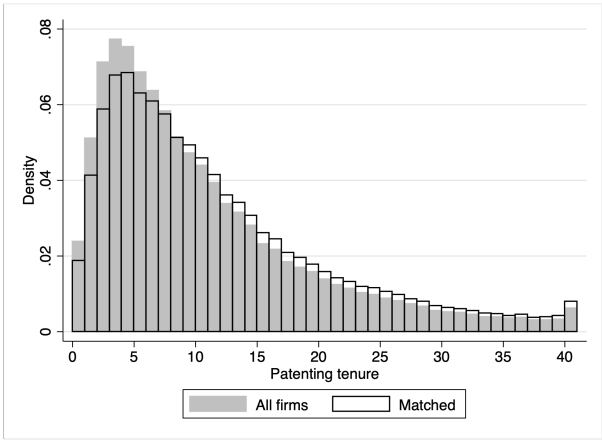


Figure I5: Matched sample representativeness: tenure



APPENDIX J

TECHNOLOGY CLUSTER GROWTH AND FIRMS ENTRY RATE

The decomposition of the annual changes in technology cluster size demonstrates the importance of between firms reallocation for technology development. This implies that technology cluster dynamism is closely related to firm dynamism. The decrease in the dynamics of technology clusters documented in Section 1.3 echoes a decrease in firm dynamics that we have observed over the last decades.[46] If young small firms are an important driver of new technology cluster formation, a decline in startup's entry rate will inevitably push technology entry rate down as we have seen before.

A question remains over how important the declining firm entry rate is for technology cluster *growth* rates, and what kind of clusters grow faster on average. Figures 1.14 and 1.15 imply that reallocation between continuing firms is more important for technology growth than reallocation between startups and firms that are leaving the marker. However, mature firms are usually larger and have more resources to expand their business activity than young firms do. As a result, it is not surprising that their entry to a technology cluster gives this cluster a bigger boost in growth. In this section, I explicitly focus on the contribution of firm entry rate to technology cluster growth. In particular, I distinguishing *new* firms entry rate from the overall entry rate in order to elaborate on the difference between the role of new firms and continuing firms.

The level of analysis in this exercise is technology cluster and year. The regression specification that I am using is as follows:

$$Technology\_growth_{ct} = \beta Technology\_char_{ct} + \gamma Firm\_char_{ct} + \varphi_t + \alpha_c + \varepsilon_{ct} \qquad \text{(J1)}$$

46. See Haltiwanger et al. (2015) for a comprehensive summary of basic facts about the decline in business dynamism in U.S.

where $Technology\_growth_{ct}$ is annual growth rate of technology cluster size, which is defined as before: share of patents submitted to technology cluster c in year t relative to all patents submitted in year t. The vector of technology cluster characteristics $Technology\_char_{ct}$ includes entry rate of firms, entry rate of *new* firms, technology cluster age and its degree in a global network. Firm entry rate is measured as the number of firms that have entered a technology cluster c in year t relative to all firms (startups and continuing firms) that are active in technology c in year t. The new firm entry rate focuses only on brand new firms that just have entered innovation market, i.e. firms that submit their first ever patent in year t.

In some specifications, I also include a vector of firm characteristics $Firm\_char_{ct}$. It summarizes various characteristics of an average firm that is active in technology cluster c in year t. These include average firm size measured by employment, average firm age and average concentration of a firm in technology cluster c. The latter is the share of *accumulated* patents that a firm has submitted in cluster c by time t relative to all patents in firms portfolio as of time t (i.e. average of z_{ict} for a given cluster c from decomposition (9)).

The main results are summarized in Table J1. The first two columns show that higher firm entry rates in a technology cluster is associated with faster patenting growth in this cluster. If we look at the entry rate of new firms only (i.e. firms that submit their first patent in this year) we also see a positive association. However, when we study the explanatory power of these two entry rates together in a regression, the sign on a new firm entry rate reverses. An increase in firm entry rate by 10 percentage points *decreases* technology cluster growth rate by 0.5 percentage points, while an increase in overall firm entry rate of the same magnitude *increases* technology cluster growth rate by 2.7 percentage points. Given that an average technology cluster grows at an annual rate of 5.4%, this is an impressive 50% increase in growth rate.

