Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

Economics

Ciclo 33

Settore Concorsuale: 13/A1- ECONOMIA POLITICA

Settore Scientifico Disciplinare: SECS-P/01- ECONOMIA POLITICA

Essay in Empirical Economics: Intangible Economy, Innovative Firms and Institution of Innovation

Presentata da: Andrea Greppi

Coordinatore Dottorato

Maria Bigoni

Supervisore

Alireza Naghavi

Co-supervisore

Tommaso Sonno

Esame finale anno 2022

Abstract

The present thesis is made up by three separate chapters covering topics related to international trade, intellectual property right, business groups and knowledge flows and finally, an attempt to identify promising and innovative young firms.
The first paper focuses on the role of institutions, and it shows how different types of institutions are important to determine comparative advantage for countries at different stage of development. The papers finds that intellectual property rights (IPR) protection changes export composition of OECD countries toward IP-intensive sectors, whereas contract enforcement is a driver of export of relation-specific inputs in non-OECD countries. However, better IPR quality encourages technology transfer by redirecting non-OECD imports toward IP-intensive industries. The second chapter studies how subsidiaries of Business Groups interact between each other. In particular, the paper highlights productivity gains that affiliates enjoy from intangible assets developed by other firms within the same group. The analysis shows that a key element to consider in order to understand these flows, is to take into account the hierarchical links between subsidiaries. This crucial step allows to show that within Business Groups knowledge flows upwards, i.e. subsidiaries in lower layers share their knowledge to subsidiaries in upper layers. The third chapter presents a novel dataset assembled during my experience at the OECD on innovative start-up. Combining information from two different data provider, Crunchbase and Dealroom, and implementing several cleaning and matching procedure, we managed to gather this dataset which covers almost the universe of innovative start-ups all over the world. This data are a key element that will be exploited in several work at the OECD, for example to study the determinants of start-ups success (innovation, scale-up) and how relevant are Killer Acquisitions in start-ups.

Contents

List of Figures

List of Tables

Chapter 1

Institutions, Development, and Patterns of Trade[1]

Abstract

This study investigates how easing international transactions through improved legal institutions can result in divergent trade patterns for different economies. We provide evidence that the level of development governs the relevance of intellectual property rights (IPR) institutions in determining a country's comparative advantage. While IPR protection changes the composition of OECD exports towards IP-intensive sectors, contract enforcement is the key driver of specialization of non-OECD exports in relation-specific inputs. We extend the analysis to a bilateral framework to show assess in a unique framework the predictions on the pattern of trade, confirming the results. The findings suggest a concentration of innovation activities in the OECD, with non-OECD countries serving as potential outsourcing destinations.

[1]This paper is a joint work with Alireza Naghavi. We are grateful to Pol Antràs, Stefano Bolatto, Enrico Cantoni, Emanuele Forlani, Marco Grazzi, Olena Ivus, Bohdan Kukharskyy, Antonio Minniti, Gianmarco Ottaviano, Alessandro Sforza, Tommaso Sonno, Farid Toubal, and Francesco Venturini for helpful comments.
Andrea Greppi, University of Bologna, Department of Economics. E-mail: andrea.greppi2@unibo.it
Alireza Naghavi, University of Bologna, Department of Economics. E-mail: alireza.naghavi@unibo.it

1.1 Introduction

A worldwide wave of trade agreements and improvements in legal institutions has facilitated international transactions over the last decades (Antràs, 2016). The issue of intellectual property rights (IPRs) has in particular gained importance in both bilateral as well as multilateral trade talks. This has especially been true when parties at the talks include both advanced (OECD) and developing (non-OECD) economies and technology is at center stage. A proliferation of regional trade agreements with strict IPR provisions has fostered technology transfer from developed to developing countries (Santacreu, 2021a,b). Trade literature has in fact recognized IPR enforcement as a source of comparative advantage (Maskus and Yang, 2018). Previous related works have associated the quality of alternative institutions with comparative advantage, for example when contractual frictions create distortions in transactions between firms and their relation-specific input suppliers (Nunn, 2007; Levchenko, 2007).

Taking a step back and looking at OECD and non-OECD countries separately, an evident observation is the technological superiority of the former and the role of the latter as outsourcing locations for the procurement of intermediate inputs.[2] Protection of IPRs has been viewed as a key determinant of success in the race for latest technologies and efficient operation in IP-intensive sectors. The question we pose in this study is whether this role of IPRs holds generally for all countries, or if different institutions determine comparative advantage depending on a country's stock of knowledge or absorptive capacity. Using the same premise, we are additionally interested in exploring how IPRs influence the *direction* of trade in IP-intensive goods.

In this paper we carry out a systematic investigation to explain the alternative patterns of specialization across countries as an outcome of the quality of different institutions. We aim to shed light on whether differences in production structure, stage of development, or technological capability play a role in deciding which institutions determine a country's comparative advantage. The findings reveal a remarkable contrast in the institutional source of comparative advantage between OECD and non-OECD countries. In the former, IPR protection drives comparative advantage in IP-intensive industries, whereas better quality rule of law institutions regulate the patterns of specialization in the latter by rendering them attractive as outsourcing locations for highly relation-specific inputs. The reasoning follows the logic that IPRs shield

[2] Ten countries account for more than 80% of global spending on R&D and, with the exception of China, they are all developed countries (http://data.uis.unesco.org). On the other hand, the share of intermediate goods produced by developing countries has raised from 33% in 2005 to about half of the world production in intermediate goods in 2014 (https://wits.worldbank.org).

knowledge and create incentives in innovative industries, and contract enforcement encourages efficient supplier investment in the customization of relationship-specific inputs.

The first contribution of the analysis to the literature on the institutional sources of comparative advantage is to show that the quality of tangible and intangible property rights protection have different and mutually exclusive effects for countries at different levels of development. Acknowledging possible endogeneity of a country's IPR regime, we then address reverse causality concerns by making use of information on the timing of IPR reforms in a difference-in-difference framework, and as an instrumental variable. The results persist in a dynamic setting and reinforce our conjecture on the effect of the quality of institutions on trade patterns across countries and industries over time. The core results are robust to a host of additional checks and to a panel specification, which enables us to account for time fixed effects and time-varying country specific variables.

While the outcome may initially question the role of IPR policy as a tool to stimulate innovation in the developing world, we shift focus to imports to examine whether it can still play a role in development without affecting the export composition of these countries. Testing the standard notion of IPR in the literature as a tool to attract technology through imports, the findings suggest that IPRs encourage technology transfer by directing the import structure of developing countries toward IP-intensive goods. Looking at the effects of IPRs on both import and export patterns allows us to highlight how the same institution can have a different impact on the structure of trade for countries with dissimilar underlying characteristics: on the one hand, it is a source of comparative advantage for technologically advanced countries; on the other hand, it is an effective instrument to trigger imports in IP-intensive industries for developing countries.

This feature leads us to also look at bilateral trade flows between country pairs to investigate when and to what extent the patterns of trade of an exporting country may also be influenced by IPRs in the importing country. The results reveal a complementarity between the protection of intangible capital in the source and destination markets for promoting trade in high-tech industries. IPR institutions are an important determinant of the structure of exports (imports) for OECD (non-OECD) countries and increase trade in IP-intensive sectors from OECD to non-OECD countries. Interestingly, the bilateral analysis confirms our main findings accounting for the standard gravity factors and control for pairwise country characteristics, providing a further robustness checks for our core results.

The paper is organized as follows. The next section discusses the related literature. Section 1.3 describes the methodology and section 1.4 the data. Section 1.5 provides preliminary evidence and reports the baseline OLS estimates. Section 1.6 conducts robustness checks, shows a panel analysis and exploits aseries of IPR reforms to address reverse causality. Section 1.7 shifts focus to technology transfer and introduces import patterns and the bilateral set-up. Section 1.8 concludes.

1.2 Literature

With the world economy witnessing substantial changes in the structure of international trade, new sources of comparative advantage have come to light. A direction taken by literature seeks to establish that the standard determinants of trade patterns driven by Ricardian efficiency and Heckscher-Ohlin factors are themselves an outcome of deeper political and economic processes, broadly identified as the concept of "institutions". These studies emanate from the empirical methodology introduced in Rajan and Zingales (1998), interacting industry and country-specific characteristics to show for example that countries with more developed financial markets tend to export relatively more in industries that require large amounts of external finance (Beck, 2003). Some key contributions in this category highlight that countries with better rule of law specialize in the production of more institutionally dependent goods (Levchenko, 2007) and in goods with a higher share of relationship-specific inputs (Nunn, 2007; Ma et al., 2010).[3]

A similar approach has been adopted to study the role of IPRs in the pattern of comparative advantage. Also drawing on variation in effective patent rights across countries and varied impact across industries within a country, Hu and Png (2013) finds that stronger patent rights are associated with faster growth in more patent-intensive industries. More recently, Maskus and Yang (2018) demonstrates the positive effect of domestic patent rights on export performance in high-R&D goods. Following the same rationale, we introduce IPRs next to other types of institutions to highlight how their impact on comparative advantage varies across countries. Doing so reveals interesting insights as OECD and non-OECD economies host different production processes based on innovative and input provision activities. Consequently, the

[3] Other related papers using this technique look at factor proportions and trade (Romalis, 2004), credit constraints (Manova, 2008), gains from division of labor and specialization (Costinot, 2009), and flexibility of labor markets (Cuñat and Melitz, 2012). Chor (2010) provides a model of comparative advantage generated from the interaction of industry and country characteristics and tests the predictions in joint presence of several sources identified in the literature. See also Nunn and Trefler (2014) for an exhaustive literature review on institutions and comparative advantage.

process of specialization for these groups may be determined by different institutional sources of comparative advantages, namely the protection of intangible versus tangible property rights. Traditional IPR literature has however highlighted the role of the patent protection as an instrument to attract technology by encouraging imports.[4] Strengthening IPRs could promote technology diffusion to developing countries by increasing exports in patent-sensitive industries into those markets and facilitating access to new foreign technologies (Ivus, 2011, 2015). In a similar vein, Delgado et al. (2013) finds that trade in knowledge-intensive goods increased relative to other types of goods after the implementation of TRIPS. Looking at both cross section as well as firms' responses to six IPR reforms in a difference-in-differences framework, Lin and Lincoln (2017) show that IPR protection attracts imports of high-tech goods from technologically advanced countries. They are also the first to consider firm patenting in a gravity equation framework.

We take this path to provide a systematic analysis of the import patterns across countries. Distinguishing between the effect of IPR on OECD vis-à-vis non-OECD countries, we show that IPRs are only an important factor for the latter to attract technology-intensive goods. We also explore the significance of the interaction between the IPR policies of a source and a destination country in determining the patterns of trade using bilateral data.[5] We explicitly consider the IPR quality of the exporting country and conduct an industry-level analysis rather than aggregate levels of trade flow and development. We then simultaneously look at the IPR regime in the importing country to see if it contributes to attracting technology-intensive goods through trade among OECD countries, among non-OECD countries, or between the two regions.

1.3 Conceptual Framework and Methodology

It is well-known that better domestic IPR protection stimulates innovation (Qian, 2007; Chen and Puttitanun, 2005), attracts inflow of FDI (Javorcik, 2004) and boosts international technology transfer and domestic R&D (Branstetter et al., 2006, 2007). These channels sum up to the notion that better quality IPR institutions encourage exports in industries more intensive in IP, as the latter are more sensitive to the protection of their intangible assets (Maskus and Yang,

[4] See e.g. Maskus and Penubarti (1995); Smith (1999, 2001); Rafiquzzaman (2002); Co (2004); Awokuse and Yin (2010).

[5] The only paper to our knowledge that touches upon the issue in a bilateral setting is Shin et al. (2016), who finds that as importing countries adopt a more stringent IPR regime, the impact on the bilateral exports of the partner nation is negatively related to the level of technology of the exporting country. They argue that IPR acts as an export barrier to trade, especially discouraging exports from developing countries that are in a catching-up phase.

2018).

Less-advanced economies that lack the initial intellectual capital are typically used as outsourcing destinations for cost reduction purposes. Nonetheless, suppliers must customize inputs to meet the required standards, i.e. engage in relationship-specific investments. In a world of contractual incompleteness, the hold-up problem leads suppliers to underinvest. Strong rule of law institutions can partially resolve the friction through contract enforcement. This issue is more crucial the higher are the specific needs because the supplier's outside option is lower and underinvestment a bigger problem. Countries with better rule of law hence have a comparative advantage in the production of goods that use intensively inputs that require relationship-specific investments (Nunn, 2007).

By offering foreign firms a minimum level of protection against expropriation, IPRs can however be an important determinant of the import patterns of developing countries and encourage an inflow of technology-intensive goods (Ivus, 2011, 2015; Delgado et al., 2013). Improved IPRs also prompt multinational firms to transfer more intangible capital to their affiliates in host countries (Branstetter et al., 2006). The effect of IPRs on technology diffusion can also be due to responses in arm's length exports and unaffiliated licensing (Ivus et al., 2016, 2017; Lin and Lincoln, 2017). This phenomenon could build the intellectual capital required for domestic innovation over time and eventually spur industrial development (Branstetter et al., 2011).

To summarize, we expect a significant and positive effect of IPR institution on specialization in IP-intensive industries within OECD countries and we expect non-OECD countries with a better rule of law to export relatively more in industries that require higher levels of relationship-specific investments. At the same time, we expect IPR institution to be an important tool for non-OECD countries to attract foreign technologies by increasing imports in IP-intensive sectors from OECD countries.

We test these hypotheses by estimating the following equation:

$$\begin{aligned} log(exp_{i,c}) = \alpha + \beta_1(IPint_i * IPR_c) + \beta_2(h_i * log(H_c)) + \\ \beta_3(k_i * log(K_c)) + \beta_4 * (z_i * RL_c) + \delta_i + \delta_c + \epsilon_{i,c} \end{aligned} \tag{1.1}$$

where $log(exp_{i,c})$ is the natural log of export in industry i from country c to the rest of the world, IPR_c is a measure of the quality of protection of intangible capital in country c, $IPint_i$ is a proxy for the contribution of IPR to the production process of each industry i, z_i is a measure of the importance of relationship-specific investments in industry i; RL_c is a measure of the quality of contract enforcement in country c; H_c and K_c denote the endowments of skilled

labor and capital of country c, and h_i and k_i are the skill and capital intensities of production in industry i. Finally, the specification also incorporates country (δ_c) and industry (δ_i) fixed effects, that capture the overall level of trade and control for unobserved country and industry characteristics. We used robust standard errors as generally applied for this methodology. Throughout the paper, we will call IPR interaction the term $IPint_i * IPR_c$, Nunn's interaction $RL_c * z_i$ and skill interaction and capital interaction the other two products from equation (1.1) involving human capital and physical capital. In the baseline exercise, we use export values of 2014, human and capital stocks of 2012, rule of law and Park index for 2010. We lag the institutional variables by four years with respect to trade flows to reflect the fact that legal changes likely take time to influence technological activity.[6]

This specification, introduced by Rajan and Zingales (1998) and brought in the trade literature by Beck (2003) and Romalis (2004), is particularly appealing because it allows to control for country and industry fixed effects that explain the total volumes of trade, and to focus on the mix of exports in each country: these interaction terms capture the relative difference in the export values across industries and countries and for this reason provide a complete map of specializations across countries. As an example, assume that two countries are similar in every aspect but their IPR quality. A positive coefficient β_1 would be an evidence of comparative advantage because it suggests that countries with higher-quality IPR institution tend to export relatively more in IP-intensive industries. The same reasoning applies for the interpretation of the other coefficients.[7] We use the same mechanism to study the interaction of industry and country-specific characteristics on imports, i.e. whether better IPR protection in a country induces more imports in IP-intensive sectors. Finally, we explore also two alternative specifications, a bilateral framework and a panel set-up, to deepen the analysis and address new questions, as we will discuss in greater details in the next sections.

1.4 Data

A key variable that lies at the center of our analysis is the data for the contribution of IP at industry level. We obtain this measure, $IPint_i$ in equation (1.1), from the report "Intellectual property

[6] Results are not sensitive to the choice of lag, which is made as an initial attempt to mitigate reverse causality concerns (see the section dedicated to this issue for a more comprehensive analysis).

[7] The underlying idea is that for each industry the dependence on a country variable, either a stock or an institutional quality, is a technological feature and so it is constant across countries; country features that satisfy better the needs of specific industries offer a more suitable environment for efficient operation of those industries. As a consequence, countries specialize in industries whose production needs are best matched with their factor endowments and institutional strengths.

rights intensive industries and economic performance in the European Union, Industry-Level Analysis Report, October 2016 Second edition" provided by EUIPO (European Union Intellectual Property Office). The intellectual property rights considered in the European report are trademarks and patents applied at EUIPO, EPO (European Patent office) and CPVO (Community Plant Variety Office) during 2006-2010 and subsequently granted. The unit of analysis of the report is at industry level, as defined by NACE 4-digit revision 2 classification and it provides the number of IP issued for 1000 employees. We take this measure as the importance of IP to the production process of each industry. [8] Table 1.1 provides a descriptive summary of this variable at industry level and a list of the three most and least IP-intensive industries.

Table 1.1: Intellectual Property Rights statistics

Descriptive statistics

Variable	Mean	Std. Dev.	Min	Max
trademark	7.55	6.54	0.47	38.80
patent	3.30	9.98	0	109.74
sum	10.85	12.99	0.47	116.92

Top Industries	N of IP
Manufacture of power driven hand tools	116.92
Manufacture of instruments and appliances for measuring and testing	70.89
Manufacture of basic pharmaceutical products	66.38

Lowest Industries	N of IP
Manufacture of ready-mixed concrete	0.47
Manufacture of prepared meals and dishes	0.88
Processing and preserving of poultry meat	1.06

All other data are from standard sources. Other industry variables are obtained from US manufacturing database maintained by the National Bureau of Economic Research and US Census Bureau's Center for Economic Studies. These variables are updated up to 2011 and are classified under the NAICS 1997 system, that we converted to NACE 4-digits. [9] We define

[8] A frequent critic for the use of industry data in this setting is that it uses information on one country and assumes that industry characteristic is constant across all other countries with the argument that technology is a structural feature and hence production requires the same process regardless of its location. Even if the data we use on IP intensity is an average of all the EU countries (and so less prone to this critique), some caveats are worth mentioning. Our identification does not require that industries have exactly the same IP intensity levels in every country, but it does rely on the ranking of sectors remaining relatively stable across countries. We implement a sensitivity checks in Table A.1.1 of the Appendix by using a dummy to split industries into high and low IP-intensive ones with respect to the median value and by looking separately at patents and trademarks to measure IP-intensity.

[9] In order to convert all the industry variables according to the NACE 4-digits classification, we match NAICS 1997 to NAICS 2007 categories and then convert this system with NACE 4-digits through an official concordance

capital intensity as one minus the share of total compensation in value added in each industry, whereas skill intensity is given by the share of non-production workers relative to overall employment multiplied by the share of labor compensation in value added. Regarding z_i, Nunn's webpage directly provides the share of input that are relationship specific in each NAICS 1997 industry and, following the same procedure as previously described, we convert these data in NACE 4-digit classification. Throughout the paper, for each industry we consider the share of input that are neither reference priced nor sold in organized exchange as relationship-specific investment (Nunn, 2007).

Data on capital stocks and GDP per capita are from IMF and converted in 2011 US dollars; data on human capital stocks are from Penn tables Feenstra et al. (2015) and are defined as the average years of schooling for the population aged 25 or above. As a primary measure of rule of law, RL_c in equation (1.1), we use Kaufmann et al. (2009) to follow more closely Nunn (2007). It is a weighted average of a number of variables that measure individuals' perceptions of the effectiveness and predictability of the judiciary and the enforcement of contracts in each country. Since the previous variable starts from 2000, when we need older values of rule of law, we use an alternative commonly used proxy from Gwartney et al. (2008). Data on IP enforcement quality IPR_c are from Park (2008), an updated version of Ginarte and Park (1997) index, the most widely used proxy in the IPR literature. The index is updated every 5 years and ranges from 0 to 5.[10] In Table (1.2), we report the mean values and correlation between these variables. It is straightforward to see that the country level variables are highly correlated, but industry characteristics much less: the industry-country match can generate comparative advantage because institutional and endowment conditions affect production in different industries in alternative ways depending on characteristics of the industry.

Finally, trade flows disaggregated at HS12 6-digit level are provided by COMTRADE and available from 1989 to 2014; also in this case, data were converted to match NACE 4-digits system.[11] Overall, we have data for 82 countries, 33 OECD members and 49 non-OECD members, as

table provided by Eurostat. All the concordance between different versions of the NAICS classification are available at: https://www.census.gov/eos/www/naics/concordances/concordances.htm. Conversion from NAICS2007 to NACE is available from the Eurostat web page RAMON - Reference and Management of Nomenclatures. When the issue was many to many or one to many, to be more conservative, we have dropped that industry.

[10] It is the unweighted sum of five separate scores that can take value up to one and each of them consists of several binary conditions which, if satisfied, indicate a stronger level of protection in that category. The five variables include several conditions to account for the degree of: coverage (inventions that are patentable), membership in international treaties, duration of protection, absence of risks of forfeiting the patent rights (for example, due to compulsory licensing or revocation of patents), enforcement of patent rights in case of an infringement.

[11] We match this classification with NACE system through a concordance table provided by ISTAT (Italian statistical Office). Every time the cross-walk from HS to NACE is not unique, we exclude the trade flow in that industry, but the number of excluded HS industries remain negligible.

specified in the Appendix.

Table 1.2: Means and correlations of stocks and industry variables

Country Variables	**mean**		**correlations**		
IPR 2010	3.58	1.00			
Human capital	2.14	0.817	1.00		
Physical capital	4.10	0.763	0.775	1.00	
Rule of law	0.31	0.754	0.690	0.765	1.00
Industry Variables	**mean**		**correlations**		
IP int.	9.84	1.00			
Skill int.	0.81	0.031	1.00		
Cap. int.	0.72	0.189	-0.686	1.00	
Relat. Specific	0.47	0.160	0.552	-0.367	1.00

1.5 Empirical Results

1.5.1 Raw data Analysis

To get a broad picture of the importance of institutions for comparative advantage, we start with a preliminary analysis of raw data. We compute an industry i's share of total export in each country c and multiply it by the IP-intensity of the industry: this gives us the average IP-intensity of export by country. It is calculated as $\tilde{IP}_c = \sum \phi_{i,c} * IPint_i$, where $\phi_{i,c} = \frac{exp_{i,c}}{exp_c}$. This average IP-intensity of export is highly correlated with the IPR quality of the country, as highlighted by the significant and positive standardized beta coefficient in the first element of the first column of Table (1.3); if we do an equivalent exercise for rule of law and contract intensity, we also find a positive and significant relationship (second element of the first column). Hence, both institutions tend to matter for specialization and the structure of trade. More interestingly, however, if we split our sample between OECD and non-OECD countries, we find that the protection of intellectual capital is only decisive for advanced economies, whereas rule of law only matters for less developed countries. This initial evidence stresses the idea that countries at different stages of development, on average, have different production structures. OECD countries with better IPR institutions specialize and export in more innovation-oriented industries. On the other hand, the underlying mechanism that drives non-OECD countries' comparative advantages originates from contracting institutions that assure their full involvement in relation-specific investments.

Table 1.3: Average IP intensity of export and IPR protection level

	Whole sample	OECD	non-OECD
IPR	0.218**	0.367***	0.085
	(0.395)	(1.29)	(0.654)
Judicial Quality	0.240***	-0.287**	0.538***
	(0.280)	(0.025)	(0.019)
Number of obs.	82	33	49

The dependent variable is the average IP intensity of exports of each country in the first row and the average contract intensity of export in the second row. Standardized beta coefficients are reported, with robust standard errors in brackets. *** indicates significance at the 1 percent level.

1.5.2 Estimation results

The basic hypothesis we want to test is whether, other things equal, export volumes in IP-intensive sectors increase with the strength of IPR enforcement across countries. Table (1.4) shows our baseline regression. In the first column we include only our main interaction of interest, for which we have data on 231 industries and 82 countries.[12] The estimated coefficient for the IPR interaction is positive and statistically significant. In the second column, we also include the standard factors endowments, and our main variable of interest remains positive and significant, reinforcing the essential role of IPR protection in production and exports of technologically intensive goods. Column III replicates Nunn (2007) and is consistent with its findings. The result shows that contract enforcement is a determinant of comparative advantage and drives specialization in contract-intensive industries. The fourth column is our preferred baseline specification and also hints at the growing link between institutions and comparative advantage with respect to the classic factors of human and physical capital.[13] Countries with better IPR protection and rule of law export relatively more in industries highly intensive in IP and in industries with a relatively higher share of relationship specific investments, respectively. Parallel to Maskus and Yang (2018), this result confirms that the protection of IPR is an effective tool to increase innovation and R&D, thus leading to specialization in sectors in which IP play a substantial role in the production process. [14]

[12] Note that we exclude from the analysis missing observations and observations with trade value equal 0, which totals to 1466 observations. Considering positive exports implies that we implement an analysis conditional on a country exporting in an industry, and try to assess whether country characteristics explain the observed difference in trade performance across industries rather than the decision to enter and trade in an industry.

[13] The mitigating role of physical capital is also in line with related literature such as Levchenko (2007) or Maskus and Yang (2018).

[14] All four coefficients have very similar magnitudes in absolute terms, ranging from a 0.07 to 0.09 change in our dependent variable following a one standard deviation change in one of the interactions. Consider for example

Table 1.4: Determinants of Comparative Advantage: baseline specification

Variable	(I)	(II)	(III)	(IV)	(V) OECD	(VI) NON-OECD
IPR int.	0.0108***	0.0079***		0.0055***	0.0139***	0.0035
	(0.0018)	(0.0019)		(0.0019)	(0.0047)	(0.0034)
Skill int.		8.361***	3.797**	3.416*	11.87***	0.852
		(1.596)	(1.780)	(1.785)	(3.875)	(2.276)
Capital int.		-0.397**	-0.183	-0.293	-0.773	0.0718
		(0.196)	(0.215)	(0.220)	(0.549)	(0.315)
Nunn int.			0.673***	0.635***	-0.387*	1.026***
			(0.108)	(0.110)	(0.207)	(0.221)
Observations:	17476	10621	7893	7893	3317	4576
R-squared:	0.740	0.778	0.780	0.781	0.766	0.704
Country FE:	Yes	Yes	Yes	Yes	Yes	Yes
Industry FE:	Yes	Yes	Yes	Yes	Yes	Yes

The dependent variable is the natural log of exports in industry i by country c to all other countries. In all regressions, robust standard errors in brackets are reported. *, ** and *** indicate significance at the 10, 5 and 1 percent level.

