Discovering pre-entry knowledge complexity with patent topic modeling and the post-entry growth of Italian firms

Marco Guerzoni[°] and Massimiliano Nuccio[*§]Federico Tamagni[‡]

[°]University of Milan-Bicocca, Italy
[§]University of Venice-Ca' Foscari, Italy
[‡]Scuola Superiore Sant'Anna, Pisa, Italy

Abstract

Innovation studies have largely recognized the role of knowledge in fostering innovation and growth of entrants. Previous literature has focused on entrepreneurial and managerial capabilities and education and knowledge incorporated in material and immaterial resources. We assume that new firms need to possess different pieces of knowledge, but beyond diversity, business performance relies also on knowledge distinctiveness. In other words, the complexity of a knowledge base is not simply the recombination of homogeneous pieces of knowledge but it also depends on the specific nature of each of them. This paper develops a new complexity indicator able to capture the complexity of the knowledge base by applying a topic modeling approach to the analysis of patent text. We explore the empirical relation between pre-entry complexity of knowledge, as measured by our complexity index, and post-entry growth performance of a sample of Italian firms entering the market in 2009-2011, which we then follow over the period 2012-2021. Baseline results show a significant and positive association between knowledge complexity and growth, even after controlling for firm characteristics and year, sector and region fixed-effects. Robustness analysis reveal this positive effect is stronger in the medium-long run while relatively weaker for innovative SMEs.

Keywords: pre-entry knowledge base, complexity, text analysis, patents, firm growth, post-entry performance

JEL: D22, O30

[*]Corresponding author. Email: massimiliano.nuccio@unive.it

1 Introduction

Empirical evidence on new firms and startups agrees on few stylized facts. First, they are small and numerous when they enter the market. Second, only very few firms survive, experience a solid growth pattern and contribute significantly to job creation.

Academic research on entrepreneurship has struggled for the secret formula which can explain the success of new, especially high growth firms. Although different factors have been suggested, most of them can be traced back to some form of knowledge. Knowledge can be embodied in the capabilities of entrepreneurs, managers and employees being the result of both their experience and formal education. Knowledge can also be embedded in material and immaterial assets which constitute the set of resources of the new firm, like machines, software, and patents. Finally, knowledge can spring out of the network of relationships and exchange that firms build with stakeholders and customers.

If most of the literature agrees that some combination of this heterogeneous knowledge can foster innovation and therefore growth, innovation scholars argue on what kind of knowledge is behind the success of startups and what proxy measure can better capture the relationship between knowledge and growth.

This paper is placed in this debate offering two related contributions. First, we propose a new measure of knowledge complexity by applying an unsupervised machine learning approach to the analysis of the text of patent abstracts. Second, we explore the empirical relation between pre-entry complexity of knowledge, as measured by our complexity index, and post-entry growth performance of a sample of Italian firms entering the market in 2009-2011, which we then follow over the period 2012-2021.

Our approach is grounded in the Economic Complexity Index introduced by **?**, which recently surged as a new paradigm for eliciting the competitive advantages of countries and regions (**?**). The core of the complexity approach rests in the Smithian intuition that the knowledge specialization associated with the division of labour is responsible not only for static efficiency gains, but also for the creation of new and advanced knowledge, which in turn generates more competitive advantages and allows for a further division of labour and specialization. The persistent wealth of a nation is, thus, explained by two self-reinforcing mechanisms: more developed countries are more diversified and more diversification generate new idea and further progress.

The creation of a new indicator of knowledge complexity presents two methodological challenges in our context. The first one consists of extracting from patent abstracts the information that can proxy for the different pieces of technological knowledge which firms recombine together. The second challenge is to exploit this information to compute a complexity index moving from patents to firms, thus measuring unobservable firm capabilities. As described in the next paragraph, we deal with the first challenge with an exercise of topic modeling, a generative model which can both elicit the various and unobserved pieces of technological knowledge present in the dataset and describes each patent as a combination of these basic units of knowledge. Secondly, we show how to aggregate the results of topic modeling to compute a complexity index both at the patent level and at the firm level.

The paper is organized as follows. In the next section we place this paper in the debate about post-entry performance and we introduce the idea of complexity of the knowledge base as a possible determinant of firm growth. In section 3 we present the data. The empirical strategy is split into sections 4 and 5: in the former we apply our empirical approach to compute the complexity index at

patent and firm level, while in the latter we present the econometric estimates of the relation linking pre-entry knowledge complexity and post-entry growth.

2 Knowledge complexity and post-entry growth performance

After the rising interest of entrepreneurship studies, the debate on the post-entry performance of firms has also progressively increased in the search of the factors of success and growth of new born companies. Seminal works in the field were collected in a special issue on post-entry performance (**?**) which identified different antecedents and among them the pre-entry experience. **?** summarize the results focusing on pre-entry knowledge base, which can be defined as resources or competencies.

The notion of pre-entry knowledge base is often captured by the pre-entry experience of the entrepreneur or the founding teams. For example, **?** track the career of the managers by looking at the previous sectors of activities, years of activities, education and managerial experience, showing that pre-entry knowledge of the business activity and pre-entry management experience affect positively different leaning activities of the business process. More recently **?**, p.870 reviews the role of entrepreneurial team pre-entry experience in a start-up's initial strategy choice In fact the debate has often questioned whether it is better to focus on core competencies (i.e. specialization) or to collect more heterogeneous skills (i.e. diversification), and whether it is more valuable to exploit the existing assets of the organization, building on acquired knowledge and experience, or to explore new frontiers beyond the firm's comfort zone. Different papers have measure the relationship between characteristics of the pre-entry career of the entrepreneurial teams and performance, **?** finds that founding teams with a more diversified working experience are encouraged to exploration while if members come from a similar experience, their firm tends to exploit the common knowledge. More importantly when founding teams have both common and diverse prior company affiliations, then firms are more likely to grow.

In this realm, the analysis of patent data is a fruitful stream of research because they implicitly convey the competencies and the immaterial assets of the inventors, being individuals or firms, thus grasping significant dimensions of their knowledge base and innovative capability. **?** support the view that the individual-level characteristics of inventors, namely their education, can shape the breadth of technological re-combinations of inventions and make a difference in the scope of patents' technological borders. The idea of capturing firms capabilities by leveraging on the different patterns of recombination of technological knowledge dates back to the exercise by **?**, which look at different pattern of co-occurrences of technological classes of patent data. This intuition has been widely employed by scholars in economics of innovation, which developed numerous indexes based on a patent's technological classes to capture different properties of technological knowledge such as coherence (**?**), related and unrelated variety (**?**) or novelty (**?**). For instance very recently, **?** suggest that variety measured as concentration (entropy) can explain a firm' success at its early stage.

The main contribution of this paper to this literature is to qualify in a novel way the concept of variety. We extend to the firm level the idea of the complexity index as developed by **?** and **?** for explaining the self-reinforcing mechanisms, which lead countries to a path of a sustained growth. In **?**, the Economic Complexity Index empirically measures Adam Smith's intuition that the division of labour not only implies a simultaneous division of knowledge, but also that the same division of labour

allows for further advancements in knowledge and technology:

> I shall only observe, therefore, that the invention of all those machines by which labour is so much facilitated and abridged seems to have been originally owing to the division of labour (**?**, p.15, reprint 2005).

