Three essays on the role of proximity in science and innovation

THÈSE N^{O} 6463 (2015)

PRÉSENTÉE LE 19 JANVIER 2015

AU COLLÈGE DU MANAGEMENT DE LA TECHNOLOGIE
CHAIRE EN ÉCONOMIE ET MANAGEMENT DE L'INNOVATION
PROGRAMME DOCTORAL EN MANAGEMENT DE LA TECHNOLOGIE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Stefano Horst BARUFFALDI

acceptée sur proposition du jury:

Prof. C. Tucci, président du jury
Prof. D. Foray, directeur de thèse
Prof. M. Feldman, rapporteuse
Prof. K. Frenken, rapporteur
Prof. M. Gruber, rapporteur e

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2015

Alle mie nonne

Acknowledgments

I wish to express my deep gratitude to my advisor, Dominique Foray. He gave me great autonomy while being constantly supportive. His intuitions and suggestions were crucial for the development of this thesis. The support and the resources he provided were undoubtedly extremely valuable. But his trust and help, especially when I mostly needed it, were priceless.

I am grateful to all members of my defense committee: Professor Maryann Feldman, Professor Koen Frenken, Professor Marc Gruber and Professor Christopher Tucci. I feel bound to acknowledge Professor Jacques Mairesse: during his frequent visits he has always been extremely helpful and his passion for Research has been an iconic example. I thank Professor Stuart Graham, Professor Francesco Lissoni and Professor Paula Stephan for their extremely valuable comments and help, any time I had the great chance to meet them. Thanks to Professor Paolo Landoni, who remained over these years an important collaborator and friend.

I thank all my colleagues and friends of the CEMI (extended) family, especially: Fabiana Visentin, Julio Raffo, Marianna Marino, Markus Simeth, Michele Pezzoni, Pierpaolo Perrotta, Monica Coffano, Stephane Lhuillery, and Viviana Munoz. I am particularly indebted with Julio who gave me unconditional help and whose energy, optimism and great skills were fundamental for my PhD. Markus, Marianna, Pierpaolo and Monica have been invaluable friends as well as important reference points any time I was in trouble. All discussions with them were a source of inspiration and motivation. I thank Fabiana for all the work and everything we learned together. A thought to Annamaria Conti, for the joined work on the second paper in this thesis.

I am grateful to my great friends and Pere, Giovanni Liotta and Giuseppe Chindemi, especially for their contagious positive humor (and their food). I leave them to, and aknowledge, Philipp Moser, Giada Baldessarelli and Claudia Pellegrin, with which I also spent great times. I thank Olov Isaksson for all our discussions (and his terrible humor) and Peter Vogel for his kindness since my first day in Lausanne (and for rescuing me in Zurich). Thanks also to all other colleagues and friends who worked or passed by CDM: Ariel Zeballos, Andreas von Vangerow, Melvin Haas, Victoria Nuguer, Toni Männistö, Abhik Mukherjee, Florian Lücker, Pinar Uysal, Chiara Forlati, Jana Thiel, Allan Cabello, Nancy Gonzalez, Tilo Peters, Deep Parekh, Joana Comas, Iro Katsifou, Shuangqing Liao, Carina Lomberg, Vincent Nassar, Nettra Pan, Simone Voldrich, Joana Pereira, Lorenzo Massa, Gianluigi Viscusi, Olivier Waeber, Argyro Nikiforou, Mervegül Kirci, Ekaterina Kuznetsova, Viet Anh Nguyen, Maryam Jahromi, Mohamad Razaghi, Amin Dehdarian, Reinier Verhoog, Ralf Dyllick-Brenzinger, Alessandro Palma, Marta Valsecchi, Sabrina Romeri. Special thanks to my friends from Gerzensee: my flat-Mate Marcel Probst, Veronica Preotu, Annette Harms, Sergio Galletta, Agustin Redonda and Mauro Boffa. I have also greatly appreciated the support of the administrative staff and the IT team at CDM. For their kind welcome and support at ZEW in Mannheim, thanks to Martin Hud, Sandra Gottschalk and Professor Georg Licht.

To all my one life friends, Simo, Fra, Mario, Gae, Gigi, Gigi 2, Salvo, Pedro, Micky, Matteo, Fede, and all others... thanks especially because it feels like I never left when I meet you.

Finally and most importantly: thanks to my family, especially to my parents for their love, support and the freedom they gave me. And to Patricia, for her love and understanding, and for everything: obrigado.

Abstract

This thesis proposes three studies that provide novel empirical evidence on how different types of proximity can affect innovation and science activities through various mechanisms and in different contexts.

In the first study (second chapter of this thesis), in collaboration with Julio Raffo, we analyze the relationship between geography and the likelihood of duplication in inventive activities. We argue that the uneven diffusion of knowledge means that the duplication of inventions will not be randomly distributed geographically and over time. First, as knowledge diffuses over time and competitive incentives decrease, the probability of a claimed invention duplicating an existing one will decrease in the time distance between the two. Second, for recent and upcoming inventions, competitive incentives are high, and localized knowledge flows increase the probability of duplication. Therefore, over a brief time period, the probability of duplication decreases with geographic distance. Conversely, the duplication of less recent inventions is more likely to occur at long distances as a consequence of less awareness of a technology existing due to missing knowledge flows. We test our hypotheses on European Patent Office (EPO) patent bibliographical data on patent citation categories. Geographic distance matters significantly less in sectors in which patents are known to be more effective as a source of information such as discrete technologies.

In the second study (third chapter), in collaboration with Annamaria Conti and Fabiana Visentin, we investigate the effects of professors' social proximity with external universities on the level of productivity of PhD students hired from these universities. Researchers hired from external environments tend to have high scientific productivity compared to those who completed their studies in the same institution where they are employed. In a population of 4,666 PhD students, we further study the scientific productivity of external students from professors' networks, defined as students with a master's degree from a different university from that of their PhD, and also from a university with which their supervisors' co-authors are affiliated. We find that these students are significantly more productive, both compared to other students with a master's degree from a different university, and to students with a master's degree from the same university as that of their PhD. In our analyses, we control for the heterogeneity of supervisors and the heterogeneity of institutions where the students obtained their master's degrees, including proxies for the specific relevance of these universities for a given supervisor. Thus, we conclude that professors hire students with higher scientific productivity from universities where their co-authors are affiliated. Additional analyses further suggest that the reduction of information asymmetries is the main mechanism to explain this finding.

In the third study (fourth chapter), in collaboration with Guillaume Burghouwt, we investigate the role of interregional knowledge integration as a driver of firm innovative performance. We adopt an unbalanced panel of 3,871 innovative companies in Germany between 1992 and 2010, for a total of 15,819 observations, and we study their innovative productivity. In fixed effects estimations, the interregional knowledge integration of regions, measured as the geographic dispersion of the knowledge sources of inventions developed in a region, positively affects innovative productivity of local firms. To address concerns of endogeneity due to the possibility of reverse causality and omitted time-variant variables, we exploit airline liberalization in Germany. We find that the shift from monopolistic to more competitive aeronautic markets positively affected interregional integration. Firms located in regions where the airline liberalization induced a higher level of interregional knowledge integration increased their innovative productivity significantly. We do not find strong differences across firms located in regions with low or high levels of R&D investments.

Keywords

Geography of innovation, proximity, interregional integration, duplication, scientific productivity, innovative productivity.

Sommario

La tesi presenta tre studi che forniscono nuova evidenza empirica su come diverse forme di prossimità inlfuenzano le attività di ricerca e innovazione, attraverso diversi meccanismi.

Nel primo studio (secondo capitolo di questa tesi), in collaborazione con Julio Raffo, viene analizzata la relazione tra geografia e la probabilità di duplicazione in attività innovative. Si ipotizza che la diffusione disuniforme di conoscenza comporta che la duplicazione di invenzioni non sia distribuita casualmente nello spazio geografico e nel tempo. Con il passare del tempo si diffonde conoscenza, l'incentivo a comptere si riduce e la probabilità che un'invenzione duplichi una esistente diminuisce. Secondo, per invenzioni recenti o prossime ad essere scoperte, l'incentivo a competere é elevato e flussi di conoscenza localizzati aumentano la probabilità di duplicazione. Di conseguenza, a breve distanza di tempo, la probabilità di duplicazione diminuisce al crescere della distanza geografica. Al contrario, la duplicazione di invenzioni non recenti avviene piu probabilimente a elevate distanze geografiche in conseguenza della minore consapevolezza dovuta alla mancanza di flussi di conoscenza. Queste ipotesi sono testate su dati di brevetto dell'ufficio dei brevetti europeo (EPO). La distanza geografica ha un'effetto inferiore in settori, come le tecnologie discrete, dove i brevetti costituiscono una fonte di informazione piú efficace.

Nel secondo studio (terzo capitolo), in collaborazione con Annamaria Conti e Fabiana Visentin, si studia come il livello di produttività scientifica media degli studenti di dottorato selezionati da università esterne sia determinato dalla prossimità sociale dei professori con tali università. Ricercatori assunti da ambienti esterni tendono ad avere una maggiore produttività scientifica rispetto a coloro che sono assunti dalla stessa università dove hanno completato gli studi. Abbiamo analizzato, in un campione di 4'666 studenti di dottorato, la produttività di studenti esterni provenienti dalla rete di conoscenze del supervisore: studenti che hanno completato gli studi in università diversa da quella del dottorato ma con la quale almeno un coautore del supervisore sia affiliato. La produttività di questi studenti risulta significativamente piú elevata sia rispetto alla produttività di altri studenti esterni e di studenti che hanno completato gli studi nella stessa università del dottorato. Nelle analisi controlliamo per l'eterogeneità dei supervisori, l'eterogeneità delle università dove gli studenti hanno completato gli studi, includendo proxy della rilevanza specifica di queste università per i supervisori. Quindi concludiamo che i professori assumono studenti con maggiore produttività scientifica dalle università con quali i loro coautori sono affiliati. Ulteriori analisi suggeriscono che la riduzione di assimmetrie informative sia il principale meccanismo alla base di questo risultato.

Nel terzo studio (quarto capitolo), in collaborazione con Guillaume Burghouwt, viene studiata l'integrazione regionale di conoscenza come determinante della produttività innovativa delle imprese. Si utilizza un campione di 3,871 imprese innovative in Germania tra il 1992 e il 2010, per un totale di 15,819 osservazioni. Con modelli ad effetti fissi, l'integrazione regionale della conoscenza, misurata come la dispersione geografica delle fonti di conoscenza delle invenzioni sviluppate in una regione, influenza positivamente la produttività innovativa delle imprese localizzate nella regione. Al fine di limitare il possibile problema di endogenità dovuto a causalità inversa e variabili omesse che variano nel tempo, viene sfruttata la liberalizzazione delle linee aere in Germania. Troviamo che il passaggio da mercati monopolistici a mercati piú competitivi nell'industria aereonautica ha determinato un aumento del livello di integrazione interregionale della conoscenza. Imprese localizzate in regioni dove la liberalizzazione ha determinato tale aumento vedono un incremento della loro produttività innovativa. Non troviamo differenze significative tra regioni con livelli diversi di investimento in ricerca e sviluppo.

Parole chiave

Geografia dell'innovazione, prossimità, integrazione interregionale, duplicazione, produttività scientifica, produttività innovativa.

Contents

Chapter 1 Introduction

1.1 Thesis motivation and structure

Knowledge is widely recognized to be the engine of economic growth (Griliches, 1979; Grossman and Helpman, 1993; Romer, 1986, 1990). Perhaps the most peculiar feature of knowledge, as compared to other economic goods, is its tendency to generate strong externalities (Foray, 2004). As a consequence, the productivity of investments in knowledge is a complex function that includes external existing knowledge as input (Cohen and Levinthal, 1990; Jaffe, 1986, 1989). Therefore, understanding knowledge diffusion and its impact on innovation performance is of primary importance in order to uncover the mechanisms leading to economic growth.

Importantly, knowledge does not diffuse freely in the economy (Cowan et al., 2000; Gertler, 2003). The recognition that geography, human interactions and mobility strictly shape the way knowledge diffuses is the fundamental pillar of a large body of literature in economic geography (Boschma and Frenken, 2005; Breschi and Lissoni, 2001; Feldman and Kogler, 2010; Keller, 2004; Krugman, 1998, 1991). Proximity and location have been dominant concepts in the debate. The former can be summarized by the notion that a certain proximity along different dimensions is a prerequisite for knowledge diffusion among economic agents. The latter concerns the location decisions in the geographic space of economic agents, which in turn determine their relative proximity. The geographical concentration of human capital, institutions and firms is often provided as the evidence of the positive effect of knowledge externalities (knowledge spillovers) generated by geographic proximity. These principles have inspired numerous theoretical and empirical contributions as well as policies. Examples such as the Silicon Valley, well known as the "archetype of industrial high tech cluster" (Feldman and Kogler, 2010), have been for decades the model for policy interventions aimed at fostering regional economies (McCann and Folta, 2008).

However, the debate on several aspects regarding the real benefits of proximity, and the underlying mechanisms explaining its effects on knowledge production, is still open. An exhaustive review of the literature is not within the scope of the discussion here. Existing reviews provide an excellent overview of the main findings and propose avenues for future investigations (Boschma and Frenken, 2005, 2011; Breschi and Lissoni, 2001; Cruz and Teixeira, 2009; Feldman and Kogler, 2010; Frenken et al., 2014; McCann and Folta, 2008). In particular, three lines of research can be identified. First, various forms of proximity, beyond geographic proximity, interact and affect innovation processes differently (Boschma, 2005). Second,

different mechanisms, beyond a simple learning process, can explain the role of proximity and of knowledge diffusion, affecting the level and direction of technological progress both positively and negatively (Bloom et al., 2013; Boschma, 2005). Third, increasing attention has been given to the importance of combining the exploitation of close knowledge with access to distant knowledge sources (Bathelt et al., 2004; Saxenian, 2007).

In the following chapters of this thesis I present three studies that contribute theoretically and empirically to these debates. Each study provides new evidence on how different forms of proximity and access to geographically distant environments might affect innovative and scientific performance. In the next paragraph of this introductory chapter I summarize the main previous findings and current lines of research that constitute a common reference framework for the studies presented in the thesis. The last paragraph provides a more detailed overview of the studies proposed in the thesis, highlighting the main contributions within the framework identified. Each chapter of the thesis includes a section that further extends the discussion to the literature on each specific topic of the chapter and presents in detail the results of the empirical analyses.

1.2 Literature framework

1.2.1 Forms of proximity

Perhaps the first evidence regarding the effects of uneven knowledge diffusion, first pointed out by the seminal work of Marshall (1891), is the spatial concentration of economic activities. Subsequent evidence further demonstrated that innovative activities and knowledge flows are disproportionally concentrated geographically as compared to the distribution of production (Audretsch and Feldman, 1996; Jaffe et al., 1993). Therefore, the literature initially focused on geographic proximity as a precondition for the diffusion of knowledge spillovers. The main assumption justifying this conclusion was that geographic proximity allows for face-to-face interactions, enabling access to tacit knowledge embedded in the individuals devoted to its production.

The geographic dimension of proximity and its effects on knowledge creation are still an ongoing area of research (Catalini, 2012; Singh and Marx, 2013). However, these and other recent studies have pointed out the fuzziness of the linkage between pure geographic proximity and knowledge diffusion. These studies have attempted to “open the black box” of knowledge spillovers (Breschi and Lissoni, 2001). Boschma (2005) revises the contributions stemming from this line of research and identifies five different dimensions of proximity: geographic, cognitive, social, organizational and institutional. While all dimensions might be positively correlated, the importance of distinguishing them, theoretically and empirically, resides in their different effects on knowledge diffusion and their different strategic implications. For example, a large

number of studies have highlighted the importance of mobility and collaboration networks as determinants of social proximity and knowledge flows (Almeida and Kogut, 1999; Breschi and Lissoni, 2009; Song et al., 2003). This evidence indicates the need of overcoming the focus on geographic proximity per se and account for all aspects related with workers mobility and networks. Furthermore, the relevance of social connections opens the discussion to the way knowledge can be diffused at high geographic distances (Agrawal et al., 2006, 2008). In general, the issue of how different proximity typologies affect knowledge diffusion and interact among each other constitutes an ongoing area of research.

1.2.2 Effects of proximity

As noted, scholars have defined knowledge spillovers as a positive externality by which agents in the economy can benefit from others' investments and existing knowledge. Due to the tacit nature of knowledge, agents close to each other should benefit disproportionally from such an externality. Various studies, making use of patent citations as measures of knowledge flows, show that such knowledge flows are more likely among agents that are close to each other, either geographically (Jaffe et al., 1993; Singh and Marx, 2013) or socially (Agrawal et al., 2008; Breschi and Lissoni, 2009). Some authors found that firms located in innovative clusters are more likely to innovate (Baptista and Swann, 1998). Also importantly, there is evidence that the productivity of firms or countries are related to the R&D investments of other firms and countries, proportionally to their proximity, either geographically, in terms of technological specialization (cognitive proximity), or trade interactions (Bottazzi and Peri, 2003; Coe and Helpman, 1995; Jaffe, 1986; Keller, 2002; Kerr, 2008; Peri, 2005). However, and surprisingly, there is no unanimous evidence that proximity per se has a positive effect on innovative performance (Boschma and Frenken, 2011; Lee, 2009). As noted by Boschma and Frenken (2011), geographical clustering can rise without any positive effect, or even with negative effects of colocation. Furthermore, other and recent empirical evidences on the effect of proximity are mixed (Boschma and Frenken, 2011; Boschma, 2005).

Part of the recent literature is attempting to solve this puzzle. First, some authors have proposed an evolutionary approach to economic geography which theorizes that product and industry life cycles determine the extent to which proximity and location in high concentrated clusters have a positive effect on performance (Audretsch and Feldman, 1996; Boschma and Frenken, 2011). In particular, emerging innovative industries would benefit the most from the geographical concentration of diverse economic and innovative activities, while mature industries are expected to profit more from distribution in smaller and specialized regions (Audretsch and Feldman, 1996; Frenken et al., 2014). Second, Boschma (2005) suggests that an optimal level of proximity can be reached. Beyond this optimal level, negative effects prevail: too much proximity of various forms might create a lack of sources of novelty and openness, the loss of economic rationales in strategic decisions and lock-in in suboptimal technological paths. Finally, part of the literature emphasizes the need to further tease out the mechanisms that might lead from knowledge spillovers

to performance, accounting for the strategic decisions of economic agents and their relative positions in the economy (Bloom et al., 2013; Breschi and Lissoni, 2001). For example, Bloom et al. (2013) distinguish and find empirical evidence of the existence of both positive technology knowledge spillovers and negative business stealing effects from R&D by product market rivals.

1.2.3 Proximity and distant knowledge

Despite the impact of information and communication technologies, the reduction of transportation costs and the increase in the geographic mobility of workers (Ding et al., 2010; Forman and van Zeebroeck, 2012; Salt, 1997), the bond between geography and knowledge production seems not to have loosened over time (McCann and Folta, 2008). Some studies suggest that knowledge flows diffuse increasingly at higher geographical distances, but the literature is not unanimous on this conclusion (Keller, 2002; Sonn and Storper, 2008). Furthermore, the tendency of innovative activities to concentrate geographically and the existence of large regional disparities in economic innovative performance are still a reality (Etherington and Jones, 2009). However, while the economic geography literature has typically emphasized the local dimension of regional economies, the importance of external linkages and external sources of knowledge have often been mentioned as an important component (Feldman and Kogler, 2010; Saxenian, 1994). Indeed, historical evidence shows that the most innovative environments have been often those presenting the best connections with external contexts and staffed with workers from all over the world (Bresnahan et al., 2001; Saxenian, 2005, 2007).