To better understand the impact of IPR reforms or other determinants of trade, it is important to consider how institutions have differential consequences that depend on the environment in which they are is established (Maskus and Ridley, 2016; Shin et al., 2016; Campi and Dueñas, 2019). We now take a step further to conduct a comparative study to test whether the impact of institutions on the composition of trade, and therefore the source of comparative advantage, varies with country-specific characteristics.[15] Doing so allows us to understand the implications of the recent global improvements in IPR standards, and whether they have been beneficial in nurturing technological capability in developing countries. Specifically, we investigate whether there is a difference between the role played by IPR institution in determining the trade structure in innovation-oriented economies and in those with lagging technologies. We also evaluate whether contract enforcement has the same significance in R&D oriented developed countries, or if they instead have a more crucial meaning in attracting multinationals into creating valuable relationship-specific outsourcing partnerships in developing countries involved in the vertical provision and hence exports of such intermediate inputs.

that if Brazil increases its IPR (3.44) up to the level in Mexico (3.88), that is about half S.D of the variation that we have for the Park index, this would lead to a 2% increase of exports of Brazil in pulp manufacturing industry (which is at 25 lowest percentile of IP intensity) whereas it would lead to a 53% increase in the manufacturing of computer and peripheral equipment (at the 75 lowest percentile of IP intensity).

[15] Maskus and Yang (2018) introduce a triple interaction multiplying the baseline IPR interaction with an indicator dummy to show that the impact is stronger for richer countries, whereas our aim is to disentangle diverse mechanisms that drive specialization for different countries.

In columns V and VI we split the sample between OECD and non-OECD countries to account for differences in production structures, organizations, innovating capabilities and the stage of development. The results highlight how the basis of comparative advantage derives from different sources: OECD countries that are on average more developed and technologically advanced form their comparative advantage based on human capital and IPR protection level; non-OECD countries, with less advanced production processes that involve tangible assets, determine their specialization with property right protection and Nunn's channel of comparative advantage.

Production of IP-intensive goods are influenced by even smallest differences in IPR levels of OECD countries endowed with intellectual capital. This implies that the protection of intangible capital is an essential tool to stimulate innovation and increase the efficiency of producing R&D-intensive goods by preventing imitation. The result is in line with Qian (2007) that shows how IPR improvements foster innovation activities in the pharmaceutical sector conditional on a minimum level of development and human capital. As expected, also human capital endowment is an important driver of specialization within OECD countries compared to our complete sample in Table (1.4). The more OECD countries specialize in innovation and high-tech activities, the more they concentrate on IP-intensive production and outsource other tangible parts of the production process to non-OECD countries. This highlights the key role of contract enforcement (rule of law) in non-OECD countries as recipients of outsourcing activities. With a more specialized focus on tangible property and the production of outsourced intermediate inputs, the level of customization required in relationships and investment by suppliers takes primary importance. As argued by Nunn (2007), tangible property right protection is therefore an essential tool in inducing foreign firms to engage in business relationships in contract-intensive industries because it increases the efficiency of production by eliminating hold-up problems and sub-optimal investment by suppliers.

1.6 Robustness checks

In this section, we perform a host of alternative specification to provide robust evidence about the validity of the results presented in the previous section. We try to address possible concerns related to omitted variable bias, through augmenting the specification with additional controls and proposing a panel specification, and reverse causality, using a series of IPR reforms.

1.6.1 Additional Controls

Before interpreting our previous results as conclusive evidence of comparative advantage, we carry out a sensitivity analysis to address several potential concerns. An immediate issue that arises is the existence of other omitted determinants of comparative advantage not included in (1.1) that may be correlated with our main variable of interest. We implement a series of robustness checks to mitigate the possibility of the observed specialization in IP-intensive industries for OECD countries being driven by other industry features or for reasons unrelated to IPR quality. To deal with this, we control for a host of alternative determinants of trade flows that, if omitted, may bias the weight played by IPR institution in shaping the observed pattern of trade. Same reasoning applies to contract enforcement and subsequent specialization in contract-intensive industries.

In order to do so, we interact several industry characteristics with the log of income per capita to control for the possibility that, for reasons other than the protection of tangible and intangible capital, high income countries specialize production of certain industries. In particular, in columns 1 and 2 of Table (1.5) we include interactions of the log of income per capita with measures of the share of value-added of each industry and the TFP growth in the last thirty years in each industry. These two interactions allow for the possibility that richer countries have a comparative advantage in more lucrative and high value-added industries or in dynamic industries characterized by rapid technological progress. In column 3 and 4, we further include interactions of human and capital intensities with log of income per capita of the country, to control for the possibility that richer countries tend to specialize in industries that are more human or physical capital intensive.

In columns 5 and 6, we augment the specification by interacting IP-intensity and contract-intensity with the log of income per capita of the country to control for the possibility that richer countries tend to specialize in these industries merely because they are more developed and not specifically due to the institutional setting. In columns 7 and 8, aware of the possibility that our proxy of IP-intensity may be correlated with other (unspecified) industry characteristics, we include industry fixed effects interacted with the country's real per capita GDP. These interactions control for the possibility that richer countries tend to produce in industries whose (unknown) characteristics are correlated with IP or contract-intensity.

Overall, a pattern consistent with the results of Table (1.4) continues to emerge throughout all robustness checks and across different specifications.[16] Between developed countries, there

[16] In Appendix A.1.2, we provide a set of controls to assess the sensitivity of our main variable, IP intensity, to

are systematic effects on trade specialization depending on the stock of human capital and on the quality of IPR institution; as for developing countries, those with better rule of law export relatively more in industries that rely heavily on relationship-specific investments. Despite changes in the magnitude of our results, our main variables of interest remain significant, reinforcing the idea that specialization of production stems from different sources in developed and developing countries. As we will show in subsequent sections, the results are robust also to alternative specifications that allow to control for different levels of unobserved heterogeneity, such as the panel framework. These are rigorous specifications that help sweep out a great amount of additional variation that could generate omitted variable bias.

1.6.2 Panel Analysis

Then, we move to a panel set-up, which enables us to control for additional time-varying country characteristics and time dynamics. Our data span from 1999 to 2014 with 4 observations per industry-country at 5 years frequencies and estimate a specification similar to equation (1.1):

$$\begin{aligned} log(exp_{i,c,t}) = \alpha + \beta_1(IPint_i * IPR_{c,t}) + \beta_2(h_{i,t} * log(H_{c,t})) + \beta_3(k_{i,t} * log(K_{c,t})) \\ +\beta_4 * (z_i * RL_{c,t}) + \beta_5 * GDP_{c,t} + \delta_{i/c/t} + \epsilon_{i,c,t} \end{aligned} \tag{1.2}$$

Compared to the static framework, we control also for a time-variant country variable such as log of GDP per capita, $GDP_{c,t}$, that can explain changes in the overall volume of trade and level of development between countries over the years. Since sector trade is correlated within a country over time, we cluster the standard errors at industry-country level. IPR index is available at 5 years frequency, from 1995 to 2010. We restrict our attention to this time horizon because we want to focus on the post-TRIPS period. Again, we allow for some delay in the effect of an IPR policy change on the trade structure because we analyze trade flows four years after any update of the IPR index; same reasoning applies to contract enforcement.

This specification represents an effective robustness check for our cross-section analysis in section (1.5) by introducing also a time dimenstion in our analysis, with the inclusion of year fixed effect and also time-varying country specific variables. In this specification, the variation that we assess is within countries across industries and over time, net of industry-specific patterns and world-wide business cycle fluctuations.

The panel results are reported in Table (1.6). Also in this new set-up, IPR institutions

alternative definitions, and also the robustness of the baseline results to alternative ways of clustering.

Table 1.5: Robustness Checks

	Control I		Control II		Control III		Control IV	
Variable	OECD	non-OECD	OECD	non-OECD	OECD	non-OECD	OECD	non-OECD
IPR interaction:	0.014***	0.004	0.04***	0.001	0.009*	-0.003	0.009*	-0.003
	(0.005)	(0.003)	(0.005)	(0.003)	(0.005)	(0.004)	(0.005)	(0.004)
Skill interaction:	11.91***	0.671	12.06***	2.09	12.75***	2.67	12.82***	2.56
	(3.86)	(2.31)	(4.62)	(3.22)	(4.62)	(3.21)	(4.50)	(3.10)
Capital interaction:	-1.05*	0.03	-2.19**	-2.14***	-2.13**	-2.09***	-2.15**	-2.06***
	(0.57)	(0.33)	(0.87)	(0.54)	(0.87)	(0.54)	(0.85)	(0.54)
Nunn interaction:	-0.427*	1.028***	-0.467**	1.071***	-0.945**	1.529***	-0.941***	1.567***
	(0.222)	(0.226)	(0.234)	(0.234)	(0.332)	(0.273)	(0.327)	(0.265)
VA*log(GDP):	7.7e-06*	7.1e-07	5.04e-06	-1.23e-06	4.99e-06	-3.50e-07		
	(4.6e-06)	(1.8e-06)	(4.74e-06)	(1.91e-06)	(4.72e-06)	(1.91e-06)		
TFP*log(GDP):	-0.923	-0.356	-0.828	-0.324	-1.478	-0.142		
	(1.569)	(0.686)	(1.595)	(0.684)	(1.605)	(0.694)		
skill int*log(GDP):			4.480**	1.490	1.770	2.377		
			(2.602)	(1.469)	(2.798)	(1.498)		
cap. int*log(GDP):			3.017***	3.579***	2.228*	3.316***		
			(1.182)	(0.690)	(1.216)	(0.7)		
IPR int*log(GDP):					0.016***	0.005**		
					(0.005)	(0.003)		
z_i*log(GDP):					1.075	-0.754**		
					(0.683)	(0.241)		
Observations:	3317	4576	3,317	4576	3317	4576	3317	4576
R-squared:	0.767	0.704	0.767	0.706	0.768	0.707	0.778	0.721
Country FE:	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry FE:	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
log(GDP)*Industry FE:	No	No	No	No	No	No	Yes	Yes

The dependent variable is the natural log of exports in industry i by country c to all other countries. In all regressions, robust standard errors in brackets are reported. *, ** and *** indicate significance at the 10, 5 and 1 percent level.

Table 1.6: Panel exercise

Variable	OECD	NON-OECD
IPR interaction:	0.0124***	0.0024*
	(0.0028)	(0.0013)
Skill interaction:	0.184	0.752
	(0.513)	(0.606)
Capital interaction:	0.046	0.103
	(0.103)	(0.125)
Nunn interaction:	0.111	0.188***
	(0.097)	(0.062)
GDP:	0.597***	0.364***
	(0.203)	(0.120)
Observations:	8530	13494
R-squared:	0.768	0.704
Country FE:	Yes	Yes
Industry FE:	Yes	Yes
Time FE:	Yes	Yes
Country-Year:	No	No

The dependent variable is the natural log of exports in industry i by country c to all other countries in year t. The panel data have a 5-years frequency and time ranges from 1999 to 2014. In all regressions, standard errors are clustered at industry-country level. *, ** and *** indicate significance at the 10, 5 and 1 percent level.

are the main source of specialization for OECD countries: improvements of the protection of intellectual capital over time have systematically affected trade structure for developed countries, leading to more exports in IP-intensive sectors. We show once more that rule of law is a key determinant of comparative advantage in more contract-intensive industries for non-OECD countries. However, we can now observe that IPR institutions start to play a marginal role when we account for variations over time.

1.6.3 IPR Reforms

Another concern besides omitted variables that invites caution when interpreting the results is the possibility that causality runs from trade flows to IPR quality. If so, the previous results would be generated by countries that specialize in IP-intensive industries having greater incentives to develop and maintain an effective system to protect intellectual capital. Although previous research has pointed toward total trade volumes affecting the development of political, economic, and legal institutions (Acemoglu et al., 2005), the question whether comparative

advantage can also affect institutions is less touched upon in the literature.[17] As the variable of interest is not at the country level, e.g. GDP, but at the disaggregated industry level, it appears less likely that a single industry can affect the institutional quality at country level. Nunn (2007) and Costinot (2009) have dealt with the presence of endogeneity regarding rule of law institutions, and Ivus (2010), Delgado et al. (2013), and Maskus and Ridley (2016) argue that the TRIPS agreement has exogenously imposed new global standards of IPR protection. While considering post-TRIPS IPR levels as exogenous may seem adequate for developing countries, it seems less reasonable if the focus is on developed countries, which were the advocates of the agreement. We thus try to address the reverse causality issue regarding IPR institutions in a more rigorous manner, exploiting a series of IPR reforms both as an IV and in a diff-in-diff set-up.[18]

The literature on IPR has extensively used a series of reforms that changed drastically the legal systems surrounding the protection of Intellectual Properties. These events have been carefully analysed by Park (2008), who have studied the evolution of the legal systems across countries and identified specific episodes of significant changes in the legal framework protecting IP. These reforms have been subsequently used also, among others, by Ivus et al. (2017) and Ivus and Park (2019). For the purposes of this section, we move to an unbalanced panel setting in which we follow the export performance of each country in a given industry over the years. In the first exercise, following Branstetter et al. (2006), Manova (2008), Delgado et al. (2013) and several other contributions, we implement a generalized diff-in-diff approach to assess how IPR changes affect the pattern of trade for a country. We estimate the following regression:

$$\begin{aligned} log(exp_{i,c,t}) = \alpha + \beta_1 * reform_{c,t} + \beta_2(IPint_i * reform_{c,t}) + \beta_3(h_{i,t} * H_{c,t}) \\ +\beta_4(k_{i,t} * K_{c,t}) + \beta_5(z_i * RL_{c,t}) + \beta_6 * GDP_{c,t} + \gamma_i + \gamma_c + \gamma_t + \epsilon_{i,c,t} \end{aligned} \tag{1.3}$$

We consider now yearly observations, from 1989 to 2015; reform is a binary variable equal 1 in the year of reform and all years afterwards, 0 otherwise. Since it is a time varying country measure, we can include it and its effect is not washed out neither by the country nor year fixed effects.[19] Standard errors are clustered at industry-country level to allow for correlation

[17] An example for such mechanism is Do and Levchenko (2009), who show that comparative advantage affects financial development: country's specialization affects its demand for external financing, which, in turn, affects subsequent financial development.

[18] Recall that throughout our analysis we also lag the IPR interaction term by four years with respect to trade flows that we study.

[19] Note that Nunn's interaction term varies at lower frequency than the dependent variable and other explanatory variables. The results on our main coefficient of interest, IPR interaction, are unaffected by the inclusion of this

over time of an industry in a country.[20] The main effect of a legal reform - β_1 - is thus identified purely from the within-country variation over time. The coefficient of interest is β_2, which express the differential impact of IPR reforms across industries depending on their IP intensity. We expect the reform to have a stronger impact on the trade performance of IP intensive sectors compared to less IP intensive sectors, since the former are more directly affected by the consequences of the reform. In addition, the result should hold only for OECD countries. In this dynamic analysis, the identification of our main interaction of interest, and similarly for other interaction terms, comes from the combination of cross-countries and time-series variation in IPR protection status across countries and cross-industry variation in IP-intensity. The exercise, reported in Table 1.7, shows that deep and exogenous legal change in the protection of IP increased exports disproportionately more in sectors intensive in IP assets only in OECD countries, suggesting that pre-reform limited IPR protection was a constraint in those countries-industries and that higher quality standards does indeed trigger a systematic change in export patterns, becoming a source of comparative advantage. Conversely, in line with the results from Section 1.5, changes in the protection system of Intellectual Properties are not sufficient to trigger improvements in the export performance of developing countries.

As a second strategy, we exploit these major IPR reforms as instruments for IPR quality. What is essential for us is that these episodes can be used as instruments because they can be considered as exogenous events and provide a random variation in todays' IPR levels. To conduct this exercise, we consider a panel set-up at industry-country level from 1989 to 2014, with five-year intervals for each industry-country observation, a choice driven from the fact that Park index is updated only every 5 years. We introduce a dummy IPR reform equal to one if a reform in the country happened in the 5-year interval between any update of the Park index and afterwards. It is a time-varying country variable that explains part of the variation in trade volume across time.[21] Also in this case, we allow lags for changes in the IPR protection system to have some effects on the trade structure because trade flows of 1989 are regressed on IPR reform of 1985 and so on. In addition, since we are working with a dynamic specification, we include in the regression log of GDP. To control for serial correlation in the export performance of an industry in a given country, we cluster at country-industry level.[22]

variable.

[20] The main results are robust to the use of clusters at country level.

[21] For example, all the countries that experienced a reform between 1982 and 1985 will have the dummy IPR reform equal one from 1985 onward, all the countries that underwent a reform between 1986 and 1990 will have the dummy IPR reform equal to one from 1990 onward.

[22] The results are unaffected from the use of robust standrad errors or clustering at the level of the exporting country.

Table 1.7: IPR reforms

Variable	All Sample	OECD	NON-OECD
IPR reform:	-0.144***	-0.123***	-0.0770**
	(0.0196)	(0.0208)	(0.0308)
IPR interaction:	0.00428***	0.00690***	0.00104
	(0.00098)	(0.00108)	(0.00134)
Skill interaction:	1.562***	0.434	0.00124
	(0.349)	(0.422)	(0.469)
Capital interaction:	-0.123	0.0876	0.0492
	(0.0710)	(0.0845)	(0.0957)
Nunn Interaction:	0.375***	0.392***	0.218***
	(0.0292)	(0.0557)	(0.0333)
GDP:	0.748***	1.403***	0.540***
	(0.0761)	(0.155)	(0.0902)
Observations:	150074	61075	88999
R-squared:	0.782	0.756	0.686
Country FE:	Yes	Yes	Yes
Industry FE:	Yes	Yes	Yes
Year FE:	Yes	Yes	Yes

The dependent variable is the natural log of exports in industry i by country c to all other countries in year t. It is a panel exercise with yearly observations, running from 1989 to 2015. IPR reform is a dummy taking value equal one in countries that experienced a structural reform of the protection of intangible capital from the year of the reform afterwards. In all regressions, standard errors are clustered at country-industry level. *, ** and *** indicate significance at the 10, 5 and 1 percent level.

In the first stage, we regress our variable of interest, IPR interaction, on the dummy IPR reform interacted with IP intensity at industry level, including again the variables described in the baseline specification (equation 1.1) plus year fixed effect since we now have a time dimension available. We exploit reforms and the timing of reform to predict IPR protection values and the instrument is highly significant. We then use the predicted values from this first stage, $I\tilde{P}R$ interaction, as explanatory variable in the second stage. The IV is relevant, as highlighted by the statistics at the bottom of Table (1.8), which are all above the critical values. The test for weak instrument rejects the null hypothesis and so we can conclude that reforms are a strong instrument.[23] The results in Table (1.8) show that also the instrumental approach confirms our main hypothesis about the importance of IPR institutions as a key determinant of comparative advantage only for OECD countries. Overall, we believe that the emergence of a consistent and stable pattern mitigates the concerns on reverse causality.[24]

1.7 Technology Transfer

1.7.1 Imports

We have just shown that the recent improvements of IPR and the global harmonization of such standards have neither helped developing countries to boost their innovation and R&D nor had any impact in their export structure. Are there any other channels through which these reforms can bring trade-related beneficial consequences for these countries, for example by technology transfer through imports? In this section we look at the other direction of trade and assess how import patterns could be affected by IPR quality. We employ the same methodology represented by equation (1.1) that guarantees a systematic study of cross-industry and cross-country differences in the sensitivity of import patterns to IPR quality.

The baseline result of our complete sample of countries is presented in column 1 of Table (1.9).[25] The findings are in line with IPR literature because it stresses how good institutions are

[23] To implement the instrumental variable approach, the Stata routine ivregress 2sls has been applied, In addiion, the post-estimation commands first and weakivtest have been used to compute the statistics in the second part of Table (1.8).

[24] In addition, we have implemented two further exercises. In Appendix A.1.3 we use IPR quality level in 1960 to instrument today's IPR values. Also, we replicated both the exercises reported in this section using a series of reforms identified by Branstetter (2006), confirming the results obtained using the reforms identified by Park (2008). We decided to focus on the reforms identified in the latter source because it provides information on a much larger set of countries, allowing for a separate analysis between OECD and NON OECD countries, which is the main interest of the paper.

[25] We have replicated the robustness checks implemented for exports above also for import flows, and all the results are confirmed qualitatively. The result is confirmed also clustering at country level, industry level and two-way clustering at country and industry level.

Table 1.8: IV Estimation

Second Stage	All Sample	OECD	NON-OECD
IPR reform:	-0.0444	0.0399	-0.0927
	(0.0588)	(0.0850)	(0.0967)
$I\tilde{P}R$ interaction:	0.00195*	0.00544***	-0.000877
	(0.00106)	(0.00165)	(0.00166)
Skill interaction:	1.432*	0.0117	0.269
	(0.829)	(0.946)	(1.060)
Capital interaction:	-0.151	-0.0549	0.123
	(0.175)	(0.218)	(0.244)
Nunn interaction:	0.501***	0.567***	0.329***
	(0.0806)	(0.149)	(0.114)
GDP:	0.611	0.623	0.382
	(0.383)	(1.103)	(0.373)
Observations:	31483	13090	18393
R-squared:	0.777	0.749	0.683
Country FE:	Yes	Yes	Yes
Industry FE:	Yes	Yes	Yes
Year FE:	Yes	Yes	Yes
First stage:			
$IPR_c \cdot IPint_i$:	1.4522***	1.3028***	1.2115***
	(0.07525)	(0.13054)	(0.10122)
Weak IV test:	367.9	96.69	40.82

The dependent variable in the second stage is the natural log of exports in industry i from country c to all other countries. It is an unbalanced panel exercise with five observations for each country-industry, running from 1989 to 2014. The first stage dependent variable is the interaction term between IP intensity at industry level and IPR reform dummy. Then, we use the predicted values, $I\tilde{P}R$ interaction, in the second stage. The bottom part of the table reports the coefficient of the IV from the first stage, together with the values of the F-test resulting from the first stage and the endogeneity test. All explanatory variables in the second stage are also included in the first stage, but to conserve space we only report the first stage coefficients for the instrumental variable. In all regressions, standard errors are clustered at country-industry level. *, ** and *** indicate significance at the 10, 5 and 1 percent level. In the IV exercise with IPR reforms, there are six observations for each industry-country variable, from 1989 to 2014 with five years of frequency.

an effective tool to increase imports, especially in industries where IP is used intensively and the risk of imitation is high. In other words, IPR protection could stimulate technology transfer. Once we split our sample into developed and developing countries, we find the sharp result that IPR institutions only affect the import structure of non-OECD countries: multinationals are concerned about exporting IP-intensive goods to developing countries with weak IPR enforcement, and use the latter as a critical factor to decide whether to enter those market. To this regard, it seems that stricter enforcement of IPR across developing countries has been beneficial because it has led to the arrival of more technologies and intangible capital into these countries. Developed countries with already high standards, on the other hand, are not perceived as a threat, hence their differences in imports across industries are not driven by IPR quality.

These findings show that the quality of IPR institutions has opposite effects on the pattern of trade based on the stage of development: for developed countries it helps boost R&D, innovation and the production in IP-intensive industries, thus leading to more export in these sectors; for developing countries it attracts imports of IP-intensive goods. In other words, what we found to be a source of comparative advantage for OECD countries also explains an opposite trade pattern in non-OECD countries: IPR protection stimulates trade in IP-intensive industries from developed to developing countries, motivating the next section of our analysis on bilateral trade.

Table 1.9: IPR quality and the pattern of imports

Variable	Whole sample	OECD	NO-OECD
IPR interaction:	0.0040***	-0.0006	0.0090***
	(0.0008)	(0.0020)	(0.0017)
Skill interaction:	-0.991	5.020***	-1.378
	(0.824)	(1.869)	(1.043)
Capital interaction:	-0.355***	-0.034	-0.298**
	(0.103)	(0.287)	(0.146)
Nunn interaction:	0.330***	0.149	0.216*
	(0.057)	(0.108)	(0.123)
Observations:	8274	3332	4942
R-squared:	0.848	0.888	0.797
Country FE:	Yes	Yes	Yes
Industry FE:	Yes	Yes	Yes

The dependent variable is the natural log of import in industry i of country c from all other countries. In all regressions, robust standard errors in brackets are reported. *, ** and *** indicate significance at the 10, 5 and 1 percent level.

1.7.2 Bilateral Trade Flows

We now move to data on bilateral trade flows, which allows us to augment the baseline exercise with gravity controls and reassess our findings in a more demanding specification.[26] Perhaps more important for our purposes, a bilateral framework makes it possible to conduct a deeper comparative analysis by further breaking up trade patterns for different countries and exploiting information for both sides of trade. In particular, we compare the exporting behavior of an OECD country with respect to a non-OECD country, and take the analysis at a more disaggregate level by observing whether or not the importing country belongs to OECD.

We run the following regression:

$$\begin{aligned} log(exp_{i,c,p}) = \alpha + \beta_1(IPint_i * IPR_c) + \beta_2(IPint_i * IPR_c * IPR_p) \\ +\beta_3(h_i * log(H_c)) + \beta_4(k_i * log(K_c)) + \beta_5 * (RL_c * z_i) + \delta_i + \delta_c + \delta_p + \delta_{c,p} + \epsilon_{i,c,p} \end{aligned} \tag{1.4}$$

where now $log(exp_{i,c,p})$ represents the natural log of exports in industry i from country c to its partner p. In this new framework, we augment the baseline model (1.1) to include importer country fixed effects δ_p and also country pair-wise fixed effects $\delta_{c,p}$ that should control for all the standard gravity controls. We cluster standard errors at exporter-industry level to allow for correlated shocks in the export performance of specific industries across several destination markets, but clustering at exporter level leaves the results unchanged.