Thus, the development of countries generates an increasing division of knowledge which leads to more advanced knowledge and further growth. Along this line, we expect to observe the occurrence of new and advanced pieces of knowledge in highly and already diversified environments. Growth is not only about an increasing variety of different pieces of knowledge within an economy, but also about the generation of new and advanced ones.

Borrowing from this idea, we propose a complexity index at the firm level to explain the conundrum of the persistent heterogeneity of post-entry firm performance. The complexity of a firm knowledge base is not simply the recombination of homogeneous pieces of knowledge, but it also depends on the different nature of each of them. If the creation of new knowledge is the cumulative result of previous specialization, new advanced knowledge is likely to appear within firms with an already highly diversified knowledge base. Thus, the same dynamic increasing returns, which ultimately explain differential growth rates for countries and regions, in a similar way are at work at the firm level. It is not by chance that **?** begin their line of reasoning with reference to a company like Google, not to a country, as the paradigmatic story of the success-breeds-success hypothesis and, similarly, Adam Smith mentioned a pin factory as clarifying example.

At the theoretical level, this approach is in line with the Resource Base View to explain competitive advantages at the firm level in the knowledge base economy (**?**). The main ingredient of the secret formula to explain firms success consists of valuable knowledge (**?**) determined by its scarcity, critical quality, and inimitability (**?**). According to the complexity approach we expect to observe the emergence of such valuable knowledge pieces in entities with an already diversified knowledge base.

On this basis and differently from previous approaches, we do not measure the knowledge base of a firm by looking at its variety only. We also qualify variety by looking at the distinctiveness of each knowledge component, where we conceive this distinctiveness of a great value, when the knowledge component appears in highly diversified firms. Specifically, we define the complexity of the knowledge base of a firm as a function of the distinctiveness of its knowledge component and the distinctiveness of each knowledge component as the a function of the complexity of the knowledge base in which knowledge components are observed. These definitions, which mathematically translates in a system of coupled equations to be jointly determined, catch the intrinsic feedback mechanisms between the diversity of competencies and the ability to generated new valuable knowledge. Our main working hypothesis is that the resulting increasing returns can explain differential growth rates of new entrants.

3 Data sources and working sample

The analysis of this paper exploits two types of data sources, patent data and firm-level balance sheet-like data on firms' financials and performance.

Patent data are central to build our newly proposed measure of knowledge complexity, which in fact exploits the text of the abstract of patents. Patent data are primarily retrieved from PATSTAT (version Global 2021), a well-known and widely used dataset published and maintained by the European Patent Office (EPO), collecting the internal databases of patent documents of the EPO and other patent offices around the world. It is one of the most comprehensive data sources for studying patent empirics, encompassing over 100 million patent records and over 200 million legal status records from 90 patent authorities around the world. (see **?**) Given the goal to examine the complexity of the knowledge base of Italian firms, from PATSTAT we retrieved all patents filed by and granted to Italian firms in the 20 years 1990-2009, i.e. before we measure entry.[1]. After applying customarily cleaning to remove potential duplicate patents (document covering same invention filed at multiple patent offices around the world), we identify patent documents assigned to Italian firms in the considered period. For 65075 of these patents, the English text of the patent abstract was available, either from PATSTAT or from the ORBIS-IP dataset, maintained by Clarivate-Bureau Van Dijk. This corpus of text forms the basis for our analysis of knowledge complexity. In fact, as we detail in Section 4 below, English text is required by topic modeling algorithms we implement in the construction of knowledge complexity measures.

The source of information about entry and entrants' characteristics is the AIDA (Analisi Infomatizzata delle Imprese Italiane) dataset, which is the Italian section of the well-known AMADEUS and ORBIS datasets, all maintained by Clarivate-Bureau Van Dijk. AIDA includes essentially the universe of limited liability Italian firms. This type of firms have to make their balance sheet public and deposit them at the local Chamber of Commerce, which are also responsible for the formal initial declaration of a new business in the business register. Bureau Van Djik essentially collects, clean, harmonize and structure these data and offer them commercially for research and or financial purposes. Much like its European and worldwide relatives dataset AMADEUS and ORBIS, subscription to AIDA allows to access data over a 10 years rolling window. For this study, we had access to the late 2021 release, providing us yearly data on balance sheets and profit/loss accounts over the period 2012-2021. Out of this initial sample, a key issue for us was to identify firms that were new entrants at the beginning of the observed period. AIDA helps in this, since it reports information on year of incorporation, that is the year a given firm or entity is enrolled as a new business in the official register. We thus took the set of firms with incorporation date in 2009, 2010 or 2011, i.e. the three years before we can observe the financial data. This allows to have a larger initial sample than we would have had by taking firms with incorporation date in just 2011, the year before the date in which financial data start. Also, the 3-years window allows to smooth year-specific idiosyncratic factors that may affect the incorporation patterns in a given single year.

The initial sample of potential entrants with incorporation date in 2009-2011, encompassed about 4000 firms. However, a key step is then to move from incorporation to identification of "genuine entry". Indeed, new registrations just signal a new entity from a legal viewpoint, but they may not correspond to a new entity in economic terms. In fact, a number of firm demography events, such as changes of names and legal status (e.g. from a simple limited firm to a limited firm listed in the stock exchange), as well as mergers, acquisitions and the like, are possibly resulting into the registration of a new entity. We therefore implemented a number of checks to identify the newly registered entities which were actually not genuine entrants. First, we dropped from the sample all the entities for which another firm with the same name and legal address, but different identifier (the so-called Bureau Van DijK identifier, BvDID) exist in AIDA. Then, on the remaining set, we performed manual checks of names and founding dates. Some firms were removed because they were clearly not new, but rather quite large

[1]There were essentially no patents with Italian assignee before this period

and/or very well known in the country. To give examples, this was the case of the newly constituted but clearly not genuinely new "ENEL INGEGNERIA E RICERCA S.P.A", a branch of the state-controlled electricity giant ENEL. And the same happened with "NUOVO PIGNONE S.P.A", which is reported as newly incorporated in 2009 after separating from General Electric, but is actually a very old firm, founded in Florence around 1840. Other firms were removed as "fake entrants" after checking their actual foundation year in their websites. The sample reduced at that point to about 3600 firms entering between 2009 and 2011, for which we could then observe growth-related variables and other firm-level characteristics over the period 2012-2021.

Table 1: Working sample composition

		N. of firms	%
AVG.SIZE	Micro (Empl.≤10)	1313	69.40
	SMEs (Empl.> 10 and ≤250)	558	29.49
	Large (Empl.> 250)	21	1.11
SECTOR(NACE)	Agric., Forestry, Fish. (A)	5	0.26
	Mining (B)	0	0
	Manufacturing (C)	662	34.99
	Utilities (D-E)	29	1.53
	Construction (F)	109	5.76
	Services excl. Finance (G-J, L and N)	608	32.14
	Finance (K)	107	5.66
	Profess. & Scientific Serv. (M)	339	17.92
	Public; Health; Educ. (O-Q)	16	0.85
	Other Services (S)	17	0.89
REGION	North-West	665	35.14
	North-East	581	30.71
	Center	388	20.51
	South	199	10.52
	Island	59	3.12

Notes: Number and % of firms in the working sample of 2009-2011 entrants, by firm size (average number of employees over the sample period 2012-2021), by sector of activity (NACE Rev.2, broad sections) and geographical macro-areas (NUTS 1 aggregation). Total number of firms is 1892.