The results of this exercise with covariates that characterize an average firm in a technology cluster are reported in Table J2. One caveat to this extended group of regressors

Table J1: Technology cluster growth and firm entry rate

	Technology growth rate			
	(1)	(2)	(3)	(4)
Firm entry rate	0.256***		0.272***	0.276***
	(0.0132)		(0.0133)	(0.0155)
New firm entry rate		0.0625***	-0.0488***	-0.0373**
		(0.0139)	(0.0138)	(0.0148)
Technology cluster age	-0.0047***	-0.0060***	-0.0045***	
	(0.00053)	(0.00051)	(0.00054)	(0.00237)
Technology cluster degree	0.00027	0.00028	0.00015	-0.00051
	(0.00081)	(0.00082)	(0.00082)	(0.00299)
Year FE	Yes	Yes	Yes	Yes
Technology FE	No	No	No	Yes
Observations	71,270	71,270	71,270	71,270
R-squared	0.009	0.003	0.009	0.044

Notes: Firm entry rate is the ratio of number of firms that enter technology cluster c in year t relative to number of all firms that are active in technology cluster c in year t. New firm entry rate has only a subsample of firms with their first patent in year t in the nominator. Mean of the dependent variable is 0.054. Standard errors are clustered on year level. *** p<0.01, ** p<0.05, * p<0.1

is that the data on firm's employment and age is available for only 30% of firms and thus within-cluster averages may not be representative of a typical firm in the technology cluster. With this caveat in mind, we see even large coefficients on firm entry rate in columns 3 and 4. The coefficient on new firm entry rate remains positive but not significant. As far as firms characteristics are concerned, we see that technology clusters with bigger and more mature firms show high growth rate.

Overall, these results indicate the importance of incumbents in boosting technology cluster growth rate. While it is young small firms that introduce new technologies to the market, it is market incumbents that contribute mostly to their growth afterwards. Elaborating on the example of smartphone technology cluster, Apple was not the first company that invented smartphones. Nor was it a complete newcomer to the innovation market – it had a profound experience in personal computers and MP3 players. But it was the company that

Table J2: Technology cluster growth and firm entry rate: extended list of controls

	Technology growth rate			
	(1)	(2)	(3)	(4)
Firm entry rate	0.433***		0.423***	0.390***
	(0.0261)		(0.0242)	(0.0288)
New firm entry rate		0.280***	0.0630	0.103**
		(0.0434)	(0.0424)	(0.0415)
Technology cluster age	-0.0102***	-0.0123***	-0.0102***	
	(0.00076)	(0.00083)	(0.00076)	(0.00979)
Technology cluster degree	-0.00153*	-0.00153*	-0.00156*	-0.00499
	(0.00082)	(0.00080)	(0.00081)	(0.00444)
Average firm employment	0.0048***	0.0039***	0.0048***	0.0114***
	(0.0011)	(0.0011)	(0.0011)	(0.0015)
Average firm age	0.0155***	0.0127***	0.0155***	0.0219***
	(0.0039)	(0.0039)	(0.0039)	(0.0049)
Average firm concentration	0.152***	0.0810**	0.123***	0.361***
	(0.0276)	(0.0373)	(0.0369)	(0.0466)
Year FE	Yes	Yes	Yes	Yes
Technology FE	No	No	No	Yes
Observations	43,741	43,741	43,741	43,741
R-squared	0.025	0.013	0.025	0.107

Notes: Firm entry rate is the ratio of number of firms that enter technology cluster c in year t relative to number of firms that are active in cluster c in year t. New firm entry rate has only a subsample of firms with their first patent in year t in the nominator. Average firm employment is logarithm of average employment of active firms in technology cluster c in year t. Average firm age is logarithm of average age of active firms in cluster c in year t. Average firm concentration is an average share of patents that firms operating in technology cluster c have submitted in year t relative to all patents they have submitted in year t. Mean of the dependent variable is 0.054. Standard errors are clustered on year level. *** p<0.01, ** p<0.05, * p<0.1

brought the technology cluster of smartwatches to another level after the introduction of iPhone in 2007.