The bilateral analysis is key because we have shown that IPR institutions play a role on both sides of a trade transaction, as they affect the pattern of trade both for the origin and the destination country. It allows us to combine these predictions in a more comprehensive manner as we can directly assess the impact of IPR institution of an importing country on the export patterns of its trading partner. We would expect more trade in IP-intensive industries not only with higher IPR quality of the exporting country, but also that of the importing country since in some cases better institutions serve as an important tool to attract intangible capital. We therefore introduce a triple interaction $IPint_i * IPR_c * IPR_p$ in our specification that takes into account also the IPR strength in the importing country and tells us whether or not the effects of the baseline IPR interaction are stronger for higher quality IPR in the destination. Considering the findings in previous sections, we expect the triple interaction to not play a role when the importing country is an OECD country, whereas it should be an important determinant of trade flows when the importing country belongs to the non-OECD group.

The results are reported in Table (1.10) and are consistent with all our previous findings, con-

[26] See Chor (2010) and Cai and Stoyanov (2016) for a bilateral set-up of our original baseline framework.

trolling also for importer country fixed effects and pair-wise country fixed effects. Our main interest lies in the sign of the triple interaction, to highlight the effect of IPR quality of the importing country on the export pattern of other countries. As expected from the aggregate import analysis, the composition of imports is affected by IPR policy in a developing country because multinational firms, particularly technology-oriented ones, require certainty with regards to the protection of their intangible capital before exporting to that market. This is especially true when flows to a developing country originate from a developed country as these transactions on average involve a higher content of technology, the stronger is the IPR regime in the exporting country. Nevertheless, importing country IPR also shifts the balance of trade between non-OECD countries toward more IP-intensive transactions. As expected, the triple interaction terms in which the importing country is a developed nation are not different from zero as entering these markets is not perceived as a threat for foreign firms due to strong IPR. Our analysis has shed light on the positive effects of IPR improvements on trade in IP-intensive industries both for developed and developing countries, in one case affecting export patterns and in the other through imports.

Table 1.10: Bilateral Trade Flow analysis

Variable	(I) O-O	(II) O-NO	(III) NO-O	(IV) NO-NO
IPRinteraction:	0.0236***	0.0211***	0.0077	0.0043
	(0.005)	(0.004)	(0.004)	(0.0034)
Triple interaction:	0.0003	0.0010***	0.0005	0.0011**
	(0.0002)	(0.0002)	(0.0006)	(0.0004)
Skill interaction:	11.80**	12.99**	2.96*	3.76**
	(4.194)	(3.349)	(1.664)	(1.665)
Capital interaction:	0.162	-0.285	1.073***	0.589**
	(0.561)	(0.500)	(0.300)	(0.263)
Nunn interaction:	-0.444**	0.135	1.139***	0.726***
	(0.202)	(0.165)	(0.206)	(0.185)
Observations:	67641	83040	44237	47271
R-squared:	0.650	0.580	0.507	0.469
Exporting Country FE:	Yes	Yes	Yes	Yes
Importing Country FE:	Yes	Yes	Yes	Yes
Pair-wise Country FE:	Yes	Yes	Yes	Yes
Industry FE:	Yes	Yes	Yes	Yes

The dependent variable is the natural log of export in industry i from country c to country i. In all regressions, standard errors are clustered at industry-exporter country level and are reported in brackets. *, ** and *** indicate significance at the 10, 5 and 1 percent level. A constant term is included but not reported. Each column refers to a different sample, identified in the first row. O refers to OECD countries, NO to non-OECD; the first letter(s) identifies the exporting country, the second letter(s) the importing country.

1.8 Conclusion

Recent contributions in trade literature have emphasized the role of institutions as a source of comparative advantage. In particular, IPR protection and rule of law have been shown to systematically affect the patterns of trade. We provide an empirical assessment of how these different legal institutions shape the patterns of specialization depending on the level of economic development. We split the sample to perform a parallel analysis for OECD and non-OECD countries. We find that in OECD countries better IPR institutions drive exports in IP-intensive industries, whereas rule of law is a determinant of exports in institutionally dependent industries for non-OECD countries. This finding is consistent with the evidence that developed countries possess the initial intellectual capital necessary to engage in innovation activities; on the contrary, rule of law in developing countries that predominantly host foreign outsourced activities attract contracts for the production of relationship-specific inputs that are exported upon completion. After a preliminary cross-sectional analysis and related robustness checks, we further test the validity of our results using IPR reforms both as an instrumental variable and in a difference-in-difference framework. In addition, the results are confirmed using a panel set-up, which allows to control for additional dimensions of unobserved heterogeneity. Implementing a symmetric framework to examine import flows reveals a different potential role for IPR institutions in developing countries. The findings provide evidence that better IPR institutions allow non-OECD countries to attract the technology embodied in IP-intensive goods by protecting foreign firms' intangible assets. Given that our study stresses the importance of IPRs for both export and import patterns, we supplement the predictions with a bilateral trade setting and reveal a complementarity between the role of IPRs in determining OECD exports and non-OECD imports of technology-intensive goods. Domestic IPRs lead OECD countries to specialize in IP-intensive industries and destination IPRs direct the trade of these goods towards non-OECD locations with strong IPR institutions. Progresses made in enhancing the IPR regime could thus be a driver of technology diffusion to developing countries as a first episode of specialization in IP-intensive sectors. An avenue of future research is to investigate whether with time this could eventually lead to a reversal in their source of comparative advantage, making IPR reform a relevant institutional feature to induce domestic innovation.

Chapter 2

BG and Knowledge flows[1]

Abstract

In this paper, we study how subsidiaries of Business Groups (BGs) interact between each other. In particular, the paper highlights productivity gains that affiliates enjoy from intangible assets developed by other firms within the same group. The analysis shows that there are two key elements to consider in order to understand the interactions between subsidiaries: the hierarchical links between them and the intangible asset rather than on other firm's level characteristics. These key steps are crucial to highlight knowledge spillovers within the boundaries of BGs and understand the direction that intangible asset follows in groups. Interestingly, we show that within Business Groups knowledge flows upwards, i.e. subsidiaries in lower layers share their knowledge to subsidiaries in upper layers. Taking into account other firm-level variables or other ways to unfold the structure of the group, lead to the puzzling finding already discussed by previous literature about the lack of interactions between subsidiaries.

[1]This paper is a joint work with Tommaso Sonno. We are grateful to Carlo Altomonte, Marco Grazzi, Alireza Naghavi, Gianmarco Ottaviano, Vincenzo Scrutinio, Alessandro Sforza and Francesco Venturini for helpful comments. This paper also benefited from participants' comments at the Unibo DSE internal seminar.
Andrea Greppi, University of Bologna, Department of Economics. E-mail: andrea.greppi2@unibo.it
Tommaso Sonno: University of Bologna, Department of Economics, E-mail: tommaso.sonno@gmail.com

2.1 Introduction

The biggest and most valuable firms derive a significant market share from their intangible asset, as the value creation has generally shifted from capital to knowledge (Crouzet and Eberly, 2019). For example, Citibank, one of the US big four banks, employs more programmers than Microsoft. The development of a software - key element of intangible asset - for online banking has provided customers with 24/7 financial services, massively reducing labor cost in retail banking and leading to important productivity and revenue gains for the bank. In fact, it has been well established the link between R&D expenditure or managerial practices, two other important components of intangible asset, and firm's productivity (Hall et al. (2005); Bloom et al. (2016)).

At the same time, the biggest corporations in the world are composed by sets of legally dependent firms, Business Groups (BGs)[2], and the pace of M&A activity has been further increasing in recent years. The literature has put forward several reasons that drive firms' choice about integration. They can be related to the choice between "make it or buy it", in which firms optimally decide whether to integrate intermediate inputs' producers or outsource those goods (Antràs and Helpman, 2004). M&A may reflect companies' need to reorganize or enlarge their factors of production in order to remain competitive through economies of scale and scope. As the importance of intangible components in the production process is more and more important, also technology-related reasons explaining the surge in M&A activity have been investigated. In increasingly complex and uncertain technological environments, innovation becomes a critical source of strategic competitive advantage (Cassiman and Golovko, 2011) and firms willing to improve their activities may wish to acquire companies to benefit from their technological abilities.[3]

Thus, there might be substantial knowledge exchanges that happen within the boundaries of a group. In this respect, intangible asset is characterized by the 4-S features: sunkness, scalability, spillovers and synergies (Haskel and Westlake, 2017). These specific characteristics of intangible asset are likely to make these investments more valuable within the boundaries

[2] According to Fortune 500 (https://fortune.com/fortune500/), the world's largest businesses by consolidated revenue, as well as the top 100 multinational enterprises (as listed by UNCTAD) are all organized as BGs. These firms not only dominate the economic and trade activities, but also the innovation activity, being reliable for about 90% of innovation investments in the US (National Science Board 2014) and equally high shares of intangible investments.

[3] For example, recent evidence has emphasized that integration is a tool to leverage knowledge on additional production processes and to create synergies (Atalay et al., 2014). From another perspective, Sevilir and Tian (2012) show that acquirers increase their innovation outcomes following M&As, consistent with the view of Holmstrom and Roberts (1998) that many acquisition transactions are made to "source innovation".

of a group. For example, innovations developed by each of these firms may have an impact on other affiliates, as they can exploit synergies in the know-how and economies of scope from the competencies that each firm holds. Therefore, understanding how knowledge is organized in these groups and exchanged across subsidiaries, given the importance of BGs to the economic activity, their contribution to the creation of knowledge and the importance of knowledge in determining the success of firms, is an essential step to comprehend the performance of such entities.

In this paper, we want to understand the organization of knowledge in BGs, how it depends on the hierarchical organization of the group and how subsidiaries share their knowledge between themselves. Since interactions between affiliates crucially depend on their hierarchical links (Altomonte et al., 2021b), we bring the organizational structure at the center of our analysis, in the spirit of the within-firm setting (Caliendo et al., 2020). Specifically, does the allocation of knowledge across subsidiaries in a group exhibit systematic regularities related to the hierarchical position? If at all, how do subsidiaries share knowledge? Is it really the case, as suggested by previous findings (Bilir and Morales (2020); Belenzon and Berkovitz (2010)), that to understand the productivity performance of affiliates is necessary to look at the firm itself and the HQ, completely neglecting other subsidiaries?

To begin with, we provide a descriptive analysis to uncover patterns in the allocation of knowledge linked to the hierarchical position of affiliates in the group. We look within each BG and compare subsidiaries placed in different hierarchical levels and we show that firms that are farther down in the hierarchical structure have less and less intangible endowment, i.e., BGs exhibit a hierarchical organization of knowledge. Second, to understand how knowledge is shared in BGs, for each firm we decompose the intangible asset developed by other subsidiaries of the group on a hierarchical basis (knowledge developed by firms in upper/same/lower layers) and we explain the productivity of each subsidiary with the intangible asset coming from these different parts of the group. We find a robust pattern suggesting that knowledge flows upward, i.e., intangible asset developed by subsidiaries at the bottom of the group systematically provide efficiency gains to firms in upper layers.

To address these questions, we exploit a very rich database obtained combining two Bureau van Dijk datasets, Orbis and Historical Ownership. The former source provides financial variables at firm level, while the latter offers information on the ownership of all firms. The baseline information on the ownership links given by the data provider are processed applying a novel algorithm developed by Sonno (2020) that constructs the worldwide ownership link for all

firms. For the purpose of this paper, we select for each year between 2007 and 2014, the 1000 biggest BGs in terms of number of subsidiaries from the US and the 1000 from Europe and we follow these groups and their global subsidiaries over this time span.

With our analysis, we provide evidence that the two milestones through which we are able to open the boundaries of a BG and uncover systematic interactions therein are related to intangible asset and hierarchical organization. Only taking into account jointly how subsidiaries are hierarchically linked and the intangible asset of these firms, we are able to retrieve robust evidence of interactions between subsidiaries. The idea that the key way to read the behavior of a BG comes from the hierarchical organization and allocation of knowledge has been well emphasized by the literature that studies the within-firm organization as knowledge hierarchies (Garicano and Rossi-Hansberg (2006), Caliendo et al. (2020)). They show theoretically how a firm is organized in hierarchical levels based on knowledge skills and empirically how the internal hierarchical organization has real consequences on firms' productivity and profitability. We change the unit of analysis and shift our attention to within BGs between subsidiaries interactions to show how the hierarchical organization is key to understand the performance these firms.

A seminal paper in our analysis is Altomonte et al. (2021b), the first source to put the emphasis on the idea that BGs are shaped for an efficient management of knowledge and the hierarchical organization is designed to transmit more easily knowledge. As in Garicano (2000), the hierarchical structure allows to use knowledge efficiently, reducing the communication costs and facilitating supervision activities between layers. We provide a systematic study of the distribution of knowledge in BGs that confirms their theoretical foundation since we highlight a hierarchical pattern in the allocation of knowledge. In our analysis, we highlight a hierarchical systematic pattern in the allocation of knowledge, providing evidence on their theoretical foundation. As expected, the huge bulk of knowledge comes from the headquarter, that on average holds 80% of the overall intangible asset of a group. Then, as we move along the hierarchical organization, there is a systematic reduction in the intangible holdings of subsidiaries. The decline from layer to layer, even if small in absolute terms given the high concentration of intangibles at the level of the headquarter, is economically meaningful when comparing it to the average intangible asset of subsidiaries.

Then, we focus on knowledge interactions between subsidiaries. An attempt to address this question has been done by Bilir and Morales (2020), whose main finding is that US headquarters' R&D has a significant effect on foreign affiliates' productivity. On top of this main

effect, they also pool all the R&D done by the rest of the group in an additional control variable, that does not seem to be significant. This result is consistent with Belenzon and Berkovitz (2010), who provide evidence that affiliates of a BG are more likely to innovate with respect to standalone firms but do not find any evidence of knowledge spillovers between affiliates. This is a puzzling result because it gives a picture of BGs in which subsidiaries do not communicate between them and so, in order to understand the performance of each affiliate, all we have to do is to look at the HQ and the firm itself, neglecting other subsidiaries.

Motivated by Altomonte et al. (2021b), who relate the hierarchical structure of BGs to the organization of knowledge, and Caliendo et al. (2020), who show the consequences of within firm organization, we unfold interactions between subsidiaries based on hierarchical links. In particular, we unpack the intangible asset of the rest of the group into that developed by firms in upper, horizontal and bottom layers respectively. Providing knowledge the flexibility to follow different path in a group, we uncover interactions that where hidden in the aggreagte analysis. We show a systematic upward flow of knowledge according to which the intangible asset developed by bottom firms leads to TFP improvements of firms in upper layers. At first glance, the result can be surprising because usually the literature considers affiliates as "pure recipients" of technology. Nevertheless, already from Arrow (1975), this view has been challenged by different perspectives of integration as a way to "source innovation" (Holmstrom and Roberts, 1998). For example, rather than implementing R&D and develop productivity enhancing activity, firms can directly buy subsidiaries to source these competencies (Phillips and Zhdanov, 2013). Our findings, highlighting an upward flow of knowledge, are fully consistent with this view and go one step further, showing how subsidiaries within BGs interact and leverage their knowledge between themselves.

As several policy reports and papers stress, the largest BGs are always more involved in the M&A activities (Gautier and Lamesch, 2021), especially targeting start-ups.[4] Our results are consistent with a business development in which BGs acquire small and young entities, with promising new technologies not developed yet (hence with small absolute values of intangible asset), that are then developed and exploited by other subsidiaries. This conjecture would add nuance to the debate triggered by Cunningham et al. (2021). They show that incumbent firms acquire innovative targets exclusively to discontinue the target's innovation projects, so to preempt future competition. While we do not provide evidence of killer acquisition nor that incumbent disrupt their target's projects, we show that incumbents benefit from target's

[4] https://www.ft.com/content/e2e34de1-c21b-4963-91e3-12dff5c69ba4

technologies.

A key caveat to emphasize is that our exercise highlights purely descriptive correlations rather than causal interactions because the organizational choice of a BG is a strategic choice done by the HQ and it is likely driven exactly by the possibility to enjoy interactions and synergies between subsidiaries (Altomonte et al., 2021b). So, we are aware that our exercise only tells us how BGs behave and become more efficient. Nevertheless, we want to be sure that the highlighted patterns are indeed knowledge flows and are the result of unfolding the knowledge of the group through hierarchical lenses. To this regard, we implement several robustness checks and exercises to provide convincing evidence that these are indeed the relevant dimensions and measures to consider. Overall, we confirm that the keystone to understand how subsidiaries interact is to look at the knowledge developed by the rest of the group through the lens of the hierarchical organization. Unpacking the intangible developed by the group according to different criteria or studying interactions of different variables according to the hierarchical structure does not uncover alternative patterns of knowledge sharing, reinforcing the puzzle about the "lack" of interactions within the boundaries of a BG.

We hope our research can contribute to several strands of literature. Firstly, we add to Bilir and Morales (2020) highlighting additional knowledge sharing that are important to determine affiliates' performance. Since the organizational structure is endogenous (Altomonte et al., 2021b), it should be considered in any analysis on the behavior of affiliates. Our study also refers to the literature on within firm structure (Garicano (2000), Garicano and Rossi-Hansberg (2006), Caliendo et al. (2020)), applying the predictions about the real consequences of hierarchical organization also to internal organization of BGs across subsidiaries.

Our result is exactly in line with previous findings that look at the effect of M&A on acquirers. For example, Phillips and Zhdanov (2013) show that large firms optimally may decide to purchase smaller innovative firms and conduct less R&D themselves. Also, Bena and Li (2014) find that acquirers are low R&D intensive but have large portfolio of innovation while target firms are very active in innovation activities but did not convert yet their R&D expenses into patents at the time of integration. Finally, Sevilir and Tian (2012) provide evidence on a positive association between M&As and acquirers' post-merger innovation outcomes, suggesting knowledge flows from the acquired firm to the acquirer, in line with the upward flow we find.

Finally, in the last two decades there have been relevant structural transformations in the economy, possibly all related to the surge of intangible technologies. Mark-ups for the top decile of global firms have been soaring (De Loecker et al. (2020); Calligaris et al. (2018)), concentration

has increased significantly (Bajgar et al., 2019), productivity for firms at the top of the distribution has increased steadily while for the laggard firms has stagneted (Berlingieri et al., 2020), as a result of much lower technology diffusion across firms (Andrews et al., 2015). Overall, our evidence calls for the importance to consider BGs as a unique entity and understand better the interactions that happen within those boundaries, as they may significantly affect these market trends. For example, Bajgar et al. (2019) are the first source that provides evidence of a substaintial increase in market concentration also in Europe because they change the unit of analysis and look at BGs rather than single firms. It is therefore key to open these black box and understand the mechanisms that drive the success of these entities. Our paper is a preliminary attempt to describe some interactions that happen therein and that are potential drivers of BGs performance.

The paper is organized as follows. Section 2.2 introduces the data and the main variables used in the analysis and section 2.3 provides some motivating descriptive statistics. Section 2.4 explains and carries out the main empirical exercises of the paper, while section 2.5 checks the robustness of the main results of the paper to alternative channels and mechanisms. Finally, section 2.6 concludes.

2.2 Data and Measurement

The empirical analysis requires three key ingredients: ownership structures of BGs, firm level measures of intangible asset and productivity. We now discuss each of these elements.

Data Sources and Sample Definition The main data used in the paper are obtained combining two products provided by Bureau van Dijk (BVD). The first key component is Historical Ownership Orbis, that enables us to reconstruct the hierarchical structure of business groups. This database provides the panel information, for each company, of all shareholders and global ultimate owner. Starting from these data, the algorithm developed by Sonno (2020) retrieves the network of ownership for each business group, relying on the definition of direct or indirect majority ($> 50.01\%$) of the voting rights provided by Bureau Van Djik and consistent with the international standards for multinational corporations (UNCTAD, 2009). The algorithm, based on the ownership links, derives the hierarchical structure of BGs, by ascending the ownership structure. For further details and a deeper description of the methodology, see Sonno (2020). For the purpose of this paper, we select for each year between 2007 and 2014, the 1000 biggest BGs in terms of number of subsidiaries from the US and the 1000 from Europe and we follow all these groups over this time span. In total, we follow the ownership structure of 4,576 BGs

and 403,522 subsidiaries. As described in Altomonte et al. (2021b), BGs usually are organized as inverted pyramidal structures, meaning that upper hierarchical layers are more densely populated of subsidiaries with respect to the lower ones. At the same time, most of BGs, even the largest - as the ones we decided to focus on in our analysis-, do not have deep structures and on average have 3 layers. For this reason, we unify all levels higher than 6 under this category.[5] We complement ownership information with subsidiary level annual balance sheet and income statements sourced from Orbis, a commercial dataset widely used nowadays for academic research. A number of steps are required to make the dataset suitable for economic analysis, including ensuring comparability of nominal values across years and countries (by deflating with industry-level PPP) and extensive cleaning and filtering to net out the influence of measurement error and extreme values in the analysis. We follow closely the procedure suggested by Kalemli-Ozcan et al. (2015) and OECD standard methodology (Gal, 2013).

Key variables. The main explanatory variable of interest in our analysis is a measure of knowledge of each firm. The concept of knowledge is very broad and it has been expanding over the years. While the literature initially mainly focused on R&D, always a larger set of expenditures and investments for non-physical goods are important to determine the success of firms and are a key driver of economic growth, such as data, proprietary software and human and organizational capital (Andrews et al., 2015). A key characteristic of intangible asset is that this kind of capital needs high initial investments, often sunk because are very targeted to the needs of the firm, but then has very little reproduction costs. Consequently, as several researches have already highlighted, in recent years the cost structure of firms has shifted from marginal to fixed costs (De Ridder, 2019). We follow Altomonte et al. (2021a) and measure intangibles as total firm expenditure on fixed costs, which are defined as net revenues minus operating profits, both of which are directly available from income statements. The results are fully robust to the use of intangible asset directly provided from firms' balance sheet.[6]
The baseline dependent variable of the paper is total factor productivity (TFP). The parameters of the value added based Cobb-Douglas production function are estimated econometrically at the firm-level using the Ackerberg et al. (2015) control function approach. This is a two-stage

[5] As it will be clearer afterwards, this choice is necessary in order to have enough variation in our analysis and a larger sample available. In the robustness checks that we implement, we test this somehow arbitrary choice and all the results remain qualitatively unchanged to alternative thresholds.

[6] This measure, even if it has already been used in academic research (Crouzet and Eberly, 2019), is often considered to be mismeasured due to accounting standards. Broadly speaking, accounting rules treat intangibles as assets if they are purchased and as expenses if they are internally generated. While exceptions to this rule exist, they tend to be rare. For example, internally generated software or R&D spending can be treated as asset investment under special circumstances, essentially when such spending is on a proven process, such as the last development stages of an already-proven R&D project or software tool.

estimation in which all parameters are obtained in the second stage. We relate value added to number of employees and real capital stock as inputs and materials as the proxy variable. An important caveat to bear in mind is that Orbis contains variables in nominal values and no additional separate information on firm-specific prices and quantities are available. Even though we deflate these measures by country-industry-year level deflators (at the two-digit detail), differences in measured (revenue) productivity across firms within a given industry may still reflect both differences in technology as well as differences in market power.[7] Trying to address some of the well-known challenges in estimating production functions when only monetary variables are available, we augment the production function including also year and country fixed effect and we estimate separately the production function for each two-digit industry.

2.3 Preliminary Evidence

The way to organize the BGs is a key choice done by the HQ, not only in terms of number of subsidiaries but also relative to their position in the group. Thus, it is important to understand how the organization of BGs is linked to some observable characteristics of the affiliates, with a particular emphasis on the role played by the knowledge of each firm, to improve our understanding of the behavior of these entities. To this regard, we want to provide, through a descriptive exercise, a detailed analysis of the distribution of knowledge in BGs to understand whether there are systematic patterns in the allocation of intangibles along the boundaries of the group.

As reported in Table 2.1, the bulk of knowledge is concentrated at the level of the headquarter, who is accountable for almost 80% of the group's investments. The aim of our research is to understand, in a more robust way, if the high level of heterogeneity that intangible asset can get across firms hides more systematic dynamics within each group and across layers. Consistently with the idea of Altomonte et al. (2021b) and the findings of Atalay et al. (2014) and Ramondo et al. (2016), we expect a hierarchical allocation of knowledge within a BG: firms in upper layers have systematically more knowledge than firms in lower layers. Table 2.1 suggests a pattern that is consistent with this prior because it seems to emerge a hierarchical pattern in the amount of intangible asset and a decaying dynamic jumping from upper to lower layers.

[7] Note that each variable has been deflated by specific country-industry (two-digit) deflators. Specific deflators are available for: value added, gross output, capital goods and material expenditures. When the deflator was not available at two-digit level, we used the same deflator at higher aggregation level.

Table 2.1: Summary Statistics

layer	N firms	N firms w Int	mean Int
0	15,391	9,511	1,342
1	157,824	46,555	58
2	146,561	47,202	37
3	104,887	33,920	27
4	55,627	16,770	30
5	26,884	8,304	28
6 +	29,581	10,011	20
Total	533,846	172,273	111

The variable Int stands for Intangible asset. Values are in thousands of Euros at 2010 price level.

To provide evidence of more systematic and robust dynamics on the allocation of knowledge within BGs, we regress the knowledge associated to each affiliate on the position of the same subsidiary on the hierarchy, controlling for parent fixed effect, industry and country fixed effect (of each affiliate):

$$log(Int_{f,HQ,c,i,t}) = \sum_{l \in L} \beta_l * layer_{f,HQ,t} + FE_{HQ} + FE_c + FE_i + FE_t + \epsilon_{f,HQ,c,i,t} \tag{2.1}$$

where HQ, f, c, i denote respectively firm, parent company, country and industry in which the subsidiary is located and active, and l indicates the hierarchical level at which the affiliate is placed (L is the maximum hierarchical level of the BG). FE_{HQ} is the headquarter fixed effect, FE_i and FE_C identify fixed effects for the country in which the affiliate is located and the industry in which the firm is active. We employ a demanding set of fixed effects that capture significant differences in knowledge investments across firms that are common to industry, country and year dynamics, and that explain a relevant difference in endowments across firms. More important, we look within a BG, to control for important unobservable characteristics of the group and common to all subsidiaries. The idea is to look at variation of knowledge across firms within the same BG and explain it with the layer to which each affiliate belongs: according to the theory of Altomonte et al. (2021b), we should expect the magnitude of the coefficients to be decreasing as layers increase.