There is a general consensus regarding the fact a combination of “local buzzes” (local knowledge spillovers) and “global pipelines” (external linkages enabling access to distant sources of knowledge), is beneficial for clusters’ innovative performance (Bathelt et al., 2004). From a dynamic perspective, the inflow of external knowledge might sustain innovative performance once opportunities arising from the exploitation of local knowledge decrease (Boschma, 2005). There is some evidence that regions with external linkages and better access to geographically distant environments have higher growth and innovative performance (Eisingerich et al., 2010; Redding and Sturm, 2008). Other studies show that breakthrough innovation is largely dependent on geographically dispersed knowledge (Phene et al., 2006). Finally, the relevance of geographic openness can be traced back to the contribution of immigrants and returnees (in particular, highly skilled) with respect to the innovation performance of regions and the creation of new innovative clusters (Bresnahan et al., 2001; Saxenian, 2005). The relevance of these dynamics is also related to the political debate and existing policies of each region promoting interregional integration (Chessa et al., 2013; Crescenzi et al., 2007). Nonetheless, deeper theoretical and empirical investigations on the importance of external linkages and sources of knowledge for innovative performance are a rather recent development, requiring further evidence.

1.3 Overview and contribution of the thesis

Table 1.1: Overview and contribution of the thesis

	First study (2nd chapter)	Second study (3rd chapter)	Third study (4th chapter)
Data and context	1. Patent citations data and patent citation categories (PATSTAT 2012). 2. Geo-localization of the inventors (REGPAT 2012)	1. Dataset of PhD students and supervisors from the Ecole Polytechnique Fédérale de Lausanne and the Swiss Institute of Technology of Zurich. 2. Curricula 3. Publication scores from SCOPUS	1. Mannheim Innovation Panel (MIP) on innovative firms in Germany. 2. Region level patent data (PATSTAT 2014, REGPAT 2014). 3. Airlines market entry information from AOG historical flight status data.
Forms of proximity	Geographic proximity: we measure the geographic distance among inventors of citing and cited patents.	Social proximity: we consider the supervisor network in terms of the presence of a coauthor in other universities.	Geographic (transportation costs): we consider the level of interregional knowledge integration (defined as the region's degree of access to and adoption of knowledge developed in other geographically dispersed regions) and the effect of the airline liberalization in Europe.
Effects of proximity	Proximity affects the likelihood of duplication of inventions. Proximity can increase the likelihood of duplication when there are incentives to compete. It otherwise decreases the probability of duplication.	PhD students hired from universities where the supervisor had a coauthor have higher scientific productivity	Access to geographically distant knowledge is mediated by airline liberalization (lower transportation costs).
Role of external links	Based on our theory, the lack of knowledge flows from geographically distant environments put inventors at risk of duplicating inventions already existing in these locations.	The results suggest that social proximity facilitates sharing information with geographically distant environments, thus reducing information asymmetries and affecting the capacity to attract external human capital.	Increased access to the external knowledge of a region determines higher innovative productivity of firms located in the region.

1.3.1 First study: geographic proximity and duplication of inventions

Recent economic theory suggests that decreasing returns to scale in R&D investments, determined by the possibility of duplicated inventions, might affect economic growth (Gómez, 2011; Jones, 2009). The second chapter of this thesis first provides a theoretical discussion of how geographic proximity, affecting the diffusion of knowledge, is expected to impact the likelihood of duplicated inventions. The hypotheses derived from the theoretical discussion are tested, making use of patent data. The data allowed us to observe and localize geographically claimed inventions that are not novel, according to the opinion of a patent

examiner, when compared to an existing patent document. The form of proximity we consider is geographic proximity; however, we acknowledge that its effect may be mediated by other forms of proximity.

The main contribution of the chapter relies on the implication for the debate on the effects of proximity. In particular, we contend that, because of localized knowledge flows, geography affects the rate of duplication. In our theory, proximity has opposite effects depending on the existence of incentives for inventors to compete on the same technological path, or instead, to avoid duplication. In particular, proximity and knowledge spillovers increase the risk of duplication in the first case. In the latter case, inventors instead run the risk to duplicate existing inventions of which they were not aware because they were located in distant locations. Therefore, we suggest that proximity and knowledge spillovers should not be simply conceived as components of a learning process allowing cumulative innovations, but that the strategic behaviors of economic agents determine the extent to which innovative efforts are indeed cumulative. At the same time, we contend that a lack of knowledge flows does not simply reduce knowledge inputs at the disposal of innovators but puts them at risk of duplicating existing inventions.

1.3.2 Second study: social proximity and productive PhD students

Universities are fundamental institutions in the production of knowledge (Black and Stephan, 2010). To the extent that the diffusion of knowledge produced in universities is also subject to the effects of proximity (Jaffe, 1989; Mowery and Ziedonis, 2014), local universities can spur local innovation and economic growth (Feldman and Kogler, 2010). Therefore, the scientific productivity of universities has also attracted the attention of scholars in the field of economic geography. There is extensive evidence showing that universities benefit, among other factors, from the inflow of external personnel, in particular foreign and foreign educated researchers and students (Black and Stephan, 2010; Levin and Stephan, 1999). As such, universities also constitute a vehicle for the attractiveness of a region with regard to international human capital and for the construction of international networks. However, little is known as to how universities attract productive researchers. Indeed, hiring processes, especially for highly skilled workers, are affected by strong information asymmetries (Arrow, 1972; Granovetter, 1995). As such, hiring from external environments may be difficult and institutions may be biased toward internal candidates, about whom they know more information (Horta et al., 2010).

The third chapter compares the productivity of PhD students coming from other universities where their supervisors have co-authors with the average productivity of other PhD students. The empirical analysis isolates the relevance of the supervisors' networks as a measure of social proximity, for the capacity of the university to attract productive students. Importantly, the effect of supervisors' networks may be positive, in situations where they help to reduce information asymmetries, as well as negative, in situations where social proximity leads to favoritism. We find that students coming from other universities where their supervisor

have co-authors are largely more productive, thus implying a positive effect. Therefore, we suggest that social proximity facilitates information sharing with geographically distant environments, affecting the capacity to attract external human capital.

1.3.3 Third study: interregional knowledge integration and innovation

The economic geography literature has empirically investigated the relationships between the economic and innovative performance of firms and the local characteristics of the region or cluster where they are located. Less evidence exists regarding the effects of the connections and knowledge flows across regions. However, the advent of Information and Communication Technologies and the reduction of transportations costs make this aspect increasingly important in understanding firm performance (Tranos, 2013). In addition, whether efforts to increase interregional integration lead to higher innovative performances and how they affect the distribution of innovation activities remains an open area of research (Cappelli and Montobbio, 2013; Chessa et al., 2013; Crescenzi et al., 2007)

The fourth chapter of the thesis investigates whether a region's increased access to external sources of knowledge has an impact on the innovative productivity of firms located in that region. We define interregional knowledge integration as a region's degree of access to and adoption of knowledge developed in other geographically dispersed regions. We exploit airline liberalization in Europe as a source of an exogenous shock to transportation costs, and consequently, an incentive for regions to access external knowledge. We find that firms located in regions where airline liberalization induced a higher level of interregional knowledge integration significantly increased their innovative productivity. The results can be traced back to the effect of increased proximity, under the form of lower transportation costs, on new sources of knowledge. The specific underlying mechanisms leading to innovative performance deserve further attention in future research. However, this evidence contributes to the discussion on the importance of the access to external sources of knowledge for regional and firm performance, with implications for firms' strategy and policy makers' initiatives.

Chapter 2 The geography of duplicated inventions

(With Julio Raffo)

2.1 Introduction

Galileo claimed the invention of the thermometer circa 1592, but this invention was subsequently also claimed by Van Guericke and Porta in 1606, Drebbel in 1608, Sanctorious in 1612, and Paul and Fludd in 1617. Sir Joseph Swann and Thomas Edison both solved the problem of electric light. Other examples involve the invention of the telegraph, the telephone, electro-magnetic clocks, the typewriter, the discovery of oxygen, the periodical classification of the chemical elements, the Diesel engine, jet propulsion, and numerous others[1]. Similar examples lead Merton to the provocative hypothesis that "far from being odd or curious or remarkable, the pattern of independent multiple discoveries in science is in principle the dominant pattern" (Merton, 1961: 477).

The duplication of inventions and multiple discoveries are natural outcomes of scientific and technological progress. However, from an economic perspective, duplication in research and innovation may be a matter of concern (Bonaccorsi et al., 2009; Dasgupta and David, 1994; Jorde and Teece, 1990; Scotchmer, 1991). Overlapping R&D outcomes lead to diminishing returns on R&D investments (Gómez, 2011; Jones, 2009, 1995; Jones and Williams, 2000; Kortum, 1993; Venturini, 2012). On the one hand, several factors, such as the technological progress in communication technologies, might decrease the rate of duplication (Brannigan and Wanner, 1983). On the other hand, the probability of duplication increases with the density of inventors and the cumulating stock of knowledge that makes it more difficult for future generations of inventors to propose novel innovations and discoveries (Jones, 2009). Notably, Bessen and Meurer (2008) find that the number and cost of patent lawsuits has consistently increased over the last 30 years and conclude that "..a significant and growing number of very expensive lawsuits occur each year

[1] For a broader historical discussion of these and other examples, refer to Merton (1961) and Lamb and Easton (1984). Importantly, Constant (1978) and Elkana (1971) revise some of these examples claiming that in some instances the level of similarity of the inventions involved has been overestimated. See Bikard (2012) for a recent literature review and discussion.

because firms have invested millions of dollars for the research, development, and commercialization of technology that is allegedly owned by others" (Bessen and Meurer, 2008: 121).

Prior research has generally addressed the characteristics and determinants of this phenomenon. Adopting patent data, we address the question of how temporal and geographic distance affects duplication. For this purpose, we refer to the stream of the literature that has discussed how proximity affects the diffusion of knowledge (Breschi and Lissoni, 2001; Jaffe et al., 1993). Several contributions in this literature have demonstrated a close connection between the distribution of economic innovation activities and innovation performance and identified leading technological geographical clusters as successful examples (Audretsch and Feldman, 1996; Baptista and Swann, 1998; Delgado et al., 2010; Porter, 1998; Saxenian, 2007). However, this research has primarily conceptualized localized knowledge flows as an element of a general learning process that enables other inventors to build on existing knowledge, generating positive externalities. We further contend that because of localized knowledge flows, geography affects the rate of duplication.

We propose to distinguish between two different mechanisms leading to duplication. On the one hand, duplication may arise from imperfect knowledge flows. As such, laggard agents duplicate inventions without being aware of the existence of the original ones. In other words, these uninformed inventors simply 'reinvent the wheel'. On the other hand, high knowledge flows favor the duplication of upcoming or very recent inventions. This is the case for agents competing for and investing in the same technological solutions (Dasgupta and Maskin, 1987). As knowledge flows increase over time but are bounded geographically, we propose the following hypotheses: first, the probability of duplication decreases over time; second, geographic distance decreases the probability of the duplication of recent inventions (close in time); third, for inventions that are not recent (distant in time), duplication becomes more likely at greater geographic distances.

We find evidence for our hypotheses in patent application data. A duplicated invention can be captured in patent data whenever a patent application is filed for an invention that is not novel. Recent patent bibliographical data from the European Patent Office (EPO) make it possible, through patent citations, to identify whenever, according to an EPO examiner, the cited patent document compromises the novelty of the citing patent application (Criscuolo and Verspagen, 2008; Guellec et al., 2012). In practical terms, including patent level fixed effects, we analyze the likelihood of observing this type of citation relative to the probability of observing a citation describing the state of the art but not threatening the novelty of a patent application. In so doing, we control for unobserved heterogeneity in the inventions and, importantly, we examine the occurrence of duplications relative to the geographical distribution of the related industrial and innovative activities (Alcacer and Gittelman, 2006; Jaffe et al., 1993).

We implicitly rely on the assumption that the knowledge disclosed in patent documents and patent protection is less than perfect, rendering it likely that inventors are not fully aware of the entire existing patent literature and partly rely on geographically close sources of knowledge (Atal and Bar, 2010; Feldman and Kogler, 2010; Walsh et al., 2007). However, empirical evidence is mixed, and various studies have found the patent literature to be an effective source of information and knowledge for inventors (Graham et al., 2009). We find support for the hypothesis that applicants and inventors who are revealed to be aware of the existence of a patent protecting a given technology (based on previous citations of the same patent) are less likely to duplicate the related invention. Furthermore, we account for the possibility that the effectiveness of knowledge disclosure in patent documents might differ across sectors. In particular, patents are expected to be more relevant for relatively established sectors, in which inventions can be more easily described in written form due to the specific characteristics of the technology and/or its stage of development. In keeping with this intuition, we further test our hypotheses for different sectors and for complex and discrete technologies separately and find that geographical proximity is less relevant for the duplication of discrete technologies and, in particular, chemistry-related technologies.

2.2 Theory and Hypotheses

2.2.1 Knowledge, competition and duplication

We define duplication as claiming an invention that is not partially or completely novel compared to an existing one and has consequently a lower (or null) value relative to the initial expectations of its inventors. This broad definition captures the notion that inventors involved in duplications rarely invent precisely the same product but component and qualitative differences are often present (Collins, 1992; Dasgupta and Maskin, 1987). Moreover, duplication corresponds to socially sub-optimal levels of investment in R&D – at least ex post - and to a net loss for the "losing" inventors at the private level.

Early literature in sociology and economics situates the phenomenon of duplication of discoveries (regarding both inventions and scientific discoveries) within the natural dynamic of scientific and technological progress (Kuhn, 1996; Lamb and Easton, 1984; Merton, 1961, 1979). Discoveries are not, or not exclusively, a product of the intellect of an individual (Merton, 1961). Rather, they are developed as a result of the process of the accumulation of knowledge, which enables further developments (Jones, 2009; Murray and O'Mahony, 2007). Incremental discoveries proceed along pre-established paths, where following steps are based on the coherent implications of previous ones. More radical breakthroughs are often the consequence of contradictions or limitations of the previous paradigm that become evident at some point (Kuhn, 1996). When "the time for an invention has come," more than one individual can reach the same result. These authors also note that the systematic presence of this phenomenon is implicit in science

and innovation practices (Lamb and Easton, 1984; Merton, 1979; Merton, 1961). Indeed, reward is based on priority, which scientists and inventors constantly rush to demonstrate (Dasgupta and David, 1994; Stephan, 1996). This perspective is useful to understand why duplication is a probable outcome in research and innovation. However, it leads to the conclusion that individuals facing these identical preconditions are exposed to a certain probability of duplicating the same effort, making the phenomenon substantially random.

Merton noted first that "unnecessary multiples" result from imperfections in communication channels (Brannigan and Wanner, 1983; Merton, 1961; Niehans, 1995). Unawareness of the existence of a certain technology may be the cause of the duplication of research efforts (Brannigan and Wanner, 1983). Access to the knowledge related to the state of the art in a particular sector is crucial in this respect. "Accumulation of knowledge is not inherent to the innovation process" (Murray and O'Mahony, 2007). Inventors, firms and innovative regions struggle to reach and maintain the technological frontier and seek to avoid the involuntary duplication of research investments (Archibugi, 1992; Jorde and Teece, 1990; Ziedonis, 2004). An agent with access to this type of knowledge is expected to be aware of the existing technologies and able to formulate more accurate evaluations of future opportunities. Failures in accessing this type of knowledge may lead to duplicative efforts in pursuit of an invention, driven by the false belief that it does not exist (Bessen and Meurer, 2008). This case describes an independent duplication (Lamb and Easton, 1984): i.e., an invention that is duplicated without an awareness of the risk of replicating others' research efforts. Independently duplicated inventions do not need to be simultaneous, depending on the accumulated state of knowledge in several locations and cultures in which they appear (Merton, 1961: 486).

Under perfect knowledge flows, inventors would be fully aware of existing technologies. However, for upcoming or recent inventions, priority might have yet to have been established, and information regarding intermediated results is not typically disclosed prior to a certain level of completion. For such technologies, duplication might nevertheless occur and is more likely in the presence of knowledge flows if inventors decide to compete in the same technological space. Competition, in the presence of knowledge flows, is therefore a source of duplication that "encourages rivals to select overly similar (i.e., correlated) projects … leading to an excessive occurrence of duplications" (Dasgupta and Maskin, 1987: 594). Several related dynamics have been discussed in the literature. Patent races involve inventors in competitions for the same technology to anticipate the potential patents of competitors (Hall and Ziedonis, 2001) in which the second inventors inevitably receive a lower return on their investment. Similarly, an inventor might want to attempt to "invent around" an existing invention that has the potential for profitable improvements or to erode the technological advantage of a competitor (Guellec et al., 2012). This might lead to some degree of innovative

outputs but also to excessively small improvements, marginal changes and overlapping contributions. The second inventor might fail to obtain any valuable property right[2].

2.2.2 Duplication and proximity

Based on the mechanisms discussed above, we argue that the probability of duplication varies over time and geographic distance to the extent that these two dimensions affect knowledge flows and incentives to compete on a given invention. The hypotheses and underlying arguments are summarized in Figure 2.1. On the one hand, the more time that passes after an invention, the greater the extent of knowledge flows regarding its existence and contents that are transferred through firms, regions or social networks. Moreover, time allows for the distribution of goods and services that exploit the technology in question, and this mechanism is known to be a critical knowledge diffusion channel (Keller, 2004). On the other hand, incentives to compete to complete a given invention decrease over time, as it becomes more difficult to catch up with the leader and the (potential) market for technologies declines[3]. Therefore, both awareness and reduced competitive incentives are aligned with respect to time, where:

H1: Time distance decreases the probability of duplication.

Second, we argue that geographic distance affects duplication. It has been shown that knowledge flows among inventors are related to geographic distance. Social and professional networks, through which knowledge flows more efficiently, are to a large extent locally based (Almeida and Kogut, 1999; Breschi and Lissoni, 2009; Song et al., 2003). Geographical proximity increases the likelihood of informal and face-to-face contacts, serendipitous information flows, low opportunity cost interactions and business relationships (Audretsch and Feldman, 1996; Catalini, 2012; Jaffe et al., 1993; Jaffe, 1986; Mowery and Ziedonis, 2001). Finally, competing inventors can monitor one another more closely and actively when they are geographically proximate, which facilitates the search for information that would not be otherwise available or that competitors are not willing to disclose[4]. Nonetheless, the effect of knowledge flows on duplication is expected to be twofold. Depending on the incentives to compete for the same invention, knowledge flows

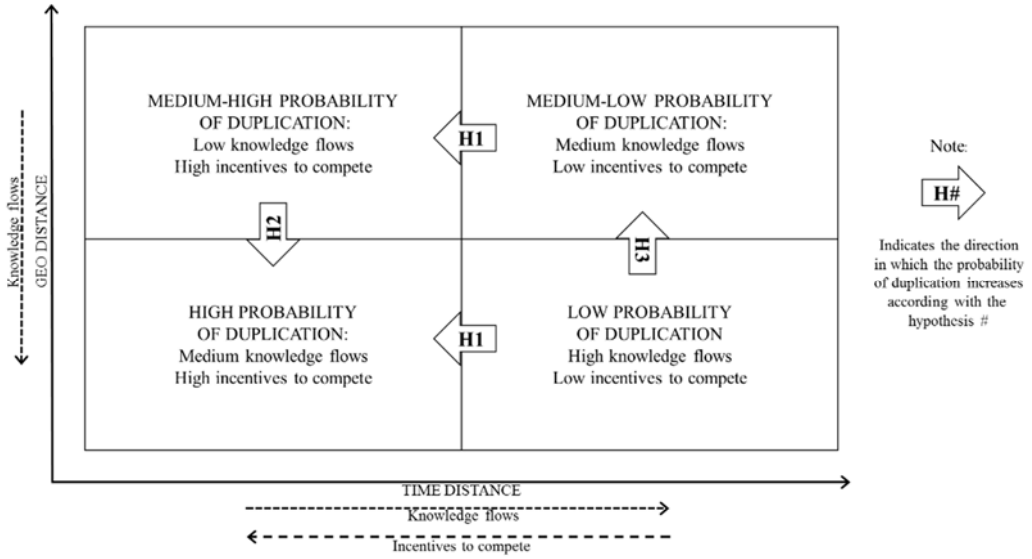
[2] As a representative example, Bessen and Meurer (2008) discuss how Kodak invested a considerable amount of resources to invent around Polaroid instant photography patents but failed. Eventually, Kodak was required to pay Polaroid $900 million and exited the market for instant photography.