We then had to gather information on the patenting activity of this set of entrants identified in AIDA. In fact, we can associate an index of complexity only to those firms which do patent, as our complexity measure relies upon topic modeling of patent text. The patent-to-firm matching is done through ORBIS-IP. Technically, we searched for which of the BvDID codes identified as potential entrants from AIDA, are also the assignee of a patent or the global ultimate owner of an assignee of a patent (i.e., considering worldwide ownership structure within group affiliations). This step further reduced the working sample considerably, for two reasons. First, only a subset of the entrants identified in AIDA have patents. This is in line with the observation that patenting at or just before entry is possible, but not a widespread phenomenon. Arguably it is even less so in the Italian context, where patenting in general, not only among entrants, is less common than in other comparable countries. Second, for the entrants identified in AIDA which do have one or more patents, in order to include them in the analysis we also need

that the matched patent documents report an abstract written in English. Otherwise those patents would be automatically dropped from the analysis, as they cannot be read by the algorithms we use for topic-modeling in the construction of knowledge base complexity (see details in Section 4).

In the end, we are left with 1892 firms entering in 2009-2011 and having at least one patent with English abstract, so that we can measure their knowledge base complexity and examine how the latter relates to post-entry growth performance over the period 2012-2021. Table 1 reports basic information about the composition of the working sample of 2009-2011 entrants. As one could expect, the vast majority of them is quite small in size. In fact, taking their average size over the sample period, the 69.4% are actually micro firms, i.e. firms with up to 10 employees. They are active in all sectors of activity, although they are mostly concentrated in Manufacturing or Services. Their geographical distribution is unbalanced toward regions in the North and Center of the country, in line with the expectation that more business dynamism characterizes the more advanced areas of the Italian economy, whereas entry is arguably more difficult and less frequent in South and island regions.

4 Topic modeling and the complexity index

The first attempt to exploit the textual content of patents to measure knowledge diversification, going beyond traditional measures based on technological classes, is by **?**, who look at the frequency distribution of trigrams, that is a group of three consecutive words, defined by the authors as triplets and elected as the unit of analysis. In more recent years, advancements in Natural Languages Processing led to development of specific algorithms with the purpose of organize texts in categories. **?** introduced in economics the use of topic modeling as an effective tool to elicit from a corpus of documents the unknown group of categories which can best define the content of the entire corpus and of each single document. The topic modeling exercise they carried on is now widely accepted in the discipline (among the many, see **????**).

4.1 LDA

Topic modeling is an unsupervised statistical modeling technique which, based on the co-occurrence pattern of words in documents, infers a latent structure of patterns of words occurring together. These patterns are defined topics and they can summarize the content of an entire documents collection (**?**). Topics are described by a frequency distribution over the words used in the documents and each documents can be thereafter summarized by a frequency distribution over the topics.

Among the various topic modeling techniques we employ Latent Dirichlet Allocation as introduced by **?** and vastly employed in economics. The input of any LDA exercise is a group of documents, in the bag-of-word format, that is without considering the position of words in the text, but their frequency distribution only. In our case, the corpus consists in English abstracts of the 65075 patents granted to Italian firms in the 20 years 1990-2009 and to the 25056 patents owned by new entrants in the years 2009, 2010, and 2011 for a total of 90131 patents. We apply the standard data cleaning procedure which consists of removing non-latin and non-ASCII characters, non informative words such as stop-words, very common and very rare words (present in more or less than 97% and 3% of the documents, respectively). After cleaning, we end up with a total of 3018548 word occurrences over a dictionary

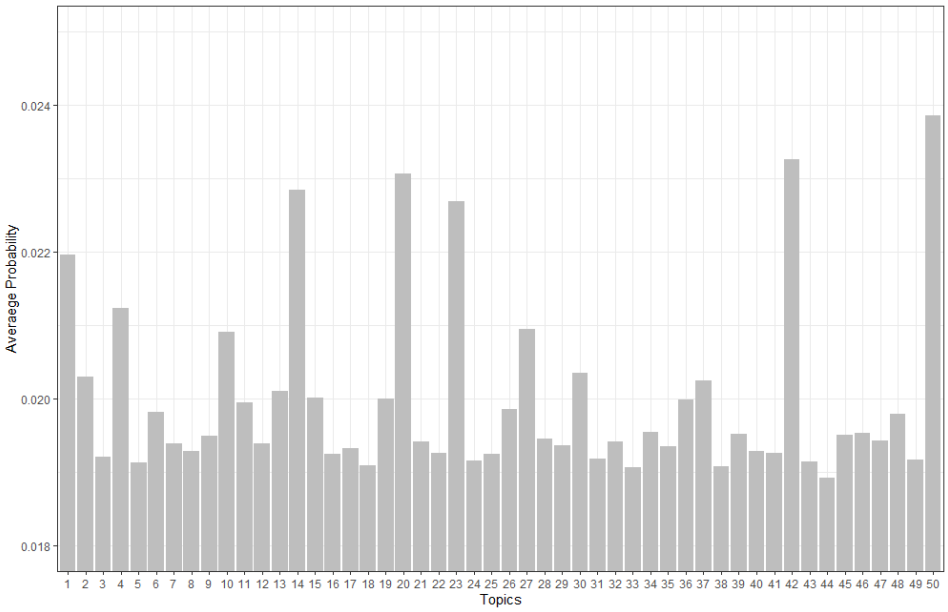


Figure 1: Topics importance by average posterior probability

of 271 unique terms from the initial 212430. On this corpus, we run the topic modeling exercise with 50 topics. Since LDA does not allow for exact inference (**??**) we use the Gibbs sampling method (**?**). The exercise is run with the package *topicmodels* (**?**) for the language R (**?**). As output we retrieve the 50 topics, each of them characterized by a probability distribution over the 271 unique terms, and a distribution of probability over the 50 topics for each of the patent.

The topics capture different aspects of the textual content. Some of them describe specific technologies, such as Topic 20, characterized by the words {**chamber, flow, fluid, valv, air**}, or Topic 22{ **machin, assembl, sheet, transfer, automat**}. Others, such as Topic 10 {**process,product, step,obtain**}, instead capture word patterns relating to the description of the inventive step of the patent.[2] Also, not all topics are equally distributed in the collection of documents, as we show in Fig. 1 reporting the relative importance of each topic in the collection as measured by their average posterior per document. For instance, Topic 42 {**group, compound, acid, formula** }, occurs more often than Topic 18 {**lower, upper, edg, later**}. Key to notice is also that each patent abstract can be seen as combination of topics. For instance the patent EP06425530A "A fuel injector for internal combustion engines" granted in 2006 to Magneti Marelli Powertrain S.p.A. is characterized by the topic distribution in Fig. 2: Topic 19 {**posit, move, movabl, lock, close** } and Topic 20 { **chamber, flow, fluid, valv, air**} display the highest posterior probability. A way of representing the full information contained in LDA output is to conceive it as a large 2-modes weighted network, as in Fig.3, or as its adjacency matrix A with dimension $90250x50$, where patents are the 90250 rows, topics are the 50 columns and the posterior probability are the values of each cell. This matrix is the input of the next step of this methodology, as explained here below.