The results are reported in a table and, for the ease of interpretation, are also represented graphically. Consistently with the prediction, we observe a declining pattern as we move farther away from the headquarter, and a significant difference (overall and not always layer by

Table 2.2: Allocation of Intangible Asset across layers within a BG

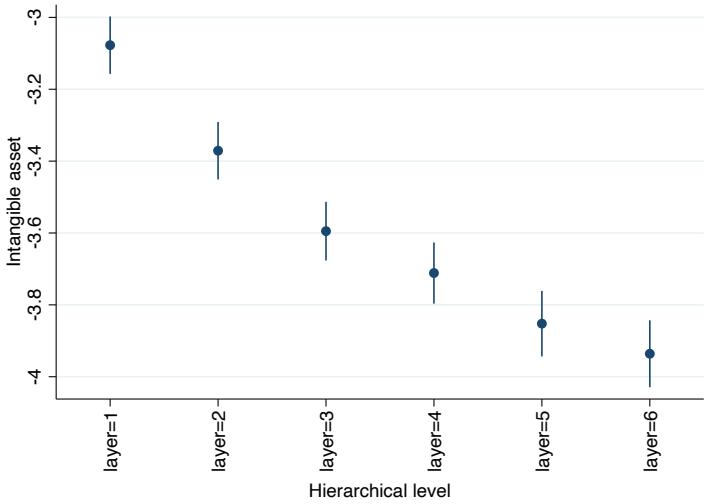
	log(Int)
layer 1	−3.077^a
	(0.0408)
layer 2	−3.371^a
	(0.0409)
layer 3	−3.595^a
	(0.0416)
layer 4	−3.711^a
	(0.0434)
layer 5	−3.852^a
	(0.0465)
layer 6 +	−3.936^a
	(0.0475)
Observations	165,886
R- squared	0.592
Subs Country FE	Yes
Industry FE	Yes
Year FE	Yes
Parent FE	Yes

Note: The dependent variable is log of intangible asset expressed in thousands of Euros at 2010 price level. Note that level 0 is taken as comparison group and that layer 6 + refers to all subsidiaries placed in layer 6 or above. Robust standard errors are reported in brackets. a, b and c indicate significance at the 10, 5 and 1 percent level.

layer) in the amount of knowledge assigned to each layer. It tells us that firms in upper layers of a BG tend to have more knowledge than firms in the bottom layers.

In line with the summary statistics from table 2.1, the huge bulk of knowledge stands on the shoulders of the HQ. The omitted category is layer 0 and these coefficients tell us that, on average, headquarters have 95% more intangible asset than firms in the first layer and 99% more than firms in the last layer. This means that the relative difference in intangible endowment between firms in different layer with respect to the asset of the headquarter are very small (4% maximum, equivalent to 50k) but still are relevant with respect to the mean values of those firms' asset (58-20k). For this reason, the regularities in the knowledge endowment across

Figure 2.1: Intangible Assets by Hierarchical layers



layers are economically relevant.[8]

To explore the sensitivity of the results to alternative ways of clustering, we replicate the analysis on a more synthetic specification in which we divide our sample of subsidiaries into two categories: those located on a hierarchical level higher than the mean layer of their BG, and those closer to the parent (Table 2.3). In particular, for each year-BG, we classify each subsidiary with a dummy variable equal to one if it is located at a level lower than the mean hierarchical level of the BG.[9] The dummy, that identifies subsidiaries closer to the headquarter, is always positive and statistically significant, suggesting that subsidiaries at higher hierarchical levels tend to have higher levels of intangible asset.

Overall, the dynamic is qualitatively the same throughout all the exercises: firms in upper layers have more knowledge than firms in lower layers and the allocation of knowledge within a BG across layers follows a hierarchical pattern. The jump from an upper to a lower layer is not systematically associated to a reduction in the intangible asset but overall, there is a clear declining pattern. This empirical regularity is consistent with the idea of BGs as knowledge hierarchies (Altomonte et al., 2021b) and provides a preliminary description on where knowledge is located in these black boxes that had been BGs until now. In addition, this evidence emphasizes one additional consideration that has never been put at the center of the analysis so far: integration is not a unique phenomenon and an important source of

[8] In Appendix A.2.1, we replicate the previous specification trying different alternative robustness checks to assess the relevance and stability of the highlighted pattern of intangible asset. All the checks confirm the hierarchical patterns in the allocation of knowledge.

[9] The results remain qualitatively the same if we consider the median layer of subsidiaries within a group rather than the mean value.

Table 2.3: Allocation of Intangible Asset and mean layer within a BG

	log(Int)	log(Int)	log(Int)	log(Int)	log(Int)	log(Int)
Mean layer	0.490[a]	0.490[a]	0.490[a]	0.490[a]	0.490[a]	0.490[a]
	(0.0108)	(0.0264)	(0.0180)	(0.0391)	(0.0224)	(0.0283)
Observations	165,886	165,886	165,886	165,886	165,886	165,886
R- squared	0.574	0.574	0.574	0.574	0.574	0.574
Subs Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Parent FE	Yes	Yes	Yes	Yes	Yes	Yes
SE Robust	Yes	No	No	No	No	No
Cluster GUO	No	Yes	No	No	No	No
Cluster firm	No	No	Yes	No	No	No
Cluster subs country	No	No	No	Yes	No	No
Cluster subs industry	No	No	No	No	Yes	No
Cluster year	No	No	No	No	No	Yes

Note: The dependent variable is log of intangible asset expressed in thousands of Euros at 2010 price level. Mean Value is a group specific dummy equal one for those subsidiaries that are above the mean layer of the group. Standard errors are reported in brackets. a, b and c indicate significance at the 10, 5 and 1 percent level.

heterogeneity between integrated firms comes from the position in which each affiliate is placed in the hierarchy. In fact, our exercise suggests that there are systematic differences, at least in terms of knowledge stock involved and also stock of knowledge to which each firm is exposed, between integrated firms depending on the hierarchical position in which they stand. This consideration is crucial, and it motivates the importance of taking into consideration the hierarchical links when trying to uncover interactions between subsidiaries, our main topic of interest.

2.4 Results

2.4.1 Empirical Specification

As already discussed in the Introduction, previous literature has already studied knowledge interactions between subsidiaries in a BG without finding any systematic relationship (Bilir and Morales (2020); Belenzon and Berkovitz (2010)). In this section, we turn to the main object of our analysis to determine the presence of knowledge flows by accounting for whether the knowledge developed in different parts of the rest of the group, to which each firm is exposed differently, does contribute in a distinct way to the dynamics of the productivity of each firm. This step can be crucial to detect the presence of those knowledge flows that so far had not

been possible to highlight within BGs. If knowledge tends to follow specific directions in BGs, and each firm is exposed differently to these patterns based on its hierarchical position, then this heterogeneity will be important to uncover them. Thus, we extend the analysis of Bilir and Morales (2020) and assess the gains that firms belonging to BGs experience due to their internal knowledge flows taking into account the hierarchical structure. In particular, we want to decompose the contribution of intangible asset developed by the whole BG into several components, to assess separately the importance of each of them. Our aim is not only to assess the presence of knowledge flows within BGs, but also their direction and intensity across different layers of the group.

We run the following specification:

$$\begin{aligned} log(TFP_{f,HQ,c,i,t}) = log(Int_{f,HQ,c,i,t}) + log(Int\_HQ_{HQ,c,i,t}) + log(Int\_Upper_{f,HQ,c,i,t}) + \\ log(Int\_Horizontal_{f,HQ,c,i,t}) + log(Int\_Bottom_{f,HQ,c,i,t}) + \\ FE_{HQ} + FE_c + FE_i + FE_t + \epsilon_{HQ,f,c,i,t} \end{aligned} \tag{2.2}$$

where we try to disentangle the direction of knowledge flows within a group by considering the position of the firm inside the structure of the BG. We explain the productivity performance of each firm looking separately at the contributions of intangible asset developed by firms from upper layers, firms belonging to the same layer, and firms from bottom layers.[10] Given our interest in understanding how subsidiaries interact within a BG, and to control for several unobserved factors that are correlated across subsidiaries, we employ BG fixed effect. The variation that we exploit is mainly given by the hierarchical position of each firm that determines the different exposure to upstream, downstream and horizontal stocks of intangible asset developed by other subsidiaries. All the results are clustered at BG level, but other ways of clustering do not affect the findings. Importantly, for each BG, Int_Upper and Int_Bottom are not defined for firms in layer 1 and in the last layer respectively. Hence, most the analysis will be restricted to firm from the second to fifth layer.[11]

[10] The results do not change if we lag, either of one or two years, the explanatory variables. This step can be important to allow intangible asset developed by firms to actually have an impact on firms' performance. Given the relatively short time-span of our data, we prefer to use contemporaneous variables in our baseline specification.

[11] See Appendix A.2.2 for a discussion of additional exercises to assess the stability of our results for firms belonging to different layers and belonging to groups with different structures.

2.4.2 Empirical Results

To emphasize our different approach with respect to previous literature and show how significant it can be, we begin replicating a simplified version of the baseline specification used by Bilir and Morales (2020). As anticipated, they are mainly interested in the effects of R&D of the HQ and of each affiliate, the two main variables of their analysis, on the productivity of each affiliate. As an additional control, they also aggregate the R&D done by other affiliates in a third variable.

We implement their baseline specification with our own data and look at a sample as close as possible to theirs. That is, we look only at BGs headquartered in US and we study only the TFP performance of foreign subsidiaries (column I). Following their specification, we employ country, year and industry (defined at 2-digit level) fixed effects. Interestingly, our sample confirms the insignificant role of knowledge developed along the boundaries of a BG in explaining the productivity of subsidiaries. This result suggests that affiliates do not interact between them, and the only relevant sharing of knowledge happens between the headquarters and each affiliate. Column 2, that expands the sample to BGs based in Europe and looks at worldwide subsidiaries, does change the result and hints at a positive contribution of the knowledge developed by other subsidiaries of the group. However, the result is weak and it is not stable to the inclusion of BG fixed effect, that is crucial in our opinion. In fact, since we want to dig deep into each entity and understand within each group how subsidiaries relate to each other, we have to control for unobserved heterogeneity that affects all firms of the group.

Therefore, to better understand the interactions between subsidiaries and the patterns that intangible asset follows within the boundaries of a group, we unpack the knowledge developed by the rest of the group into several components, grouped based on hierarchical classification. This step is crucial to allow interaction between subsidiaries to be different based on their hierarchical link and let knowledge to follow different paths within the groups. The idea that the key way to understand the behavior of a BG comes from the hierarchical organization and allocation of knowledge has been well emphasized by the literature that studies the within-firm organization and its organization as knowledge hierarchies (Garicano and Rossi-Hansberg (2006)). In addition, Caliendo et al. (2020) also shows how the within firm organization has real consequences on firms' productivity. Since also interactions between affiliates crucially depends on hierarchical links (Altomonte et al., 2021b), we bring the organizational structure at the center of our analysis to understand how subsidiaries interact between themselves. The results, reported in table 2.5, show that this decomposition highlights dynamics that were

Table 2.4: Drivers of TFP and knowledge flows

	log(TFP)	log(TFP)	log(TFP)
Int	0.0259^a	0.0426^a	0.0397^a
	(0.00286)	(0.00338)	(0.00317)
Int HQ	0.0630^a	0.0470^a	0.0160^b
	(0.00387)	(0.00683)	(0.00691)
Int Rest Group	0.00236	0.0198a^a	−0.00495
	(0.00398)	(0.00735)	(0.00578)
Observations	18,014	41,205	41,139
R-squared	0.728	0.720	0.783
HQ from	US	US and EU	US and EU
Country FE:	Yes	Yes	Yes
Industry FE:	Yes	Yes	Yes
Time FE:	Yes	Yes	Yes
HQ FE:	No	No	Yes

Note: The dependent variable is log of TFP at firm level. The explanatory variables are all in log and refer respectively to: intangible asset of the firm, intangible asset that comes from the HQ and intangible asset developed by the rest of the group. All the values of intangible asset are in thousands of Euros at 2010 price levels. Robust standard errors are reported in brackets. a, b and c indicate significance at the 10, 5 and 1 percent level.

covered in the aggregate analysis. A stable but surprising results stands out: knowledge sharing seems to flow upward and a key role is played by the knowledge developed by firms in bottom layers.

On top of the importance of the HQ, which significantly contributes to the efficiency of its subsidiaries, as it is well established in the literature by now (Bilir and Morales, 2020), we detect additional important patterns. Our decision to look at interactions through a hierarchical perspective is key to let emerge additional interactions between subsidiaries that were hidden in a pooled analysis. This is because knowledge flows only in specific directions in a group. Headquarters, in order to enjoy these flows, structure the groups in specific ways (Altomonte et al., 2021b) and expand vertically in order to acquire technologies and knowledge that are then leveraged backward in the structure. This bottom-up hierarchical pattern is consistent with relevant findings that emphasize this direction of spillovers upon integration. For example, Phillips and Zhdanov (2013) show that large firms optimally may decide to purchase smaller innovative firms and conduct less R&D themselves. Also, Bena and Li (2014) find that acquirers are less R&D intensive but have larger portfolio of innovation while target firms are active in innovation, not yet converted their R&D expenses into patents at the time of

Table 2.5: Drivers of TFP and knowledge flows by hierarchical links

	log(TFP)	log(TFP)	log(TFP)	log(TFP)	log(TFP)	log(TFP)	log(TFP)	log(TFP)	log(TFP)
Int	0.0501^a	0.0434^a	0.0473^a	0.0502^a	0.0519^a	0.0404^a	0.0444^a	0.0459^a	0.0447^a
	(0.00240)	(0.00317)	(0.00287)	(0.00261)	(0.00288)	(0.00399)	(0.00332)	(0.00379)	(0.00530)
Int HQ		0.0170^b				0.0230^b	0.0193^a	0.0258^a	0.0328^a
		(0.00659)				(0.00920)	(0.00734)	(0.00821)	(0.0124)
Int Upper			−0.00604^b			−0.00853^b			−0.00875
			(0.00288)			(0.00434)			(0.00601)
Int Horizontal				0.0109^a			0.00730^b		0.00858^c
				(0.00243)			(0.00317)		(0.00513)
Int Bottom					0.0157^a			0.0123^a	0.0163^a
					(0.00246)			(0.00340)	(0.00441)
Observations	118,367	42,706	72,896	106,559	87,572	25,170	38,577	31,770	16,601
R-squared	0.840	0.832	0.849	0.837	0.840	0.834	0.828	0.830	0.833
Country FE:	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry FE:	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time FE:	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
HQ FE:	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: The dependent variable is log of TFP at firm level. The explanatory variables are all in log and refer respectively to: intangible asset of the firm, intangible asset that comes from the HQ and intangible asset developed by other subsidiaries of the group that are in upper/same/lower layers. All the values of intangible asset are in thousands of Euros at 2010 price levels. Clustered standard errors at BG level are reported in brackets. a, b and c indicate significance at the 10, 5 and 1 percent level.

integration. Finally, Sevilir and Tian (2012) find a positive association between M&As and acquirers' post-merger innovation outcomes, suggesting knowledge flows from acquired to acquirer, that is in line with the upward flow we find. Our result confirms these predictions looking at the relationship between subsidiaries and focusing on productivity outcome rather than innovation performance. The idea of vertical expansion used for specialization and technological acquisition is consistent also with other findings with a closer focus on value chain, who show that suppliers undertake substantial R&D investments (Calzolari et al., 2015) and that firms integrate technologically important supplier, such as Berlingieri et al. (2021). The emergence of such patterns and the presence of relevant interactions with other subsidiaries emphasizes the importance to look at the whole structure of a BG and not only HQ-affiliate relationship to understand the performance of a firm.

2.5 Robustness and Sensitivity Analysis

We now test the sensitivity and robustness of the baseline estimates. A key consideration to bear in mind is that our analysis is a fully descriptive exercise. We cannot claim that we capture a systematic effect of having knowledge from lower firms rather than from subsidiaries placed in other positions. The organizational structure of a group is a strategic choice of HQ, that places a subsidiary in a given position likely because it expects some interactions. Therefore, our estimates do not provide any causal effect, rather they highlight correlations that suggests to us how BGs behave and become more efficient. Having this caveat in mind, we want to provide convincing evidence that what we are capturing are indeed knowledge flows and that are the result of unfolding the knowledge of the group through hierarchical lenses.[12]

2.5.1 *Robustness to alternative variables:*

First of all, we test whether the baseline result is robust to a series of alternative definitions of intangible asset and TFP, reported in Table 2.6. In our baseline specification, the knowledge to which each subsidiary is exposed is computed using mean values of the stock of knowledge coming from firms in upper/same/lower layers. The choice to consider mean values is dictated from the fact that a very relevant number of firms does not report a value of intangible asset. Since this does not mean that firms have actually zero value of intangible endowment, we prefer to use mean values because considering the sum would imply assigning zero values to a firm's intangible asset when not available. In alternative checks (columns 2,4 and 7), we also consider

[12] In addition to the few exercises proposed in the next subsections, further tests are discussed in Appendix A.2.2.

the sums of these stocks. As additional checks (columns 3 and 4), we also consider intangible asset directly provided by the balance sheets of firms and a non-parametric estimation of productivity based on log of labor productivity. Qualitatively, the results are never affected by these alternative definitions of our main variables of interest. Finally, the pattern does not depend on the somewhat arbitrary restriction we impose on the maximum number of layers of a group. If we change this number to a maximum of 8 layers, the results do not change.

2.5.2 *Is it the hierarchical organization that matters?*

To evaluate whether our exercise is indeed capturing the importance of hierarchical organization of the group, we propose alternative ways to decompose the aggregate structure of a group. For example, BGs are entities active in very different business and it could be that the organizational structure is systematically related to a pattern in the industries in which firms are active. In particular, we want to understand whether there is a specific regularity in the organization of firms that is related to industry of activity and might affect our baseline finding. For example, subsidiaries in bottom layers might be more concentrated in specific industries, and this concentration triggers the positive interactions from bottom firms. If that is the case, then we do not know whether the effect of downward knowledge comes from the hierarchical link or it is confounded by the industry concentration channel. To address this concern, we unfold the intangible asset developed by the rest of the group according to a different criterion. In particular, we completely neglect hierarchical considerations and for each firm we aggregate the intangible endowments of other affiliates active respectively in the same 4-digit industry, in the same 2-digit industry (excluding the ones in the same 2-digit) and finally out of the same 2-digit industry. This gives, for each firm, three variables that pool different part of the overall knowledge of the group. Since focusing on 4-digit industry might be too disaggregated, we also try to aggregate intangible asset developed by firms of the group that belong to the same 2-digit industry in a unique category, including also subsidiaries from the same 4-digits. So, in column 2 of Table 2.7, on top of intangible developed by each firm, we consider the contribution coming from affiliates that belong to the same 2-digit industry and subsidiaries

Table 2.6: Robustness Analysis

	log(TFP)	log(TFP)	log(TFP)	log(TFP)	labor prod.	log(TFP)	log(TFP)
Int	0.0447^a	0.0452^a	0.0295^a	0.0293^a	0.0257^a	0.0443^a	0.0452^a
	(0.00530)	(0.00543)	(0.00256)	(0.00259)	(0.00558)	(0.00483)	(0.00512)
Int HQ	0.0328^a	0.0331^a	−0.00117	−0.00147	0.0311^a	0.0324^a	0.0352^a
	(0.0124)	(0.0137)	(0.00533)	(0.00584)	(0.0119)	(0.0118)	(0.0131)
Int Upper	-0.00875	-0.00591	0.00362	0.000791	−0.0118^c	-0.00786	-0.00679
	(0.00601)	(0.00482)	(0.00348)	(0.00327)	(0.00624)	(0.00543)	(0.00450)
Int Horizontal	0.00858^c	0.00580	0.00310	-0.00177	0.0148^a	0.0108^b	0.00525
	(0.00513)	(0.00442)	(0.00333)	(0.00326)	(0.00531)	(0.00497)	(0.00426)
Int Bottom	0.0163^a	0.0154^a	0.00944^a	0.00832^a	0.0216^a	0.0147^a	0.0143^a
	(0.00441)	(0.00345)	(0.00284)	(0.00278)	(0.00465)	(0.00417)	(0.00309)
Observations:	16,601	16,128	23,056	22,536	17,029	18,556	17,255
R-squared:	0.833	0.833	0.809	0.811	0.853	0.830	0.832
Int variable:	mean	sum	mean BS	sum BS	mean	mean & 8 layers	sum & 8 layers
Country FE:	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry FE:	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time FE:	Yes	Yes	Yes	Yes	Yes	Yes	Yes
HQ FE:	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: The dependent variable is log of TFP at firm level. The explanatory variables are all in log and refer respectively to: intangible asset of the firm, intangible asset that comes from the HQ and intangible asset developed by the rest of the group. All the values of intangible asset are in thousands of Euros at 2010 price levels. In column I, reports the same estimates than column IX of Table 2.5, while in column II the explanatory variables use the sum of the stocks of intangible. In column III and IV, intangible asset from balance sheet are used and in column V we use labor productivity instead of TFP. In column VI and VII we build the stocks of intangible asset, both for mean values and sums, considering the maximum number of layers of a group to 8. All the values of intangible asset are in thousands of Euros at 2010 price levels. Clustered standard errors at BG level are reported in brackets. a, b and c indicate significance at the 10, 5 and 1 percent level.

Table 2.7: Knowledge Flows pooling Knoweldge of the group based on industry classification

	log(TFP)	log(TFP)
Int	0.0497^a	0.0515^a
	(0.00412)	(0.00303)
Int same 4-dig ind.	−0.0011	
	(0.00263)	
Int same 2-dig ind (excl 4-dig)	-0.00362	
	(0.00327)	
Int same 2-dig ind.		-0.00352
		(0.00241)
Int other industries	0.00485	0.00298
	(0.00489)	(0.00354)
Observations:	53,510	88,879
R-squared:	0.835	0.839
Country FE:	Yes	Yes
Industry FE:	Yes	Yes
Time FE:	Yes	Yes
HQ FE:	Yes	Yes

Note: The dependent variable is log of TFP at firm level. The explanatory variables are all in log and refer respectively to: intangible asset of the firm, intangible asset that comes from firms in the same 4-digit industry, same 2-digit industry (excluding the same 4-digit industry in column I) and finally intangible developed by firms in other 2-digit industries. All the values of intangible asset are in thousands of Euros at 2010 price levels. Clustered standard errors at BG level are reported in brackets. a, b and c indicate significance at the 10, 5 and 1 percent level.

that are in different 2-digit industries.[13] [14]

Importantly, if we augment this specification with the variables from our baseline equation 2.2, it confirms the main findings that knowledge flows upward. The result of Table 2.7 is important because it tells us that the keystone to understand how subsidiaries interact is to look at the group through the lens of the hierarchical organization and that there is not a pattern in industry activity of bottom firms that is confounding this baseline result. Unpacking

[13] This result is confirmed by an alternative exercise. For each firm we compute the fraction of firms in bottom layers, relative to the number of firms in bottom for which intangible asset is available, that belong to the same 2-digit industry. We then create a dummy to distinguish between firms that have an index above/below the mean to identify firms that have a fraction of bottom firms in the same 2-digit industry that higher than for the average firm. We replicate our baseline exercise interacting the knowledge coming from bottom firms with this index and we find that the contribution of bottom knowledge is not statistically different between the two groups.

[14] In a third alternative exercise, we build an index for each firm that looks at the concentration of bottom firms over different industries. With this index we try to capture how specialized is the bottom knowledge of a firm in a specific industry, no matter whether it is on the same or different industries of activity of the firm itself. This alternative specification does not affect the result, implying that there are not specific features of industry concentration in bottom layers that make knowledge flows upstream.

the intangible developed by the rest of the group shows that the presence of "spillovers" is not related to the industry closeness of firms and the puzzle about the lack of interactions within the boundaries of a BG persists (Bilir and Morales (2020); Belenzon and Berkovitz (2010)).

2.5.3 *Are these Knoweldge Flows?*

The focus of our exercise, as motivated by the interest of previous literature on knowledge hierarchies (Altomonte et al. (2021b); Garicano (2000)) and by our descriptive statistics, is to capture knowledge interactions between subsidiaries. To achieve this goal, we implement a placebo exercise in which we replace intangible asset with other firms' characteristics and we study how interactions are affected. This exercise allows us to consider whether we are indeed capturing knowledge flows between affiliates or rather other kind of interactions, simply related to baseline firms' characteristics, that might drive the observed patterns. In practice, we do a similar exercise to our baseline but rather than decomposing intangible asset of the rest of the group, we focus on employment interactions; for each firm, we try to explain its productivity performance with the contribution coming from the employment of firms in upper, lower and horizontal layers. The result is important to show that our analysis on knowledge flows is indeed related to intangible asset and was not driven by other confounding factors, such as the size of other affiliates. This concern is particularly relevant within BGs, since affiliates often have supply chain relationships and exchange several inputs. In addition, since our dependent variable measures TFPR and we lack information of within-BG shipments, it is very hard for us to disentangle between actual knowledge flows and other exchanges that might trigger productivity gains for firms. For example, if bottom firms are suppliers of input for upper firms and are more knowledge intensive, they are likely to be more efficient and provide higher quality input to upper firms, and this might lead to the spurious TFP gains observed. If our baseline specification was capturing other source of interaction, for example more efficient input provision or positive demand shock that increases the profitability, it should be present also in this specification.

At the same time, this exercise helps us to mitigate a different potential concern according to which future affiliate performance shocks may be known to other subsidiaries at the time of intangible's investment choice, providing them incentives to increase their investments and size (reverse causality). We do not think this scenario represents a concrete threat in our set-up because each firm has several firms in upper layers (the typical BG has a structure that resembles an inverted pyramidal organization) and so, if the incentive to invest for a firm was

driven exclusively by future positive affiliate shock of only one upper firm, it would be very unlikely to lead to the observed results.