[3] In this sense, we assume competitive incentives with respect to a new technological opportunity to be linear or monotonically decreasing. We acknowledge that the relationship between time and competitive incentives might be more complex than this and related to the technology and industry life cycle. However, our argument is based on the notion that on average, the economic return to investment to anticipate a competitor on a given specific technology (up to the risk of duplicating the same invention) is decreasing over time: first, the probability of obtaining priority on a given invention decreases substantially; second, the competitor is more likely to develop further mechanisms for protecting its technology.

[4] Note that our hypotheses rely on the well-documented inverse relationship between geographic distance and knowledge flows. However, we do not directly measure knowledge flows. Additionally, we generally refer to knowledge flows without further distinction in the paper among different typologies of knowledge flows mechanisms such as pure or pecuniary knowledge spillovers, personnel mobility, and the distribution of products (Breschi and Lissoni, 2001).

can either increase or decrease the probability of duplication. We exploit the asymmetry of incentives to compete with respect to time distance to disentangle the effect of proximity according to competition and unawareness.

Figure 2.1: Conceptual summary of main hypotheses



Over brief periods of time, incentives to compete are high, and investing in a technology that is also being developed by a competitor could be rational. Essentially, once a technological opportunity is identified, agents sharing the same pool of information face the choice of racing for it or searching for an alternative option. Their relative position in the technological space – i.e., with respect to competitors – and search and switching costs will be strong factors influencing the decision. Arguably, in this context, knowledge flows concerning an upcoming invention increase the opportunity cost to search for alternative investments relative to the possibility of anticipating other inventors for a given technology. Therefore, intermediate knowledge flows can both cause and exacerbate a race, allowing competitors to catch up to one another (Encaoua and Ulph, 2005). Similarly, it has been argued that proximity – beyond a certain optimal level – might have a negative lock-in effect, which reduces the ability of nearby individuals to engage in original efforts (Boschma, 2005; Martin and Sunley, 2006). In this sense, "local buzzes" (regarding a technological opportunity) might cause more than one inventor to pursue the same invention (Boschma, 2005). Although they might not be completely aware of one another, we can nevertheless argue that they would be simultaneously competing for the same technological solution. For instance, a simple conference or an informal conversation that enables two individuals to identify an opportunity could end with two inventions that would not be independent because they were inspired by a "common spark" (Breschi and Lissoni, 2005). For upcoming and recent technologies, the rate of duplication should increase in geographic proximity. Thus we hypothesize the following:

H2: Geographic distance decreases the probability of duplication.

In contrast, with distance in time, regarding a technology that is no longer recent, the incentive to invest in it declines unless there is unawareness of its existence. Local knowledge flows provide the opportunity to abandon valueless efforts, improve existing inventions rather than replicating the same effort, or specialize in complementary and differentiated technologies (Dasgupta and David, 1994). Empirical evidence has demonstrated that inventors make substantial use of local information or knowledge to create novel products and processes (Feldman and Kogler, 2010; Giuri et al., 2007). For a given existing technology, local knowledge flows are advantageous for proximate inventors relative to more distant ones: proximity makes it possible to avoid duplicating an invention when it would no longer constitute a promising investment but rather a piece of prior art suitable for further developments. In other words, we can argue that if duplication occurs for a technology that is not recent, this is more likely when there is a lack of knowledge flows and hence is most probable far from the location where the technology was originally developed. Therefore:

H3: As time distance increases, geographic distance increases the probability of duplication.

2.2.3 Duplication of patents and geography

The knowledge disclosure requirement in patent documents also relies on the assumption that given a certain existing technology, someone will replicate it at some point if the information regarding its existence and its propriety is not made public (Denicolo and Franzoni, 2003; Kitch, 1977). The disclosure of knowledge in patent documents is therefore a mechanism to avoid duplication. However, the evidence on the potential of patents as a source of information and knowledge for inventors is mixed. Graham et al. (2009) find that a considerable share of inventors uses the patent literature as a source of information. Similarly, Cohen at al. (2002) find patents to be a more effective mechanism for intra-industry knowledge flows in Japan than in the USA, and their results further suggest that patent documents can be a successful tool for knowledge diffusion. Other authors have noted that the disclosure of knowledge is not a sufficient condition for cumulative innovation (Murray and O'Mahony, 2007) and that codified knowledge has limited power as a source of knowledge for inventors at geographic distance (Feldman and Kogler, 2010).

While evidence regarding the effective knowledge disclosure potential of patent documents is not unanimous, one can argue that the existence of a patent application does not guarantee that information regarding the invention is immediately available to each inventor in a certain sector. Notably, Jaffe et al. (2000) find that inventors were fully aware of less than one-third of the citations regarding their own patents. In a survey of academic inventors, although for a specific sector, Walsh et al. (2007) find that a very small percentage (5%) of the respondents reported being aware of patents relevant to their research, and the majority of the respondents often proceed without considering relevant patents. Moreover, patent examiners are responsible for a substantial and increasing share of citations (Criscuolo and Verspagen, 2008). Those citations of the original patent application might involve other patents, of which inventors or applicants only

became aware after the completion of their invention. Furthermore, other evidence questions the capacity of patents to provide complete information on the boundaries of intellectual property (Bessen and Meurer, 2008; Sternitzke, 2009). Finally, one must consider that within the lead-time of an R&D project, from its concept stage to the first patent publication and ultimately to the granting of a patent, many concurrent and similar projects may be in process and reach an advanced stage before any document is disclosed.

Based on this discussion, we can expect that technologies seeking patent protection can be subject to duplication and that the probability of this outcome varies over time and geographical distance, in accordance with our hypotheses. In other words, our first hypotheses rely on the assumption that the knowledge disclosure generated by patent documents and patent protection is less than perfect (Atal and Bar, 2010; Walsh et al., 2007) and that knowledge diffusion over time and geographic distance can compensate for this (Feldman and Kogler, 2010). Ideally, we would like to directly measure the awareness of the inventors regarding the existence of a given technology or research activity of a potential competitor. However, to determine whether an inventor was aware of the existence of a given patent while developing an invention is an empirical challenge, and it is virtually impossible to be certain whether the inventor had full or partial unawareness. Nonetheless, it is possible to identify cases in which the inventor is revealed to be aware (or is very likely aware) of the existence of a patent, simply by considering whether one of her previous patents (or, alternatively, the applicant) cited the same patent before (Lampe, 2007). In these cases, irrespective of the origin of the citation, the inventor and the applicant should be aware of the existence of the patent. In such cases, we expect that the probability of independently duplicating the same invention would be nearly zero. Moreover, the competitive incentive to obtain property rights to the same invention should be low because it might already be protected, and in any case, it would no longer be recent. Therefore, we hypothesize that:

H4: The duplication of a given invention is less likely when the inventor (or the applicant) is revealed to be aware of the existence of a patent protecting that invention.

Finally, the effectiveness of patents as sources of information might vary dramatically across different sectors. The literature has widely discussed the greater potential for knowledge disclosure from patents on discrete technologies. In sectors such as chemistry and pharmaceuticals, patents define clearer property boundaries, and inventors exhibit a much greater awareness of the existing prior art in the patent literature. Inventors can better avoid independent duplication by consulting the patent literature and are also able to monitor competitors more easily at a distance. Understandably, patents in these sectors are, on average, of higher value and are less likely to be subject to litigation (Bessen and Meurer, 2008; Graham et al., 2009). Therefore, we expect geographic distance to have a considerably different impact across sectors, and more generally, we expect it to have a greater impact in the case of discrete technologies relative to complex ones. For this reason, we also consider our hypotheses for different sectors and separately for discrete and complex

technologies. We expect geographic distance to be more relevant (H2 and H3) for complex than for discrete technologies.

2.3 Data and methods

2.3.1 Patent citations and duplicated inventions

Identifying duplications is a remarkably difficult task (Bikard, 2012). Past examples have been discovered through historical investigation, even decades later, and there are reasons to believe that many others remained undiscovered (Lamb and Easton, 1984; Merton, 1961; Ogburn and Thomas, 1922; Simonton, 1979). Furthermore, several authors argued that the level of similarity of the cases identified is often limited and that related discoveries might be perceived as duplicating each other when substantial differences are instead present (Bikard, 2012; Constant, 1978; Elkana, 1971)[5].

Recent data on patent citations by EPO provide information suitable for our purpose. Indeed, a patent application constitutes an invention claim. Typically patent applications report a description of the technology and a list of one or more features of the technology (defined as "claims") that are required to be novel, on which the applicant desires to obtain intellectual property protection. Whenever a patent application is filed for an invention that is not novel, this should be captured by the patent system. Examiners must verify the novelty of an invention relative to existing state of the art in the public domain. Whenever an examiner considers a piece of knowledge as proof of a lack of novelty for the claimed invention, this prior element – typically a document, but not necessarily so – must be cited in the search or examination report.

Traditionally, patent citations have been used as a proxy for knowledge flows occurring among inventors. However, Jaffe et al. (1993) noted how this indicator could be noisy due to the presence of examiner citations and citations added for different scopes. Moreover, the recent debate on the use of patent citations has acknowledged that not all patent citations are appropriate indicators of knowledge flows (Alcácer et al., 2009; Alcacer and Gittelman, 2004; Criscuolo and Verspagen, 2008). In particular, Breschi and Lissoni (2005) directly highlight the possibility that patent citations refer to duplicative efforts. Recent EPO data allow us to identify the original source of a citation– i.e., application, search report, examination, opposition, etc. – and what it represents. The EPO examiner is always the individual responsible for categorizing citations, regardless of whether the citation was already in the original applications. Therefore, the

[5] See Bikard (2012) for a thorough discussion of theoretical and methodological issues related with the identification of multiples discoveries and inventions. The author also proposes an objective and replicable methodology to identify simultaneous multiple discoveries from scientific publications. However we cannot adopt his methodology especially because it is limited to the identification of cases of duplication very close in time and because it cannot be easily extended to EPO patent documents. The methodology we describe in this chapter relies on the patent examiner's expertize to establish duplication both close and far in time.

duplication of inventions can be recorded in patent documents as citations to the original invention when the examiners have categorized the citations accordingly.

Table 2.1: Patent citation categories

Categories	Description
A	Documents defining the state of the art and not prejudicing novelty or inventive step
Y	Particularly relevant documents when combined with another document, such a combination proving the lack of an inventive step.
X	Citations classified under this category are such that when considered alone, a claimed invention cannot be considered novel or cannot be considered to involve an inventive step.
E	Any patent document relevant to novelty (same as X citation) bearing a filing or priority date earlier than the filing date of the application searched for but published after that date.
D	Documents cited in the original application (usually referred as to "applicant or inventor citations").

The EPO provides its examiners with precise guidelines on how to distinguish citations in several categories[6] (Michel and Bettels, 2001). The most relevant for our study are summarized in Table 2.1. Category A corresponds to the typical citation, which describes the state of the art relevant and embedded in the citing patent document without compromising the novelty or inventive step requirements. In contrast, Y X and E citations refers to citations affecting the patentability of the citing application. Y-cited documents differ from X- and E- cited documents, as they exclusively refer to the lack of an inventive step and must be combined with at least one other citation. However, each X or E citation is sufficient to challenge the patentability or validity of a claim in the citing document. The only difference between E and X citations is that the former links documents that are highly proximate in time, where the citing application was filed between the filing and the publication dates of the cited one. As such, we consider X and E categories as the main indicators of duplication, where the citing application is assumed to replicate the X- or E-cited patent document. Finally, it is possible to distinguish citations present in the original document from those added by the examiner. As mentioned above, only examiners categorize citations, making all citations relevant for our analysis regardless of their origin. However, as it might be expected, the large majority of X and E citations are added by the examiner.

Certain comments are in order here. First, many inventions are not patented or published (Arundel and Kabla, 1998). In this case, it is virtually impossible to identify duplications in a systematic manner. However, one could also contend that the duplication of non-patented technology might be fully rational and less problematic since the second inventor can still freely exploit the duplicated invention. Second, a patent application also must be filed for the replicating invention in order to observe duplication. Above, we noted

[6] See "EPO guidelines for Examination in the European Patent Office", http://www.epo.org/law-practice/legal-texts/guidelines.html

that the duplication of a patented invention is not unlikely, as knowledge disclosure through patent documents might be imperfect. Nonetheless, in principle, the existence of a patent should discourage the second inventor from filing her patent even after having completed her (not novel) inventive effort. However, as we have noted, inventors are unaware of the majority of patents cited in their own patent publications, and examiners add a large number of citations. Moreover, incentives for the inventor to perform a patent search before and after developing her invention could be low because the cost of this search could exceed the cost of allowing the examiners to identify relevant patent literature (Atal and Bar, 2010). Third, if two inventors arrive at the same invention within a relatively brief period of time, it is likely that the patent application for the first invention would have yet to have been published (the EPO process entails 18 months from the filing date to publication). Furthermore, it is even more likely that the first patent will have yet to have been granted. Therefore, it is likely that if an inventor has developed a technology with the intention to patent it, she will file a patent application regardless of the existence of a similar patent (regardless of whether she is aware of the risk of duplication and especially if she is not aware of it). Whenever an examiner identifies the prior patent, we are able to observe a citation linking the two inventions. However, to summarize, patent citation categories are by no means an exhaustive indicator of duplicative efforts, and duplicated inventions might, to a large extent, be unobservable.

As a final concern, it is important to note that a patent citing an existing patent with an X or E citation can nevertheless be granted. To some extent, this might conceal the fact that X and E citations are likely a noisy measure of duplicative efforts and might also capture different phenomena. Nonetheless the fact that patents with X and E citations are often granted largely reflects that X and E citations typically refer to several but not necessarily all claims related to a single patent. Therefore, they generally correspond to a decline in the number of claims, which significantly decrease the value of the patent relative to what was originally claimed by the inventor (Tan and Roberts, 2010). Furthermore, the presence of X and E citations certainly increases the probability that the patent application will be rejected (Guellec and van Pottelsberghe de la Potterie, 2000).

In conclusion, X and E citations correspond to a claimed invention that is ultimately not novel compared to an existing one, based on the informed opinion of an expert, the patent examiner, precisely dedicated to the identification of such instances. As each patent claim is costly for the applicant, we assume that in the majority of cases, these citations will correspond to a research effort by an inventor which did not realize any economic value due to the presence of a precedent patent. In this sense, two patents linked by an X or E citation match our definition of duplication. We conducted few unstructured interviews with patent attorneys and patent examiners with two objectives: to ascertain the correctness of our interpretation of patent citation categories and to verify the plausibility of our hypothesis and results. The interviewees were generally in line with our interpretation of the data. They warn us on the possibility that the difficulty to sharply distinguish between citation categories and the process of negotiation between examiners and applicants (and in

particular with patent attorneys) is likely to introduce considerable noise in the data. However we consider that this issue can only downward bias the significance of our estimates. Finally, the interviewees pointed to existing documentation and provided us with anecdotal examples in line with our methodology and hypotheses. To further ease the understanding of the data and the methodology, we discuss in appendix to this chapter an example of a citing patent reporting both a citation categorized as A and a citation categorized as X and their descriptions.

2.3.2 Data and model

The sample is constructed using the patent citations data from EPO's Worldwide Patent Statistics Database (PATSTAT, September 2010) and inventor location information from the OECD's REGPAT Database (December 2010). Additionally, each NUTS 3 region (the third level of the Nomenclature of Territorial Units for Statistics from EUROSTAT) has been geo-localized to construct a measure of the distance between citing and cited patents. Unfortunately, PATSTAT primarily contains citations categorized for EPO patent documents only. Similarly, REGPAT only contains the location information of inventors from EPO and PCT patent documents. This means that our sample must be restricted to EPO patent documents citing EPO patent documents (EP-EP). We do not consider inventor self-citations because an inventor (or group of inventors) can only reasonably duplicate the research efforts of other inventors. Of these patents, we selected those that received at least one X or E citation and at least a different citation (primarily A). The final sample has 302,156 EP-EP citations pairs corresponding to 108,229 EPO citing patents published between 1982 and 2007. We select this period because the patent citation category are more reliable.

Our main hypotheses represent a prediction of the actual location of an invention (citing) in time and space relative to another invention (cited), as the former represents claims that are not novel due to the existence of the latter. For this reason, we focus on citing patents found to have at least one not novel claim (at least one X or E backward citation). We adopt a methodology similar to that of Alcacer and Gittelman (2006), who studied the distribution of examiner citations compared to inventor citations. Therefore, we include fixed effects for the citing patent application. In practical terms, we analyze the probability of observing an X or E citation with respect to observing a different one (primarily A citations) for each citing patent application. In so doing, we avoid the use of artificial counterfactual citation pairs, the validity of which might be highly sensitive to the method by which they are constructed (Thompson, 2006; Thompson and Fox-Kean, 2005). Furthermore, the location of a duplicated invention (XE cited) is not compared with a generic distribution of similar technologies but relative to the location of an invention that is directly relevant to the focal invention (primarily A-cited patents). The sample is also restricted to patents with at least one

non-XE citation to maintain a consistent reference point for the location of the duplicated invention within a given group of citations of a citing patent[7].

Citing patent fixed effects address also the issue of the heterogeneity across citing patent applications. Fixed effects estimation will not only control for sector and other patent-specific heterogeneity but also for any trend in the probability of duplication with respect to filing date. Furthermore, endogeneity concerning the possibility that the reciprocal distance between two inventions might be correlated with their relative characteristics, such as the innovative capacity of the inventors or applicants, are limited because the patent level fixed effects control for any characteristic of the citing invention, including time-variant characteristics of the inventors and applicants at the moment of filing.

Accordingly, the model is specified as follows:

$$
\begin{aligned}
P\left(XE_{ij}=1 \mid X_{ij}, 0<\textstyle\sum_j XE_{ij}<J\right) & \\
&= \beta_0+\beta_1 Same\ applicant_{ij}+\beta_2 Tec\ similarity_{ij} \\
&+\beta_3 \log(Time\ distance_{ij}+1)+\beta_4 \log(Geo\ distance_{ij}\ +1) \\
&+\beta_5 \log(Time\ distance_{ij}+1)\times \log(Geo\ distance_{ij}\ +1) \\
&+\beta_6 Applicant\ aware_{ij}+\beta_7 Inventor\ aware_{ij}+a_i+u_j+\varepsilon_{ij}
\end{aligned}
$$

The left-hand side represents the probability that a citation from patent document *i* to patent document *j* is categorized as X or E conditional on a set of independent variables (*Xij*) pertaining to the pair of patent documents. Therefore, the dependent variable (*XE*) is a binary variable taking value one if the citation linking two patents is an X or E and zero otherwise[8]. The model has been specified as a linear probability model (LPM) and estimated accordingly, although the alternative of a conditional logit has also been considered. The linear probability model is not inferior to a probit or logit model, provided that the "proper" non-linear model is unknown (Angrist and Pischke, 2008). Furthermore, it allows for a direct interpretation

[7] Note that, at the EPO, inventors have no obligation regarding the citations included in their patent applications, and examiners follow a rule of parsimony in the number of citations added to a patent document. This justifies cases in which examiners only include citations relevant to patentability (XE) and no other citations are mentioned. Despite not being observable in patent citations, it is unreasonable to assume that the invention had no precedents; including cases in which only XE citations are reported would mechanically downward bias our estimates.