[2]Fig. 6 in Appendix A shows the posterior probability of the most characterizing words for each topic.

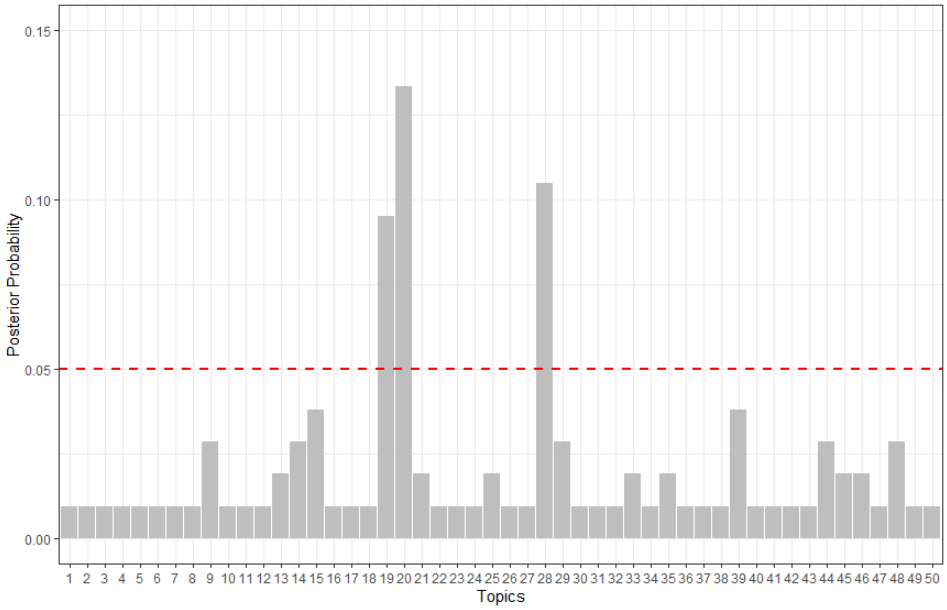


Figure 2: Topics distribution for Patent EP06425530A. Red line is the median posterior probability in the sample

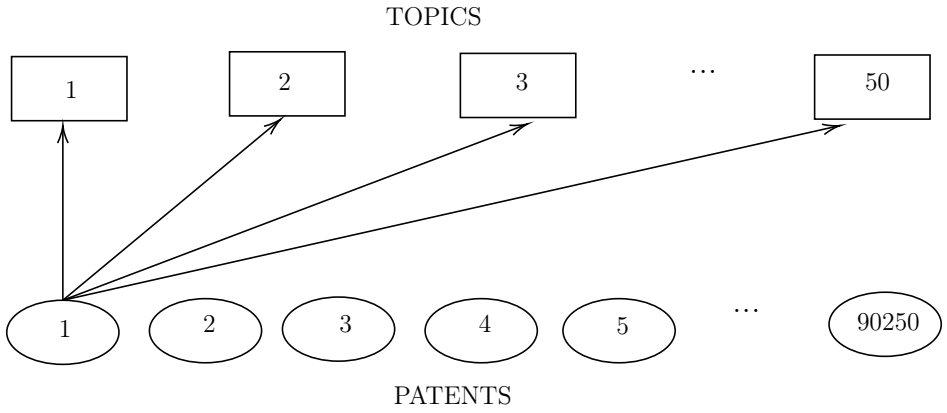
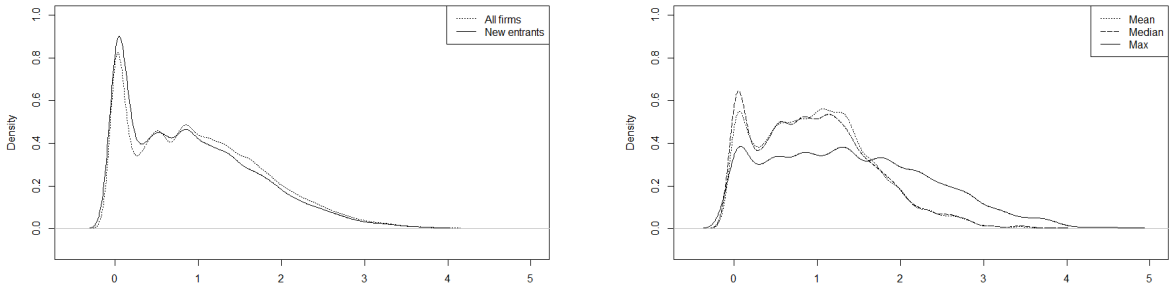


Figure 3: two-modes network representation of topic modeling output

4.2 Knowledge complexity at the firm level

The essence of the notion of complexity combines the diversification of the knowledge base of a firm, which we capture with the topics, with the distinctiveness of the topics themselves. Under the assumption that more complex knowledge bases are associated with more distinct topics, and vice-versa, we need to jointly determine:

1. Complexity of the knowledge base= f(Variety of knowledge components and distinctiveness of knowledge components)
2. Distinctiveness of the knowledge components= f(Variety of uses of knowledge components and Complexity of the knowledge base)



(a) Patent complexity distribution (b) Firms' complexity distribution

Figure 4: Kernel density of patent complexity

The standard complexity index at the country level is measured on non-weighted two modes country-product network, in which an edge exists if the normalized revealed technological advantage index takes a value larger than 0. We have as input of the process a patent-topic two modes network as the adjacency matrix A. We simplify A into the adjacency matrix M considering the value of each cell equal to 1 if its value is larger than the the median value 0.05 of the topic distribution, and zero otherwise. Thus, in the matrix M, the generic element $M_{i,j}$ informs about the presence of an edge between the patent i and topic j.

There are two different approaches to the empirical calculation of complexity. We follow **?** which allows for non-linearity:

$$\widetilde{PC_i^n} = \sum_j M_{i,j} * TD_j^{(n-1)} \tag{1}$$

$$\widetilde{TD_j^n} = \frac{1}{\sum_i M_{i,j} * \frac{1}{PC_j^{(n-1)}}} \tag{2}$$

The complexity of a patent i (PC_i) is the sum of the single links with a topic ($\sum_j M_{i,j}$), weighted by the distinctiveness of a topic j (TD_j). In the ECI approach (**?**) the topic distinctiveness is the symmetric sum over the columns ($\sum_i M_{i,j}$), weighted by the patent complexity. However, **?** suggest that the distinctiveness of the knowledge component does not necessarily follow the same linear relation and we account for the non-linearity as in Equation 2. Calculations are made with a two-steps iteration process with the package *economiccomplexity* (**?**) for the language R (**?**).

As a result, we obtain the variable complexity, which is the complexity of each single patent. Then, to move from patent-level to firm-level complexity, since the post-2009 entrant firms which we study might have more than one patent, for each firm we compute the average complexity of its patent portfolio, and also its median and maximum value as further controls. Fig. 4a shows the density distribution of complexity for the pre-2009 patents and for the patents owned by the new entrants after 2009. Post-2009 patents owned by new entrants do not show a significant difference from the complexity of pre-2009 patents. Fig. 4b depicts the aggregation of patent complexity at the firm level for the new entrants (Mean, Median and Max). Compared to the distribution of patent-level complexity, the aggregation by firm does not qualitatively change the shape of the distribution but it considerably shifts the probability mass to the right. The shift is particularly marked when we consider the patents with

the highest complexity (max). This evidence derives from the working assumption of the empirical model which characterizes with a higher complexity those patents occurring in firms with a more diversified knowledge base. Table 2 shows the most complex and least complex patents. All the most complex patents describes invention which combine knowledge in mechanics and electronics, while the least complex one are in mechanics or in very specialized niches such as drug production and pet care services.