The result in Table 2.8 shows that knowledge flows are the main source of interactions between subsidiaries while employment does not seem to lead to stable productivity gains across subsidiaries. If the same patterns had emerged also with other firms' characteristics, it might have been possible that our baseline results were actually capturing more general interactions between each affiliate and other bottom subsidiaries. Interestingly, the significance of employment is not robust to more stringent specification in which we include also knowledge interaction, providing support for the fact that our channel really happens via knowledge interactions.

2.5.4 *Role of the Headquarter*

The information available to us allow also to investigate how deep and uniform is the influence of the knowledge of the headquarter along the boundaries of the group. This step is interesting not only to understand the role and influence of the HQ for its affiliates, also those hierarchically far away from it, but also to draw the attention towards an alternative dynamic that might drive our baseline results. Since firms that are closer to the HQ experience more interactions with the HQ and are also mechanically more exposed to bottom knowledge, we want to be sure that there is not interference between these two mechanisms. So, we interact the endowment of the HQ with a dummy that indicates in which layer a firm is active to assess whether the influence of the HQ spreads uniformly along all the boundaries or if the interactions depend on the hierarchical distance.

In column I of Table 2.9, we assess how homogeneous and broad is the influence of the HQ. We believe this exercise is very interesting and it reinforces our claim that integration is not a unique phenomenon and the organizational structure of BGs can lead to different outcomes upon integration. As we see, subsidiaries enjoy systematically stronger productivity gains interacting with the HQ the closer they are. The variable Int HQ represents the baseline effect of the intangible developed by the HQ on the TFP of affiliates placed in the first layer, and all other dummies identify the differential impact for firms places in the indicated layer relative to firms in the first layer. Despite the influence exerted by the GUO decades as firms are farther away in the hierarchy, it remains always positive.

To assess whether the non-homogeneous role of the HQ can impact our baseline results, we augment this specification controlling also for the usual terms of our baseline specification.

Table 2.8: Drivers of TFP pooling employment by hierarchical links

	log(TFP)	log(TFP)
Int	0.0462[a]	0.0459[a]
	(0.00322)	(0.00548)
Emp HQ	−0.00787[c]	−0.0160
	(0.0356)	(0.00402)
Emp Upper	0.00505	0.0224[a]
	(0.0044)	(0.00845)
Emp Horizontal	0.00883	0.0145[c]
	(0.00572)	(0.0086)
Emp Bottom	0.0115[c]	−0.0126
	(0.0052)	(0.0098)
Int HQ		0.0312[b]
		(0.0137)
Int Upper		−0.0145[c]
		(0.00741)
Int Horizontal		0.0051
		(0.00513)
Int Bottom		0.0180[a]
		(0.00547)
Observations:	39,741	16,044
R-squared:	0.859	0.834
Country FE:	Yes	Yes
Industry FE:	Yes	Yes
Time FE:	Yes	Yes
HQ FE:	Yes	Yes

Note: The dependent variable is log of TFP at firm level. The explanatory variables look at log of intangible asset and employment of the group and decompose the in the component developed by the HQ and by firms in upper/same and bottom layers. All the values of intangible asset are in thousands of Euros at 2010 price levels. Clustered standard errors at BG level are reported in brackets. a, b and c indicate significance at the 10, 5 and 1 percent level.

From column 1 to column 2, since we include in the specification also the knowledge developed by bottom, horizontal and upper layers, we exclude firms in the top and bottom layer of each group, since for these firms upper and bottom knowledge are not defined. So, the effect of the intangible of the HQ refers to the effects of this variable on the productivity of firms in layer 2, while the rest of the dummies tell us the differential impact of the intangible of the HQ on firms belonging to different layers. The reduction in the influence of the HQ is now less pronounced, but it is still present and suggests that the HQ shares more knowledge with firms hierarchically closer. More important for us, the importance on knowledge developed by bottom firms is confirmed also allowing the effect of the HQ to be flexible and to depend on the hierarchical distance, guaranteeing that our baseline specification was not affected by the decaying influence of the HQ on its subsidiaries along the hierarchical structure.

2.6 Conclusion

In this paper, we study how subsidiaries of Business Groups interact between each other, trying to address a puzzling result highlighted by previous literature that has shown the lack of significant interactions between affiliates (Bilir and Morales (2020), Belenzon and Berkovitz (2010)). The analysis shows that there are two key elements to consider in order to understand these interactions: the hierarchical links between them and the intangible asset rather than on other firm's level characteristics. These key steps, that build on the literature studying within-firm hierarchical organization based on knowledge (Garicano, 2000) and recently also applied to BGs (Altomonte et al., 2021b), are crucial to highlight knowledge spillovers within the boundaries of BGs and understand the direction that intangible asset follows in groups. Interestingly, we show that within BGs knowledge flows upwards, i.e. subsidiaries in lower layers share their knowledge to affiliates in upper layers, leading to productivity gains for these latter.

Table 2.9: Drivers of TFP, knwoeldge flows and the role of HQ

	tfp	tfp
Int	0.0427^a	0.0444^a
	(0.00319)	(0.00531)
Int HQ	0.0186^a	0.0346^a
	(0.00658)	(0.0125)
Int HQ layer 2	−0.00201^b	
	(0.000805)	
Int HQ layer 3	−0.00289^a	-0.00120
	(0.00100)	(0.00118)
Int HQ layer 4	−0.00384^a	−0.00357^a
	(0.00121)	(0.00138)
Int HQ layer 5	−0.00347^b	−0.00295^c
	(0.00136)	(0.00151)
Int HQ layer 6	−0.00553^a	
	(0.00149)	
Int upper		-0.00798
		(0.00597)
Int horizontal		0.00769
		(0.00512)
Int bottom		0.0135^a
		(0.00440)
Observations:	42,706	16,601
R-squared:	0.832	0.833
Country FE:	Yes	Yes
Industry FE:	Yes	Yes
Time FE:	Yes	Yes
HQ FE:	Yes	Yes

Note: The dependent variable is log of TFP at firm level. The explanatory variables are all in log and refer respectively to: intangible asset of the firm, intangible asset that comes from the HQ and intangible asset developed by other subsidiaries of the group that are in upper/same/lower layers. Int HQ layer 2,3,4,5,6 refer to an interaction term between the intangible of the HQ and a dummy indicating in which layer the firm is placed. All the values of intangible asset are in thousands of Euros at 2010 price levels. Clustered standard errors at BG level are reported in brackets. a, b and c indicate significance at the 10, 5 and 1 percent level.

Chapter 3

The OECD Start-ups database[1]

Abstract

This paper describes the development of the OECD start-up database, a unique dataset that aims to provide comprehensive micro-level data on start-ups across 75 countries. The database combines several sources of detailed microdata on start-ups, venture capital deals, background information on entrepreneurs (such as education and gender), and has been combined with patent data from the global Patstat database and other sources. The OECD start-up database contains information on almost 900 000 start-ups founded between the years 2000-2020, of which more than 160 000 have received venture capital financing and more than 50 000 have been granted or have applied for at least one patent. The coverage of the data has been cross-validated with a number of external sources at the international and national level. These sources include other proprietary vendors (Pitchbook, Preqin, PWC) as well as national and regional venture capital associations (Latin America Venture Capital Association and Israel's IVC Research Center). The database will be used as the foundation for several projects aiming to explore the whole life cycle of start-ups (from creation to scale-up and exit) and analyse the role of public policies for supporting innovative entrepreneurship.

[1] This paper is a joint work with Milenko Fadic. The authors would like to thank Roman Arjona, Matej Bajgar, Anna Correia, Chiara Criscuolo, Antoine Dechezleprêtre, Helene Dernis, Kohei Kitazawa, Julie Lassebie, Carlo Menon, Nathalie Scholl, Mariagrazia Squicciarini, Paolo Veneri, and Andrew Womer for their comments, and Lukasz Wielogorski, Gligor Micajkov, Xander Pelgrom and Alan Liang for their support in downloading the raw data.
Fadic Milenko, OECD. E-mail: milenko.fadic@oecd.org
Andrea Greppi, University of Bologna, Department of Economics. E-mail: andrea.greppi2@unibo.it

3.1 Introduction

Start-ups have been identified as an important driver of innovation, jobs and economic growth (Grimaldi et al. (2011), Calvino et al. (2018), Haltiwanger et al. (2013), Decker et al. (2014)). For this reason, creating an ecosystem for innovative start-ups has been a priority across OECD countries, as evidenced by the growing number of implemented policies designed to support new and young firms. To empirically assess the effectiveness of existing policies aimed at fostering start-ups and help implement new policies to support these firms, timely micro-level data that is representative of the population of innovative start-ups is essential. However, obtaining comparable cross-country data on innovative start-ups is challenging. In fact, the very definition of what a start-up is generally differs based on the context in which it is being analysed. Therefore, statistical offices, researchers, and policymakers often use slightly different definitions when analysing start-ups.

This paper describes the development of the OECD start-up database, an unprecedented database that aims to provide comprehensive micro-level data on start-ups across OECD countries, BRICS, and other partner countries. The database, which builds on previous OECD work (Tarasconi and Menon (2017)), combines several sources of detailed microdata on start-ups, venture capital deals, background information on entrepreneurs (such as education and gender), and has been combined with patent data from the global Patstat database and other sources.[2] This paper describes the original data sources used to build the OECD start-up database, notably Crunchbase and Dealroom, two commercial data providers of firm-level data, and the steps implemented to clean, standardise, harmonise, match, and cross-validate the different original sources in a consistent and comprehensive framework.

In total, the OECD start-up database contains information on almost 900 000 start-ups founded between the years 2000-2020, of which more than 160 000 have received venture capital financing and more than 50 000 have been granted or have applied for at least one patent. The data has been cross-validated with a number of external sources at the international and national level, allowing for the identification of coverage issues. These sources include other proprietary vendors (Pitchbook, Preqin, PWC) as well as national (Turkish Start-up Ecosystem, Estonian Startup Database, and Israel's IVC Research Center) or regional venture capital associations (Latin America Venture Capital Association). The database will be used as the

[2] The results presented by previous OECD work (Breschi et al. (2018), Tarasconi and Menon (2017), Lassébie et al. (2019)) use Crunchbase and Patstat as the only source of data. The OECD start-up database includes the updated version of these sources and also additional sources for firms, investors, and universities. The updated database also leverages updated algorihtms for data cleaning and processing.

foundation for several projects. For example, it will be used to gather evidence on the recent dynamics of innovative ecosystems across countries, notably in the context of the COVID-19 crisis, focusing particularly on start-ups operating in areas related to the digital and green transitions. The OECD start-up database will also be used to explore how public research is leveraged by entrepreneurs by examining, among other topics, the commercialization of public research by academics through university spinoffs, patent collaboration between start-ups and universities, and the role of private and government venture capital on supporting academic entrepreneurship. The database will also explore how mergers and acquisitions of innovative start-ups affect subsequent innovation and scaling-up of advanced technologies.

The rest of the paper is structured as follows. Section 3.2 discusses the methods used by previous literature to identify innovative start-ups. Section 3.3 describes the different data sources used to compile the OECD start-up database. Section 3.4 explains the cleaning and matching steps that have been undertaken. Section 3.5 provides some descriptive statistics and key features of the database. Section 3.6 compares the database to a number of external sources used as a benchmark. Section 3.7 concludes.

3.2 Obtaining data on start-ups and identifying innovative firms

Although start-ups generally refer to young firms, there is no standardised definition used in the literature. Researchers and policymakers alike use different criteria to define start-ups which vary according to the question at hand. In the context of analyses of business dynamics and job flows, start-ups generally refer to all new firms entering the market, whereas in the context of innovation studies, start-ups are understood as young firms with high-growth potential in high-tech sectors. Because of their potential for economic growth and productivity, the aim of the OECD start-up database is to identify innovative start-ups rather than all new firms. Broadly speaking, the literature has taken two approaches to identify innovative start-ups. Most studies identify innovative start-ups by combining firm age with criteria such as sector of operation (e.g. high-tech sectors), identity of firm's founder or founding team (e.g. academics), support by venture capital or business angel financing, or patent applications or filings (Bertoni et al. (2015), Gaddy et al. (2017), Dechezleprêtre and Fadic (2022)).[3] For example, Colombo et al.

[3]Firms founded by academics can be considered innovative as many of them are based on academic researched that has been commercialised (Etzkowitz (2003)). Similarly, start-ups that have received venture capital finance provide a sample of firms that investors clearly consider having high growth potential and promising business models. Venture capital firms spend considerable resources filtering start-ups, evaluating the competitive environment, and analysing the qualifications of a start-up's founders (Kaplan and Lerner (2010)). Patents provide a signal, though imperfect (Griliches (1998)), of a firm's innovative efforts and success and provides information on

(2010) use the Research on Entrepreneurship in Advanced Technologies database developed by the Politecnico di Milano to identify a set of 1 974 Italian start-ups operating in high-tech sectors in the manufacturing and services industries using age of the firm, status (independent or subsidiary), and level of technology. Some more recent studies have used different criteria such as open-source collaborations (Conti et al. (2021)) or Twitter announcements (Cripps et al. (2020)) to identify these firms.

A second approach taken by a number of studies relies on the use of commercial databases that focus on tracking start-ups and consider all firms in the dataset as innovative. Kaplan and Lerner (2010) provide an overview of the wide number of data sources that track innovative start-ups—particularly those focusing on venture capital investments. VentureXpert and Venture Source are two of the most established companies in the area and started collecting data in 1961 and 1994, resepectively. As the VC industry have increased in size and popularity, there have been a number of new data providers on VC and start-ups have emerged in recent years. Burgiss Private I, Cambridge Associates (CA), Pitchbook, Preqin, Crunchbase and Dealroom.[4] Given the earlier data gap in the VC-finance data (Kaplan and Lerner (2010)) and the timeliness and coverage of these databases, these commercial data sources have been extensively used in academic research to identify innovative start-ups across industries and countries.[5]

It is worth highlhgting that geographically, the majority of the previous studies are focusesd mainly in the United States and/or a single country. In Europe, the VICO (Financing Entrepreneurial Ventures in Europe : Impact on innovation, employment growth, and competitiveness) project collected a database on young high-tech entrepreneurial companies operating in seven European countries (Belgium, Finland, France, Germany, Italy, Spain, and the United Kingdom) using VC-backed companies and the second a control group of non-VC backed companies (Bertoni et al. (2015)). It is important to consider the advantages and limitations of each approach, especially when using the findings to provide policy recommendations. The first strand of the literature, i.e. those base on the combination of selected criteria and firm age, often focuses on firms operating in selected sectors or on very specific types of firms and tends to have smaller sample sizes, making it challenging to generalise the results. Some of the indicators used may also exclude particular firms, such as innovative firms that do

the firms' technological field, innovative capabilities, and originality.

[4] Other providers include AVCJ, BarclayHedge, CBinsights, and Cobalt.

[5] For instance, Howell (2021) use Pitchbook, CB Insights, and Capital IQ to compare patenting activity between VC-backed and non VC-backed firms. Breschi et al. (2018) uses Crunchbase to present cross-country descriptive evidence on innovative start-ups. Cannone and Ughetto (2014) use Crunchbase to obtain a global list of firms operating in the ICT and electronics sectors to study the internationalization of high-tech start-ups. den Besten (2020) provides a list of 78 papers that use Crunchbase and have been published in peer-reviewed journals.

not use patents to protect their innovations but rather trade secrets or copyrights (Hall et al. (2013)), or firms that are innovative but do not receive VC as they do not fit the "venture capital playbook" (Catalini et al. (2019)). In contrast, studies that consider all firms listed by private vendors as innovative typically cover a larger number of start-ups, sectors and countries than the aforementioned studies. While this larger sample might help address the external validity concerns, Da Rin et al. (2011) argue that these databases do not follow a strict methodology to identify innovative start-ups. Therefore, the unfiltered sample of firms from these databases likely includes a non-trivial number of non-innovative firms, leading to important sample selection issues.

The underlying data for the OECD start-up database comes from Crunchbase and Dealroom, two commercial providers of data of firms, venture capital deals, and exits. Upon inspection of the micro-level data, it is clear that both Crunchbase and Dealroom also include in their sample firms that are clearly not innovative. By combining these sources with external data, it is possible to use additional indicators (such as sector of operation, VC funding, patent filings, contributions to open source projects) to estimate their innovative potential. Ongoing work is also developing alternative methods to identify innovative firms using available metadata (such as description of activities).

3.3 Data Sources

The OECD start-up database provides a comprehensive micro-level dataset on start-ups across countries. The following sections describe the different sources of data used in the database. Firm-level data on startups, venture capital deals, and background information on entrepreneurs come from Crunchbase and Dealroom, two leading vendors of data on startups and tech ecosystems.[6] Information on the alma mater of founders from these two sources is disambiguated and harmonised using the Global Research Identifier Database (GRID). These data are supplemented with data on corporate and government venture capital entities (Dechezleprêtre and Fadic (2022)) , firm-level patent applications and their quality (Squicciarini et al. (2013)), indicators on university quality and data on exchange rates and price levels.

[6]The data on start-ups comes from a unique database assembled under the KnowInn project (From Knowledge to Innovation: Building the evidence base to inform innovative entrepreneurship and risk-finance policies) aims to analyse the relationship between innovative entrepreneurship, public research, and government policies. It is financially supported by the European Commission under the Horizon 2020 programme. Data on GovVC entities was created from a variety of sources and validated by National Delegates to the OECD Committee on Innovation, Industry, and Entrepreneurship (CIIE) as well as national experts.

3.3.1 Data on start-ups, founders, and investors

Crunchbase

The main data provider on start-ups is Crunchbase (www.crunchbase.com). Crunchbase was created in 2007 and focuses specifically on covering innovative firms. Crunchbase has some specific features that provide major advantages over other commercial databases covering similar information. In particular, it contains cross-linked information on companies, funding received, funders, and staff. Moreover, the information is updated continuously, thereby providing more up-to-date information than other, more traditional, sources such as census of business establishments. Crunchbase has been previously used in the academic literature and has also been an important resource for policy analysis (Breschi et al. (2018), Cannone and Ughetto (2014)). For instance, Chen et al. (2020) use Crunchbase to understand whether interactions of inventors from acquiring firms and acquired start-ups lead to positive innovation outcomes. To assess the coverage quality, the authors compare Crunchbase with data from two different sources. The first source is the Israeli data source of venture capital-backed startups. The second source is SDC Platinum, which covers technology acquisitions across countries. The authors conclude that Crunchbase has a very high level of coverage relative to the national and commercial database.

Dealroom

The second main data provider on start-ups is Dealroom (https://dealroom.co/). Like Crunchbase, Dealroom is a database that aims to identify and monitor innovative companies around the world. Dealroom tracks over one million companies, 85,000 investors and includes more than 270,000 transactions and funding rounds since the year 2000. Dealroom sources its data through a network of crowd-sourced contributors (founders, VCs, accelerators, governments, and tech journalists), automated data feeds from social media, curated media, analytics providers and web crawlers. Dealroom is the official platform used by a number of cities and national agencies to track entrepreneurial activities in their districts. For example, LaFrenchTech is a platform that tracks the evolution of start-ups in France, following new entities from their creation, to potential VC financing and exit strategies. Other relevant examples are the cities of Berlin and Madrid, and the national agency of Lithuania.[7] To ensure its data quality, Dealroom data is manually checked and curated by Dealroom's internal research team. One major advantage of Dealroom is that is focuses on a different geographical market than Crunchbase.

[7] For a lists of other institutional users of Dealroom, see Ecosystem Platform | Dealroom.co

While Crunchbase has high coverage of North American new ventures, Dealroom's information are mostly focused on European markets. Thus, by combining these two different yet complementary sources of data, the OECD database provides a unique, unprecedented, and comprehensive global database on start-ups and VC activity.

3.3.2 Corporate venture capital and government venture capital entities

Across OECD countries, there are a number of Corporate Venture capital (CVCs) and government venture capital (GovVC) entities, each organised through a variety of complex legal structures. Because of this, categorizing an entity as a CVCs or GovVCs is difficult and many authors used different methodologies based on their specific research question (Röhm et al. (2020)). Yet, CVCs and GovVC entities play a critical role in the start-up ecosystem, providing finance to start-ups seeking funding. In an effort to provide the most comprehensive and accurate list of CVCs and GovVC, the OECD embarked in two parallel projects to collect and validate CVCs and GovVCs entities. The sources used to create the CVC lists were: Crunchbase, Thomson Reuters, CB Insights, Bureau Van Dijk (BvD)'s Zephyr and ORBIS databases. The sources used to create the list of GovVC were Crunchbase, Dealroom, Pitchbook, Preqin, and data from Invest Europe. Delegates of OECD member countries to the Committee on Innovation, Industry and Entrepreneurship (CIIE) through an online survey validated all GovVC entities. Section 3.4.5 provides detailed information on the methodology used to compile both lists.

3.3.3 University Data

Link between start-ups and universities

Crunchbase and Dealroom contain detailed information on the educational background of founders and key company staff members. However, their data does not include a unique university identifier that can be cross-linked with other datasets. For this reason, names of universities and research institutions in the database are disambiguated, harmonised and matched to the Global Research Identifier Database (GRID).[8] GRID is an open access database on institutions associated with academic research, founded and maintained by Digital Science & Research Solutions Ltd. GRID provides descriptive metadata on research institutions, including website, Wikipedia page, and aliases of the research institution. This information is particularly important as it facilitates the matching between different sources (for example

[8] https://www.grid.ac/

by noting that the research institution École Normale Supérieure is also known as ENS Paris). Importantly, GRID also provides the relationship between different research institutions that allows to map an institution to their "parent" (i.e. mapping Harvard Medical School to Harvard University). GRID is a comprehensive database sourcing their information from research funding grants and research paper affiliations. The September 2021 version of GRID includes 101,637 institutions from 217 countries/territories.[9]

University Quality: Leiden rankings

To obtain a (partial) measure of university quality, the GRID database is matched with ranking data from the Centre for Science and Technology Studies (CWTS) Leiden Ranking. The CWTS Leiden Ranking is a worldwide university ranking based entirely on bibliographic indicators for the year 2016.[10] The Leiden ranking builds on the Science Citation Index Expanded, the Social Sciences Citation Index, and the Arts & Humanities Index from the Web of Science (Woos) database. This ranking is based on bibliographic data from the Web of Science database produced by Clarivate Analytics for the period 2011-2014 and can be used to assess the research performance of universities. CWTS provides several different measures of university quality. One measure often used is the proportion of a university's publications that, compared with other publications in the same field and in the same year, belongs to the top 10% (or top 1, or top 50%) most frequently cited. This measure is independent of university size, but size-dependent measures are also available. These data also contain indicators of scientific collaborations (based on number of co-authored publications, with the possibility to focus on international collaborations). The different indicators are computed for all fields, and by field (Biomedical and health sciences, Life and earth sciences, Mathematics and computer science, Physical sciences and engineering, Social sciences and humanities).

3.3.4 Patent Data

Firms in the database are matched with patent application data from the spring 2021 version of the EPO Worldwide Patent Statistical Database (Patstat). Patstat provides information on patent filings from all major patent offices in the world. Patstat includes name and address of applicants, geographical coverage of the patent (international patent families), grant status, inventor names, IPC and CPC classification codes, application, publication and grant dates.

[9] The last release of the GRID database is scheduled for Q4 of 2021. Following that date, the GRID database will be maintained by the Research Organization Registry (ROR).

[10] https://www.leidenranking.com/downloads

The database also includes citations to other patents and the academic literature. As a result, it provides a unique source to examine the links between innovative start-ups, explore the relation between industry and academia. Squicciarini et al. (2013) and Squicciarini and Dernis (2013) provide additional details on the contents of the Patstat database and additional indicators of patent quality, scope, reach, and impact.

3.3.5 Consumer price indices and exchange rates

To provide internationally comparable price levels in real monetary terms, the database includes consumer price indices and exchange rates obtained from the OECD data portal. The consumer price indices (CPIs) used is the National CPIs at the most aggregate level: CPI All Items with 2015 as the year of reference. Data on exchange rates are obtained for each country and currency. Exchange rates refer to the yearly average and are expressed in National currency per US dollar.

3.4 Methodology

The OECD start-up database is stored as a relational SQL database, maintained in a secured OECD server. The database is accessible only to authorized and licensed users working in research related to the CIIE Programme of Work. The data have been checked for statistical anomalies, data integrity, and have been aggregated at regional and national levels to cross-validate it with sources from National Statistical offices and private vendors. This section presents the methodology to extract, clean, and transfer the original comma delimited files to the database. Each record is stored using a primary key to guarantee the uniqueness of the individual company/business records.

3.4.1 Processing of data on start-ups and founders

This subsection explains the main steps undertaken to clean the information from Crunchbase and Dealroom. A description of the raw data available from these sources is available upon request, and the document in composed of two sections. The first part describes the steps taken to clean firm-level data, including adding missing country information, adding geographical coordinates, standardizing cities and countries, and cleaning links to social media and websites. The second part describes the steps taken to clean individual-level data, including standardizing founder's gender and identifying job titles.

Cleaning of entities

Given the global coverage of the database, the raw data contains records with characters from different writing languages. The first step to clean the entities files from Crunchbase and Dealroom is to import the raw data and convert its encoding to UTF-16, the coding used by Microsoft SQL Server. This is needed for a proper import of the data into the database.
The next step is to address records where the field for country of firm operation is missing. Three algorithms are used to obtain this information. The first algorithm derives country codes from a firm's phone number.[11] The second algorithm selects the country reported by individuals working in the organization and, if the country is unique, assigns it to the firm. In case there are multiple countries reported, the algorithm does not assign a country to the firm. The third algorithm derives country information using a firm's website address (such as websites ending in: .jp, .fr, and .it). All firms whose country of operation are identified using these algorithms are labelled. This procedure is applied to firms from Crunchbase and Dealroom.
Once the missing country information is added, all country names are harmonised to meet the OECD guidelines. The guidelines are provided by the OECD legal department and include a list of names of countries and territories. This step allows to harmonise country names across datasets (such as properly naming Vietnam as Viet Nam) and properly categorizing territories (for instance assigning Réunion to France). The names of cities are also harmonised using the OECD's functional urban definition.[12]
The next step is to add the geographical coordinates (GIS) of the firm's location. This is done only for firms in Crunchbase as Dealroom already includes that information. The algorithm to obtain the GIS is based on Google Maps API using firms' addresses. Once all GIS coordinates are obtained, they are cross-verified against the firm location to check for geographic consistency. This step verifies that latitude and longitude are within the reported country and city. Following this step, all hyperlinks of a firm are cleaned. This process goes through the different URLs in the database to check that they meet the length restrictions of the database and are valid hyperlinks. The final step is to clean the firm's description to ensure it meets the size limitations of the database and does not contain any hypertext mark-up language (HTML).