[8] We do not consider Y citations in the dependent variable. The fact that Y-cited documents only reveal the absence of an inventive step if combined with other documents creates ambiguity. Nonetheless, our results are robust to the inclusion of Y citations in the dependent variable or if they are completely excluded from the sample.

of the coefficients, especially due to the presence of interaction terms, and facilitates comparisons of the effects across different specifications and samples.

Therefore, on the right-hand side, we parameterize the model as linear function of X_{ij}, where β_k are the parameters of interest, a_i and u_j are the fixed error terms at the citing and cited patent levels (where we initially only control for the former - a_i - through fixed effects estimation) and ε_{ij} is the idiosyncratic error term. Within X_{ij}, we consider the following set of variables. As control variables, we consider whether the citation links two patents that have at least one applicant in common. Second, we consider the share of common International Patent Classification (IPC) codes as a measure of technological proximity. *Time distance* is the number of years between the priority dates of the two patents. Then, we specify the geographic distance in different manners for two main model specifications. First, we operationalize it as a continuous variable measuring the minimum great-circle distance – in units of 10 kilometers – between all possible pairs of NUTS3 regions where inventors of the citing and cited patent are located (*Geo distance*). However, the results were consistent when considering the average or maximum distance of such pairs. We use the logarithmic transformation (plus 1) of *Geo distance* (when continuous) and *Time distance*. This functional transformation is adopted under the assumption of a non-constant and decreasing marginal effect of distance[9]. Additionally, the interaction term of these two variables is included.

In a second model specification, we consider geographic distance as a set of four dichotomous variables indicating whether at least one of the possible pairs of inventors of the citing and cited patents come from the same NUTS3 region, a different NUTS3 region but the same NUTS2 region, a different NUTS2 region but the same country, and a different country (we will refer to this measure as "*Discrete distance*"). In this case, the *same NUTS3* variable is excluded as the reference category and the other three dummy variables are jointly included with their interaction with time distance. We use this alternative specification to show robustness of the results to different measures and to acknowledge the importance of regional and political borders as factors creating discontinuities in knowledge diffusion with respect to geographical distance (Singh and Marx, 2013; Thompson and Fox-Kean, 2005). Finally, we employ two dummy variables as indicators of revealed awareness: *Applicant aware* indicates whether the cited patent was previously cited in at least one patent of at least one applicant of the citing patent (but not in a patent of any inventor of the citing patent); *Inventor aware* indicates whether the cited patent was previously cited in at least one patent of at least one inventor of the citing patent. All variables used are summarized in Table 2.2. Table 2.3 also reports the 10th, 25th, 50th, 75th, and 90th percentiles of *Time distance* and *Geo distance*.

[9] Similar results are obtained when considering a specification without transforming the variables or adopting a second-order polynomial functional form.

Table 2.2: Variable description and descriptive statistics

Variable	Description	Obs	Mean	Std. Dev.	Min	Max
XE	Equal to 1 if the citation is an X or an E, 0 otherwise. (Dependent variable)	302,156	0.476	0.499	0	1
Same applicant	Equal to 1 if at least one applicant is the same in the citing and cited document, 0 otherwise.	302,156	0.121	0.327	0	1
Tec similarity	Share of common ipc codes between the citing and the cited patent	302,156	0.851	0.270	0	1
Time distance	Number of years between the priority dates of the two patents	302,156	6.145	4.906	0	39.86301
Geo distance	Minimum distance in units of 10 km among all pairs of NUTS3 regions where inventors of the citing and the cited patent are located.	302,156	377.322	421.125	0	1,960.58
Discrete geo distance:						
Same NUTS3	Equal to 1 if at least one inventor of the citing patent and one inventor of the cited patent are in the same NUTS3 region, 0 otherwise.	302,156	0.144	0.351	0	1
Different NUTS3 - Same NUTS2	Equal to 1 if no pairs of inventors of the citing and cited patent is in the same NUTS3 region and at least one pair of inventors is in the same NUTS2 region, 0 otherwise.	302,156	0.057	0.233	0	1
Different NUTS2 - Same country	Equal to 1 if no pairs of inventors of the citing and cited patent is in the same NUTS2 region and at least one pair of inventors is in the same country, 0 otherwise.	302,156	0.201	0.401	0	1
Different Country	Equal to 1 if no pairs of inventors of the citing and cited patent is in the same country, 0 otherwise.	302,156	0.597	0.490	0	1
Applicant aware	Equal to 1 if the cited patent was previously cited in at least one patent of at least one applicant of the citing patent but not in a patent of any inventor of the citing patent, 0 otherwise.	302,156	0.103	0.304	0	1
Inventor aware	Equal to 1 if the cited patent was previously cited in at least one patent of at least one inventor of the citing patent, 0 otherwise.	302,156	0.221	0.415	0	1

Table 2.3: Time and geographic distance percentiles

Variable	p10	p25	p50	p75	p90
Time distance (years)	1.56	2.55	4.67	8.37	13.05
Geo distance (10km)	0.00	15.02	82.45	862.36	958.14

2.4 Results

2.4.1 Full sample

Table 2.4 reports the results of the primary models for the continuous measure of geographic distance. In Model 1, we first include our control variables and *Geo distance*, *Time distance* and their interaction. In Model 2, we add the awareness measures. In Table 2.5, we consider the alternative measure of geographic distance, namely *Discrete distance*: three dummies are included indicating citations to an invention from a different NUTS3 region but within the same NUTS2 region (*Different NUTS3 – Same NUTS2*), in a different NUTS2 region but in the same country (*Different NUTS2 – Same country*) and in a different country (*Different country*). Interactions between these dummies and *Time distance* are included. Again, in Model 2, we add the awareness measures.

Table 2.4: LPM with continuous geographic distance

	Model 1	Model 2
Same applicant	-0.059***	-0.056***
	(0.006)	(0.006)
Tec similarity	0.117***	0.118***
	(0.007)	(0.007)
Time distance (log)	-0.255***	-0.250***
	(0.005)	(0.005)
Geo distance (log)	-0.020***	-0.019***
	(0.002)	(0.002)
Geo distance (log) * Time distance (log)	0.015***	0.015***
	(0.001)	(0.001)
Applicant aware		-0.033***
		(0.005)
Inventor aware		-0.020***
		(0.005)
Constant	0.800***	0.799***
	(0.010)	(0.010)
Citing patent FE	Yes	Yes
Observations	302,156	302,156
Number of citing patents	108,229	108,229
F	1402	1013

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 2.5: LPM with discrete geographic distance

	Model 1	Model 2
Same applicant	-0.061***	-0.059***
	(0.006)	(0.006)
Tec similarity	0.116***	0.118***
	(0.007)	(0.007)
Time distance (log)	-0.278***	-0.272***
	(0.006)	(0.006)
Different NUTS3 - Same NUTS2	-0.077***	-0.075***
	(0.018)	(0.018)
Different NUTS2 - Same country	-0.108***	-0.104***
	(0.013)	(0.013)
Different Country	-0.166***	-0.161***
	(0.012)	(0.012)
Different NUTS3 - Same country * Time distance (log)	0.038***	0.036***
	(0.010)	(0.010)
Different NUTS2 - Same country * Time distance (log)	0.072***	0.069***
	(0.007)	(0.007)
Different Country * Time distance (log)	0.119***	0.115***
	(0.006)	(0.006)
Applicant aware		-0.031***
		(0.005)
Inventor aware		-0.019***
		(0.005)
Constant	0.839***	0.837***
	(0.012)	(0.012)
Citing patent FE	Yes	Yes
Observations	302,156	302,156
Number of citing patents	108,229	108,229
F	798.8	660.3

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

As expected, we find a significantly negative effect of *Time distance*, thereby supporting H1. From the sign of the coefficients on *Geo distance* and the interaction effect coefficient, we obtain initial evidence supporting both H2 and H3. Specifically, the coefficient is negative for *Geo distance* alone, meaning that the effect is negative for citations close in time. In other words, when the time lag between the filing dates of patents is brief, geographic distance decreases the probability of observing duplication. In contrast, the coefficient on the interaction effect is positive, indicating that the effect of *Geo distance* increases at higher values of time distance. A coherent picture is obtained by considering the discrete measure of distance; for recent technologies, duplication is less likely across different NUTS3 regions, different NUTS2 regions and different countries (Table 2.5). Conversely, the probability of overlapping claims increases across different regions, especially different countries, at high levels of time distance.

Finally, as shown in Model 2 in both Table 2.4 and Table 2.5, we find that duplication is less likely when either inventors or applicants of the citing patent are found to be aware of the existence of the cited patent. This result provides support for H4. The inclusion of these variables only marginally affects the results

concerning with time and geographic distance, indicating the robustness of the previous results. Finally, the two control variables have a predictable effect. Duplication is less likely among inventions from the same applicant, which might be due both to a higher diffusion of knowledge and reduced competitive incentives among inventors within a single company, regardless of time and geographic distance. Further, duplication is more likely between inventions with a higher share of IPC codes in common.

While the signs of the main coefficients and that on the interaction term are supportive of hypotheses 1 to 3, they are insufficient to conclude that the effect of time distance is always negative, as implied by H1, or that the effect of geographic distance becomes positive, as implied by H3, within meaningful ranges of variable values. Furthermore, they are not directly useful for determining the significance and magnitude of the effects. We address this issue in two ways. First, we compute the marginal effects of one of the two variables (*Geo distance* and *Time distance*) for different values of the other (*Time distance* and *Geo distance*). Second, we graphically plot the predicted probabilities of observing a XE citation and its confidence intervals as a function of both variables at different values of the other variable. Marginal effects and predicted probabilities, and their standard errors and confidence intervals, are computed using Delta method estimation, based on the estimates from Model 2 reported in Table 2.4 and Table 2.5.

Table 2.6: Time distance marginal effects for continuous geographic distance percentiles

Time distance margins	
Geo distance	Time distance (log) mfx
25th percentile	-0.210***
	(0.003)
50th percentile	-0.186***
	(0.002)
75th percentile	-0.151***
	(0.003)
Observations	302,156

Delta method estimation

*** p<0.01, ** p<0.05, * p<0.1

Table 2.7: Time distance marginal effects for discrete geographic distance values

Time distance margins	
Descrete geo distance	Time distance (log) mfx
Same NUTS3	-0.272***
	(0.006)
Different NUTS3 - Same NUTS2	-0.236***
	(0.009)
Different NUTS2 - Same country	-0.204***
	(0.005)
Different Country	-0.157***
	(0.003)
Observations	302,156

Delta method estimation

*** p<0.01, ** p<0.05, * p<0.1

Table 2.6 and Table 2.7 report the marginal effects of *Time distance* at different values of geographic distance. Table 2.6 considers the 25th the 50th and the 75th percentiles of the continuous variable *Geo distance*, while Table 2.7 considers the four different categories of *Discrete geo distance*. This estimation further confirms H1, demonstrating that geographic distance only attenuates the effect of time, which consistently remains significantly negative at meaningful geographic distances in our sample. This trend is to some extent coherent with the notion that the probability of duplication decreases over time but decreases more rapidly at short distances due to the effect of knowledge diffusion.

Second, we compute the marginal effects of geographic distance as a function of time distance. Table 2.8 reports the marginal effect of geographic distance – measured as continuous or discrete – for time distances ranging from 0 to 14 years between the cited and the citing patent documents. This range corresponds to more than 90% of our sample values. These estimations confirm the results discussed above. The effect of geographic distance, when measured as continuous, is significantly negative at a time distance from 0 to 2 years, is not significant at 3 years and thereafter is positive and significant. Regarding the discrete measures of distance, it is noteworthy that within a single NUTS2 region but different NUTS3 regions, the probability of duplication is significantly lower at lags of up to 4 years and only becomes significantly higher at high values of time distance. The marginal effects of the other two categories, citations among different NUTS2 regions and different countries, exhibit a similar pattern to that observed for continuous geographic distance. We find a particularly strong country border effect.

Table 2.8: Geographic distance marginal effects for values of time distance

	Continuos geo distance margins	Discrete geo distance margins		
Time distance	Geo distance (log) mfx	Different NUTS3 Same NUTS2 mfx	Different NUTS2 Same country mfx	Different Country mfx
0 years	-0.019***	-0.075***	-0.104***	-0.161***
	(0.002)	(0.018)	(0.013)	(0.012)
1 year	-0.009***	-0.049***	-0.056***	-0.081***
	(0.001)	(0.012)	(0.009)	(0.008)
2 years	-0.003***	-0.035***	-0.029***	-0.034***
	(0.001)	(0.009)	(0.007)	(0.007)
3 years	0.001	-0.024***	-0.009	-0.001
	(0.001)	(0.008)	(0.006)	(0.006)
4 years	0.004***	-0.016**	0.006	0.024***
	(0.001)	(0.007)	(0.006)	(0.006)
5 years	0.007***	-0.009	0.019***	0.045***
	(0.001)	(0.008)	(0.006)	(0.006)
6 years	0.009***	-0.004	0.030***	0.063***
	(0.001)	(0.008)	(0.007)	(0.006)
7 years	0.011***	0.001	0.039***	0.078***
	(0.001)	(0.009)	(0.007)	(0.006)
8 years	0.013***	0.005	0.047***	0.092***
	(0.001)	(0.010)	(0.007)	(0.007)
9 years	0.014***	0.009	0.054***	0.104***
	(0.001)	(0.010)	(0.008)	(0.007)
10 years	0.016***	0.013	0.061***	0.115***
	(0.001)	(0.011)	(0.008)	(0.008)
11 years	0.017***	0.016	0.066***	0.125***
	(0.001)	(0.012)	(0.009)	(0.008)
12 years	0.018***	0.019	0.072***	0.134***
	(0.001)	(0.012)	(0.009)	(0.008)
13 years	0.019***	0.021*	0.077***	0.143***
	(0.001)	(0.013)	(0.010)	(0.009)
14 years	0.020***	0.024*	0.082***	0.151***
	(0.001)	(0.014)	(0.010)	(0.009)
Observations	302,156	302,156	302,156	302,156

Delta method estimation

*** p<0.01, ** p<0.05, * p<0.1

An effective way to summarize these results is to graphically represent the predicted probabilities obtained from the models. Figure 2.2a depicts the expected probability as a function of *Time distance*, at the 25th, 50th and 75th percentiles of *Geo distance*. Figure 2.2b depicts the predicted probability as a function of *Geo distance* considering *Time distance* at five different percentiles: 10th, 25th, 50th, 75th and 90th. Figure 2.3a and Figure 2.3b are similar and display the three discrete distance measures instead of the continuous measure.

Figure 2.2: Predicted probabilities as a function of time distance and continuous geographic distance

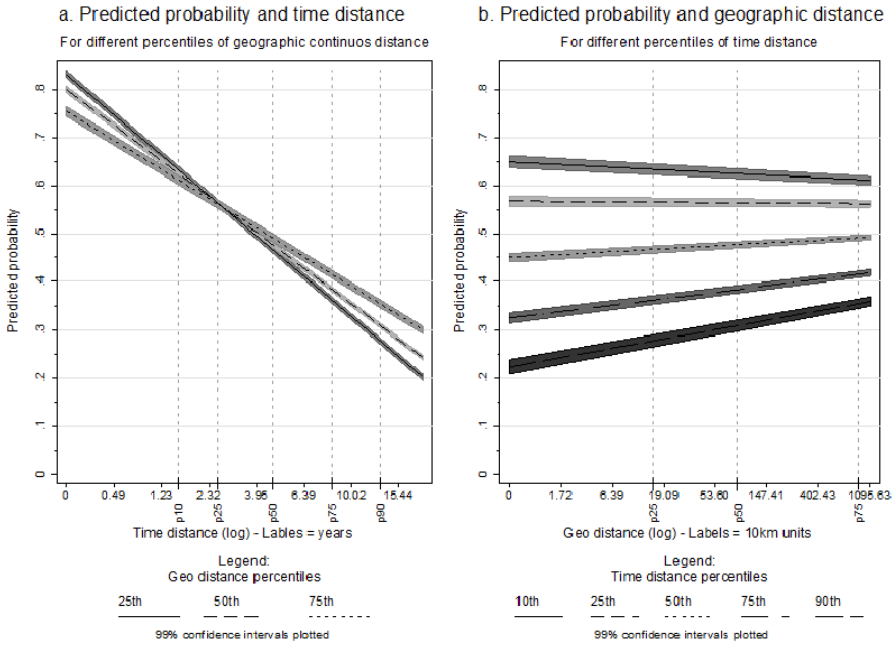
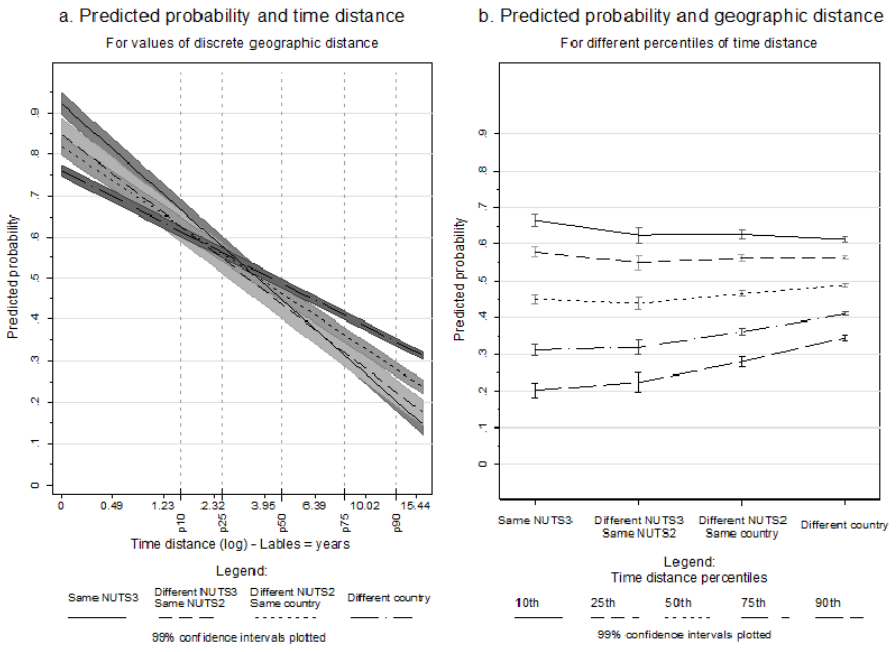


Figure 2.3: Predicted probabilities as a function of time distance and discrete geographic distance



The graphs further confirm the pattern implied by the marginal effects analysis and facilitate the discussion of the magnitude of the effects observed[10]. Here, we discuss selected representative examples. In Figure 2.2a, we observe that, when *Geo distance* is set at its 25th percentile (equal to 150 km), the predicted probability declines by approximately 82% (from 83% probability to 28%) when *Time distance* moves from 0 years to 13 years (the 90th percentile of time distance). This decline is attenuated but remains pronounced at approximately 75%, when *Geo distance* is set at its 75th percentile (equal to 8,623 km). Therefore, the effect of *Time distance* is consistently strongly negative. Figure 2.2b, we can observe that the effect of *Geo distance* is negative when time distance is set at its 10th percentile value (1,6 years), although the effect is not strong: the probability of observing an XE citation falls from 65%, for a *Geo distance* close to 0[11] to 61% for a *Geo distance* of 8,623 km (75th percentile). Conversely, when *Time distance* is at its 75th or 90th percentile, the effect of *Geo distance* is strongly positive. Figure 2.3 leads to a similar conclusion and reveals significant differences, especially for citations linking inventors in the same NUTS3 region, inventors in different NUTS2 regions and in different countries. From Figure 2.3b, when *Time distance* is set at its 10th percentile value, the predicted probability is approximately 67% within the same NUTS3 region, it is significantly lower, at an average of 62% across different NUTS2 regions in the same country and is on average 61% across different countries. When *Time distance* is set at its 90th percentile value, the predicted probability within the same NUTS3 region is 20%, while it increases to 28% across different NUTS2 regions and to 35% across different countries (an increase of 40% and 75%, respectively).