Most complex patents	New Entrants
Improvements introduced in the remote switches	NO
Circuit breaker for low-voltage electric circuit	NO
A fastening device for rails comprising a rigid structure	YES
A docking system for space modules	YES
Least complex patents	
The invention concerns the use of pramipexol	NO
A basin for washing and drying pets combined with a basin support	NO
Glass fabric produced with zero-twist yarn	NO
Jacquard selection in a textile machine characterized by a selection jack	NO

Table 2: Most and least complex patents

5 Complexity and post-entry growth: empirical analysis

We then move to explore the empirical relation between pre-entry complexity of knowledge, as measured by the complexity index, and post-entry growth performance of the sample of 2009-2011 Italian entrants we can then follow over the period 2012-2021. We recall these are 1892 firms for which we have financial data from AIDA-Bureau Van Djick and have at least one patent with English abstract at entry. In fact, obviously, we need a firm has at least one patent in order to associate the complexity index to a firm. As explained above, if they have just one patent, then the complexity of the firm's knowledge base is proxied by the complexity index of that single patent. For firms having more than one patent, we define knowledge complexity of a firm by aggregating the complexity indexes associated to the patents the firm has, taking the mean, the median and the maximum of the complexity indexes associated to those patents.

We take the simple log-difference of annual sales S to define the growth rates

$$g_{i,t} = log(S_{i,t}) - log(S_{i,t-1}) \tag{3}$$

of firm i in year t. Despite other proxies of size-growth dynamics are routinely considered in firm growth empirics, the choice of sales allows us to focus on the association between complexity and post-entry success in the market. This is in fact our research question. Conversely, size proxies such as employment or tangible assets look at the input side of growth processes and are therefore more suited to describe growth of production capacity, relating to labour and investment dynamics.

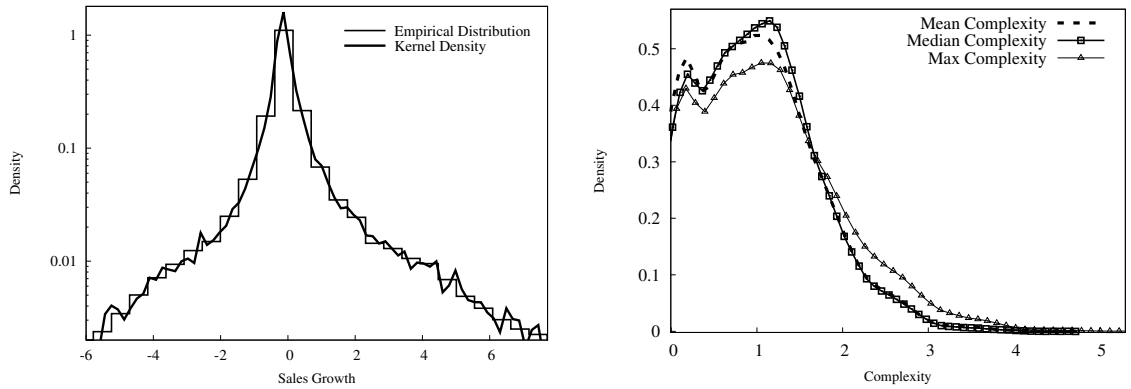


Figure 5: *Left*: Empirical distribution and kernel density of sales growth, pooling all years. *Right*: Empirical density of Mean, Median and Max Complexity of patents per firm. Figures obtained on the working sample of 2009-2011 entrants.

Table 3: Descriptives of main variables and controls

	Mean	Std.Dev.	Min	Max
PANEL A – Main Variables:				
Sales Growth	0.1840	1.3265	-10.4578	11.1291
Mean Complexity	1.0037	0.6823	0	3.9507
Median Complexity	0.9818	0.7023	0	3.9507
Max Complexity	1.4190	0.9756	0	5.2906
PANEL B – Controls:				
N. of Employees	18.2114	52.0615	1	1519
Productivity	394.2518	1,564.6748	0	66,693.5859
ROS	3.5731	12.0989	-49.87	30
Liquid Assets	3.6221	2.6097	-6.9078	13.1520
Leverage(%)	32.6384	31.4121	-98.7710	100
N. of patents	5.9413	22.5600	1	734
Intangibles	548.0454	3,831.002	1	10,1385.2

Panel A of Table 3 shows basic descriptive statistics of complexity measures and sales growth, computed across the firms in the working sample and pooling over the period 2013-21, i.e. the years over which we can compute the yearly growth rates. Overall, growth is quite skewed, while we also observe, in Figure 5, that our data confirm the fat-tailed, Laplace-like behavior of the growth rates distribution which has been established as a robust stylized fact in the firm-growth literature (see **??????**). The descriptive statistics of complexity measures show that Mean and Median Complexity are quite similar to each other, whereas Max Complexity is likely to discriminate more across firms than the other two measures. This is in line with Figure 4b, showing that the distribution of Max Complexity, compared to the other two complexity measures, is sensibly shifted to the right, less concentrated for low values and more concentrated at relatively high values.

Table 4: Baseline regression results

	(1)	(2)	(3)	(4)	(5)	(6)
Mean_Complexity	0.0083**			0.0134***		
	(0.004)			(0.004)		
Median_Complexity		0.0091**			0.0120***	
		(0.004)			(0.004)	
Max_Complexity			-0.0036			0.0114***
			(0.002)			(0.003)
$\ln(Empl_{t-1})$				-0.0291***	-0.0288***	-0.0289***
				(0.003)	(0.003)	(0.003)
$\ln(LabProd_{t-1})$				-0.0399***	-0.0402***	-0.0399***
				(0.003)	(0.003)	(0.003)
ROS_{t-1}				0.0010***	0.0010***	0.0010***
				(0.000)	(0.000)	(0.000)
$\ln(Liquid_{t-1})$				0.0075***	0.0077***	0.0074***
				(0.001)	(0.001)	(0.001)
$Leverage_{t-1}$				-0.0004***	-0.0004***	-0.0004***
				(0.000)	(0.000)	(0.000)
$\ln(\#patents)$				0.0018	0.0019	-0.0037
				(0.003)	(0.003)	(0.003)
$\ln(Intang_{t-1})$				0.0063***	0.0062***	0.0063***
				(0.001)	(0.001)	(0.001)
Constant	0.0105**	0.0098**	0.0242***	0.3692***	0.3655***	0.3688***
	(0.004)	(0.004)	(0.005)	(0.048)	(0.048)	(0.048)

Notes: LAD estimates, robust standard errors in parenthesis. Regressions in columns 4-6 also include year, sector (macro-sections of NACE Rev.2) and region (20 NUTS-2 areas) fixed-effects. Asterisks denote significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

5.1 Baseline analysis

Our baseline model to examine the growth-complexity relationship is the following regression

$$g_{i,t} = \alpha + \beta \text{ Complexity}_i + \delta \mathbf{X_{i,t-1}} + \epsilon_{i,t} \tag{4}$$

where g is sales growth defined above, Complexity is alternatively the Mean, Median or Max Complexity index obtained by aggregating at the level of each firm i the complexity of firm i's patents, and the vector $\mathbf{X}$ is a set of firm-level controls, lagged one year to at least partially avoid simultaneity. Recall that the time period available for estimation is 2012-2021 and that complexity indexes are measured once, before entry (i.e., they do not vary over the estimation period, hence the omission of the t subscript in the equation). Thus, the coefficient of main interest, β, is to be interpreted as capturing the role of complexity for post-entry growth patterns of the 2009-2011 entrant firms.