[11] The Python library used to process phone numbers can be found in the following link: https://github.com/daviddrysdale/python-phonenumbers
[12] For a list of global FUAs, see http://www.worldcitiestool.org/

Cleaning of founder information

The cleaning of founder information consists mostly of standardizing the different founder-level data columns within datasets and harmonizing these columns between Crunchbase and Dealroom. The first step is to standardise founder's genders. While Dealroom has three gender categories (male, female, missing), Crunchbase contains more than 50 gender categories. The algorithm to standardise gender categories looks at variations of records which likely refer to the same gender (such as male, man, men, hombre, homme, etc…). Gender categories not identified by the algorithm as either male or female are not standardised (such as non-binary, androgynous, and transgender). Following the standardization, genders are re-classified as 1) female, 2) male, 3) other, and 4) missing.

The second step—particularly important as it helps identify start-up founders—is to clean job titles. For this step, the algorithm uses regular expression (regex) to standardise each job title by analysing the different variations of texts. For example, the titles 'Founder', 'foundr', 'founding member', 'Co-founder' are all standardised to 'founder'. This algorithm captures common variation and misspellings, which is particularly relevant as the data type of the title field in Crunchbase is free text (which allows any type of character). Individuals whose job title contain the words 'founder', 'founding', 'co-founder partner', and 'CEO', are tagged as start-up founders. For the words 'partner' and 'CEO', an additional restriction is imposed: the individual must hold this job title since the company's creation.

The third step is to clean the data on the different degrees of founders. A similar algorithm to the one used to standardise each job title is employed using a degree-specific dictionary. For example, MBA, EMBA, Master of Business Administration, M.B.A. are all labelled as MBA. This algorithm produces an additional variable indicating the following degree(s) of individuals: high school degree, bachelor (undergrad), master, PhD, Juris doctor, or MBA. The fourth and final step is to match the universities listed in the founder's educational information to the Global Research Identifier Database (GRID). The list of universities reported in Crunchbase and Dealroom is not standardised and therefore there are variations in the name of organizations within and between datasets. The matching between the source data and GRID is done using four algorithms. The first algorithm matches universities using their exact name. The second algorithm uses variations of universities' names, including aliases and acronyms, to search for a match (for instance using ENS Paris instead of École Normale Supérieure). The third algorithm uses information on parent institutions (eg Harvard Medical School, Harvard Kennedy School belong to Harvard University) to perform the match. The algorithm replaces the child organiza-

tion with the parent organization to perform the matching (replacing Harvard Business School with Harvard University). The fourth algorithm uses a series of string-matching algorithms to provide pairs of universities that are likely the same institution. These pairs are reviewed manually. Virtually all universities in Crunchbase and Dealroom are matched to an entity in GRID.

3.4.2 Combining Crunchbase and Dealroom: steps for de-duplication

Before combining Crunchbase and Dealroom, it is first important to check each source independently to identify firms present in the data multiple times. For this purpose, the names of start-ups within each data source are de-duplicated using some key variables such as names, addresses and website information. Potential duplicates are flagged with a specific variable, together with an indicator to identify, out of the duplicates, the records with the most information available. Following this step, firms from the two sources are combined and de-duplicated. This step prevents the double counting of start-ups. Although Dealroom's metadata indicates if a record from its platform is also available in Crunchbase, it is incomplete in some cases. To ensure that the remaining firms are not duplicates, a disambiguation procedure is implemented to standardise names across the two source and check for potential duplicates. Start-ups that are considered as duplicates are identified with an indicator variable and the duplicate observations are used to complement potential missing data. For example, this criterion is used to complement information on investments received, and on the employment history and educational background of founder. Only one observation per firm is kept in the database, with information coming from the various duplicate observations.

3.4.3 Matching with Patstat

Start-ups in the OECD database are matched with the names of patent applicants in Patstat. This matching is implemented using a set of algorithms in the Imalinker system (Idener Multi Algorithm Linker) developed for the OECD by IDENER (Squicciarini et al. (2013)). The Imalinker system matches companies to patent applicants in Patstat using a matching procedure designed to maximise the number of correct matches. While the method attempts to minimise both "false positive" and "false negative" errors, priority is given to the minimisation of false positives. The matching procedure is implemented through two key steps. First, the names of firms in the start-up database and the names of patent applicants are separately harmonised using country-specific dictionaries. These dictionaries help harmonise legal entity denomina-

tion (e.g. “Corp” and “Corporation”), common names and linguistic rules that might affect the spelling of names and how names are written. This step is important because it allows accounting for features of the data, such as shortened expressions, that might prevent the matching. Second, a series of string-matching algorithms are used to compare the harmonised names from the two datasets and provide a matching accuracy score for each pair.[13] To ensure a high precision rate, the algorithm selects only pairs of company and patent applicant's names for which the accuracy score is above a minimum threshold. Matched pairs with a high accuracy score are considered as exact matching, while matched pairs with a score between the minimum threshold and the high accuracy score are reviewed manually—only around 200 firms of which were classified as matches.

3.4.4 Identifying government and corporate venture capital entities

Government and corporate venture capital entities are important players in the venture capital ecosystem as they provide financing and other forms of support to innovative start-ups. Although both types of entities have been studied extensively in the academic literature, there is no standardised definitions of what constitutes a private, government, or mixed fund as different authors use different definitions depending on the context of the study, the disaggregation of the data, and the research question. This section describes the methodology to identify these entities.

4.4.1. Government VC entities

Broadly speaking, government venture capital entities (GovVCs) have been defined using two distinct approaches. The first approach defines a fund as GovVCs if it is fully owned and/or managed by the government. The second approach is based on indirect government financing, where the government is involved as a limited partner (see for example Alperovych et al. (2018) and Brander et al. (2015)). The GovVC entities listed in the OECD start-up database are those that are fully owned and/or managed by the government.[14] The methodology to identify GovVCs is based on a two-step approach. In the first step, the Productivity, Innovation,

[13] Levensthein (Levenshtein et al. (1966)) and Jaro-Winkler (Jaro (1989), Jaro (1995), Winkler (1999)) distances are used to compare the harmonised names from the two datasets and provide a matching accuracy score for each pair.

[14] The other possible definition (involvement of the government as a Limited Partner in an otherwise Private fund) has important limitations. First, it is not always possible to determine the share of the government's involvement in private funds. This would lead to classify as GovVC a private fund with even a marginal government involvement. Second, this broad definition does not capture other important dimensions of GovVCs such as the autonomy of the fund's managers to make investment decisions. Hence, this definition can classify as GovVC a fund where the government's involvement is purely financial.

and Entrepreneurship (PIE) division within the STI Directorate identified a list of potential government venture capital investors. The list was assembled using the sources listed in section 3.3.2. In the second step, delegates of OECD member countries to the Committee of Innovation, Industry and Entrepreneurship (CIIE) validated these results through a survey. The survey allowed delegates to review the institutions, correct any information, and add funds that were not included in the previous list. The survey was completed by 37 out of 38 OECD countries. More information can be found in Dechezleprêtre and Fadic (2022).

4.4.2. Corporate VC entities

Corporate venture capital (CVC) is a form of venture capital where companies invest in young start-up firms for financial and non-financial reasons, such as acquiring new technologies, searching for business synergies, or entering emerging markets. In contrast to independent venture capitals (VC) and private equity (PE) firms, CVCs are wholly owned and funded by a single parent company, whose main field of business is usually different from investment itself. The CVCs listed in the OECD start-up database consists of entities that are legally separated from their parent company. The methodology to identify these CVCs is based on a three-step approach. First, a list of potential CVCs was created using the following sources: Crunchbase, Thomson Reuters, CB Insights, Bureau Van Dijk (BvD)'s Zephyr and ORBIS databases. The potential CVCs from these three sources are then harmonised and de-duplicated to obtain a list of unique CVCs. In the next step, each matched entity needs to be verified for whether it corresponds to a proper CVC and not to a private equity firm or an entire company. Finally, each CVC is then manually verified through an inspection of their official websites to ensure that the list of CVCs contains only entities that are legally independent from a parent company. It also allows to determine if it is a genuine CVC rather than a private equity firm, VC firms, or government entity.

3.4.5 4.5. Cross-validation and consistency checks

Following the cleaning and processing steps, a final cross-validation step is performed to check for potential inconsistencies in the data. This includes comparing the year of the patent applications and the date of funding rounds to the foundation year of the firm. In case of discrepancies between these dates (for instance in cases where the filing of the first patent is decades before the foundation of the firm), the matches are checked manually. Finally, implausible values in funding rounds and/or repeated funding rounds are manually assessed.

Number of start-ups in the OECD start-up database, by year of foundation

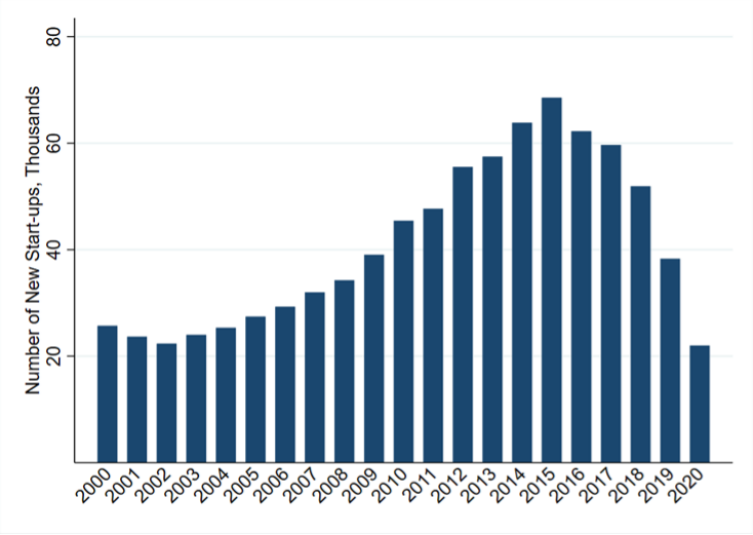


Figure 3.1: The figure shows the total number of start-ups founded between 2000 and 2020 in the OECD Start-up database. It includes firms from OECD countries plus 37 partners and non-member countries.
Source: OECD Start-up database.

3.5 Description of the OECD start-up database

The OECD Start-up database contains information on around 900 000 start-ups founded between the years 2000-2020. Figure 3.1 shows that the number of start-ups available in the database has increased steadily over the period 2000-2015. Following this period, the number of new start-ups available in the database declines, with more recent years experiencing a more pronounced drop.

One potential explanation for the drop in recent years is the delay in the collection of information available from the data providers. Data providers use (among others) administrative and web scraping techniques to gather their data. While web crawling algorithms tend to run continously, administrative data is updated at less regular intervals. Therefore, it is possible that start-ups that are "active" in the web—such as those that launch a new product or received VC—are identified earlier by the data providers than startups identifed through administrative sources. Figure 3.2 examines this potential explanation by reporting the number of VC deals obtained each year distinguishing between deals received by start-ups younger or equal than 5 years old and those firms older than 5 years of age. The trend across time, also in most recent years, is slightly increasing for young start-ups while it is flatter start-ups founded more than 5 years ago, suggesting that the lack of coverage in the number of start-ups in the last years,

Number of VC deals for young start-ups and older firms, by year of deal

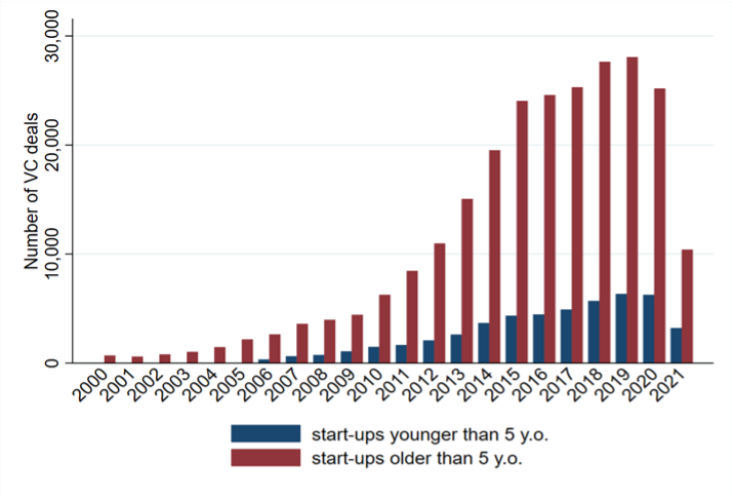


Figure 3.2: Number of venture capital funding deals for each year for the period 2000-20.The figure includes deals on start-ups from OECD countries plus 37 partners and non-member countries. Venture capital deals include: pre-seed, seed, angel, series funding, growth funding, late stage funding. Deals are distinguished between those involving start-ups founded less (or equal) and more than five years before the event of the deal.
Source:OECD Start-up database.

highlighted in Figure 3.1, is mostly driven by non-VC funded start-ups.

Table 5.1 provides descriptive statistics of firms in the database. The majority of start-ups in the database have their status as "operating". Around 6 percent of start-ups have successfully exited the market, either through IPO or acquisition. In total, 23 percent of start-ups received some type of financing, most of which is VC financing, and 6 percent have been granted or have applied for patents. Approximately one quarter of firms have information available on their founder(s), which shows – based on this sub-sample of firms – that almost one start-up out of six has at least one female founder. A total of 74 361 firms have information on the educational background of their founders. Of those firms, around 9.5% of founders have a PhD degree and 18% of founders have an MBA.

VC financing is often used as a common indicator of the innovativeness potential for new firms as venture capital funders spend considerable resources filtering start-ups, evaluating the competitive environment, and analysing the qualifications of a start-up's founders (Kaplan and Lerner (2010)). For this reason, and for data limitation, most studies on start-ups have focused only on VC-backed firms. Table 3.2 provides descriptive statistics for this sub-sample of firms. Start-ups that access funding receive their first deal, on average, at the age of two. On average, each firm receives almost three rounds of VC financing, and the value of each round

Table 3.1: Descriptive statistics of firms in the OECD start-up database

	Number	Share(%)
Companies	855 573	
Received financing	197 735	23.10
VC backed	163 166	19.00
Patents	51 710	6.04
Firms with founder information	209 593	25.0
Firms with founder's educational background	74 361	35.0*
At least one female founder	34 032	16.5*
Closed	20 315	2.37
IPO or Acquired	55 802	6.50

Note: Descriptive statistics on start-ups from OECD countries plus 37 partners and non-member countries. The variable financing refers to the number of start-ups that have received any type of funding event, while VC-backed consider start-ups that have received at least one VC financing. Patents consider start-ups that have already been granted or have applied for at least one patent. Closed and IPO or Acquired refer to the operating status of the start-up. *The share of firms with founder's educational background and at least one female founder are expressed as a share of firms with founder information.
Source: OECD Start-up database.

is on average worth almost USD 7.3 million. The distribution of VC funding is highly skewed and few start-ups have had a significant number of rounds (up to 29 rounds for a single firm).

Table 3.2: Descriptive statistics of VC financing

	Mean	Min	Max	Standard dev.
Age at first funding	2.2	0	21	2.87
Funding per deal	$ 7 330 397	$ 121	$1 090 000 000	$43 300 000
Number of VC deals	2.85	1	29	2.12

Note: Key statistics on Venture capital funding for start-ups from OECD countries plus 37 partners and non-member countries that have received at least VC financing. Venture capital deals include pre-seed, seed, angel, series funding, growth funding, late stage funding. Excludes deals that do not have information on funding amounts. Values of deals are expressed in constant 2005 USD dollars.
Source: OECD Start-up database.

Start-ups play a key role in promoting new technologies and fostering economic growth, and for this reason their patenting performance, as a proxy for their innovation abilities and potential, is a relevant measure to consider. In total, 51 710 start-ups have filed 367 989 different patents, for a total of 973 841 patent applications. Looking at this sub-sample of patenting start-ups, the median firm has filed 2 applications.[15] In total, 14% of VC backed firms have

[15]The mean number of patents for firms that have applied for or been granted a patent is 7. Note that due to the structure of the OECD start-up database, this statistic includes corporate spin-off companies associated with bigger and well-established firms. Consequently, in some cases those spin-off companies are assigned also all the patents filed by the parent entity. This makes the distribution of patents in the sample is highly skewed and

received or applied for at least one patent (23 256 start-ups), a share that is more than twice higher than for the full sample of start-ups. Interestingly, almost 2/3 of VC-backed patenting firms have applied for a patent before receiving funding. On average, start-ups apply for their first patent 3 years after their foundation.

As reported in the upper panel of Figure 3.3 a disproportionate amount of patents are filed in the United States patent office. Although around one third of companies are located in the United States, almost half of the patents applications in the OECD Start-up database are filed in the United States Patent and Trademark Office (USPTO). When examining the patent filings of non-US firms, the USPTO received more filings than the World Intellectual Property Organization (WIPO) and the European Patent Office (EPO), suggesting that the disproportionate amount of patents filed in the USPTO is driven by a higher number of patents filed by both US and non-US start-ups in the USPTO. The lower panel of Figure 3.3 reports the number of patents by technological class and shows that almost a third of total patent are related to Information and Communication Technologies (ICTs) with other important technologies being medical and environment-related.

The OECD start-up database collects information on new ventures from virtually all countries in the world. Table 3.3 reports the geographical coverage of the database by country of the firm. The last two columns report respectively the mean value per year of VC financing and the total VC financing over the years 2015-2020. Almost 310 000 firms are based in the United States, by far the most represented country not only for the number of start-ups but also for the total VC financing. On average, between the years of 2015-2020, firms based in the United States received approximately 47 percent of global VC financing. In Europe, most start-ups are located in Germany, United Kingdom and Netherlands. Interestingly, more than 10% of the sample are located in Brazil, Russia, India, China (People's Republic of) and South Africa (BRICS) and these firms have received more than 30 percent of total VC funding, with a substantial role played by start-ups from China, ranked second in terms of total VC financing.

Firms in the OECD start-up database operate in all major sectors of the economy. Figure 3.4 present the number of firms operating in each sector and the total VC (all years available) received by sector, respectively. Most start-ups in the OECD database operate in Professional services, which include firms working in: advertising, marketing, content and publishing, data and analytics, design, jobs recruitment, legal, security, sales and marketing. Not surprisingly,

warrants caution in the analysis of the patent portfolio of companies. For example, while firms at the 99th percentile have 76 patents, the firm with the highest number of patent has 4 696.

Number of Patents by Application Authority and by technological class

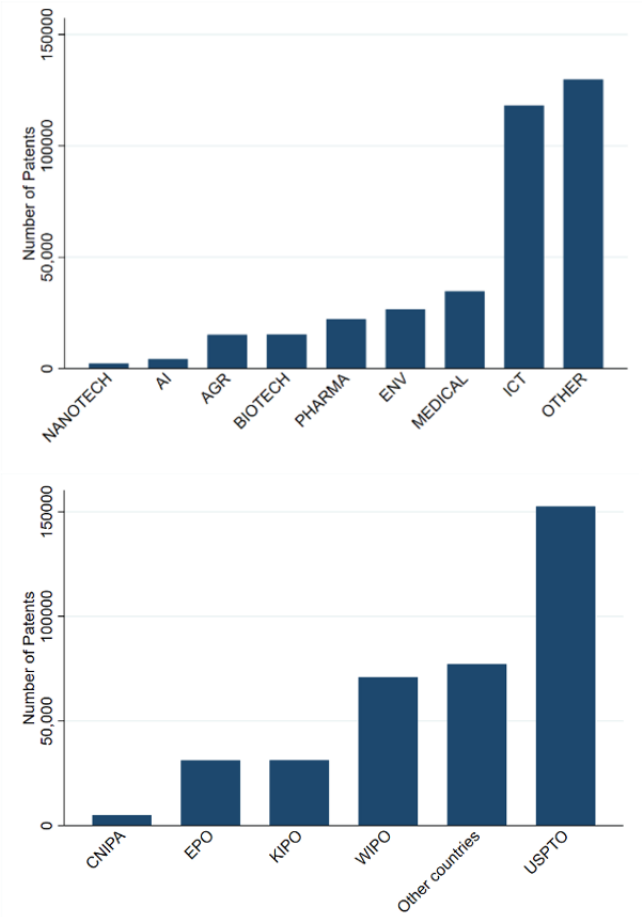


Figure 3.3: Note:Statistics on patents filed by start-ups from OECD countries plus 37 partners and non-member countries. The upper panel shoes the number of patents filed respectively in Chinese National Intellectual Property Administration (CNIPA), European Patent Office (EPO), Korean Intellectual Property Office (KIPO), World Intellectual Property Office (WIPO), authorities across other countries and United States Patent and Trademark Office (USPTO). The lower panel shoes the number of patents by technological class, where the category NANOTECH refers to patents related to nanotechnologies, AI to artificial intelligence, AGR agriculture, PHARMA to pharmaceutical, ENV to environment, MEDICAL to medical, ICT to information and communication technologies, and OTHER includes patents not identified in any of the previous categories.
Source: OECD calculations.

Table 3.3: Geographic Coverage of the OECD Start-up database, by country

	Start-ups	VC-backed	Mean VC financing	Total VC financing
Australia	22 154	2 348	957	5 740
Austria	2 340	486	144	865
Belgium	6 373	726	354	2 123
Brazil	11 284	2 129	1 195	7 172
Canada	32 494	6 334	2 362	14 171
Chile	1 556	640	43	256
China*	24 688	16 400	46 154	276 921
Colombia	1 462	371	299	1 794
Costa Rica	313	42	1	8
Czech Republic	2 526	261	21	126
Denmark	6 224	883	1 037	6 223
Estonia	1 959	367	99	596
Finland	4 960	905	415	2 489
France	30 547	4 619	2 520	15 122
Germany	39 976	4 256	3 153	18 918
Greece	1 355	148	14	87
Hungary	1947	389	42	253
Iceland	429	92	58	349
India	47 807	7 487	6 795	40 771
Ireland	4 980	1 087	436	2 616
Israel	11 794	2 810	2 486	14 918
Italy	15 964	1 268	265	1 590
Japan	19 689	3 390	2 096	12 575
Korea	9 831	1 292	905	5 429
Latvia	690	186	18	107
Lithuania	1 517	212	59	354
Luxembourg	636	121	136	815
Mexico	3 036	813	293	1 758
Netherlands	32 848	1 910	762	4 572
New Zealand	6 091	308	106	635
Norway	5 028	568	213	1 275
Poland	5 421	853	111	666
Portugal	3 576	485	62	370
Russia	5 134	1 361	172	1 031
Slovak Republic	1 208	92	9	55
Slovenia	848	89	9	51
South Africa	3 667	475	131	786
Spain	17 717	2 406	679	4 073
Sweden	10 457	1 802	826	4 954
Switzerland	9 546	1 386	987	5 924
Turkey	4 569	866	215	1 287
United Kingdom	71 451	10 226	6 522	39 130
United States	308 422	68 929	78 408	470 446
Other countries	61 048	10 802	6 713	40 280

Note: Key statistics on Venture capital funding for start-ups from OECD countries plus 37 partners and non-member countries that have received at least VC financing. Venture capital deals include pre-seed, seed, angel, series funding, growth funding, late stage funding. Excludes deals that do not have information on funding amounts. Values of deals are expressed in constant 2005 USD dollars. China stands for China (People's Republic of)
Source: OECD Start-up database.

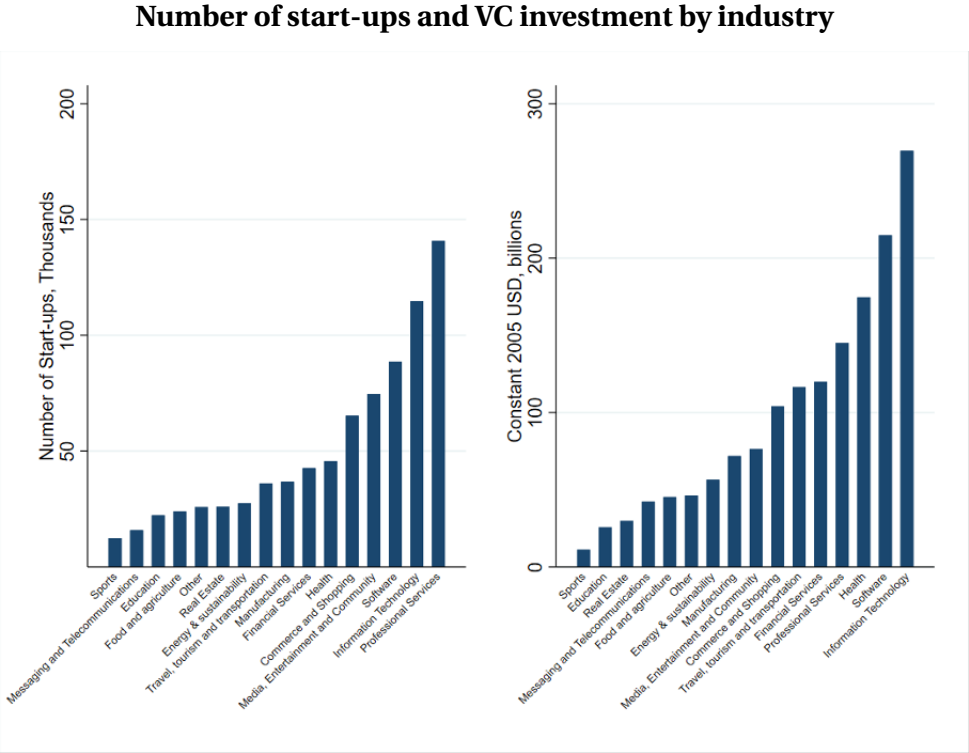


Figure 3.4: The figures above include start-ups from OECD countries plus 37 partners and non-member countries. The left panel shows the number of start-ups in the OECD start-up database, by industry. The right panel shows the total VC financing (billions 2005 USD) by industry. Venture capital deals include pre-seed, seed, angel, series funding, growth funding, late stage funding. It excludes deals that have not information on funding amount.
Source: OECD calculations.

start-ups operating in the IT and software sectors received most of the VC funding. Interestingly, although start-ups working in the health sector rank 6th in terms of number of stat-ups, they are the ranked third in terms of VC received, suggesting that these firms receive larger funding rounds and/or more deals. Finally, Figure 3.5 shows the changes in VC financing by industry between the periods 2011-2015 and 2016-2021, by industry. Interestingly, VC investments into start-ups working in real state, health, and education has almost doubled between those periods.