While active entry of firms contributes to the technology cluster growth, it is unclear how stable this growth is after the short-run boost that new-entering firms are generating. The quantile decomposition of the average change in technology cluster size (Figure 1.11) shows

that across-clusters dispersion in this indicator is going down over time. If we look at time trend of the cross-sectional standard deviation of technology cluster growth rate, we also see the same decline (Figure A9 in Appendix A). This decrease in the variance of technology cluster growth can also be a symptom of the plummeting firm dynamism in high-tech sector. With less firms entering, there is less selection in and out of the market and as a result, less volatility on a technology cluster level as well.

Table J3: Variation of technology growth and firm entry rate

	Standard deviation of technology growth rate			
	(1)	(2)	(3)	(4)
Average firm entry rate	0.615***		0.692***	0.422***
	(0.0161)		(0.0158)	(0.0200)
Average new firm entry rate		0.128***	-0.209***	-0.0822***
		(0.0208)	(0.0208)	(0.0215)
Technology cluster age	-0.0152***	-0.0183***	-0.0144***	
	(0.00034)	(0.00052)	(0.00032)	(0.00473)
Technology cluster degree	-0.0031***	-0.0030***	-0.0036***	-0.0103***
	(0.00030)	(0.00036)	(0.00032)	(0.00159)
Year FE	Yes	Yes	Yes	Yes
Technology FE	No	No	No	Yes
Observations	55,632	55,632	55,632	55,632
R-squared	0.066	0.038	0.068	0.404

Notes: Average firm entry rate is 5-year average of the ratio of number of firms that enter technology cluster c in year t relative to number of firms that are active in cluster c in year t. Average new firm entry rate has only a subsample of firms with their first patent in year t in the nominator. Mean of the dependent variable is 0.766. Standard errors are clustered on year level. *** p<0.01, ** p<0.05, * p<0.1

To shed more light on this hypothesis, I modify regression (J1) to explore what factors can explain volatility of technology cluster growth and hence its sustainability in a longer run. The dependent variable is standard deviation of a technology cluster growth rate based on the last 5 years of observations. That means, in order to measure volatility of a technology cluster c in year t, I am computing standard deviation of the technology growth rates in years $t-4$, $t-3$, ..., t. For consistency, the main regressor – firm entry rate – is also averaged across

the last five years of observations. Table J3 reports the results without characteristics of an average firm in a technology cluster as covariates while Table J4 provides the analysis with full set of regressors. As one would expect, there is a strong positive association between firm entry rate and volatility of a technology cluster.

Table J4: Variation of technology growth and firm entry rate: extended list of controls

	Standard deviation of technology growth rate			
	(1)	(2)	(3)	(4)
Average firm entry rate	0.777***		0.840***	0.531***
	(0.0157)		(0.0196)	(0.0242)
New firm entry rate		0.239***	-0.257***	-0.0892**
		(0.0338)	(0.0393)	(0.0434)
Technology cluster age	-0.0152***	-0.0191***	-0.0149***	
	(0.00046)	(0.00077)	(0.00045)	(0.00062)
technology cluster degree	-0.00298***	-0.00267***	-0.00311***	-0.0127***
	(0.00023)	(0.00032)	(0.00024)	(0.00222)
Average firm employment	-0.0071***	-0.0088***	-0.0071***	0.0020***
	(0.00042)	(0.00043)	(0.00043)	(0.00054)
Average firm age	-0.00064	-0.00457*	-0.00005	0.00183
	(0.00270)	(0.00264)	(0.00266)	(0.00190)
Average firm concentration	-0.178***	-0.190***	-0.102***	0.0558***
	(0.0215)	(0.0228)	(0.0208)	(0.0178)
Year FE	Yes	Yes	Yes	Yes
Technology FE	No	No	No	Yes
Observations	38,194	38,194	38,194	38,194
R-squared	0.108	0.070	0.110	0.509