Adding a considerably rich set of controls is particularly important, since standard panel techniques to control for time-invariant unobserved heterogeneity cannot be applied in our context for the simple reason that the Complexity measures are time-invariant. The set of controls X includes: the number of employees ($Empl$) to proxy for firm size; a measure of labour productivity ($LabProd$), computed as sales per employee; a proxy of profitability generated by the operational activity of firms, defined as

the Returns on Sales ratio (*ROS*); a leverage index (*Leverage*) computed as total assets over equity, accounting for financial structure of firms and also often used as an indirect proxy of access to credit; the amount of cash and liquid assets (*Liquid*), which is also a proxy of access to credit; and two measures of the innovative activity of firms, that is the number of patents and the value of intangible assets (*Intang*). Summary statistics of the controls are in Table 3, pooling over the entire sample period. We also add a full set of year, sector (broad NACE sections, see Table 1) and regional (20 NUTS-1 regions) fixed-effects.

In Table 4 we report our baseline estimates. As a benchmark providing information on unconditional correlations, columns 1-3 report estimates of a model without controls, while estimates of the model of Equation 4 are in columns 3-6. Notice that here, as well as in the rest of the paper, the estimation results are obtained through the LAD estimator. This estimator is standard in firm-growth literature, as it copes with the well-documented non-normality in the distribution of growth rates which we also observe in our data (recall Figure 5).

The results suggest a significant and positive association between knowledge complexity and growth, even after controlling for firm characteristics and fixed-effects. This finding is essentially invariant to the three alternative measures of complexity. The estimated coefficients on the control variables are broadly in line with the expectations. Growth negatively associates with size, while it positively associates with profitability, liquidity and investment in intangibles. Negative sign on productivity is in line with earlier studies showing that efficiency is not a strong driver of firm growth, suggesting that market selection does not work by strongly rewarding more efficient firms with faster sales growth, in particular in Italy (**??**). The negative sign on Leverage implies faster growth for firms having comparatively more debt (less assets) per unit of equity. This in line with the notion that accessing external finance is key for growth of new firms in the early phase after entry.

5.2 Robustness checks and extended analysis

In Table 5 we examine robustness of results with respect to inclusion of additional regressors. First, we extend the baseline model to also include the square of the complexity measures. This helps capturing possible non-linear effects of complexity, that may result from strongly skewed and asymmetric empirical distribution of complexity measures (recall Figure 5). The estimation results, in columns 1-3 of Table 5, confirm a positive association between complexity and growth rates, although at weaker levels of significance compared to the baseline estimates. The non-linear effects turn out as not statistically significant.[3]

Second, we test whether our baseline findings survive when we add to the set of controls a measure of knowledge specialization. This is defined as the Herfindahl–Hirschman (HH) which measure whether knowledge in a patent is concentrated in few topics or rather homogeneously distributed among the many. Formally, we compute

$$HH_i = \sum_{j}^{50} p_{i,j}^2 \tag{5}$$

[3]Estimated coefficients on the control variables replicate the baseline results. We thus do not comment further on these patterns. The same holds true for all the robustness checks and extensions presented in the rest of the section.

Table 5: Robustness checks

	Complexity Squared			HH-index			Innovative SMEs		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Mean_Complexity	0.0179*			0.0082*			0.0133***		
	(0.011)			(0.004)			(0.004)		
Mean_Complexity^2	-0.0020								
	(0.004)								
Median_Complexity		0.0197*			0.0078*			0.0133***	
		(0.010)			(0.004)			(0.004)	
Median_Complexity^2		-0.0030							
		(0.004)							
Max_Complexity			0.0155*			0.0072**			0.0107***
			(0.009)			(0.004)			(0.003)
Max_Complexity^2			-0.0014						
			(0.003)						
HH index				0.8401***	0.8297***	0.8508***			
				(0.267)	(0.268)	(0.250)			
Innov_SME							0.0900***	0.0888***	0.0896***
							(0.019)	(0.019)	(0.021)
ln $Empl_{t-1}$	-0.0290***	-0.0287***	-0.0289***	-0.0284***	-0.0281***	-0.0290***	-0.0252***	-0.0258***	-0.0255***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
ln $LabProd_{t-1}$	-0.0397***	-0.0400***	-0.0401***	-0.0406***	-0.0404***	-0.0409***	-0.0386***	-0.0392***	-0.0383***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
ROS_{t-1}	0.0010***	0.0010***	0.0010***	0.0010***	0.0010***	0.0009***	0.0012***	0.0012***	0.0011***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
ln $Liquid_{t-1}$	0.0076***	0.0076***	0.0075***	0.0078***	0.0077***	0.0080***	0.0074***	0.0076***	0.0077***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$Leverage_{t-1}$	-0.0004***	-0.0004***	-0.0004***	-0.0004***	-0.0004***	-0.0004***	-0.0004***	-0.0004***	-0.0004***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
ln $\#patents$	0.0016	0.0017	-0.0032	-0.0055	-0.0051	-0.0081**	0.0009	0.0011	-0.0040
	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)
ln $Intang_{t-1}$	0.0062***	0.0062***	0.0064***	0.0056***	0.0056***	0.0056***	0.0050***	0.0052***	0.0048***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Constant	0.3658***	0.3661***	0.3687***	0.3490***	0.3477***	0.3506***	0.3495***	0.3530***	0.3483***
	(0.046)	(0.048)	(0.049)	(0.043)	(0.043)	(0.045)	(0.043)	(0.041)	(0.041)

Notes: LAD estimates, robust standard errors in parenthesis. All the regressions also include year, sector (macro-sections of NACE Rev.2) and region (20 NUTS-2 areas) fixed-effects. Asterisks denote significance levels: $^*p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

where, $p_{i,j}$ is the posterior probability of topic j in patent i. To aggregate at firm level, for firms with more than a patent, we consider the mean value computed over the patents in the firm patent portfolio. The results, reported in columns 4-6 of Table 5, corroborate the baseline finding of a positive and significant association between complexity and post-entry growth. The HH index also has strongly significant positive coefficient, meaning that coherence is indeed a strong driver of growth. Notice that the HH is proxy for the traditional approach to variety since it does not consider the distinctiveness of each topic.

Next, we try to improve upon the ability of the regression models to account for innovativeness of firms. The AIDA data do not report information on other standard proxies of innovative activity, such as product/process innovation, while R&D is measured but largely missing for our set of entrant firms. We exploit here information on whether firms comply with the definition of "Innovative SMEs", a special status recognized in 2015 by the Italian regulation which allow firms to access fiscal, financial and administrative benefits. Innovative SMEs have the obligation to comply with some specific innovation requirements regarding specific patents, staff ownership and volume of investments in research and development. In our working sample of 2009-2011 entrants, we have 89 firms that comply with the definition. Accordingly, we add to the baseline model a dummy "$Innov\_SME$". The estimates in columns 7-9 show that these firms grow faster than the rest of the sample. We anyway confirm that complexity has a positive association with growth, no matter the complexity measure considered.