2.4.2 Sectors and discrete and complex technologies

In this section, we present the results of our analyses for different sectors based on the sector of the citing patent.[12] Specifically, we distinguish five sector categories: chemistry, electrical engineering, instruments, mechanical engineering and other fields. In addition, we consider the distinction between complex technologies and discrete technologies.[13] In each category, we only consider patents that report IPC codes corresponding to sectors assigned to that category, excluding from the analysis those assigned to both categories.[14] These series of analyses are depicted in Table 2.9, where Models 1-5 present results for the five selected categories: chemistry, electrical engineering, instruments, mechanical engineering and other fields. In addition, Models 6 and 7 present results for complex technologies and discrete technologies, respectively.

[10] Note that the values discussed refer to the probability that a citation is an X or E citation in our sample, where we only consider patents with at least one X or E citation and an A citation. These values show that the likelihood of observing an X or E citation varies significantly relatively to the sample average (48%). However, importantly, they cannot be interpreted as ratios of duplicated inventions in absolute terms.
[11] Note that continuous geographic distance is 0 when at least two inventors are located in the same NUTS3 region.
[12] We assign sectors following the World Intellectual Property Organization (WIPO) IPC-Technological field concordance table.
[13] This classification follows Von Graevenitz, G., Wagner, S., and Harhoff, D. (2013), "Incidence and Growth of Patent Thickets: The Impact of Technological Opportunities and Complexity". *The Journal of Industrial Economics*, 61(3): 521–563.
[14] These correspond to less than the 15% of our sample.

For the sake of brevity, we only report models based on the continuous measure of geographic distance. The results remain significant for all sectors. However, as expected, the magnitude of the coefficients of geographic distance is considerably lower for chemistry (Model 1). Conversely, the estimate for electrical engineering (Model 2) exhibits the highest values. As expected, complex technologies exhibit higher coefficients than do discrete technologies.

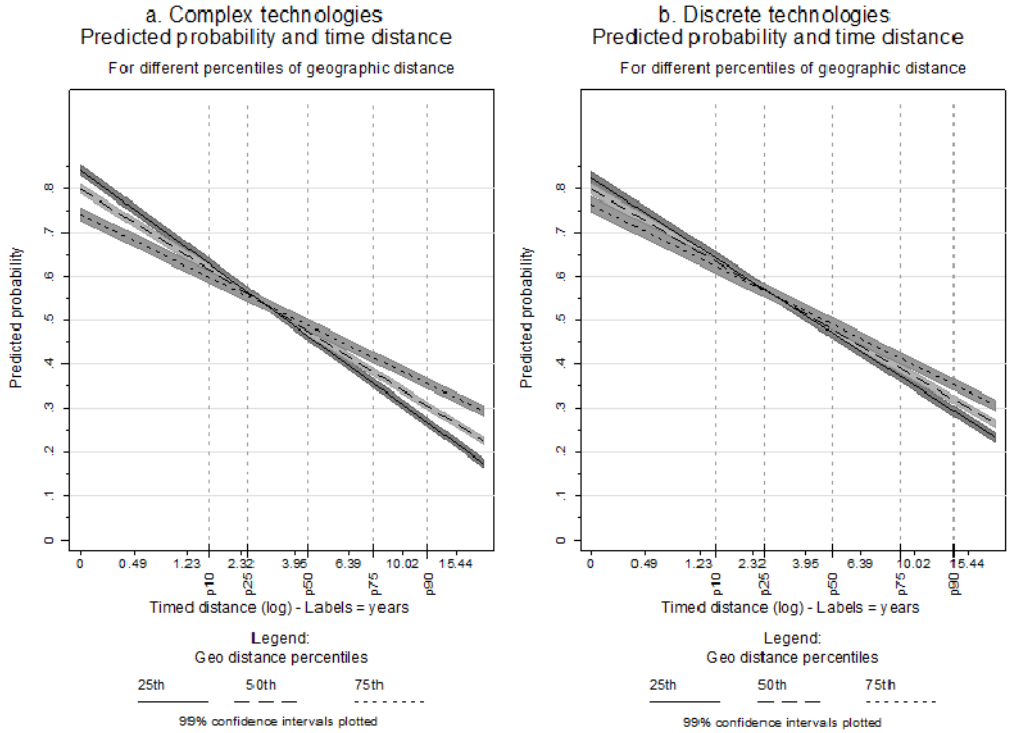
Table 2.9: Sectors and discrete and complex technologies

	Model 1 Chemistry	Model 2 Electrical engineering	Model 3 Instruments	Model 4 Mechanical engineering	Model 5 Other fields	Model 6 Complex technologies	Model 7 Discrete technologies
Same applicant	-0.016*	-0.080***	-0.091***	-0.057***	-0.119***	-0.082***	-0.030***
	(0.009)	(0.013)	(0.016)	(0.010)	(0.025)	(0.009)	(0.009)
Tec similarity	0.052***	0.175***	0.167***	0.129***	0.146***	0.155***	0.061***
	(0.011)	(0.015)	(0.019)	(0.012)	(0.026)	(0.010)	(0.011)
Time distance (log)	-0.229***	-0.314***	-0.318***	-0.222***	-0.213***	-0.279***	-0.222***
	(0.008)	(0.011)	(0.013)	(0.008)	(0.019)	(0.007)	(0.007)
Geo distance (log)	-0.008***	-0.035***	-0.028***	-0.017***	-0.026***	-0.026***	-0.014***
	(0.003)	(0.004)	(0.005)	(0.003)	(0.008)	(0.003)	(0.003)
Geo distance (log) * Time distance (log)	0.009***	0.022***	0.021***	0.014***	0.015***	0.019***	0.011***
	(0.002)	(0.002)	(0.003)	(0.002)	(0.004)	(0.001)	(0.001)
Applicant aware	-0.046***	-0.030**	-0.009	-0.033***	-0.036	-0.022***	-0.048***
	(0.009)	(0.012)	(0.016)	(0.010)	(0.024)	(0.008)	(0.009)
Inventor aware	-0.051***	0.028**	0.022	-0.027***	-0.009	0.012	-0.048***
	(0.008)	(0.013)	(0.015)	(0.009)	(0.022)	(0.008)	(0.008)
Constant	0.827***	0.825***	0.837***	0.751***	0.764***	0.793***	0.826***
	(0.017)	(0.024)	(0.030)	(0.019)	(0.044)	(0.016)	(0.017)
Citing patent FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	88,233	65,597	39,361	89,774	19,191	151,499	94,535
Number of groups	29,525	24,326	14,582	32,552	7,244	56,630	31,570
F	340.5	246.9	178.8	257.6	49.20	516.1	332.7

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

To provide further evidence of the significance of this difference across different technologies, Figure 2.4 reports the predicted probabilities as a function of *Time distance* at different percentiles of continuous geographic distance for complex and discrete technologies (in Figure 2.4a and Figure 2.4b, respectively). The average effect of time is qualitatively identical for complex and discrete technologies, albeit slightly lower for discrete technologies. For complex technologies, the effect of *Geo distance* remains equivalent, if not stronger, relative to those previously discussed for the entire sample. On the contrary, for discrete technologies, the magnitude of the effect is lower, and for most of the values of *Time distance* in the sample, the confidence intervals of the predicted probability at different levels of geographic distance are partly overlapping, meaning that the difference can only be considered significant at extreme values of *Time distance* and *Geo distance*.

Figure 2.4: Predicted probabilities as a function of time distance for complex and discrete technologies



Interestingly, the effects of the revealed awareness also differ substantially across sectors: they are negative and high in magnitude for chemistry but lower and in certain cases not significant for other sectors. Surprisingly, the effect of an inventor having previously cited the focal cited patent is positive in electrical engineering. This result is difficult to interpret, but it could be due to the high level of competition in this sector and the lower effectiveness of patents in communicating property boundaries. The strong effect of the awareness variables in chemistry, in contrast, is consistent with the hypothesis that in this sector, not only are patents exploited as a source of information to a greater extent but they also effectively describe the property boundaries of the applicant, reducing both the risk and opportunity of resulting in overlapping claims (Bessen and Meurer, 2008).

2.5 Robustness

In Table 2.10, we check the robustness of the reported results to a series of potential concerns. First, we also include fixed effects for the cited patents. In order to include both levels of fixed effects we employ the Mundlack procedure, hence including means for all regressors at both the citing patent and the cited patent

levels (Mundlak, 1978). Standard errors are clustered at the citing patent level[15]. We do not adopt this model as our main specification because the inclusion of cited patent fixed effects automatically downward biases the estimate values, as cited patent fixed effects perfectly predict the outcome variable for patents that only appear once as cited patents. However, the results remain significant and qualitatively equivalent (Model 1).

Second, we consider the distinction between citations from the original patent application and those added by the examiner. The main rationale behind this distinction is to investigate the possibility that inventors and applicants add or omit citations for strategic reasons, which might bias our results. It is worth noting that we intentionally avoid interpreting these two categories as a sharp distinction between citations to patents of which the inventor was or was not aware, respectively. This is especially critical in the case of EPO patent applications because inventors do not have any obligation concerning citations to prior art. Therefore, the examiner adds the large majority of patent citations during the search and examination procedures. Moreover, there is no guarantee that the citations present in the original application refer to patents known by the inventor when the invention was in development: citations can be added later, by either the applicant or patent attorney. More important, it is unlikely that inventors add citations to recently filed patents, as they might not yet be public. Similarly, even if they are aware of such patents, inventors have incentives not to cite inventions from competitors, as they might jeopardize the patentability of their own inventions. In Model 2, we only consider examiner citations and the results for the time distance and geographic distance remain essentially unchanged. Interestingly, citations added by the examiner, which the inventor is found to be aware of, are more likely to indicate duplication. This evidence is coherent with the lack of incentives for the inventor to report citations to such patents even if she is aware of them.

Finally, we check the robustness of our results across countries based on the inventors' location. In Models 3 to 8 in Table 10, we separately consider the top five countries in our sample according to patent counts (France, Germany, Italy, Japan, the USA) and aggregate all remaining countries into an additional category (Other countries). Patents are assigned to the geographic area where the majority of inventors are located. The rationale for this test is that part of the results might be driven by country-specific characteristics or countries being covered differently in EPO patent data (De Rassenfosse et al., 2014). Regarding the statistical significance and direction of the effect, the results on time and geographic distance variables hold for all countries; while those on the awareness variables are relatively less robust. Arguably, the differences observed might relate to country differences in technological specialization, institutional frameworks and other context-specific characteristics.

[15] Significance levels are identical when adopting bootstrapped standard errors or block-bootstrapped standard errors with citing patents as clusters.

Table 2.10: Robustness

	Model 1 FE citing and cited	Model 2 Examiner citations	Model 3 France	Model 4 Germany	Model 5 Italy	Model 6 Japan	Model 7 Usa	Model 8 Other countries
Same applicant	-0.054***	-0.051***	-0.090***	-0.057***	-0.146***	-0.027***	-0.071***	-0.077***
	(0.008)	(0.006)	(0.026)	(0.011)	(0.034)	(0.010)	(0.015)	(0.013)
Tec similarity	0.148***	0.117***	0.144***	0.120***	0.116***	0.103***	0.094***	0.137***
	(-0.010)	(0.007)	(0.029)	(0.012)	(0.033)	(0.013)	(0.016)	(0.014)
Time distance (log)	-0.247***	-0.253***	-0.218***	-0.209***	-0.216***	-0.303***	-0.286***	-0.228***
	(0.007)	(0.005)	(0.024)	(0.008)	(0.025)	(0.008)	(0.013)	(0.011)
Geo distance (log)	-0.014***	-0.022***	-0.022**	-0.011***	-0.022**	-0.023***	-0.021***	-0.016***
	(0.002)	(0.002)	(0.009)	(0.004)	(0.011)	(0.003)	(0.004)	(0.004)
Geo distance (log) * Time distance (log)	0.011***	0.015***	0.018***	0.013***	0.014***	0.015***	0.015***	0.014***
	(0.001)	(0.001)	(0.005)	(0.002)	(0.005)	(0.002)	(0.002)	(0.002)
Applicant aware	-0.043***	-0.024***	-0.020	-0.045***	-0.053	-0.014	-0.046***	-0.030**
	(0.006)	(0.006)	(0.032)	(0.010)	(0.036)	(0.010)	(0.013)	(0.012)
Inventor aware	-0.022***	0.012**	-0.111***	-0.052***	-0.001	0.019*	0.018	-0.024**
	(0.007)	(0.006)	(0.024)	(0.009)	(0.030)	(0.011)	(0.013)	(0.011)
Constant	0.511***	0.807***	0.754***	0.733***	0.772***	0.868***	0.862***	0.747***
	(0.006)	(0.011)	(0.054)	(0.020)	(0.059)	(0.018)	(0.028)	(0.025)
Citing patent FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Cited patent FE	Yes	No	No	No	No	No	No	No
Observations	302,156	267,040	16,148	81,836	12,704	75,709	50,996	64,763
Number of groups (citing patents)	108,229	97,342	5,899	29,309	4,638	26,460	18,543	23,380
F	347.22	882.4	36.99	244.9	31.22	399.5	206.0	175.8

Model 1: Cluster robust standard errors in parentheses; Model 2-8: Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

2.6 Conclusion

In this paper, we discussed the determinants of duplicated inventions and empirically examined how this phenomenon is distributed geographically and over time, using data on patent citations. We argued that over time, knowledge flows increase and competitive incentives fade for a given invention. For this reason, we expect that the time elapsed between inventions should negatively affect the probability of duplication. Concerning the geographic distribution, we argued that knowledge flows on emerging technologies and localized opportunities encourage inventors to compete on the same technological path, which makes duplication for recent and upcoming technologies more likely when the inventors are geographically proximate. Finally, we argue that failures in knowledge flows at high geographic distances may also cause independent duplications for not recent inventions, such as that at high time distances, we expect to observe duplication at high geographic distances.

We exploit patent citation information from EPO patents to map duplicated inventions with respect to the original inventions (i.e., X and E citations) and to other related inventions that do not threaten the patentability of the invention in question (i.e., A citations). In other words, our results are based on comparing the geographical, time and technological differences of the pair of duplicating-original patent applications to the pair of duplicating-related patent applications. In accordance with our hypotheses, when more proximate in time, we observe that original patents are more likely located at shorter distances and within the same region compared to related patents. Conversely, the effect of distance is reversed as time passes between the original and duplicating patents. As such, “relatively old” patent applications present

overlapping invention claims when inventors are geographically distant, especially in different countries. While several economic models have considered duplication as a random outcome (Gómez, 2011; Jones and Williams, 2000), our results suggest that the likelihood of duplication is unevenly geographically distributed.

Based on this evidence, we can also discuss potential further implications. First, we infer that proximity and knowledge flows involve more detailed underlying dynamics than would be implied by a simple learning process. Proximity allows for learning from existing technologies and the identification of valuable opportunities, hence avoiding investments in existing technologies. Therefore, a potential crucial advantage of inventors located in technological clusters is the possibility to identify technological opportunities in a timely manner while, simultaneously, avoiding duplication. However, knowledge flows relating to upcoming technologies might encourage different inventors to pursue the same idea. These dynamics, combined with strong competition and the lack of alternative sources of knowledge and opportunities, might reduce the innovative capacity of a region (Audretsch and Feldman, 1996; Martin and Sunley, 2006). Our results also suggest that inventors in distant locations are exposed to the risk of replicating R&D efforts performed even years before. Therefore, these considerations suggest conceptualizing knowledge flows in accordance with a broader taxonomy of outcomes: 1) internalized knowledge flows that enable cumulative innovation; 2) rival knowledge flows exploited to anticipate competitors that might lead to a certain degree of duplicative efforts; 3) imperfect knowledge flows that cause inventors to miss opportunities for innovation; and 4) imperfect knowledge flows that cause redundant (duplicative) innovation efforts.

Second, we can reinterpret our empirical findings in the context of information disclosure through patent documents. Arguably, if the information disclosed in patent documents were sufficient to ensure perfect diffusion, we should not have observed any significant pattern in duplication with respect to geographic distance. It is worth noting that this does not necessarily imply that patent documents fail to disclose information but rather that knowledge simply is not easily diffused. It remains possible that patent information is more disclosed within the patent system than outside of the system. Moreover, our results suggest that in certain sectors – especially those related to discrete technologies – patents are a more successful channel for knowledge diffusion through geographic regions. In a similar vein, we also found that inventors and applicants are less likely to duplicate inventions protected by patents that they have cited in the past. To some extent, this suggests that inventors avoid duplicating patents of which they are aware. However, again, this result was not stable across different specifications and especially not across different sectors.

Finally, our methodology also contributes to the debate on the meaning of patent citations. Our interpretation is primarily based on the definitions of the patent citation categories. In our framework, X or E citations are by definition an indication (based on the opinion of an examiner) of an overlap between two inventions that might or might not be the result of knowledge flows. Similar to previous studies, our results are consistent with the understanding of knowledge flows being geographically localized but cast further

doubts on the use of patent citations as direct indicators of knowledge flows. In this sense, we confirm that examiner citations can present significant geographical patterns with respect to inventor location (Alcacer and Gittelman, 2004). As a consequence, it remains highly ambiguous which patent citation category is more appropriate as a direct indicator of knowledge flows. Moreover, our results contribute to the concern that certain citations, particularly when added by the examiner, might actually reflect a lack of knowledge flows.

A limitation of our analysis is that we cannot fully distinguish independently duplicated inventions from those issuing from competitive behavior. This is due to the impossibility to perfectly measure the awareness of inventors. Similarly we do not dispose of a measure of competitive incentives, holding constant the age of a technology, among inventors. Our evidence is only consistent with both the existence of inventions independently duplicated because of missing knowledge flows and inventions duplicated as a result of rival R&D strategies fostered by local knowledge flows. Another limitation concerns our having implicitly assumed the existence of welfare costs of duplication, issuing from suboptimal investments in inventive efforts. On the contrary, duplication might be, for example, the necessary side effect of virtuous competitive dynamics within technological clusters. More generally, considering the sub-optimality of a given level of duplication would require accounting for the incentives to innovate, the amount of resources allocated to a given inventive effort and the costs of information access and knowledge diffusion. In this sense, also independent duplication is only socially sub-optimal if the cost of accessing information and knowledge diffusion are assumed to be lower than the cost of producing redundant inventions. All this information cannot be included in our framework.

2.7 Appendix

In this appendix we compare a citing patent application with one patent cited with a citation categorized as X (X-cited) and one patent cited with a citation categorized as A (A-cited). The comparison is useful in order to understand the methodology we adopt which is based on two fundamental characteristics of the data: 1. the invention claimed in the X-cited document has to overlap in some substantial manner with the invention claimed in the citing patent application, providing the indication of duplication; 2. the invention claimed in the citing patent application is very close to the invention claimed in the A-cited document but presents some substantial novelty compared to it. The example presented is taken and summarized from the Patent Teaching Kit provided by EPO[16] (pg. 248). Discerning different categories of citations in general requires a deep knowledge of the technology and a detailed analysis of the patent claims. The description provided in the Patent Teaching Kit EPO helps identify the peculiar characteristics of an X-cited document, compared to an A-cited document. Table A1 reports the descriptions of the citing patent, the X-cited patent and the A-cited patent.