Notes: Average firm entry rate is 5-year average of the ratio of number of firms that enter technology cluster c in year t relative to number of firms that are active in cluster c in year t. Average new firm entry rate has only a subsample of firms with their first patent in year t in the nominator. Average firm employment is logarithm of average employment of active firms in technology cluster c in year t. Average firm age is logarithm of average age of active firms in cluster c in year t. Average firm concentration is an average share of patents that firms operating in technology cluster c have submitted in year t relative to all patents they have submitted in year t. Mean of the dependent variable is 0.766. Standard errors are clustered on year level. *** p<0.01, ** p<0.05, * p<0.1

APPENDIX K

COSTS OF INNOVATIONS AND TECHNOLOGY GROWTH

Section 1.7 studied the relationship between innovation market concentration and technology cluster growth in term of its size and degree. We saw that higher concentration of firms in a technology cluster is detrimental for its size growth and its centrality growth. However, this relationship has weakened after 2000, which implies that there might be an omitted variable in the regression: a factor that affects both concentration of firms in a technology cluster and its growth potential.

A plausible candidate for this factor is costs of innovations and in particular, human capital costs. Figure 1.19 shows that average number of inventors per citation-weighted patent is decreasing before 2000 but starts a rapid increase after 2000. The numerator of this indicator, team size of inventors, is going up steadily during this period as captured in Figure K1 below. But the denominator, average number of received citations per patent (truncated to 5-year horizon after patent application date) has an inverted V-shape dynamics with a peak around 2000 – Figure K2. Moreover, there could also be change in the correlation between team size of inventors and number of citations received by a patent after 2000.

Figure K1: Average team size of inventors measured as number of inventors coauthoring a patent

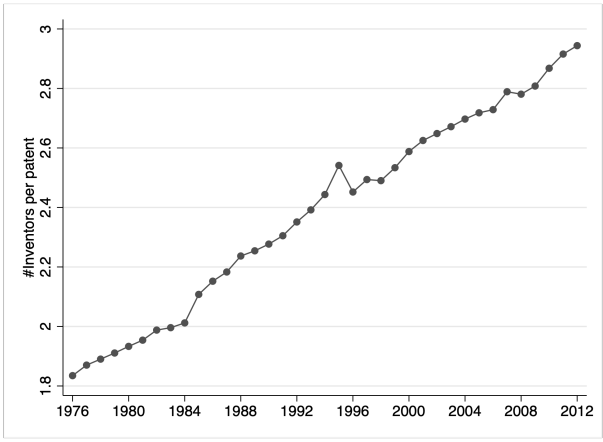
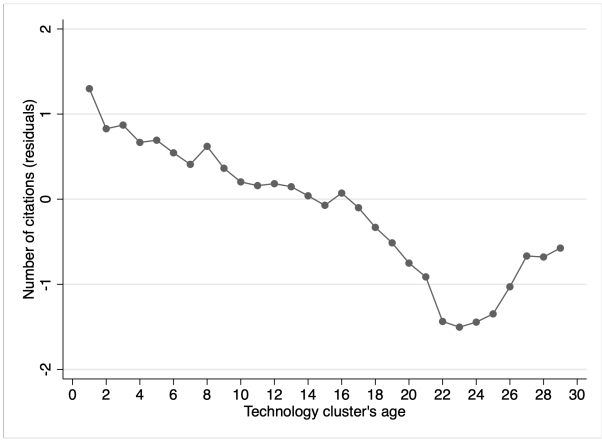


Figure K2: Average number of received citations per patent, truncated at 5-year horizon after the application date