Another issue deserving further investigation is whether the observed role of complexity may hinder a sheer size effect. In fact, it may well be that complexity increases with firm size, due to larger firms being naturally more likely to have more and more complex patents, compared to smaller firms. On the other hand, one could also argue that, for a given level of knowledge complexity, larger firms are in a better position to "manage" their knowledge complexity and translate complexity into growth outcomes, thanks to their generally deeper financial pockets and superior organizational capabilities. We do control for firm size and number of patents in the main analysis. But this is not enough to control for the combined effect of size and complexity in shaping growth patterns.

In Table 6, we report about a series of variations of the baseline regression model, where we further investigate the role of firm size and allow for size-complexity interactions. First, we account for the observed abundance in our sample of micro firms, i.e. firms with up to 10 employees (recall Table 1). We create a dummy ($MICRO$) taking value 1 if a firm is indeed a micro firm, and use this in place of the number of employees as the control for firm size. The results, reported in columns 1-3, reveal that micro firms do grow faster and confirm the positive association between growth rates and complexity measures which emerged from the baseline estimates. Next, in columns 4-6 of Table 6, we show estimates of a model including the interaction between the MICRO dummy and the complexity measures. The coefficients on the direct effect of complexity, capturing how complexity relates to growth among the firms with more than 10 employees ($MICRO$=0), are all positive and significant, in line with previous estimates. The positive coefficients on the $MICRO$ dummy confirm that very small firms in the sample grow faster, but then the un-significant interaction terms reveal no difference between $MICRO$ and other firms in the strength of association between complexity and growth. Similar results for the direct effects of size and complexity emerge also from the estimates in columns 7-9, referring to a model where we now include the $Innov\_SME$ dummy and its interaction with complexity. In this case, however, the interaction coefficient turns significant and negative. This hints that small-medium firms, even when they are innovative SMEs, may have indeed difficulties in translating complexity into growth, and may even grow less than other firms, for given complexity level.

Table 6: Firm Size effects

	MICRO firms dummy			Interaction Complexity×MICRO			Complexity×Innovative SMEs		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Mean_Complexity	0.0123***			0.0140***			0.0162***		
	(0.004)			(0.004)			(0.004)		
Median_Complexity		0.0112***			0.0130***			0.0151***	
		(0.004)			(0.004)			(0.004)	
Max_Complexity			0.0107***			0.0100***			0.0121***
			(0.003)			(0.004)			(0.003)
$MICRO_{t-1}$	0.0462***	0.0453***	0.0474***	0.0516***	0.0512***	0.0441***			
	(0.008)	(0.008)	(0.008)	(0.012)	(0.011)	(0.011)			
$MICRO_{t-1}$×Mean_Complexity				-0.0056					
				(0.009)					
$MICRO_{t-1}$×Median_Complexity					-0.0058				
					(0.009)				
$MICRO_{t-1}$×Max_Complexity						0.0026			
						(0.007)			
Innov_SME							0.1666***	0.1531***	0.1422***
							(0.036)	(0.032)	(0.032)
Innov_SME×Mean_Complexity							-0.0811**		
							(0.032)		
Innov_SME×Median_Complexity								-0.0778***	
								(0.029)	
Innov_SME×Max_Complexity									-0.0357*
									(0.019)
ln $Empl_{t-1}$							-0.0248***	-0.0251***	-0.0253***
							(0.003)	(0.003)	(0.003)
ln $LabProd_{t-1}$	-0.0355***	-0.0354***	-0.0357***	-0.0355***	-0.0353***	-0.0354***	-0.0380***	-0.0386***	-0.0379***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
ROS_{t-1}	0.0011***	0.0011***	0.0011***	0.0010***	0.0010***	0.0012***	0.0013***	0.0012***	0.0012***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
ln $Liquid_{t-1}$	0.0057***	0.0056***	0.0061***	0.0057***	0.0057***	0.0061***	0.0069***	0.0071***	0.0074***
	(0.002)	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
$Leverage_{t-1}$	-0.0004***	-0.0004***	-0.0005***	-0.0004***	-0.0004***	-0.0005***	-0.0004***	-0.0004***	-0.0004***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
ln $\#patents$	-0.0013	-0.0013	-0.0065*	-0.0015	-0.0015	-0.0062*	0.0008	0.0010	-0.0043
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
ln $Intang_{t-1}$	0.0033**	0.0033***	0.0033***	0.0031**	0.0033***	0.0033***	0.0051***	0.0051***	0.0052***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Constant	0.2913***	0.2922***	0.2969***	0.2872***	0.2857***	0.2936***	0.3416***	0.3458***	0.3384***
	(0.034)	(0.041)	(0.042)	(0.033)	(0.040)	(0.040)	(0.040)	(0.040)	(0.041)

Notes: LAD estimates, robust standard errors in parenthesis. All the regressions also include year, sector (macro-sections of NACE Rev.2) and region (20 NUTS-2 areas) fixed-effects. Asterisks denote significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 7: Time effects A

	PRE-COVID			Period 2013-2016			Period 2017-2021		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Mean_Complexity	0.0135***			0.0113			0.0099**		
	(0.005)			(0.008)			(0.005)		
Median_Complexity		0.0114***			0.0117			0.0092**	
		(0.004)			(0.008)			(0.005)	
Max_Complexity			0.0133***			0.0100			0.0103***
			(0.003)			(0.007)			(0.004)
$\ln Empl_{t-1}$	-0.0365***	-0.0359***	-0.0367***	-0.0523***	-0.0525***	-0.0522***	-0.0158***	-0.0154***	-0.0155***
	(0.003)	(0.003)	(0.003)	(0.006)	(0.005)	(0.005)	(0.003)	(0.003)	(0.004)
$\ln LabProd_{t-1}$	-0.0474***	-0.0472***	-0.0475***	-0.0668***	-0.0668***	-0.0666***	-0.0270***	-0.0271***	-0.0267***
	(0.003)	(0.003)	(0.003)	(0.006)	(0.006)	(0.006)	(0.004)	(0.004)	(0.004)
ROS_{t-1}	0.0007	0.0007*	0.0007	0.0014**	0.0013**	0.0013**	0.0015***	0.0014***	0.0015***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)
$\ln Liquid_{t-1}$	0.0073***	0.0072***	0.0074***	0.0080***	0.0081***	0.0075***	0.0079***	0.0077***	0.0078***
	(0.001)	(0.001)	(0.001)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)
$Leverage_{t-1}$	-0.0006***	-0.0006***	-0.0006***	-0.0011***	-0.0011***	-0.0011***	-0.0002	-0.0002	-0.0002
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
$\ln \#patents$	0.0022	0.0023	-0.0044	0.0018	0.0020	-0.0029	0.0017	0.0021	-0.0028
	(0.003)	(0.003)	(0.003)	(0.005)	(0.004)	(0.005)	(0.003)	(0.003)	(0.004)
$\ln Intang_{t-1}$	0.0091***	0.0089***	0.0093***	0.0113***	0.0114***	0.0118***	0.0037**	0.0034**	0.0031**
	(0.001)	(0.001)	(0.001)	(0.003)	(0.003)	(0.003)	(0.002)	(0.001)	(0.002)
Constant	0.4748***	0.4750***	0.4731***	0.5700***	0.5721***	0.5654***	0.2404	0.2388	0.2415
	(0.050)	(0.049)	(0.048)	(0.083)	(0.085)	(0.084)	(0.206)	(0.204)	(0.204)