Table 2.11: Appendix - Citation categories example

Patent	Citing	X-Cited	A-cited
Title	A washing device heated electronically*	Heating device used for a household appliance	A heating device for washing and/or drying machines for laundry
Publication n.	EP 03005120	DE 10025539	EP 0352499 A2
Description	A heater comprising a foil heating element which is attached to or integrated into the lower part of a tub, which has contact with the medium to be heated. The tub is the receptacle that contains the water and washing powder or liquid inside the washing machine. The drum rotates inside the tub around an axis, which in this example is slightly inclined. The foil heating element is attached (e.g. glued) to or integrated into the lower part of the tub. Thus, the foil heating element is also "adapted in its shape" to the bottom of the tub. Insulating strips may be used for dividing the foil heating element into different sections. The proposed heater is simple and inexpensive to manufacture, less prone to interference and provides the possibility to reduce water and energy consumption. Thanks to the simple structural design of the foil heater, the risk of calcification and linting is significantly reduced.*	Heating device comprises a ceramic-filled polymer layer arranged between a surface of the appliance to be heated and an electrically conducting heating foil. Preferred Features: The heating foil is covered on the side facing away from the polymer layer by an insulating molded body. The polymer layer has a thickness of 70-150 microns and the heating foil is an iron-chromium-aluminum alloy. The insulating molded body is made from vermiculite.	A heating device for washing and/or drying machines for laundry, com-prising a parallelepiped plate made of electrically insulating material having applied to one surface thereof, by the silk screen process or simi-lar procedures, at least one electrical resistor based on electrically conductive metal powders mixed with glass frit, the resistor having a wavy or other pattern.The plate is placed inside the vessel of the washing or dry-ing machine for laundry and the corresponding terminals of the resistor are then connected with the power supply network of the machine. One thus obtains an elevated heat radiating capacity of the plate adapted to effect rapid heating of the washing solution or the drying air of the machine in question.

* note: title and description translated and adapted from the original (German) by the authors and based on the description reported on the Patent Teaching Kit provided by EPO

The citing patent application relates to a heating element for a washing machine. The A-cited patent identified during the prior art search shows a very similar device. However, the analysis of the patent claims reveals a technical difference, which translates into a technical effect and constitutes an inventive step: “The invention as claimed allows a much lower minimum water level than is possible with the washing machine shown in EP0352499, because the foil heating element is directly fitted onto the wall of the tub” (Patent

[16] http://www.epo.org/learning-events/materials/kit/download.html

Teaching Kit EPO, pg. 251). On the contrary, the device disclosed in the X-cited patent comprises all of the additional features initially claimed in the citing patent: a heating foil with its supporting polymer layer and heat-conducting material. “There is at least one interpretation of the prior art which is conclusive and logical that shows all the features of the claim of our invention” (Patent Teaching Kit EPO, pg. 253). The citing patent could be granted only conditionally on a revision and reduction of the specific features to be protected (claims). Similar examples could be drawn from our sample that for the sake of brevity, are not shown here.

Chapter 3 The scientific productivity of PhD students from professors' networks

(With Annamaria Conti and Fabiana Visentin)

3.1 Introduction

To date, academia remains a major locus of knowledge and innovation production. Within academia, graduate students play a fundamental role both in the creation and transfer of knowledge within academia or to industry (Black and Stephan, 2010; Leten et al., 2014; Stephan, 2012; Van Looy et al., 2011). Therefore, various scholars have studied the scientific productivity of different categories of PhD students (Gaulé and Piacentini, 2013; Levin and Stephan, 1999). A growing literature shows that mobile graduate students, and more in general researchers, hired from external environments, relative to the institution in which they are employed, are highly productive (Gaulé and Piacentini, 2013; Levin and Stephan, 1999; Libaers, 2007). Consequently, hiring from other institutions and environments appears crucial to enhance performance in research institutions.

However, hiring processes are affected by strong information asymmetries (Arrow, 1972; Granovetter, 1995) and institutions might favor internal or better known candidates (Horta et al., 2010). As a consequence, a large body of literature suggests that the existence of networks crossing institutional boundaries can impact the possibility to hire productive candidates from other environments (Granovetter, 1973; Ponzo and Scoppa, 2010). Indeed, inter-institutional and international collaborations are increasingly a fundamental component of the academic profession (Adams et al., 2005; Baruffaldi and Landoni, 2012; Hoekman et al., 2010; Jones et al., 2008). From a theoretical standpoint their effect might be both positive and negative. Few studies have addressed this issue in academia (Lissoni et al., 2011; Pezzoni et al., 2012) and in other contests evidences are mixed (Antoninis, 2006; Castilla, 2005).

This paper contributes to the existing literature by empirically comparing the scientific productivity of three different categories of students: PhD students that obtained their master's degree from the same

university as the one of the PhD (hereafter "internal students"); PhD students who are hired from external universities with which no one of the supervisor's co-authors is affiliated (hereafter "external students outside professors' network"); PhD students who are hired from external universities but with which their supervisor's co-authors are affiliated (hereafter "external students from professors' network"). In particular we focus our discussion on this latter category of students. In principle, students in this category can have lower productivity if professional networks lead to favoritism (Horta et al., 2010; Prendergast and Topel, 1996) limiting the potential benefits of hiring from external institutions. On the contrary, they could have higher productivity if, reducing the information asymmetries, networks allow to hire the students with the highest expected productivity (Cornell and Welch, 1996; Saloner, 1985). Other different mechanisms, which based on the literature we define as post-hiring social effects (Fernandez et al., 2000; Yakubovich and Lup, 2006), might also explain a positive effect: for instance, the supervisor might interact more closely with the student because of the existence of a social connection.

We conduct this analysis using a novel dataset of 4,666 PhD students in science or engineering who graduated from two major Swiss technology institutes: the Swiss Institute of Technology of Lausanne (EPFL) and the Swiss Institute of Technology of Zurich (ETH). In our sample, approximately one half of the students obtained a master from a different university from the one of the PhD (external students) and approximately half of them (one quarter of the entire sample) are from the supervisor's research network. In line with previous studies (Gaule' and Piacentini, 2012), in our baseline analyses, we consider detailed controls at student and professor level, and we adopt professor fixed effect estimation. Additionally, we control for mutual characteristics of the student and the professor that might directly confound the effect of professor's network. Finally, we control for the heterogeneity of institutions where the students obtained their master's degrees, including proxies of the specific relevance of these universities for a given supervisor.

We find that external students from professors' network are significantly more productive than both internal students and other external students. We show that our models pass a placebo test where we construct our variable of interest based on the network of a different professor whose research interests are close as possible to those of the supervisor of the focal student. Thus, we conclude that the presence of a professor's coauthor in a given university has a positive effect on the productivity of students hired from that same university. In the last session of the paper, we provide a series of additional analyses that suggest that the main mechanism likely to explain the effect encountered is the resolution of information asymmetries, and we rule out some of the possible alternative explanations.

3.2 Conceptual framework

In this paper, we ask whether students with a master degree from a different university of the one of the PhD but hired from a university with which their supervisor's co-authors are affiliated (external students

from professors' network) have a different average productivity as compared to other external students and from internal students.

Faculty networks can affect productivity outcomes of hired PhD students in multiple ways. In principle, they can encourage practices, such as favoritism, that are likely negatively correlated with students' productivity outcomes. In this sense, the presence of personal connections might confound the process of objective selection of external candidates (Prendergast and Topel, 1996). Horta et al. (2010) recognize that the negative effects of institutional inbreeding (defined as the practice of having researchers hired from the same university who trained them, that typically turn out to be less productive) can be the consequence of rational behavior: since internal networks and the pre-existing information on internal candidates lower the uncertainty regarding their quality vis-à-vis external candidates, employers might prefer internal candidates regardless of a lower expected productivity. Professors' networks typically and increasingly cross institutional boundaries and the same arguments might apply to candidates coming from connected environments as compared to candidates from unconnected ones.

Indeed, hiring processes are affected by strong information asymmetries. However, the economics and sociological literature have typically identified networks to be potentially beneficial for labor market outcomes, precisely because they might decrease asymmetric information allowing to assess the "unobservable habits of action" (Arrow, 1972; Granovetter, 1995). Therefore, networks can allow for the identification of candidates with higher expected productivity. As a consequence, the average quality of the applicants hired from somehow connected environments may be high in equilibrium, higher than the quality of employees for what the employer could only rely on formal screening mechanisms (Montgomery, 1991; Saloner, 1985). Notably, networks might operate reducing information asymmetries both for the employer and the employee, not only increasing the possibility for the employer to select better candidates, but also improving the complementarity of the match between the two (Fernandez et al., 2000; Yakubovich and Lup, 2006).

These theories, leading to the hypothesis that networks positively impact labor market outcomes, do not exclude the possibility of negative effects. Granovetter (1995) firstly outlined the notion that connections not involving a strong relationship - "weak ties" -, are sufficient to gain most of the information advantage leading to improvements in the labor market outcomes. On the contrary, when "strong ties" are present, such as strong collaborations, friendship and family ties, favoritism might prevail, leading to negative outcomes. In addition, recent literature has further detailed and extended the mechanisms trough that networks might have a positive impact. Based on this literature we distinguish broadly between hiring effects and post hiring effects (Antoninis, 2006; Fernandez et al., 2000). The former can be considered as strictly related to the effect of information asymmetries, as discussed above. On the contrary, the latter concerns possible post-hiring social dynamics that affects differently individuals hired through a network. For instance, individuals

hired within one employer's network can receive better supervision or might feel more motivated to perform well.

Positive effects of networks have been found in different contexts such as low skilled immigration (Munshi, 2003), call-center employees (Castilla, 2005) start-ups selected by venture capitalists with ethnicity ties (Hegde and Tumlinson, 2011). On the contrary, other studies have found negative network effects on economic outcomes (Ponzo and Scoppa, 2010). Negative effects are especially found when either strong informal connections, rather than professional connections, are present (Ponzo and Scoppa, 2010) or, similarly when family ties are used (Sylos Labini, 2005). The role of networks in academia is a relatively new development and especially the literature on PhD students has so far overlooked this aspect. Existing evidence indeed suggests that similar dynamics might be in place in academia. Laband and Piette (1994) and Brogaard et al. (Brogaard et al., 2014) find that journal editors use social connections to identify high-impact papers for publication. Moreover, Li (2012) finds that the presence of related reviewers improves the quality of research that the NIH (National Institutes of Health) supports. Social connections between candidates and evaluators have a positive impact on candidates' promotion in academic careers, although evidence on candidate research productivity is mixed (Pezzoni et al., 2012; Zinovyeva and Bagues, 2012).

None of the above mentioned studies has considered the role of supervisors' networks in PhD student outcomes. This role might be crucial because, as many scholars have emphasized, PhD programs are predominantly populated by foreign or foreign-educated students, and information asymmetries are definitely a concern for these students (Black and Stephan, 2010; Gaulé and Piacentini, 2013; Stuen et al., 2007). As suggested by Gaulé and Piacentini (2013), supervisors tend to resolve this information problem by selecting students from a handful of selective schools. However, the supply of students from these schools is limited, and even within selective schools, there is wide variability in student quality or backgrounds. We examine the relationship between supervisor networks and student productivity, bearing in mind that networks can affect productivity outcomes of hired PhD students positively or negatively, and through different mechanisms.

3.3 Data and methods

3.3.1 Context

Our empirical context involves PhD students from EPFL, Lausanne, and from ETH, Zurich. These universities are the two Federal Institutes of Technology of Switzerland. EPFL is located in the French-speaking part of Switzerland, and ETH is in the German-speaking part. EPFL and ETH are responsible for a large portion of the research in science and engineering that is produced in Switzerland, and they host the

largest doctoral programs in these disciplines. Official statistics from 2011 reveal that EPFL and ETH hosted 60% of the PhD students in science and engineering enrolled in Switzerland[17].

From the population of PhD students, we select a sample of 4,666 PhD students who had graduated from EPFL or ETH during the 2000-2008 period. Universities' human resource departments had extensive biographical information on the students who graduated during this time window. We complement the information from human resources with that extracted from student dissertations. We match this information with student publication records using Scopus. We also collect fine-grained data on student supervisors. There are 558 professors (227 at EPFL and 331 at ETH), each of whom supervised an average of eight PhD students at the time of our sample period. In the sample, 36% of the PhD students were affiliated with EPFL, and the remainders were affiliated with ETH. For EPFL, we selected all of the PhD students who had graduated during our sample period, whereas for ETH, which is a much larger university, we considered only a sub-sample. The latter was obtained by randomly selecting a sample of supervisors in each department and including their PhD students. In this manner, we cover approximately 30% of the PhD students who graduated from ETH during the 2000-2008 period.

When classified by discipline, 14% of the EPFL PhD students are in computer science, 39% are in engineering, 4% are in life science, and the remaining students are in basic science. For ETH, 6% of students are in computer science, 42% are in engineering, 13% are in life science, and 39% are in basic science. PhD students are selected by professors through a formal interview process[18]. Once a professor hires applicants, they work with that professor for the entire duration of their PhD program. Hence, switching to another supervisor is rare.

PhD students at both EPFL and ETH generally complete their PhDs within four years. Extensions are possible, but they are typically no longer than six months. The dropout rate is approximately 10% at both universities. PhD applicants must have already obtained their master's degrees. Hence, they spend most of their time performing research rather than taking courses, given that they have already taken most courses during their respective master's programs. In general, EPFL and ETH are multi-cultural environments with PhD students originating from a variety of countries and academic institutions, partly because of the high-quality research that is pursued at these universities and the high salaries that are offered to PhD students, compared with other countries[19].

[17] These data were obtained from the Swiss Federal Statistical Office. They can be accessed at http://www.bfs.admin.ch/bfs/portal/en/index.html.
[18] Beginning in 2006, EPFL changed its rules and established that PhD applicants must submit their application to a central committee, which conducts an initial screening. Our sample does not include cohorts who joined after 2005.
[19] Data on salary differentials can be accessed at http://jahia-prod.epfl.ch/files/content/sites/acide/files/ activities/documents/sondage-ACIDE-Doctorants.pdf.

Of the PhD students in our sample, 59% obtained their master's degrees at a university other than the affiliation of their PhD (the percentage is 67% at EPFL, and 54% at ETH). Moreover, 56% of students at EPFL and 46% at ETH were previously affiliated with universities outside of Switzerland. Not surprisingly, PhD students from French universities constitute the largest foreign group at EPFL: they represent nearly 13% of all PhD students. Similarly, PhD students coming from German universities are the largest foreign group at ETH, representing almost 22% of the total students. Non-European students constitute a small minority at both universities. For instance, North American students represent approximately 1.5% of all PhD students at each school, whereas Chinese and Indian students together represent approximately 3%.

Regarding the supervisors, approximately 55% of them are foreign at EPFL, whereas the percentage of foreign professors is 58% at ETH. At both EPFL and ETH, the largest foreign group is composed of German professors, who represent, respectively, 11% and 26% of total foreign professors.

3.3.2 PhD students categories

First, we distinguish between internal students (PhD students with a master's degree from the same university of the PhD) and external students (PhD students with a master's degree from a different university). As noted above, in our sample, this latter category accounts for the 59% of students. In our empirical analysis we are interested in further distinguishing external students who are hired from a university within the network of their supervisors (external students from professors' network). We construct the latter measure as an indicator that takes a value of one if the student had obtained her master's at one of the universities from which her supervisor draws her co-authors[20]. For this purpose, we searched for the academic institutions with which supervisor *j's* co-authors are affiliated and compared them with the universities from which *j*'s students had obtained their master's degrees[21]. Data on coauthors' affiliations are available from Scopus. For each student *i*'s affiliation, the relevant comparison is with the affiliation of supervisor *j*'s co-authors, who had written scientific articles with *j* up to the student's year of entry into the doctoral program[22]. We consider only those institutions in the last three quartiles of the supervisor's distribution of coauthor affiliations. By applying this cutoff, we aim to smooth out noise[23]. 47% of the external students, corresponding to 27% of the total sample, are hired from a university within the network of the supervisor. The remaining 53% (31% of the total sample) obtained their master's from a University where their supervisors have no co-authors.

[20] This measure resembles that employed by Laband and Piette (1994), who measure author/editor connections with an indicator variable equal to one if any of the authors of a paper had received a PhD degree from the same affiliation as that of the editor.
[21] The comparison was performed by adopting a matching algorithm implemented by Raffo and Lhuillery (2009)
[22] We consider a lag of three years between the time that the coauthors begin writing an article and the time at which the article was published. The results are robust to applying other cutoffs.
[23] The results hold when we do not apply a cutoff.

Note that, taking into account previous studies (Gaulé and Piacentini, 2013; Levin and Stephan, 1999; Libaers, 2007), we might further distinguish these categories of PhD students based on their nationality or the location where they completed their studies. Nonetheless, in our sample, the category of external students coincides to a large extent to the one of foreign or foreign-educated students: about 87% of external students are also foreign or foreign-educated. Similarly, 85% of the internal students are Swiss. Therefore, we refrain from further detailing the categories considered in the analyses because this would lead to a high number of poorly represented categories. Nonetheless, the results of the analyses we present would not change if we further control for the nationality of the students.

3.3.3 PhD students productivity

We measure productivity using two different count criteria of the scientific articles that a student published during her PhD program. These measures have been frequently used in studies of scientific productivity (Ding et al., 2010). We adopt a broad definition of scientific articles and include papers that have been published in conference proceedings. This definition is adopted because an important fraction of our students are in computer science and in electrical engineering, and in these fields, conference proceedings have at least the same importance as journal articles[24]. We count a student's publications from the moment that the student enters the doctoral program until one year after graduation[25]. This specification accounts for the lags between the time at which a student completes a research project and the time at which the results of the project are published (see, for example, Arora and Gambardella, 2005).

In the main models we count all articles as retrieved from Scopus attributable to the PhD. Figure 3.1 displays the distribution of PhD students by the total number of articles that they have published. As shown, a large percentage of the students, approximately 85%, had at least one publication, and 75% had more than one publication. The average student publication count is similar for EPFL and ETH: 4.61 for EPFL and 4.75 for ETH. By discipline, the average number of student publications is 5 in computer science and basic science, and 4 in engineering and life science.

In a second set of models we consider only articles above a given threshold of quality (High quality publications). In this case, we count only those publications having received a number of citations that is higher than the median number of citations received by student articles that were published in the same year and in the same field as student *i*'s articles. We adopt this methodology rather than a simple citation count because of standard problems that are inherent to measuring research quality using citation counts. For

[24] We do not apply any weighting based on impact factor because conference proceedings rarely have an associated impact factor.
[25] In robustness analyses that are not reported here, we also count the number of publications from the year that a PhD student begins the doctoral program to two years after graduation. The signs and significance of the coefficients remain unchanged.

instance, Amin and Mabe (2000) show that citation counts vary significantly across subject fields and that they are correlated with factors such as article length or article styles (whether they are reviews or not), which are not informative of research output quality[26]. As shown in Figure 3.2, by applying this criterion, we find that the percentage of students who had not published an article increases from 15 to 30. The average student publication count is now 2.34 for EPFL and 2.78 for ETH. By discipline, the average count of student publications is 3 in computer science and basic science and 2 in engineering and life science.