For now, we will defer from the discussion about the potential forces behind Figure 1.19 and focus on how the change in the costs of innovations relates to market concentration and change in technology degree if we treat it as an omitted variable from regressions (1.16) and (1.17). The modified regressions are as follows:

$$Size\_gr_{ct} = \beta_1 HH_{ct} + \beta_2 Inventors\_pp_{ct} + \beta_3 HH_{ct} \times Inventors\_pp_{ct} + \gamma \mathbf{X}_{ct} + \gamma_t + \varphi_c + \varepsilon_{ct} \tag{K1}$$

$$Degree\_gr_{ct} = \beta_1 HH_{ct} + \beta_2 Inventors\_pp_{ct} + \beta_3 HH_{ct} \times Inventors\_pp_{ct} + \gamma \mathbf{X}_{ct} + \gamma_t + \varphi_c + \varepsilon_{ct} \tag{K2}$$

where $Inventors\_pp_{ct}$ is average number of inventors per citation-weighted patent in technology cluster c in year t. The results are summarized in Table K1 for a technology cluster's size growth and in Table K2 for its degree growth. In case of degree growth, note that the coefficient on HH index is positive now – higher concentration of firms in a technology cluster is association with higher growth rate of technology cluster degree. In case of size growth, we do not see any significant relationship between innovation market concentration and technology growth. On the other hand, there is a strong negative association between costs of innovations measured as inventors count per citation-weighted patent and size growth rate. Higher costs of innovation should suppress entry of outside firms to a technology cluster and thus slow down reallocation – the crucial force behind technology cluster growth, as we have seen in Section 1.6. That being said, we should interpret these results with caution. Costs of innovation highly correlate with firms concentration index – one standard deviation increase in $Inventors\_pp_{ct}$ is associated with an increase in HH_{ct} by 10% of its standard deviation. If the costs of innovations are themselves an outcome of a changing market concentration we may encounter the so-called "bad controls" problem.[47] In that case, the analysis of the relationship between firms concentration and technology cluster growth within certain group of innovation costs would suffer from selection bias.

47. See (Angrist and Pischke, 2008) for more details on this issue.

Table K1: Technology growth, market concentration, and cost of innovation production

	Technology size growth rate			
	(1)	(2)	(3)	(4)
HH, weighted	0.323**	-0.344	-0.752	-0.135
	(0.155)	(0.177)	(0.718)	(0.137)
Inventors per patent	-0.0896***	-0.139***	-0.134***	-0.158***
	(0.0218)	(0.0211)	(0.0115)	(0.0227)
HH×Inventors per patent	.0008	0.0856*	-0.0555*	0.152***
	(0.0606)	(0.0505)	(0.0317)	(0.0582)
Technology cluster size	75.36	399.9*	366.7***	675.2*
	(75.34)	(234.8)	(56.40)	(352.8)
Technology cluster age	-.0063***			
	(.0009)			
Sample	All	All	pre-2000	post-2000
Year FE	Yes	Yes	Yes	Yes
Technology FE	No	Yes	Yes	Yes
Observations	79,777	79,252	42,988	35,767
R-squared	0.328	0.374	0.086	0.491

Notes: Dependent variable mean is .054. Standard errors are clustered on the fixed effect level. *** p<0.01, ** p<0.05, * p<0.1

Table K2: Degree of a technology cluster, market concentration, and cost of innovation production

	Technology degree growth rate			
	(1)	(2)	(3)	(4)
HH, weighted	0.0285***	0.0434***	0.0408***	0.0273***
	(.0053)	(.0047)	(.0074)	(.0078)
Inventors per patent	.00053	.00174*	.00151	.00178*
	(.00102)	(.00089)	(.00184)	(.00102)
HH×Inventors per patent	-.0044*	-.0053**	-.0040	-.0047*
	(.0024)	(.0025)	(.0047)	(.0027)
Technology cluster size	0.235	-0.558	0.187	0.129
	(0.429)	(0.376)	(1.145)	(0.229)
Technology cluster age	-.00021*			
	(.00012)			
Sample	All	All	pre-2000	post-2000
Year FE	Yes	Yes	Yes	Yes
Technology FE	No	Yes	Yes	Yes
Observations	88,297	87,891	48,529	39,019
R-squared	0.010	0.036	0.052	0.042

Notes: Dependent variable mean is 0.014. Standard errors are clustered on the fixed effect level. *** p<0.01, ** p<0.05, * p<0.1