Notes: LAD estimates, robust standard errors in parenthesis. All the regressions also include year, sector (macro-sections of NACE Rev.2) and region (20 NUTS-2 areas) fixed-effects. Asterisks denote significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

A further qualification of the baseline findings that we explore relates to possible variation over time. This is interesting in general, as the sample time-period spans over quite peculiar macro-dynamics. Indeed, the final two years in the data (2020 and 2021) are obviously affected by the COVID19 pandemic, while the initial years coincide with the phase of unprecedented downturn following the Great Depression and the subsequent debt-crisis, leading to severe economic turmoil. However, exploring time variation of results is also interesting in relation to our specific focus on entrant firms. Indeed, it is well-known that the initial years after entry are typically characterized by peculiar patterns, entailing high instability due to processes of adaptation, trial and errors, combining fast growth and frequent failures, while firms tend to stabilize after surviving this initial phase.

To account for possible influence of COVID19 pandemic, we have re-estimated the baseline model on a reduced sample excluding the last two years (2020 and 2021). The results, in columns 1-3 of Table 7, are in line with our main findings. Next, we split the sample into two periods spanning the years 2012-2016 and 2017-2021, respectively, and re-estimate the baseline models separately over the two periods. The corresponding results are reported in columns 3-6 an columns 7-9 of Table 7, respectively. They do reveal significant variation. Indeed, we find that complexity measures loose their statistical significance in the first period, while they remain positive and significant in the second period.

Table 8: Average growth patterns

	Avg.Growth 2013-16 vs. 2012			Avg.Growth 2017-21 vs. 2016		
	(1)	(2)	(3)	(4)	(5)	(6)
Mean_Complexity	0.0018			0.0186***		
	(0.013)			(0.007)		
Median_Complexity		0.0017			0.0183**	
		(0.012)			(0.007)	
Max_Complexity			0.0016			0.0149***
			(0.010)			(0.006)
ln $Empl_{init}$	-0.0639***	-0.0641***	-0.0632***	-0.0061	-0.0064	-0.0048
	(0.009)	(0.009)	(0.009)	(0.005)	(0.005)	(0.004)
ln $LabProd_{init}$	-0.1102***	-0.1103***	-0.1099***	-0.0180***	-0.0180***	-0.0158***
	(0.009)	(0.009)	(0.009)	(0.005)	(0.006)	(0.005)
ROS_{init}	0.0007	0.0007	0.0007	-0.0005	-0.0003	-0.0005
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
ln $Liquid_{init}$	0.0084*	0.0084*	0.0084*	0.0042**	0.0042**	0.0035**
	(0.005)	(0.005)	(0.004)	(0.002)	(0.002)	(0.002)
$Leverage_{init}$	-0.0013*	-0.0013*	-0.0013*	0.0007***	0.0006**	0.0006***
	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)
ln $\#patents$	0.0129	0.0131*	0.0117	0.0033	0.0032	-0.0030
	(0.008)	(0.008)	(0.009)	(0.005)	(0.005)	(0.005)
ln $Intang_{init}$	0.0119***	0.0119***	0.0116***	-0.0020	-0.0015	-0.0022
	(0.004)	(0.004)	(0.004)	(0.002)	(0.002)	(0.002)
Constant	0.8170***	0.8276***	0.8144***	0.0470	0.0467	0.0307
	(0.135)	(0.136)	(0.137)	(0.067)	(0.069)	(0.063)

Notes: LAD estimates, robust standard errors in parenthesis. The subscript *init* on the time-varying controls indicate that they are set to their initial year values, i.e. 2012 for columns 1-3 and 2016 for columns 4-6. The regressions also include sector (macro-sections of NACE Rev.2) and region (20 NUTS-2 areas) fixed-effects. Asterisks denote significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

To further explore this result, we perform a different exercise where we examine how complexity relates to the average growth experienced by the firms in the two periods. More specifically, we run the regression

$$\text{AVG}g_{i,p} = \alpha + \beta\ \text{Complexity}_i + \delta\ \mathbf{X_{i,init}} + \epsilon_{i,p} \tag{6}$$

where $AVGg$ is the average of yearly sales growth rates g experienced by firm i over the period p (spanning the years 2012-2016 and 2017-2021, respectively), while the control variables are measured in the initial year of each period (i.e, the *init* subscript correspond to 2012 and 2016, respectively). In this way, we smooth growth patterns from possible year-by-year variability. The results, in Table 8, corroborate that the the complexity-growth association is positive and significant in the second period, while statistically not significant in the first years. Further investigation, and perhaps more detailed data, are needed to better explain these patterns. They may stem from adverse macro-economic conditions in the observed period or instability of growth patterns in the very first phase after entry, or a combination of these potential drivers. In any case, at their face value, they suggest that knowledge complexity is more likely to spur growth over medium-long run, than in the very short-run after entry.

6 Conclusion

Pre-entry knowledge is considered a factor that substantially affects the post-entry performance of new firms. Innovation, entrepreneurship and organization studies have adopted different proxies to understand this relationship. This paper conceptualizes and builds a new index of knowledge complexity based on text analysis of patents to measure the knowledge base of a firm. We assume that the complexity of the knowledge base is linked to the distinctiveness of the single pieces of knowledge of the firm and viceversa, in a self-reinforcing loop. Our measure of complexity therefore combines the degree of diversity and distinctiveness of the knowledge embedded in the patents. Exploiting data on of a sample of Italian firms born between 2009 and 2011, which we then follow over the period 2012-2021, we test the hypothesis that pre-entry complexity affects post-entry growth, controlling for size, labour productivity, profitability, financial structure of firms, access to credit, and innovative activity of firms. We find that complexity has generally a positive and significant impact on growth. Robustness analysis reveal this positive effect is stronger in the medium-long run while relatively weaker for innovative SMEs.

Policy and managerial implications are relevant for R&D strategies and business growth. The evidence presented in this paper corroborates the idea that valuable knowledge originate from a process of division of labour and shed new light on the mechanisms presented in the literature about the knowledge variety of founders in entrepreneurial literature. We claim that the presence of variety is the observable outcome of a division of labour and, at the same time, the pre-requisite for the creation of persistent competitive advantages. Firms entering in the market, inventors assessing new ventures and start-up oriented policies should carefully evaluate whether the variety of a team is the result of a coordinated knowledge specialization rather then a simple performance indicator.

The external validity of this paper is limited. Although the sample has ruled out possible fake entrants, the generalization of these results would require different sample in different countries and in different time frames. The focus on new entrants with at least one patent lead to a small sample and these results should be taken with caution since we cannot verify whether the complexity-growth association can be extended also to incumbents. However, we are confident that the idea of exploiting

the concept of variety at the firm level might be a fruitfully exploited in future research. The evidence produced in this paper points in this direction.

References

reference

Appendix

A LDA output

Figure 6: Posterior probability of top words per topic