Figure 3.1: Distribution of PhD students by their publication count

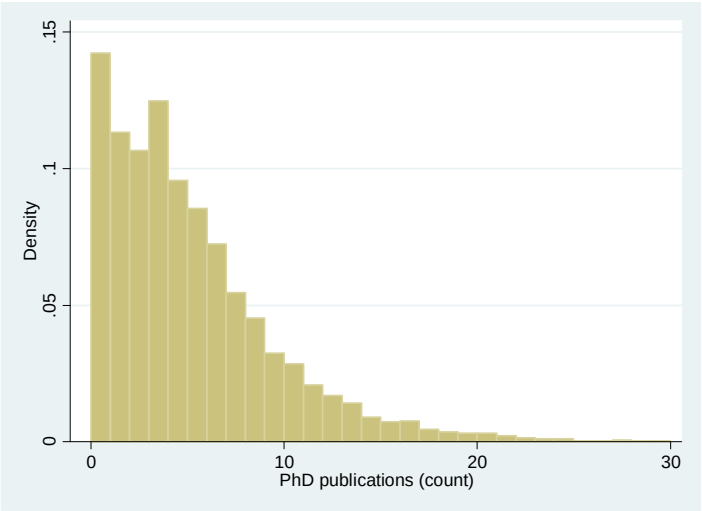
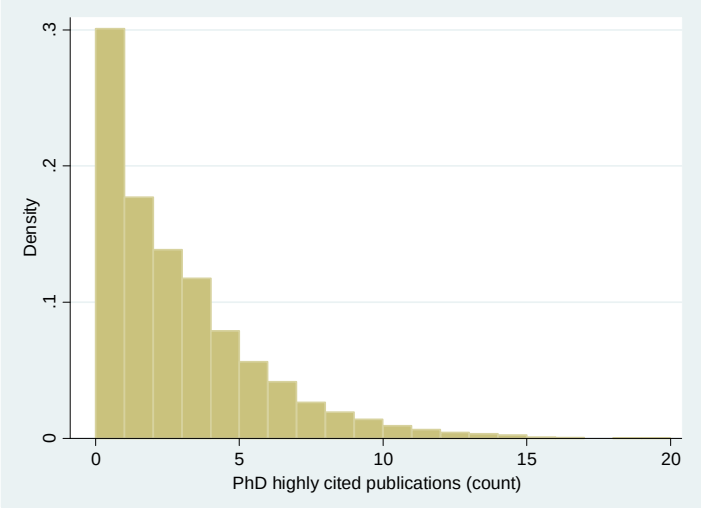


Figure 3.2: Distribution of PhD students by their highly cited publication count



[26] Furthermore, we could only gather data for the cumulative number of citations obtained for each publication up to the year 2012. Therefore, depending on the publication date, different truncation periods are applied to the count of citations, introducing additional noise in the analysis. In robustness checks that are not presented here (but are available upon request), we used the citation count as the dependent variable and obtained similar results. The coefficient on the main variable of interest was still positive and statistically significant, although at the 10% confidence level.

3.3.4 Model

We estimate a quasi-maximum likelihood estimation (QMLE) with fixed-effects Poisson model, given that our measure of student research productivity can only take positive integer values (Hausman et al., 1986). This model has several desirable properties, including consistency in the coefficient estimates and in the standard errors regardless of assumptions on the variance functional form (Wooldridge, 1997). Hence, we estimate the following equation:

$$E(\#Pubs_{ijt}|X_{ijt}) = \exp(\beta_0 + \beta_1 \textit{External student from prof's network} + \beta_2 \textit{External student outside prof's network} + \beta_3 CProx_{ijt} + \beta_4 CPhD_i + \beta_5 CProf_{jt} + \beta_4 CProfUni_{ij} + EntryYearFE_t + ProfFE_j + UniFE_i)$$

Table 3.1 summarizes the variables adopted in the analyses and their definition. The main variable of interest is *External student from prof's network,* which is equal to one if a student is an external student from the network of the supervisor. In addition, we include a dummy for external students outside their supervisors' networks: *External student outside prof's network.* These two dummies are mutually exclusive. We present results both for the entire sample of students and for external students only. Note that in the first case (full sample) the baseline category of comparison for the variable of interest *External student from prof's network* are internal students. In the second case (sample of external students) the category of comparison are other external students outside the professor network.

To capture the effect of belonging to the research network of a supervisor, we must control for factors that are correlated with our indicator variable and are likely to affect student productivity. Hence, we use three sets of controls. The first set, $CProx_{ijt}$ includes three measures of the cultural proximity between a student and her supervisor and the research group. The first measure is a dummy variable that is equal to one if the student and the supervisor have the same nationality. We denote this variable as *Prof & PhD have same nationality*. The second measure is a dummy that is equal to one if the student obtained her master's degree in one of the universities with which her supervisor was previously affiliated: *Prof & PhD come from same university*. Finally, the last measure is defined as the number of students from j's group who have the same nationality as *i*, *N of Prof PhDs with same nationality as PhD i*. The logic for including these variables is that they are likely correlated with the variable of interests and might affect at the same time scientific productivity. In particular, working with culturally proximate colleagues may reduce interaction costs, with a resulting positive impact on the productivity of the students (Fernandez et al., 2000; McPherson et al., 2001).

.

Table 3.1: Variable description

Scientific productivity	
Student publications	Count of student publications.
Student highly cited publications	Count of publications that had received a number of citations higher than the median of citations received by other student articles, published in the same year and in the same field.
PhD students categories	
External student from prof's network	Equal to 1 if a student obtained her mater's degree from a university, different from the one of the PhD, from which her supervisor draws her coauthors.
External student outside prof's network	Equal to 1 if a student obtained her mater's degree from a university, different from the one of the PhD, from which her supervisor does not draw her coauthors.
Student-Prof pair controls	
Prof & student have same nationality	Equal to 1 if a student has the same nationality as her supervisor.
Prof & student come from same university	Equal to 1 if a student had obtained her master's at one of her supervisor's past affiliations.
N of Prof students with same nationality as student	Number of supervisor *j* 's PhD students with the same nationality as student *i*.
Student main controls	
Student age	Student age.
Gender	Equal to 1 for female gender.
Pre-sample student pubs	Count of papers a student had published prior to starting her PhD.
Ranking of master's university	Ranking assigned by the QS World University Rankings to the university at which student *i* had obtained her master's.
Professor-Lab main controls	
Prof age	Professor age.
Pre-sample prof pubs	Count of articles supervisor *j* had published in the five years prior to the year in which student *i* begins her PhD.
Pre-sample prof patents	Count of patents supervisor *j* was granted in the five years prior to the year in which student *i* begins her PhD.
Size PhD group	Size of a supervisor *j* 's PhD group at the moment a PhD student *i* enters the group.
Grant amount	Amount of basic research grants (in hundred thousands of real Swiss Francs) supervisor *j* had obtained during the period in which she had supervised student *i* . This amount is averaged over the duration of the student's PhD.
Prof-University of master pair controls	
Pubs stock of master's university (by field)	Count of the publications the university from which student *i* had obtained her master's had produced in the research area of the student. We consider the following areas: physics, mathematics, chemistry, engineering, material science, life science, and computer science.
Ranking of master's university (by field)	Equal to 1 if the university from which the student had obtained her master's degree is in the top 50 universities for the research area in which the student is specialized, according to the QS World University Rankings.
N of times master's university is cited by prof	Number of times that supervisor *j* cites in her articles authors who are affiliated with the same university as the one from which student *i* obtained her master's degree.
Placebo and robustness controls	
Placebo most cited author	Equal to 1 if a student had obtained her master's degree from one of the universities from which the most highly cited scientist by supervisor j derives her coauthors
External from prof's network w/ pre-sample pubs	Equal to 1 if External student from prof's network is = 1 and if the student had published at least on paper prior to starting her PhD.
External from prof's network w/o pre-sample pubs	Equal to 1 if External student from prof's network is = 1 and if the student had not published any paper prior to starting her PhD.
Student with pre-sample pubs	Equal to 1 if the student had published at least on paper prior to starting her PhD.
Intensity of network ties	Share of supervisor *j* 's papers coauthored with scientists affiliated with student *i* 's master's university.
Share of pubs with prof	Share of student publications coauthored with her supervisor.

$CPhD_i$ is a matrix of controls related to PhD student i. The included variables are standard controls potentially correlated with the productivity of a PhD candidate. Specifically, $CPhD_i$ encompasses a student's age at entry into the PhD program and a gender dummy that is equal to one for females. Previous studies have shown that there is a negative correlation between age and research productivity (Levin and Stephan, 1991) and that there are gender differences in the production of scientific articles (Ding et al., 2006; Long, 1990). We control for aspects of a student's quality that a professor can verify by examining the student's curriculum. For instance, we count the number of the student's publications in the two years prior to the beginning of her PhD studies. This variable, which we denote as *Pre-sample student pubs*, can be considered a strong signal of the research skills of the student when she applies for a PhD. We include a measure for the average quality of the university from which a student has obtained her master's degree, *Ranking of master's*

university, which corresponds to the ranking assigned to that university by the QS World University Rankings[27].

$CProf_{jt}$ is a matrix of controls for supervisor j. We include the age of a professor when her student began the doctoral program (*Prof age)*. Moreover, we include the professor's number of publications published in the five years prior to the year in which a student begins her PhD (*Pre-sample prof pubs*). We regard this variable as a measure of the knowledge capital that a professor shares with her students and that complements her students' contributions in the production of scientific output (Waldinger, 2010). We also include a variable that indicates the five-year pre-sample count of US and European patents that a professor was granted (*Pre-sample prof patents*). This variable is intended to capture professor involvement with industry over time. When professors are involved with industry, they might be more interested in having their students work with industry partners than in publishing articles. Previous studies have shown a positive impact of grant money on scientific productivity (Ding et al., 2010; Ganguli, 2010). Consistent with these studies, we include the amount of basic research grants (in hundreds of thousands of real Swiss Francs) that a professor had obtained during the period in which she supervised student i. This amount is averaged over the duration of the student's PhD program. We denote the variable as *Grant amount*. Finally, we also control for the size of a supervisor's PhD group at the moment that a PhD student enters the group (*Size of PhD group*).

In our base-line regressions we include all controls mentioned above and we add university-department fixed effects[28], as there are differences in the publication patterns of PhD students across departments (Stephan, 2012) and across universities. Moreover, we control for year fixed effects, based on the year in which the student begins her PhD program. Following Gaulé and Piacentini (2013), in subsequent model specifications we include professors' fixed effects. Professors' fixed effects control also for specific characteristics of the subfield of specialization of the student, at a more detailed level with respect to university-departments fixed effects. Indeed, university-department fixed effects model is nested in the professor fixed effects model. More importantly, we control for any additional time invariant unobserved characteristic of the professor that, being correlated with her network and the average productivity of her students, might bias our results.

Finally, in a last set of models, we further address the issue of the quality of the university of the master's degree of the student. More generally, we consider that the network of co-authors of a professor is not random but likely correlated with the quality of the university to which a coauthor is affiliated. This fact

[27] The variable was rescaled such that the coefficient represents a change in student productivity corresponding to an increment of 100 in the university ranking.
[28] The departments we consider are the Physics, Mathematics, Chemistry, Civil Engineering, Electrical Engineering, Mechanical Engineering, Micro Engineering, Material Science, and Computer Science Departments. Each department is considered separately for ETH and EPFL.

makes the presence of a coauthor in a given university potentially endogenous to the average quality of students in that university. As a consequence, we might observe a positive effect of our variable of interest while more productive students would be drawn from that university anyway, regardless of the presence of a coauthor. Indeed, our baseline controls might not be enough to solve this issue. To overcome this concern, first, we include fixed effects for the university of the master's degree of the student.

Second, we consider an additionally set of variables ($CProfUni_{ij}$) that are meant to control for the specific relevance of a given university relative to the research interests of a professor. This is because, although student-university fixed effects are a strong control for the average characteristics of a university, these characteristics may vary significantly by research area. It is possible that a professor establishes contacts with universities that are relevant for her specific field of research and that, independently from her contacts, she hires students from these universities because they are specialized in her research field.

First, we use a dummy, *Ranking of master's university (by field)*, which is equal to one if the university from which the student obtained her master's degree is among the top 50 universities for the research field in which the student has specialized, according to the QS World University Rankings[29]. We consider the following fields: engineering, computer science, and basic sciences. Second, we construct a count of the publications that the university has produced in the student's research field. We collect this information from the Scopus publications' database. In this case, we consider a more fine-grained list of fields: physics, mathematics, chemistry, engineering, material science, life science, and computer science. We denote this variable as follows: *Pubs stock of master's university (by field)*. Finally, we constructed a variable that is defined as the number of times that supervisor *j* cites in her articles authors who are affiliated with the same university as the institution from which student *i* obtained her master's degree (*N times master's university is cited by prof*). Associating this count with each student's past affiliation provides a strong indication of the relevance for a supervisor's research of that affiliation. Descriptive statistics for all variables are reported in Table 3.2.

[29] Additional details can be found at http://www.topuniversities.com/university-rankings/world-university-rankings. We prefer to use this ranking rather than the Shanghai Jiao Tong ranking of universities because the former encompasses a more comprehensive list of European universities.

Table 3.2: Variable descriptive statistics

	All students					*External students*				
Variable	Obs.	Mean	Std.Dev.	Min	Max	Obs.	Mean	Std.Dev.	Min	Max
Student publications	4,666	4.69	4.38	0.00	30.00	2,735	4.87	4.49	0.00	30.00
Student highly cited publications	4,666	2.46	2.77	0.00	20.00	2,735	2.55	2.82	0.00	20.00
External student from prof's network	4,666	0.27	0.44	0.00	1.00	2,735	0.46	0.50	0.00	1.00
External student outside prof's network	4,666	0.31	0.46	0.00	1.00	2,735	0.54	0.50	0.00	1.00
Student-Prof pair controls										
Prof & student have same nationality	4,666	0.30	0.46	0.00	1.00	2,735	0.23	0.42	0.00	1.00
Prof & student come from same university	4,666	0.04	0.19	0.00	1.00	2,735	0.04	0.20	0.00	1.00
N of Prof students with same nationality as student	4,666	2.44	3.43	0.00	31.00	2,735	1.45	2.49	0.00	31.00
Student main controls										
Student age	4,666	26.56	2.51	21.00	40.00	2,735	26.70	2.60	21.00	40.00
Gender	4,666	0.23	0.42	0.00	1.00	2,735	0.23	0.44	0.00	1.00
Pre-sample student pubs	4,666	0.37	1.12	0.00	16.00	2,735	0.42	1.28	0.00	16.00
Ranking of master's university	4,666	1.95	2.17	0.01	6.01	2,735	3.16	2.12	0.01	6.01
Professor-Lab main controls										
Prof age	4,666	47.71	7.81	28.00	70.00	2,735	47.54	7.73	29.00	70.00
Pre-sample prof pubs	4,666	30.40	28.71	0.00	179.00	2,735	31.59	28.64	0.00	179.00
Pre-sample prof patents	4,666	0.63	1.68	0.00	16.00	2,735	0.61	1.65	0.00	16.00
Size PhD group	4,666	7.00	6.00	0.00	41.00	2,735	6.99	6.08	0.00	41.00
Grant amount	4,666	0.74	0.90	0.00	11.59	2,735	0.78	0.86	0.00	10.58
Prof-University of master pair controls										
Pubs stock of master's university (by field)	4,666	7,266	5,655	0.00	40,801	2,735	4,377	4,368	0.00	40,801
Ranking of master's university (by field)	4,666	0.50	0.50	0.00	1.00	2,735	0.15	0.36	0.00	1.00
N of times master's university is cited by prof	4,666	24.61	39.39	0.00	292.00	2,735	4.31	12.47	0.00	155.00
Placebo and robustness controls										
Placebo most cited author	4,666	0.27	0.44	0.00	1.00	2,735	0.46	0.50	0.00	1.00
External from prof's network w/ pre-sample pubs	4,666	0.06	0.24	0.00	1.00	2,735	0.11	0.31	0.00	1.00
External from prof's network w/o pre-sample pubs	4,666	0.21	0.41	0.00	1.00	2,735	0.35	0.48	0.00	1.00
Student with pre-sample pubs	4,666	0.20	0.40	0.00	1.00	2,735	0.21	0.40	0.00	1.00
Intensity of network ties	4,666	0.01	0.03	0.00	0.36	2,735	0.01	0.04	0.00	0.36
Share of pubs with prof	4,666	0.57	0.43	0.00	1.00	2,735	0.58	0.42	0.00	1.00

3.4 Results

3.4.1 Full sample

The regression results for the full sample of observations are presented in Table 3.3. The first Model displays the baseline results with university-department fixed effects. The coefficient of the main variable of interest, *External student from prof's network*, is positive and statistically significant at the 1% confidence level. The magnitude of the coefficient suggests that external students from the professor's network are 12% more productive than internal students[30]. Interestingly, external students outside the research network of the professor are not more productive than internal students. Notably, external students remain overall more

[30] Poisson estimates are interpreted as $(e^{\beta} - 1) * 100$ percentage change.

productive[31]. These results imply that the higher productivity of external students is mainly driven by the productivity of external students from the professors' networks.

We find that cultural proximity between a student and her supervisor and the research group (*Prof & PhD have same nationality*, *Prof & PhD come from same university*, and *N of Prof PhDs with same nationality as PhD i*) are not significantly different from zero. With respect to the other controls, we highlight some interesting results. For instance, the age of a student and female gender are negatively correlated with student productivity. Conversely, having scientific publications prior to beginning a PhD is positively associated with student productivity. Similarly, the stock of a supervisor's publications is positively associated with student productivity. Finally, supervisor age is negatively correlated with student productivity. The coefficient of the average quality of the university from which a student obtained her master's degree is positive, as expected, and significantly different from zero.

In Model 2, we include supervisor fixed effects. By adding supervisor fixed effects, the coefficient of the variable declines but remains highly significant. Interestingly, the sign of the coefficient for the stock of a supervisor's publications now becomes negative and is still significant at the 1% confidence level. A possible reason for this result is that, holding constant time-invariant characteristics of the supervisors, therefore over the carrier of a professor, students benefit more from the professor research activity at the beginning of the career of the professor.

In Model 3, we include fixed effects for the universities from which the students obtained their master's degree. Consequently, the variables *External student outside prof's network* and *Ranking of master's university* drop out because they do not vary across groups. The coefficient of our variable of interest remains significant at the 1% confidence level, and its coefficient increases in magnitude, from 0.086 to 0.11. The increase in the coefficient's magnitude suggests that our student-university fixed effects may capture some (initially) omitted factors that are negatively correlated with student productivity, on the contrary of what expected.

Finally, in the last Model of Table 3.3 we include the additional controls for the relative relevance of the university for the focal professor. Supervisor fixed effects and student-university fixed effects are still included. As expected, the coefficients of *N times master's university is cited by prof* and *Ranking of master's university (by field)* are positive and significant. The coefficient of *Pubs stock of master's university (by field)* it is not significant. When we include our control, the magnitude of the indicator variable *External student from prof's network* declines slightly relative to the results in Model 3 but remains highly significant.

[31] Considering a dummy variable for external students in the model (equal to the sum of the two dummy variables *External student from prof's network* and *External student outside prof's network*) shows that external students have on average a significantly 10% higher productivity.