3.6 Benchmarking

To assess its quality, coverage and representativeness, the OECD Start-up database is compared with a number of external sources at the international and national level, that represent well-established sources for start-up activity. These sources include other proprietary vendors, aggregate data from institutional sources, and national or regional venture capital associations.

Changes in venture capital between 2011-2015 and 2016-2021, by industry

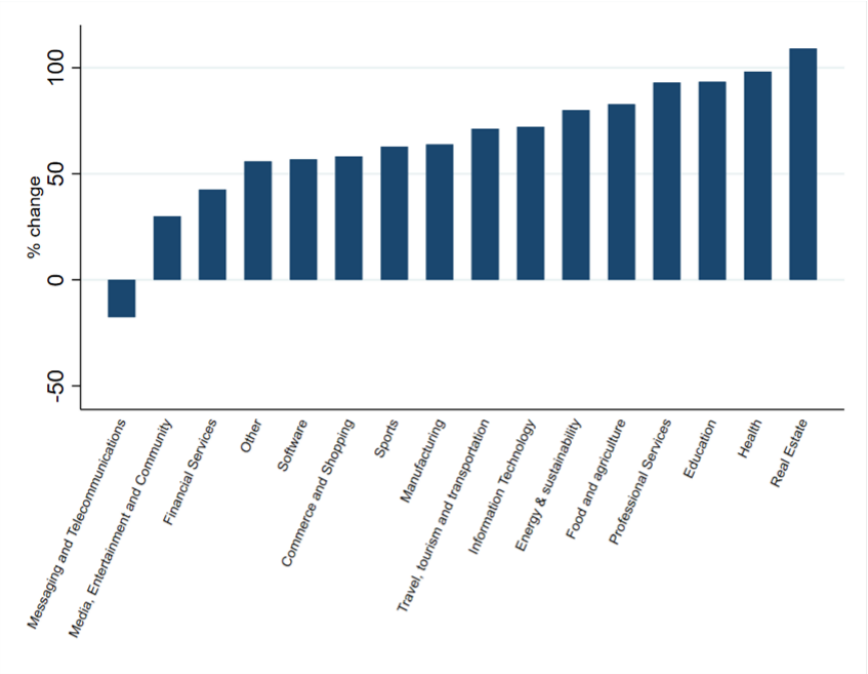


Figure 3.5: The figure above include start-ups from OECD countries plus 37 partners and non-member countries. Changes in total venture capital funding between period between 2011-2015 and 2016-2021, by industry. Venture capital deals include pre-seed, seed, angel, series funding, growth funding, late stage funding. It excludes deals that have not information on funding amount.
Source: OECD calculations.

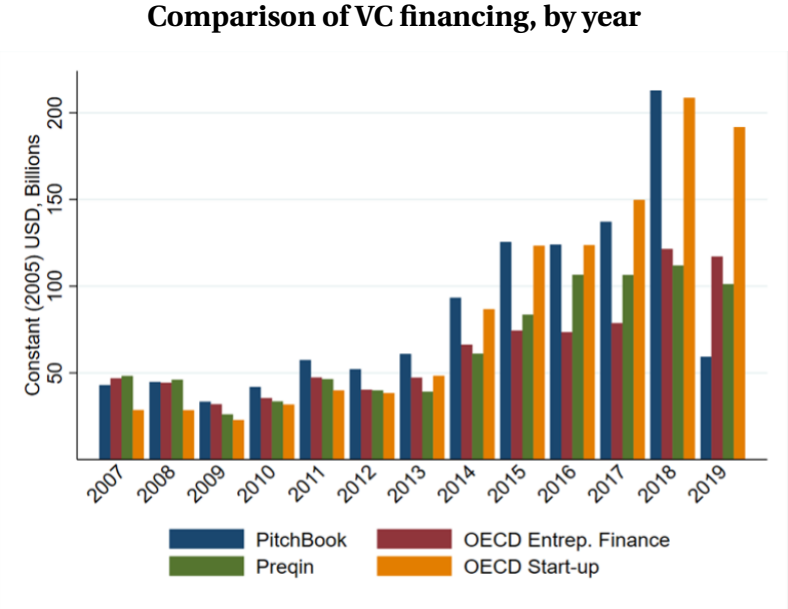


Figure 3.6: The figure reports the total VC financing per year provided by the OECD Start-up database and other data sources (PitchBook, OECD Entrepreneur Finance, Prequin). The figure includes all countries covered by the respective sources. Venture capital deals from the OECD Start-up database include pre-seed, seed, angel, series funding, growth funding, late stage funding. Values of deals are expressed in 2005 USD dollars. OECD Entrepreneur Finance includes financing only from OECD countries, while other sources use worldwide data.
Source: OECD calculations, OECD Entrepreneurship Finance database, PitchBook and Preqin.

In order to facilitate a comparison, only the years of overlap between the different sources are considered in each of the following figures. As a starting point, Figure 3.6 compares aggregate yearly VC investments from the OECD start-up database, PitchBook and Preqin (two other commercial databases), and to the OECD Entrepreneurship Finance database. Although the OECD Start-up database covers around 60% of the amounts reported by other sources for the years 2007-2012, the coverage improves markedly from 2014 and outpaces the other sources in 2017-2019.

Figure 3.7 compares VC investments by deal stage (early vs. late VC) between the OECD start-up database and Preqin. Between the years 2007 and 2013, both sources have comparable coverage, with OECD database having considerably more coverage in the years 2014 for both early and late stage deals.

Figure 3.8 compares the data from the OECD start-up database against other sources for firms located in the United States. The graph shows that the OECD database's coverage is similar to other established commercial sources that have a strong focus on the US market (PwC MoneyTree Report and Pitchbook).

The OECD Start-up database has also been compared to country and regional specific

VC financing between OECD start-up database and Prequin, by stage and year

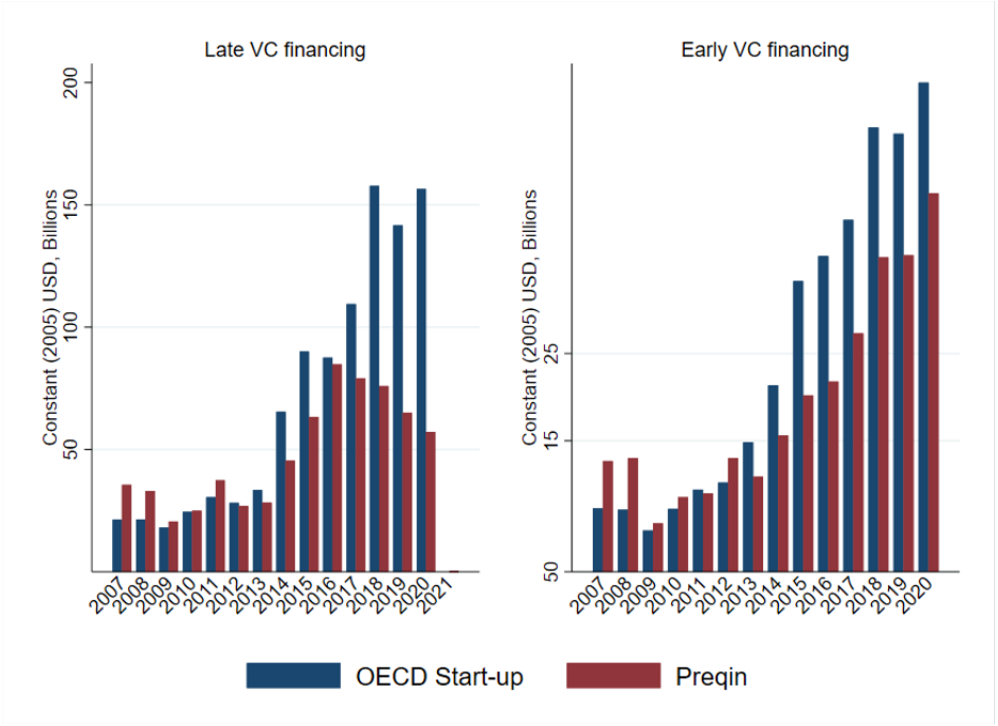


Figure 3.7: The figure reports the total VC financing per stage and year provided by the OECD Start-up database and Prequin. The figure includes all countries covered by the respective sources. Early venture capital deals from the OECD Start-up database include pre-seed, seed, angel, and series A funding. Late stage deals include: Series B-Z, growth funding, late stage funding. Early stage deals from Preqin include: Early Stage, Early Stage: Seed, Early Stage: Start-up. All other VC deals are labelled as later VC. Values of deals are expressed in 2005 USD dollars.
Source: OECD calculations and Preqin.

Comparison of VC financing in the United States, by year

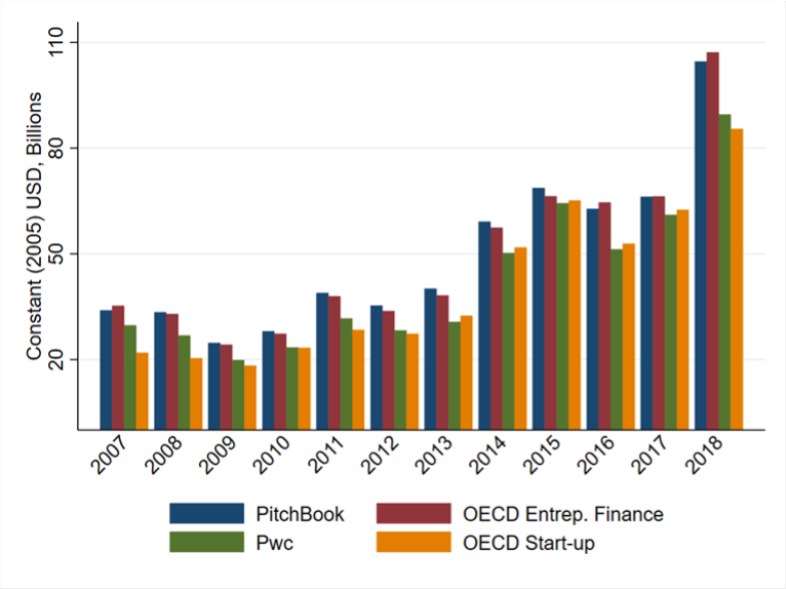


Figure 3.8: The figure reports the total VC financing in the United States per year provided by the OECD Start-up database and other data sources (PitchBook, OECD Entrepreneur Finance, Pwc). Venture capital deals from the OECD Start-up database include pre-seed, seed, angel, series funding, growth funding, late stage funding. Values of deals are expressed in 2005 USD dollars.
Source: OECD calculations, OECD Entrepreneur Finance, PitchBook and Pwc.

reports. Figure 3.8 shows the amount of VC reported in the OECD Start-up database from ventures based in Argentina, Brazil, Chile, Colombia and Mexico. These figures have been directly compared to the statistics reported by the Latin America Venture Capital Association, showing that the OECD Start-up database has a good coverage for these countries Figure 3.8 does a similar comparison for Israel. It compares the amount of VC available in the OECD Start-up database to those provided by Israeli High-Tech Funding Report, edited by IVC research center (IVC Research Center, 2020) . Until 2016, the OECD Start-up database has a good coverage relative to the statistics in the national report, while in most recent years the gap between the two sources widens. Finally, Figure 3.10 looks at the number of new start-ups founded each year in Estonia comparing the figures available from the OECD Start-up database to those obtained from the online platform start-up Estonia.

The benchmarcking exercise shows that, from a global persepctive, the OECD start-up database typically has higher coverage than many other aggregate sources and, importantly, the coverage has improved over most recent years. Nevertheless, the comparison of the OECD start-up database against national sources of data has highilighted that, for some cases and in some years, there are significant differences in the coverage between the OECD start-up

Comparison of VC financing in Latin America, by year

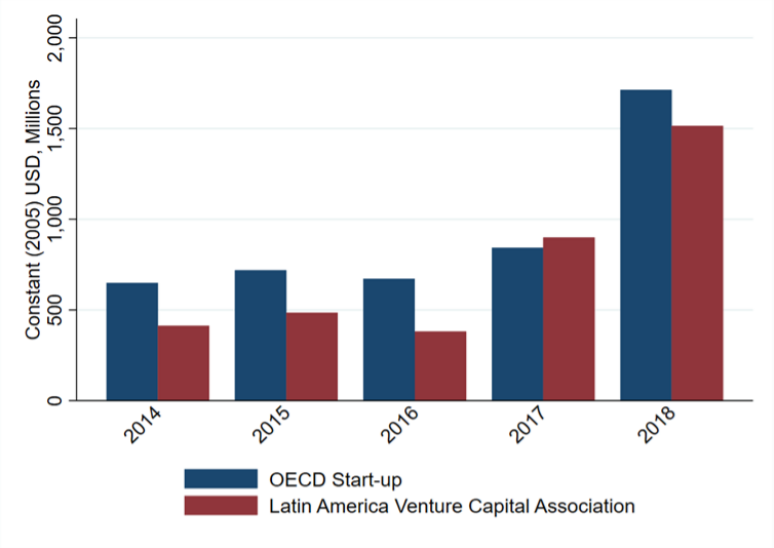


Figure 3.9: The figure reports the total VC financing per year provided by the OECD Start-up database and a report from the Latin America Venture Capital Association. Countries included in the analysis are: Argentina, Brazil, Colombia, Chile, Mexico. Venture capital deals from the OECD Start-up database include pre-seed, seed, angel, series funding, growth funding, late stage funding. Values of deals are expressed in 2005 USD dollars.
Source: OECD calculations and Latin America Venture Capital Association.

Comparison of VC financing in Israel, by year

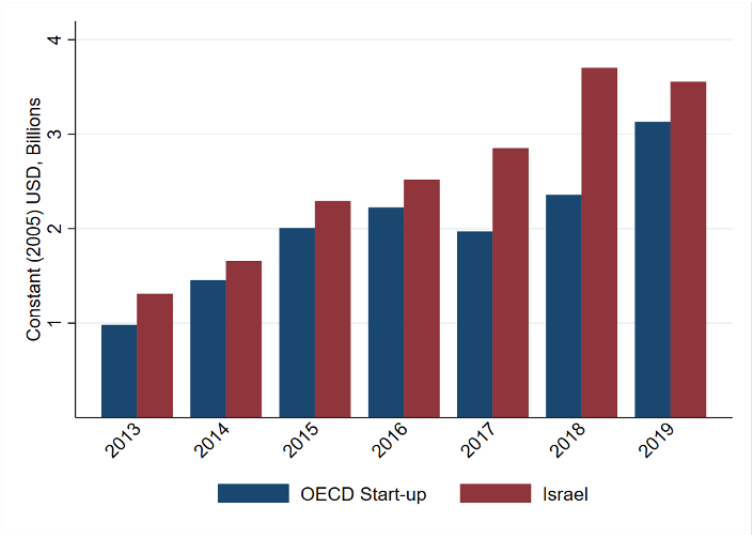


Figure 3.10: The figure reports the total VC financing in Israel per year provided by the OECD Start-up database and the report Israeli High-Tech Funding Report, edited by IVC research center. The last year from the IVC report is incomplete and only report data for the first 3 quarters of the year. Venture capital deals from the OECD Start-up database include pre-seed, seed, angel, series funding, growth funding, late stage funding. Values of deals are expressed in 2005 USD dollars.
Source: OECD calculations and IVC research center.

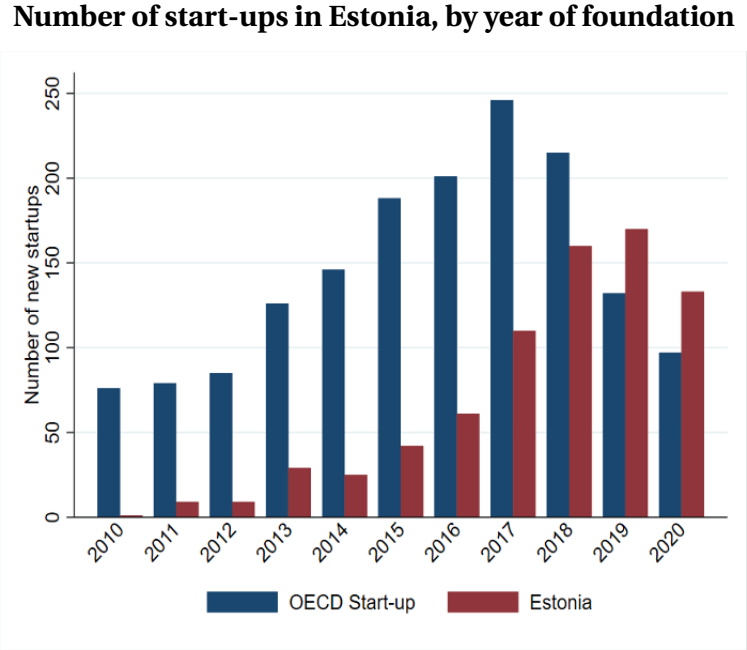


Figure 3.11: The figure shows the total number of start-ups founded each year in Estonia available in the OECD Start-up database and from the online platform start-up Estonia. Source: OECD calculations and Estonian Start-up database.

datbase and the national sources which will warrant further investagion.

3.7 Conclusions

With detailed micro-level information on approximately 900 000 companies, the OECD start-up database provides an unprecedented, timely, and comprehensive source of data on innovative start-ups and venture capital financing. It provides numerous opportunities for policy and academic research, since it represents a unique source to track innovative firms across OECD countries and to identify new start-ups that are developing technologies to address society's grand challenges, including the green transition and the digital transformation. This is particularly important in the current demographic and economic setting, in which structural shifts in the global economy have produced new challenges and opportunities for virtually all players, stakeholders, and community members of the innovative ecosystem. Therefore, the OECD start-up database can provide timely information that are essential for policymakers to assess the effectiveness of existing policies aimed at fostering start-ups and to develop new policies to address the emerging needs of these firms.

Notwithstanding the promising research opportunities provided by the OECD start-up database, there are several important limitations that must be considered when using the database for

policy recommendations. First, while the database focuses on innovative start-ups, the underlying data are built using information assembled by the data providers for commercial purposes and not economic research or policy analysis. Therefore, the underlying sample has not always been compiled applying a coherent framework and it may vary based on the context, geographic focus, and purpose of the study. This calls for caution when using the entire sample of firms available. To address this issue, previous OECD work has used selected criteria to define and identify innovative start-ups, including using indicators related to the founder's academic degree, venture capital financing, and patenting activity. Restricting the sample following these criteria, however, can only allow for an analysis that aims at explaining differences across innovative firms rather than provide evidence of what makes a firm innovative.

Second, the OECD start-up database is more likely to include more successful start-ups than start-ups that fail. This is because the underlying data sources behind the database rely on online presence, business records, and external financing as main inputs for their data. As a result, firms that do not survive or never receive funding— particularly those created before 2010—are less likely to be included in the database.

Third, depending on the question at hand, data from the OECD start-up database might require further processing. One example concerns the missing information about VC amounts, which is not available in around 20-25% of deals. Another potential issue is the lack of a direct mapping between the industries listed in Crunchbase and Dealroom and official classifications such as NACE, NAICS, or ISIC. This issue is present in all studies looking at innovative stat-ups, as traditional classifications generally cannot capture the ever-changing fields of start-ups.

Fourth, while the aggregate figures on new start-ups and venture capital are, on average, consistent with figures from other external sources, there are some important differences worth highlighting. Crunchbase and Dealroom were founded in 2007 and 2013 respectively, and therefore their coverage preceding those dates might not be as comprehensive as other more established sources, as reported in section 3.6. For the same reason, start-ups that failed and ceased operations are unlikely to have been included in the database, particularly those preceding 2007. Thus, there is a risk of spurious accelerating growth of venture capital deals, calling for caution in drawing conclusions from the examination of trends over time, especially before 2010. Finally, although the data provide a wealth of information on firms, besides patenting data, there is little information on the financial outcomes of firms. One standard and important indicator of early entrepreneurial success available in the database is the amount of funding

raised, or whether it was acquired, or if it went public. Additional information, such as the number of employees, revenues growth, and valuation, are scarce and not consistent across countries.

Future work on the OECD database aims at focusing on two streams: better identifying innovative start-ups and complement the data with additional sources to measure innovation. Current measures to identify innovative start-ups rely on financing or patent filings to identify innovative start-ups, thus excluding all innovative start-ups that do not attract VC interest or use other methods than formal intellectual property (IP) to protect their innovations. The identification can instead rely on proprietary machine learning algorithms which seek to identify innovative firms from administrative data or from firms' websites or description of products, services and activities (Kinne and Lenz (2021), Catalini et al. (2019), Guzman and Stern (2020)). The output of this work will contribute to identify, near real-time, new technological developments across countries. At the same time, current OECD work is looking at alternative methods to measure innovation, in order to complement the information provided by patent data and overcome some of its limitation, such as the inherent lag between application and publication or the fact that patents only capture part of the codified knowledge of firms. Ongoing work is examining other forms of intellectual property protection (trademarks and designs), contributions to open access projects and measures of communication activity (through e.g. mobile app downloads or Twitter followers).

Notwithstanding its limitations, the OECD start-up database, along with other sources based on representative statistics and/or distributed micro-data, provides a useful framework and data infrastructure to explore the whole life cycle of innovative start-ups from creation to scale-up and exit, and to analyse the role of public policies to support innovative entrepreneurship.

Bibliography

Acemoglu, D., Johnson, S., and Robinson, J. (2005). The rise of europe: Atlantic trade, institutional change, and economic growth. *American Economic Review*, 95(3):546–579.

Ackerberg, D. A., Caves, K., and Frazer, G. (2015). Identification properties of recent production function estimators. *Econometrica*, 83(6):2411–2451.

Alperovych, Y., Quas, A., and Standaert, T. (2018). Direct and indirect government venture capital investments in europe. *Economics Bulletin*, 38(2):1219–1230.

Altomonte, C., Favoino, D., Morlacco, M., Sonno, T., et al. (2021a). *Markups, intangible capital and heterogeneous financial frictions*. Centre for Economic Performance, London School of Economics and Political ....

Altomonte, C., Ottaviano, G., Rungi, A., and Soon, T. (2021b). Business groups as knowledge-based hierarchies of firms.

Andrews, D., Criscuolo, C., and Gal, P. N. (2015). Frontier firms, technology diffusion and public policy: Micro evidence from oecd countries.

Antràs, P. (2016). *Global Production: Firms, Contracts, and Trade Structure*. Princeton University Press, Princeton, NJ.

Antràs, P. and Helpman, E. (2004). Global sourcing. *Journal of Political Economy*, 112(3):552–580.

Atalay, E., Hortaçsu, A., and Syverson, C. (2014). Vertical integration and input flows. *American Economic Review*, 104(4):1120–48.

Awokuse, T. O. and Yin, H. (2010). Does stronger intellectual property rights protection induce more bilateral trade? evidence from china's imports. *World Development*, 38(8):1094–1104.

Bajgar, M., Berlingieri, G., Calligaris, S., Criscuolo, C., and Timmis, J. (2019). Industry concentration in europe and north america.

Beck, T. (2003). Financial dependence and international trade. *Review of International Economics*, 11(2):296–316.

Belenzon, S. and Berkovitz, T. (2010). Innovation in business groups. *Management Science*, 56(3):519–535.

Bena, J. and Li, K. (2014). Corporate innovations and mergers and acquisitions. *The Journal of Finance*, 69(5):1923–1960.

Berlingieri, G., Calligaris, S., Criscuolo, C., and Verlhac, R. (2020). Laggard firms, technology diffusion and its structural and policy determinants.

Berlingieri, G., Pisch, F., and Steinwender, C. (2021). Organizing global supply chains: Input-output linkages and vertical integration. *Journal of the European Economic Association*, 19(3):1816–1852.

Bertoni, F., Colombo, M. G., and Quas, A. (2015). The patterns of venture capital investment in europe. *Small business economics*, 45(3):543–560.

Bilir, L. K. and Morales, E. (2020). Innovation in the global firm. *Journal of Political Economy*, 128(4):1566–1625.

Bloom, N., Sadun, R., and Van Reenen, J. (2016). Management as a technology? Technical report, National Bureau of Economic Research.

Brander, J. A., Du, Q., and Hellmann, T. (2015). The effects of government-sponsored venture capital: international evidence. *Review of Finance*, 19(2):571–618.

Branstetter, L. (2006). Is foreign direct investment a channel of knowledge spillovers? evidence from japan's fdi in the united states. *Journal of International Economics*, 68(2):325–344.

Branstetter, L., Fisman, R., Foley, C. F., and Saggi, K. (2007). Intellectual property rights, imitation, and foreign direct investment: Theory and evidence. Technical report, National Bureau of Economic Research.

Branstetter, L., Fisman, R., Foley, C. F., and Saggi, K. (2011). Does intellectual property rights reform spur industrial development? *Journal of International Economics*, 83(1):27–36.

Branstetter, L. G., Fisman, R., and Foley, C. F. (2006). Do stronger intellectual property rights increase international technology transfer? empirical evidence from us firm-level panel data. *Quarterly Journal of Economics*, 121(1):321–349.

Breschi, S., Lassébie, J., and Menon, C. (2018). A portrait of innovative start-ups across countries.

Cai, J. and Stoyanov, A. (2016). Population aging and comparative advantage. *Journal of International Economics*, 102:1–21.

Caliendo, L., Mion, G., Opromolla, L. D., and Rossi-Hansberg, E. (2020). Productivity and organization in portuguese firms. *Journal of Political Economy*, 128(11):4211–4257.

Calligaris, S., Criscuolo, C., and Marcolin, L. (2018). Mark-ups in the digital era.

Calvino, F., Criscuolo, C., and Menon, C. (2018). A cross-country analysis of start-up employment dynamics. *Industrial and Corporate Change*, 27(4):677–698.

Calzolari, G., Felli, L., Koenen, J., Spagnolo, G., and Stahl, K. O. (2015). Trust, competition and innovation: theory and evidence from german car manufacturers.

Campi, M. and Dueñas, M. (2019). Intellectual property rights, trade agreements, and international trade. *Research Policy*, 48(3):531–545.

Cannone, G. and Ughetto, E. (2014). Born globals: A cross-country survey on high-tech start-ups. *International Business Review*, 23(1):272–283.

Cassiman, B. and Golovko, E. (2011). Innovation and internationalization through exports. *Journal of International Business Studies*, 42(1):56–75.

Catalini, C., Guzman, J., and Stern, S. (2019). Hidden in plain sight: venture growth with or without venture capital. Technical report, National Bureau of Economic Research.

Chen, Q., Hsu, D. H., and Zvilichovsky, D. (2020). Inventor commingling and innovation in technology startup mergers & acquisitions. *under review*.

Chen, Y. and Puttitanun, T. (2005). Intellectual property rights and innovation in developing countries. *Journal of Development Economics*, 78(2):474–493.

Chor, D. (2010). Unpacking sources of comparative advantage: A quantitative approach. *Journal of International Economics*, 82(2):152–167.

Co, C. Y. (2004). Do patent rights regimes matter? *Review of International Economics*, 12(3):359–373.

Colombo, M. G., D'Adda, D., and Piva, E. (2010). The contribution of university research to the growth of academic start-ups: an empirical analysis. *The Journal of Technology Transfer*, 35(1):113–140.

Conti, A., Peukert, C., and Roche, M. P. (2021). Beefing it up for your investor? open sourcing and startup funding: Evidence from github. *Open Sourcing and Startup Funding: Evidence from GitHub (August 25, 2021)*.

Costinot, A. (2009). On the origins of comparative advantage. *Journal of International Economics*, 77(2):255–264.

Cripps, H., Singh, A., Mejtoft, T., and Salo, J. (2020). The use of twitter for innovation in business markets. *Marketing Intelligence & Planning*.

Crouzet, N. and Eberly, J. C. (2019). Understanding weak capital investment: The role of market concentration and intangibles. Technical report, National Bureau of Economic Research.

Cuñat, A. and Melitz, M. J. (2012). Volatility, labor market flexibility, and the pattern of comparative advantage. *Journal of the European Economic Association*, 10(2):225–254.

Cunningham, C., Ederer, F., and Ma, S. (2021). Killer acquisitions. *Journal of Political Economy*, 129(3):649–702.

Da Rin, M., Hellmann, T., and Puri, M. (2011). A survey of venture capital research: Center discussion paper. *Tilburg University*.

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

De Ridder, M. (2019). Market power and innovation in the intangible economy.

Dechezleprêtre, A. and Fadic, M. (2022). Can government venture capital help bring research to the market?

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

Delgado, M., Kyle, M., and McGahan, A. M. (2013). Intellectual property protection and the geography of trade. *Journal of Industrial Economics*, 61(3):733–762.

den Besten, M. L. (2020). Crunchbase research: Monitoring entrepreneurship research in the age of big data. *Available at SSRN 3724395*.

Do, Q.-T. and Levchenko, A. A. (2009). Trade, inequality, and the political economy of institutions. *Journal of Economic Theory*, 144(4):1489–1520.

Etzkowitz, H. (2003). Research groups as 'quasi-firms': the invention of the entrepreneurial university. *Research policy*, 32(1):109–121.

Feenstra, R. C., Inklaar, R., and Timmer, M. P. (2015). The next generation of the penn world table. *American Economic Review*, 105(10):3150–82.

Gaddy, B. E., Sivaram, V., Jones, T. B., and Wayman, L. (2017). Venture capital and cleantech: The wrong model for energy innovation. *Energy Policy*, 102:385–395.

Gal, P. N. (2013). Measuring total factor productivity at the firm level using oecd-orbis.

Garicano, L. (2000). Hierarchies and the organization of knowledge in production. *Journal of political economy*, 108(5):874–904.

Garicano, L. and Rossi-Hansberg, E. (2006). Organization and inequality in a knowledge economy. *The Quarterly journal of economics*, 121(4):1383–1435.

Gautier, A. and Lamesch, J. (2021). Mergers in the digital economy. *Information Economics and Policy*, 54:100890.

Ginarte, J. C. and Park, W. G. (1997). Determinants of patent rights: A cross-national study. *Research Policy*, 26(3):283–301.

Gravelle, J. (2015). Reform of us international taxation: Alternatives. Library of Congress, Congressional Research Service.

Griliches, Z. (1998). Patent statistics as economic indicators: a survey. In *R&D and productivity: the econometric evidence*, pages 287–343. University of Chicago Press.

Grimaldi, R., Kenney, M., Siegel, D. S., and Wright, M. (2011). 30 years after bayh–dole: Reassessing academic entrepreneurship. *Research policy*, 40(8):1045–1057.

Guzman, J. and Stern, S. (2020). The state of american entrepreneurship: New estimates of the quantity and quality of entrepreneurship for 32 us states, 1988–2014. *American Economic Journal: Economic Policy*, 12(4):212–43.

Gwartney, J., Lawson, R., and Norton, S. (2008). *Economic freedom of the world: 2008 annual report*. Fraser Institute.

Hall, B., Helmers, C., Rogers, M., and Sena, V. (2013). The choice between formal and informal intellectual property: a review. *Journal of Economic Literature (forthcoming)*.

Hall, B. H., Jaffe, A., and Trajtenberg, M. (2005). Market value and patent citations. *RAND Journal of economics*, pages 16–38.

Haltiwanger, J., Jarmin, R. S., and Miranda, J. (2013). Who creates jobs? small versus large versus young. *Review of Economics and Statistics*, 95(2):347–361.

Haskel, J. and Westlake, S. (2017). *Capitalism without capital*. Princeton University Press.

Holmstrom, B. and Roberts, J. (1998). The boundaries of the firm revisited. *Journal of Economic perspectives*, 12(4):73–94.

Howell, S. T. (2021). Learning from feedback: Evidence from new ventures. *Review of Finance*, 25(3):595–627.

Hu, A. G. and Png, I. P. (2013). Patent rights and economic growth: evidence from cross-country panels of manufacturing industries. *Oxford Economic Papers*, 65(3):675–698.

Ivus, O. (2010). Do stronger patent rights raise high-tech exports to the developing world? *Journal of International Economics*, 81(1):38–47.

Ivus, O. (2011). Trade-related intellectual property rights: industry variation and technology diffusion. *Canadian Journal of Economics*, 44(1):201–226.

Ivus, O. (2015). Does stronger patent protection increase export variety? evidence from us product-level data. *Journal of International Business Studies*, 46(6):724–731.

Ivus, O. and Park, W. (2019). Patent reforms and exporter behaviour: Firm-level evidence from developing countries. *Journal of the Japanese and International Economies*, 51:129–147.

Ivus, O., Park, W. G., and Saggi, K. (2017). Patent protection and the composition of multinational activity: Evidence from us multinational firms. *Journal of International Business Studies*, 48(7):808–836.

Ivus, O., Saggi, K., and Park, W. (2016). Intellectual property protection and the industrial composition of multinational activity. *Economic Inquiry*, 54(2):1068–1085.

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406):414–420.

Jaro, M. A. (1995). Probabilistic linkage of large public health data files. *Statistics in medicine*, 14(5-7):491–498.

Javorcik, B. S. (2004). The composition of foreign direct investment and protection of intellectual property rights: Evidence from transition economies. *European Economic Review*, 48:39–62.

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

Kaplan, S. N. and Lerner, J. (2010). It ain't broke: The past, present, and future of venture capital. *Journal of Applied Corporate Finance*, 22(2):36–47.

Kaufmann, D., Kraay, A., and Mastruzzi, M. (2009). *Governance matters VIII: Aggregate and individual governance indicators 1996-2008*. World Bank.

Kinne, J. and Lenz, D. (2021). Predicting innovative firms using web mining and deep learning. *PloS one*, 16(4):e0249071.

Lassébie, J., Sakha, S., Kozluk, T., Menon, C., Breschi, S., and Johnstone, N. (2019). Levelling the playing field: Dissecting the gender gap in the funding of start-ups.

Levchenko, A. A. (2007). Institutional quality and international trade. *Review of Economic Studies*, 74(3):791–819.

Levenshtein, V. I. et al. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Lin, J. X. and Lincoln, W. F. (2017). Pirate's treasure. *Journal of International Economics*, 109:235–245.

Ma, Y., Qu, B., and Zhang, Y. (2010). Judicial quality, contract intensity and trade: Firm-level evidence from developing and transition countries. *Journal of Comparative Economics*, 38(2):146–159.

Manova, K. (2008). Credit constraints, equity market liberalizations and international trade. *Journal of International Economics*, 76(1):33–47.

Maskus, K. E. and Penubarti, M. (1995). How trade-related are intellectual property rights? *Journal of International Economics*, 39(3-4):227–248.

Maskus, K. E. and Ridley, W. (2016). Intellectual property-related preferential trade agreements and the composition of trade. Robert Schuman Centre for Advanced Studies Research Papers, Nr. 35.

Maskus, K. E. and Yang, L. (2018). Domestic patent rights, access to technologies and the structure of exports. *Canadian Journal of Economics/Revue Canadienne d'Économique*, 51(2):483–509.

Nunn, N. (2007). Relationship-specificity, incomplete contracts, and the pattern of trade. *Quarterly Journal of Economics*, 122(2):569–600.

Nunn, N. and Trefler, D. (2014). Domestic institutions as a source of comparative advantage. In *Handbook of International Economics*, volume 4, pages 263–315. Elsevier.

Park, W. G. (2008). International patent protection: 1960–2005. *Research policy*, 37(4):761–766.

Phillips, G. M. and Zhdanov, A. (2013). R&d and the incentives from merger and acquisition activity. *The Review of Financial Studies*, 26(1):34–78.

Qian, Y. (2007). Do national patent laws stimulate domestic innovation in a global patenting environment? a cross-country analysis of pharmaceutical patent protection, 1978–2002. *Review of Economics and Statistics*, 89(3):436–453.

Rafiquzzaman, M. (2002). The impact of patent rights on international trade: Evidence from canada. *Canadian Journal of Economics/Revue Canadienne d'Économique*, 35(2):307–330.

Rajan, R. and Zingales, L. (1998). Financial dependence and growth. *American Economic Review*, 88(3):559–86.

Ramondo, N., Rappoport, V., and Ruhl, K. J. (2016). Intrafirm trade and vertical fragmentation in us multinational corporations. *Journal of International Economics*, 98:51–59.

Röhm, P., Merz, M., and Kuckertz, A. (2020). Identifying corporate venture capital investors–a data-cleaning procedure. *Finance Research Letters*, 32:101092.

Romalis, J. (2004). Factor proportions and the structure of commodity trade. *American Economic Review*, 94(1):67–97.

Santacreu, A. M. (2021a). Dynamic gains from trade agreements with intellectual property provisions. mimeo, Federal Reserve Bank of Saint Louis.

Santacreu, A. M. (2021b). International technology licensing, intellectual property rights, and tax havens. mimeo, Federal Reserve Bank of Saint Louis.

Sevilir, M. and Tian, X. (2012). Acquiring innovation. In *AFA 2012 Chicago Meetings Paper*.

Shin, W., Lee, K., and Park, W. G. (2016). When an importer's protection of ipr interacts with an exporter's level of technology: Comparing the impacts on the exports of the north and south. *World Economy*, 39(6):772–802.

Smith, P. J. (1999). Are weak patent rights a barrier to us exports? *Journal of International Economics*, 48(1):151–177.

Smith, P. J. (2001). How do foreign patent rights affect us exports, affiliate sales, and licenses? *Journal of International Economics*, 55(2):411–439.

Sonno, T. (2020). Globalization and conflicts: the good, the bad and the ugly of corporations in africa.

Squicciarini, M. and Dernis, H. (2013). A cross-country characterisation of the patenting behaviour of firms based on matched firm and patent data.

Squicciarini, M., Dernis, H., and Criscuolo, C. (2013). Measuring patent quality: Indicators of technological and economic value.

Tarasconi, G. and Menon, C. (2017). Matching crunchbase with patent data.

UNCTAD, M. T. (2009). on statistics for fdi and the operations of tncs. *INDIVIDUAL STUDIES*.

Winkler, W. E. (1999). The state of record linkage and current research problems. In *Statistical Research Division, US Census Bureau*. Citeseer.

Appendix A

Appendix

A.1 Appendix Chapter 1 - Institutions, Development, and Patterns of Trade

A.1.1 List of Countries and the Year of IPR Reform

OECD countries	non-OECD countries
Australia, Austria, Belgium, Canada, Chile, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Lithuania, Luxembourg, Mexico, Netherlands, New Zealand, Norway, Poland, Portugal, South Korea, Slovakia, Spain, Sweden, Switzerland, Turkey, USA, United Kingdom	Algeria, Angola, Argentina, Bolivia, Botswana, Brazil, Bulgaria, Burundi, Cameroon, China Colombia, Congo, Costa Rica, Cyprus, Dominican Republic, Ecuador, Egypt, El Salvador, Ethiopia, Fiji, Guatemala, Honduras, India, Jamaica, Jordan, Madagascar, Malawi, Malaysia, Malta, Mauritius, Nepal, Nicaragua, Pakistan, Panama, Paraguay, Peru, Russia, Rwanda, Singapore, South Africa, Sri Lanka, Tunisia, Uganda, Ukraine, Tanzania, Uruguay, Vietnam, Zambia, Zimbabwe

Table A.1.1: Robustness exercise on IP-intensity

Variable	OECD	NON-OECD
$IPR_c * patents$:	0.0129***	0.0036
	(0.0047)	(0.0038)
$IPR_c * trademarks$:	0.0252*	0.0050
	(0.0136)	(0.0902)
$IPR_c * I_{IPdummy}$:	0.380**	0.141
	(0.167)	(0.114)
Country FE:	Yes	Yes
Industry FE:	Yes	Yes

The dependent variable is the natural log of exports in industry i by country c to all other countries. In all regressions, robust standard errors in brackets are reported. *, ** and *** indicate significance at the 10, 5 and 1 percent level. Each row of the Table represent an alternative estimate of our main interaction of interest $IPR_c * IPint$. In addition, even if not reported, each regression includes all the other variables specified in equation (1.1).

A.1.2 Sensitivity of the IP-intensity Measures

In this section, we propose some additional checks to validate the baseline results reported in Table 1.4.

First of all, we provide a further set of controls to address the sensitivity of our main variable, IP-intensity, to the use of alternative specifications. More specifically, regarding the IP-intensity measure, we first replicate our baseline specification including only the contribution of either patents or trademarks. Then, more important, in the spirit of Branstetter et al. (2006), Ivus (2010), Branstetter et al. (2011), and Delgado et al. (2013), we split the overall IP intensity in high and low IP intensity with the use of a dummy to distinguish between industries above and below the median value. Table A.1.1 illustrates these results and shows that they are unaffected when using these alternative definitions of IP intensity.

Then, in Table A.1.2 we show that the baseline results reported in Table 1.4 are robust to alternative ways of clusteting. In particular, the results remain significant, despite with lower statistical power, using respectively country, industry and two ways clustering at country and industry level.

Table A.1.2: Robustness to alternative clustering

	Country		Industry		Country and Industry	
Variable	OECD	non-OECD	OECD	non-OECD	OECD	non-OECD
IPR int.:	0.014*	0.004	0.014***	0.004	0.014*	0.004
	(0.00765)	(0.00398)	(0.0042)	(0.00302)	(0.00718)	(0.00358)
Skill int.:	11.87**	0.852	11.87**	0.852	11.87*	0.852
	(4.775)	(3.871)	(5.290)	(2.399)	(5.870)	(3.875)
Capital int.:	-0.773	0.0718	-0.773	0.0718	-0.773	0.0718
	(0.929)	(0.624)	(0.581)	(0.412)	(0.928)	(0.666)
Nunn int.:	-0.387	1.026**	-0.387	1.026***	-0.387	1.026**
	(0.374)	(0.413)	(0.290)	(0.278)	(0.417)	(0.439)
Observations:	3317	4576	3317	4576	3317	4576
R-squared:	0.766	0.704	0.766	0.704		
Country FE:	Yes	Yes	Yes	Yes	Yes	Yes
Industry FE:	Yes	Yes	Yes	Yes	Yes	Yes

The dependent variable is the natural log of exports in industry i by country c to all other countries. In the first two columns, standard errors are clustered at exporting country level, in the following two columns, at industry level, and in the last two column, two ways clustering at country and industry level are applied. *, ** and *** indicate significance at the 10, 5 and 1 percent level.

A.1.3 Instrumental Variable using old data on IPR

We propose a further IV strategy to address the concern on reverse causality. We exploit historical IPR protection values that are highly correlated to today's values. Because each country's quality of IPR in 1960 is pre-determined and unaffected by trade flows in 2014, it can be a candidate to isolate exogenous variation in today's quality of IPR institutions. At the same time, the instrument is highly related to our potentially endogenous variable, given the persistency in the quality of institutions across countries. In particular, we regress $IPR_{c,1960} \cdot IPint_i$ on $IPR_{c,2010} \cdot IPint_i$, and used the predicted values $\tilde{IPR}$ as main explanatory variable for the second stage. All additional variables specified in equation 1.1 are included in the first and second stage.[1] The instrument is relevant when we look at the all sample and for OECD countries, as highlighted by the statistics at the bottom of Table (A.1.3), which are all above the critical values, suggesting that old IPR values are a valid instrument for developed countries. The IV coefficient is positive and statistically significant for the all sample and OECD countries, providing support for the importance of IPR institution in shaping comparative advantage and mitigating the potential positive feedback effect that trade might have on IPR enforcement.

[1] In this way we control for possible influences that IPR protection in 1960 could have had on trade values other than through its direct effect on IPR protection level in 2010. In fact, a possible concern for the validity of this instrument is that IPR quality in 1960 may also affect comparative advantage through channels other than IPR quality in 2010, not satisfying the exclusion restriction. For example, IPR in 1960 can be related to other country characteristics, such as GDP, that may have a direct impact on trade flows, see Ginarte and Park (1997) and Chen and Puttitanun (2005).

Table A.1.3: IV Estimation

Second Stage	All Sample	OECD	NON-OECD
$\tilde{IPR}$ interaction:	0.00836***	0.0241**	0.00816
	(0.00297)	(0.0101)	(0.0214)
Skill interaction:	3.539*	12.98***	0.412
	(1.879)	(4.139)	(2.554)
Capital interaction:	-0.300	-1.107	0.128
	(0.228)	(0.746)	(0.405)
Nunn interaction:	0.615***	-0.522**	1.037***
	(0.115)	(0.228)	(0.239)
Observations:	7086	2812	4274
R-squared:	0.772	0.771	0.671
Country FE:	Yes	Yes	Yes
Industry FE:	Yes	Yes	Yes
Year FE:	Yes	Yes	
First stage:			
$IPR_{c,1960} \cdot IPint_i$:	0.61443***	0.33323***	0.1489
	(0.06923)	(0.0398)	(0.12407)
Weak IV test:			
Effective F-statistic	80	75	1.5

The dependent variable in the second stage is the natural log of exports in industry i from country c to all other countries. The first stage dependent variable is the interaction term between IP intensity at industry level and IPR protection quality in 1960. Then, we use the predicted values, $\tilde{IPR}$ interaction, in the second stage. The bottom part of the table reports the coefficient of the IV from the first stage, together with the values of the F-test resulting from the first stage and the endogeneity test. All explanatory variables in the second stage are also included in the first stage, but to conserve space we only report the first stage coefficients for the instrumental variable. *, ** and *** indicate significance at the 10, 5 and 1 percent level.

A.2 Appendix Chapter 2 - Business Groups and Knowledge Flows

A.2.1 Allocation of intangible asset

We implement some robustness checks to reinforce the evidence reported in section 2.3 about the presence of decaying patterns in the allocation of knowledge within BGs. We report graphically only some of these additional results, but regression analysis on the spirit of equation 2.1 is in line. The patterns have in some cases a higher volatility and the jumps from one layer to the next one are less sharp, but the stylized trend is still present.

To provide evidence about how relevant are the differences in the allocation of knowledge between subsidiaries looking only at variation across them, we replicate the analysis excluding the GUO of each group to give more emphasis at the difference between subsidiaries because we only look at variation across them. In figure A.2.1, the omitted category are now firms in layer 1 and, on the horizontal axis, layers range from 2 to 6. Again, the decaying pattern persists, suggesting that firms in higher layers have systematically more intangible asset than firms in lower layers. These coefficients tell us that firms in layer 1 have 36% more intangible asset than firms in layers 2, while for example firms in layer 6 have 70% less intangible asset than firms in layer 1.

Given the high number of firms that do not report any value for intangible asset, we consider these observations as zeros rather than missing values. The result from figure A.2.2 shows that the pattern is still present, despite less pronounced. In Figure A.2.3 we change the somehow arbitrary choice of placing a cap at the maximum number of layers of a group, trying an alterhative threshold of 8 layers, while in Figure A.2.4 we use the variable Intangible asset directly provided by balance sheet data. Finally, in Figure A.2.5 we augment equation 2.1 controlling for employment at firm level to understand whether the observed allocation of knowledge is related to a regularity in the organizational structure of a group or rather is correlated to firms' characteristics, that might otherwise drive the pattern. Qualitatively, the overall stylized pattern is weaker but still present.

Figure A.2.1: Intangible Assets by Hierarchical layers excluding the GUO.

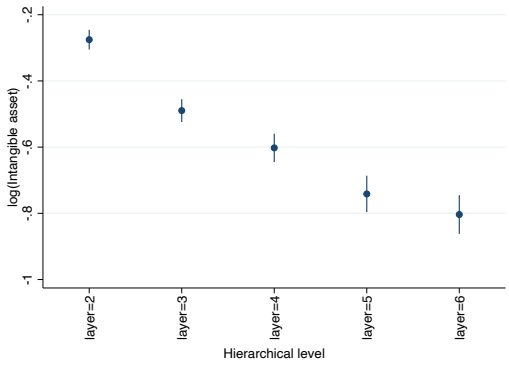


Figure A.2.2: Intangible Assets by Hierarchical layers assigning intangible asset equal to zero when missing.

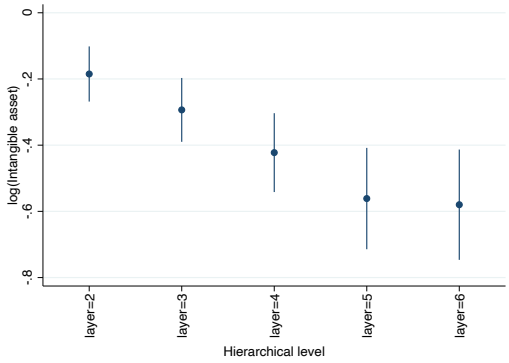


Figure A.2.3: Intangible Assets by Hierarchical layers with maximum layer equal to 8.

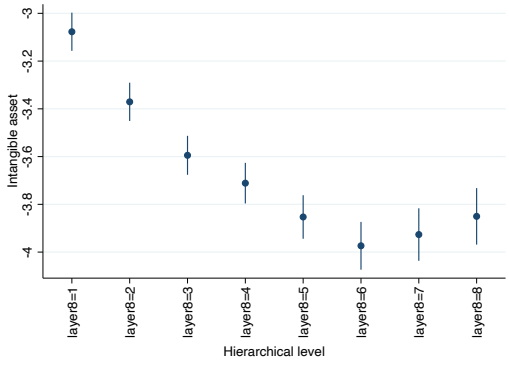


Figure A.2.4: Intangible Assets by Hierarchical layers using intangible asset directly provided by balance sheet.

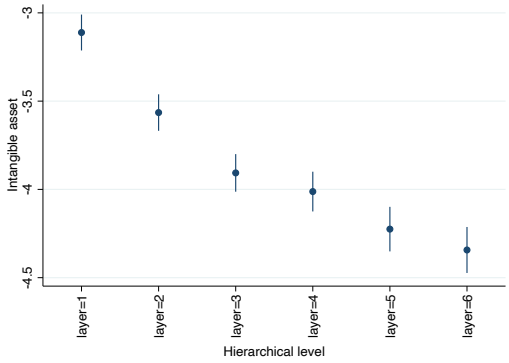
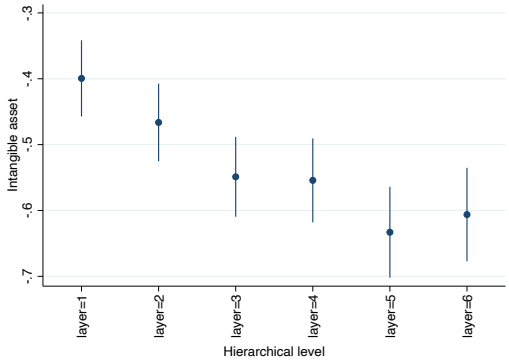


Figure A.2.5: Intangible Assets by Hierarchical layers controlling for employment at firm level.



A.2.2 Minor Robustness Checks

We implement some minor robustness to assess the sensitivity of the results of our main analysis to some potential threats that might bias our results. A potential concern is that multinationals may misreport output or innovation investment for tax purposes. To account for this possibility, we provide estimates that omit affiliates located in known tax havens (identified in Gravelle (2015) and used also by Bilir and Morales (2020). Since our baseline sample is composed only of OECD countries, this specification excludes firms located in Cyprus, Ireland, Luxembourg, Malta, Netherlands and Switzerland. The results are not affected by this sample modification, neither if we simply exclude firms from these countries from the sample neither if we also exclude these firms from the computation of knowledge stocks.

The results are not driven by specific sectors and qualitatively hold both for firms active in services and manufacturing industries. Similarly, the results holds both for domestic and foreign affiliates and for EU based and US based groups.

Finally, we assess whether the results occur only for groups structured in specific ways or rather are more generalised. Thus, we first replicate the previous specification looking only at firms from a specific layer. In alternative, we replicate the exercise as described in equation 2.2 but restricting each time the sample to firms belonging to BGs with a given maximum number of layers. In these two exercises, the sample is very different from specification to specification, making very hard any interpretation. For our purpose, we emphasize that the results are always consistent, in the sense that when we find a significant pattern of knowledge flow, it is always consistent with our baseline result.