Table 3.3: QMLE on the count of PhD student publications (All students)

	Model 1	Model 2	Model 3	Model 4
External student from prof's network	0.113***	0.086**	0.112***	0.103***
	(0.037)	(0.035)	(0.037)	(0.038)
External student outside prof's network	-0.010	-0.001		
	(0.043)	(0.038)		
Prof & student have same nationality	0.012	0.015	0.004	0.005
	(0.029)	(0.034)	(0.033)	(0.033)
Prof & student come from same university	-0.069	-0.041	-0.044	-0.053
	(0.054)	(0.077)	(0.075)	(0.074)
N of Prof students with same nationality as student	-0.004	-0.001	0.005	0.005
	(0.005)	(0.006)	(0.006)	(0.006)
Student age	-0.033***	-0.026***	-0.032***	-0.032***
	(0.006)	(0.006)	(0.006)	(0.006)
Gender	0.217***	0.192***	0.196***	0.198***
	(0.032)	(0.031)	(0.031)	(0.031)
Pre-sample student pubs	0.072***	0.058***	0.081***	0.081***
	(0.013)	(0.013)	(0.013)	(0.013)
Ranking of master's university	0.014*	0.020***		
	(0.008)	(0.008)		
Prof age	-0.010***	-0.322	-0.305	-0.302
	(0.002)	(0.250)	(0.270)	(0.268)
Pre-sample prof pubs	0.006***	-0.005***	-0.004***	-0.004***
	(0.000)	(0.001)	(0.001)	(0.001)
Pre-sample prof patents	-0.009	-0.003	-0.010	-0.011
	(0.007)	(0.012)	(0.013)	(0.013)
Size PhD group	-0.000	0.007	0.005	0.006
	(0.003)	(0.005)	(0.005)	(0.005)
Grant amount	0.132	0.464	0.349	0.353
	(0.168)	(0.420)	(0.388)	(0.384)
Pubs stock of master's university (by field)				-0.055
				(0.042)
Ranking of master's university (by field)				0.296**
				(0.138)
N of times master's university is cited by prof				0.027*
Entry year FE	Yes	Yes	Yes	Yes
University-department FE	Yes			
Professor FE		Yes	Yes	Yes
Student-university FE			Yes	Yes
Observations	4,666	4,645	4,340	4,340
Number of University-Departments	19	19	19	19
Number of Professors	558	544	535	535
Number of Student-Universities	578	576	289	289
Pr(chi2)	0.00	0.00	0.00	0.00

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

3.4.2 External students

We replicate the regression specifications that we present in Table 3.3, restricting the sample to external students: PhD students who obtained their master's degree at a university other than the affiliation of their PhD. This changes the interpretation of the result relative to the main variable of interest which now indicates the difference in scientific productivity of external students from the professor's network as compared directly to other external students. The results are displayed in Table 3.4. Model 1 presents the baseline results, Model 2 includes supervisor fixed effects, Model 3 includes both supervisor and student-university fixed effects, and Model 4 adds the additional university specific controls. Regardless of the specification that we adopt, the coefficient of *External student from prof's network* is statistically significant at the 1% confidence level. Moreover, the magnitude increases relative to the results that we present for the entire sample and that are reported in Table 3.3. For instance, when we include supervisor and student-

university fixed effects as well as the university specific controls (Model 4), the coefficient is 0.14, whereas with the entire sample it was 0.10. As a final point of interest, the coefficient of the variable *N of times master's university is cited by prof* is statistically strongly significant at the 1% level.

Table 3.4: QMLE on the count of PhD student publications (External students)

	Model 1	Model 2	Model 3	Model 4
External student from prof's network	0.156***	0.121***	0.166***	0.143***
	(0.038)	(0.036)	(0.041)	(0.042)
Prof & student have same nationality	-0.006	0.028	-0.006	-0.011
	(0.042)	(0.050)	(0.053)	(0.054)
Prof & student come from same university	-0.114*	-0.009	-0.023	-0.075
	(0.067)	(0.093)	(0.094)	(0.093)
N of Prof students with same nationality as student	-0.006	-0.006	-0.002	-0.001
	(0.007)	(0.008)	(0.009)	(0.009)
Student age	-0.028***	-0.022***	-0.031***	-0.031***
	(0.007)	(0.007)	(0.009)	(0.009)
Gender	0.218***	0.205***	0.215***	0.217***
	(0.039)	(0.039)	(0.041)	(0.041)
Pre-sample student pubs	0.069***	0.064***	0.083***	0.083***
	(0.012)	(0.013)	(0.014)	(0.014)
Ranking of master's university	0.015*	0.022***		
	(0.008)	(0.008)		
Prof age	-0.009***	0.138	0.019	0.044
	(0.002)	(0.322)	(0.388)	(0.363)
Pre-sample prof pubs	0.005***	-0.005***	-0.005***	-0.004***
	(0.001)	(0.001)	(0.002)	(0.002)
Pre-sample prof patents	-0.008	-0.003	-0.018	-0.021
	(0.009)	(0.021)	(0.024)	(0.024)
Size PhD group	0.000	0.006	0.006	0.008
	(0.003)	(0.006)	(0.006)	(0.006)
Grant amount	0.020	0.230	0.036	0.073
	(0.232)	(0.746)	(0.651)	(0.644)
Pubs stock of master's university (by field)				-0.046
				(0.048)
Ranking of master's university (by field)				0.327**
				(0.154)
N of times master's university is cited by prof				0.056***
				(0.020)
Entry year FE	Yes	Yes	Yes	Yes
University-department FE	Yes			
Professor FE		Yes	Yes	Yes
Student-university FE			Yes	Yes
Observations	2,735	2,625	2,313	2,313
Number of University-Departments	19	19	19	19
Number of Professors	522	428	408	408
Number of Student-Universities	578	567	277	277
Pr(chi2)	0.00	0.00	0.00	0.00

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

3.4.3 High quality publications

In this section, we adopt a more restrictive definition of a student's publication count applying a citations-based quality threshold. As described in the methodology section, we only count highly cited publications. Table 3.5 presents the regression results. For the sake of brevity, we only present the most complete model specifications, separately for the full sample (Model 1) and for the sample of external students only (Model 1). The results of the models not presented are coherent with those discussed. In all cases, the results are very similar to those presented in the previous sections. *External student from prof's network* continues to have a positive and statistically significant impact on student productivity. In general, we note that the coefficients on the variable of interest tend to be larger than those presented in the previous tables. As an example,

having restricted the sample to students who obtained their master's degree at a university other than their PhD affiliation, we find that the coefficient is equal to 0.17, whereas it was 0.14 with the simple publication count. These results suggest that the variable of interest has an effect also in terms of quality of the publications and, most importantly, provides confidence on the robustness of the results to different criteria of constructing the outcome variable based on quality thresholds.

Table 3.5: QMLE on the count of PhD students highly cited publications

	Model 1 All students	Model 2 External students
External student from prof's network	0.149***	0.168***
	(0.050)	(0.053)
Prof & student have same nationality	0.060	0.040
	(0.042)	(0.065)
Prof & student come from same university	0.055	0.050
	(0.084)	(0.108)
N of Prof students with same nationality as student	0.005	-0.002
	(0.008)	(0.012)
Student age	-0.043***	-0.047***
	(0.008)	(0.010)
Gender	0.204***	0.203***
	(0.040)	(0.051)
Pre-sample student pubs	0.083***	0.088***
	(0.014)	(0.017)
Prof age	-0.258	-0.066
	(0.321)	(0.286)
Pre-sample prof pubs	-0.005***	-0.004**
	(0.001)	(0.002)
Pre-sample prof patents	-0.008	-0.024
	(0.014)	(0.024)
Size PhD group	0.006	0.012
	(0.006)	(0.008)
Grant amount	0.946**	1.162*
	(0.472)	(0.665)
Pubs stock of master's university (by field)	-0.091*	-0.070
	(0.055)	(0.061)
Ranking of master's university (by field)	0.283*	0.311*
	(0.165)	(0.187)
N of times master's university is cited by prof	0.027	0.061**
	(0.022)	(0.025)
Entry year FE	Yes	Yes
Professor FE	Yes	Yes
Student-university FE	Yes	Yes
Observations	4,179	2,212
Number of University-Departments	19	19
Number of Professors	501	380
Number of Student-Universities	278	268
Pr(chi2)	0.00	0.00

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

3.4.4 Placebo tests

In this section, we implement a placebo test designed to verify further whether the positive relationship between belonging to a supervisor's research network and student productivity captures a supervisor's tendency to work with students from universities that are relevant for her research, independently from the presence of an actual connection. Specifically, we introduce a regressor that indicates whether a student had studied in one of the universities from which a researcher who is as close as possible to supervisor j, in terms of the research interests, draws her co-authors. The idea is that the researcher whom we select (placebo) and supervisor j should draw their co-authors from institutions relevant for the research of both of them.

However, in our full models we should observe an effect only for students hired from the university where the supervisor j draws her co-authors. Hence, an evidence of a higher productivity of students within the network of the placebo professor would imply that our controls do not properly account for the endogeneity of the professor network with respect to average productivity of the students in a specific research field[32].

We present results using as placebo a researcher who is the most cited researcher by supervisor *j*. The variable *Placebo most cited author* is obtained constructing our variable of interest based on this placebo research network. In analyses not presented here (showing equivalent results and available upon request) we also use as placebo a professor, within our sample, affiliated with ETH, if *j* belongs to EPFL, or with EPFL, if *j* belongs to ETH, with the highest probability of publishing in the same scientific journals as *j*. Not surprisingly, the placebo variables are significantly positively correlated with the variable of interest: 67% of the students from the network of supervisor *j* are also from the network of the most cited researcher by supervisor *j*.

Table 3.6 presents results for the placebo test. We present results for the entire sample (Models from 1 to 3) and for the sample of external students (Models from 4 to 6). In Model 1 and 4 we include the placebo variable in our baseline model specifications, instead of the variable of interest. In Model 2 and 5 we adopt our full model specification with professor and student-university fixed effects and the additional controls. Finally, in model 3 and 6 we include both the placebo variable and the variable of interest *External student from prof's network.* In our baseline models the variable *Placebo most cited author* is positive and significant. However, in our full specifications it is not significant. Also importantly, when including both variables, our variable of interest is still positive and significant. Significance is lower for the entire sample and the magnitude of the effect is slightly smaller; however, this might be simply due to the strong correlation between the variable of interest and the placebo variable. Overall, the results provide further confidence that the effect identified is attributable to the actual presence of the professor's coauthors in a university and not to unobserved quality of the university.

[32] An alternative strategy to further test the hypothesis of endogeneity of our variable of interest with respect to the average quality of students in a given university would have been to find an instrumental variable correlated with the probability of a professor having a coauthor affiliated with it and not correlated with the quality of students. Given the difficulties to find a variable with such a credible exclusion restriction we opted for the placebo test presented.

Table 3.6: Placebo tests

	All students			External students		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Placebo most cited author	0.072**	0.017	-0.010	0.075**	0.066	0.037
	(0.029)	(0.036)	(0.040)	(0.035)	(0.043)	(0.045)
External student from prof's network			0.070*			0.136***
			(0.037)			(0.044)
External student outside prof's network	-0.083**					
	(0.035)					
Prof & student have same nationality	0.013	0.010	0.006	0.013	-0.001	-0.010
	(0.029)	(0.033)	(0.033)	(0.040)	(0.053)	(0.054)
Prof & student come from same university	-0.056	-0.050	-0.059	-0.071	-0.060	-0.075
	(0.054)	(0.072)	(0.072)	(0.067)	(0.094)	(0.093)
N of Prof students with same nationality as student	-0.006	0.006	0.006	-0.005	-0.002	-0.001
	(0.005)	(0.006)	(0.006)	(0.007)	(0.009)	(0.009)
Student age	-0.032***	-0.032***	-0.032***	-0.028***	-0.030***	-0.030***
	(0.006)	(0.006)	(0.006)	(0.007)	(0.009)	(0.009)
Gender	0.216***	0.199***	0.199***	0.216***	0.217***	0.218***
	(0.032)	(0.031)	(0.031)	(0.039)	(0.041)	(0.041)
Pre-sample student pubs	0.073***	0.082***	0.082***	0.069***	0.083***	0.083***
	(0.013)	(0.013)	(0.013)	(0.013)	(0.014)	(0.014)
Ranking of master's university	0.023***			0.012		
	(0.007)			(0.008)		
Prof age	-0.010***	0.039**	0.038**	-0.009***	0.006	0.032
	(0.002)	(0.016)	(0.016)	(0.002)	(0.373)	(0.358)
Pre-sample prof pubs	0.006***	-0.005***	-0.005***	0.005***	-0.005***	-0.004***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
Pre-sample prof patents	-0.009	-0.011	-0.010	-0.009	-0.021	-0.020
	(0.007)	(0.013)	(0.013)	(0.009)	(0.024)	(0.024)
Size PhD group	0.000	0.006	0.006	-0.000	0.008	0.008
	(0.003)	(0.005)	(0.005)	(0.003)	(0.006)	(0.006)
Grant amount	0.125	0.360	0.361	0.002	0.126	0.086
	(0.167)	(0.380)	(0.382)	(0.230)	(0.633)	(0.639)
Pubs stock of master's university (by field)		-0.051	-0.057		-0.037	-0.048
		(0.041)	(0.041)		(0.048)	(0.048)
Ranking of master's university (by field)		0.298**	0.295**		0.323**	0.325**
		(0.135)	(0.137)		(0.150)	(0.154)
N of times master's university is cited by prof		0.040***	0.039***		0.065***	0.054***
		(0.014)	(0.014)		(0.020)	(0.020)
Entry year FE	Yes	Yes	Yes	Yes	Yes	Yes
University-department FE	Yes			Yes		
Professor FE		Yes	Yes		Yes	Yes
Student-university FE		Yes	Yes		Yes	Yes
Observations	4,666	4,340	4,340	2,735	2,313	2,313
Number of University-Departments	19	19	19	19	19	19
Number of Professors	558	535	535	522	408	408
Number of Student-Universities	578	289	289	578	277	277
Log likelihood	0.00	0.00	0.00	0.00	0.00	0.00

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

3.5 On mechanisms and alternative explanations

Provided that in our models we properly control for other confounding factors, our analyses implies that professors obtain students that turn out to be more productive from a university where they have coauthors. However, according to theory, different mechanisms might explain this finding. Here, we distinguish broadly between hiring effects or post-hiring effects. In terms of hiring effects, networks decrease asymmetric information allowing for a better selection of students and a better match between students and supervisor. In terms of post hiring effects, students from a research network may be more productive because, for instance,

they have lower costs of interacting with their supervisors and/or because the supervisor may feel more committed to students from her network.

The importance to separate the two typologies of mechanisms relies on the fact that they clearly have different implications. However, to properly distinguish between the two would require to control for the intrinsic quality of students and possibly to observe both applicants, and also potential applicants, to a given PhD student position[33]. Since we do not dispose of this information we cannot separate the two explanations with specifically dedicated models. However, we provide evidence in the attempt to identify the most prevalent typology of mechanisms. In addition, we try to rule out some of the most trivial possible explanations that might justify the results encountered. This series of analyses is presented in Table 3.7 and is discussed below. For the sake of brevity, we only report coefficients for the main variables, and we present results only for the model specifications with professor and student-university fixed effects and all controls. We distinguish results for the full sample (Model 1, 3, 5 and 7) and for the sample of external students (Model 2, 4, 6 and 8).

3.5.1 Interaction with publication before the PhD

First, we proceed by examining instances in which networks would lead to different outcomes, depending on whether hiring effects or post-hiring effects are the main drivers of our results. We note that variable *Pre-sample student pubs*, when is not null[34], constitute a good proxy of the quality of the student. Furthermore, the number and the content of published papers constitute a strong signal for the professor who not only can be more confident in the quality of the student but can also better estimate the potential match of the student's competences with her research interests. As a consequence, we can expect that information asymmetries are considerably lower for students with at least one publication before the PhD.

Exploiting this idea, we interact a dummy indicating if a student has at least one publication before the PhD (*Student with pre-sample pubs*) with our variable of interest, obtaining two distinct categories of students: external students from the professor network with publications before the PhD (*External from prof's network w/ pre-sample pubs*) and without publications before the PhD (*External from prof's network w/o pre-sample pubs*). The results are displayed in Model 1 and 2 of Table 3.7. We expect that if there is any post-hiring effect, the coefficient on the first category (*External from prof's network w/ pre-sample pubs*)

[33] In fact note that hiring effects imply that better students are hired from a given university, holding constant the average quality of students in a given university. On the contrary, post-hiring effects imply that holding constant the quality of each student, those within the network of supervisor turn out to be more productive.

[34] We keep in mind that many students do not have the possibility to publish before their PhD, regardless of their quality. In our sample, about 80% of students do not have any publication before the PhD. While we consider the number of publications a good proxy of quality for students having at least one publication, we rather consider this proxy simply unobserved for students without publications.

should still be significant. On the contrary, we find that it is not significant and the coefficient on the variable *External from prof's network w/o pre-sample pubs* is still significant and in the sample of external students is also higher in magnitude compared to the results in Table 3.3 and Table 3.4. This latter result implies that the effect of our variable of interest is not significant for students with publications before the PhD and is stronger for students without publications before the PhD. In other words, the effect is stronger when information asymmetries are also stronger, which suggests that hiring effects are prevalent.

Table 3.7: Robustness regressions

	Pre-sample pubs interaction		Intensity of network control	
	Model 1	Model 2	Model 3	Model 4
	All students	External students	All students	External students
External student from prof's network			0.090**	0.150***
			(0.039)	(0.043)
Intensity of network ties			1.027	-0.574
			(0.840)	(0.873)
External from prof's network w/ pre-sample pubs	0.091	0.059		
	(0.057)	(0.071)		
External from prof's network w/o pre-sample pubs	0.091**	0.159***		
	(0.039)	(0.044)		
Student with pre-sample pubs	0.167***	0.264***		
	(0.044)	(0.079)		
Controls	Yes	Yes	Yes	Yes
Entry year FE	Yes	Yes	Yes	Yes
Professor FE	Yes	Yes	Yes	Yes
Student-university FE	Yes	Yes	Yes	Yes
Observations	4340	2313	4340	2313
Number of University-Departments	19	19	19	19
Number of Professors	535	408	535	408
Number of Student-Universities	289	277	289	277
Pr(chi2)	0.00	0.00	0.00	0.00
	Pubs with prof control		Students w/o collaborative pubs	
	Model 5	Model 6	Model 7	Model 8
	All students	External students	All students	External students
External student from prof's network	0.088***	0.115***	0.086**	0.103**
	(0.034)	(0.037)	(0.039)	(0.040)
Share of pubs with prof	-0.163***	-0.287***		
	(0.040)	(0.056)		
Controls	Yes	Yes	Yes	Yes
Entry year FE	Yes	Yes	Yes	Yes
Professor FE	Yes	Yes	Yes	Yes
Student-university FE	Yes	Yes	Yes	Yes
Observations	3729	2032	2225	1824
Number of University-Departments	19	19	19	19
Number of Professors	506	380	401	360
Number of Student-Universities	289	277	280	272
Pr(chi2)	0.00	0.00	0.00	0.00

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

3.5.2 Weak and strong networks

We measured membership in a research network using a dummy that takes a value of one if a student has studied at one of the universities from which a supervisor draws her co-authors. Having at least one coauthor from a given university is supposed to be sufficient to attenuate information asymmetries, while stronger connections should play no role, or even have a negative effect (Granovetter, 1973). On the contrary, if post-hiring effects were the main driver of the positive effect of networks in our analyses, we might expect stronger ties to have an additional impact: for instance, professors would commit more to the supervision of students sent by co-authors with whom they collaborate more. Hence, we add the share of a supervisor's publications with co-authors from student i's past affiliation *Intensity of network ties* to our regressions. The results are displayed in Model 3 and 4 of Table 3.7. We find that the coefficient of *Intensity of network ties* is statistically insignificant regardless of the sample definition, whereas the coefficient of *External student from prof's network* remains statistically significant at the 1% confidence level. This evidence is again in line with a hiring effect interpretation of the results.

3.5.3 Publications coauthored with the supervisors

In this section we worry about the possibility that students hired from a professor network might result to be more productive because they are more directly involved in the research activity of the professor. Notably, in our sample, a high percentage of publications (57%) of the students are coauthored with the supervisor and 82% of the students with at least one publication has coauthored at least one paper with the professor. However, we observe some variance across students. Therefore, we construct a variable equal to the share of publications of a student coauthored with the supervisor (*Share of pubs with prof*). This variable is meant to capture post-hiring dynamics between a student and her supervisor and, in particular the strength of their interaction. Note that this variable is a "bad control" (Angrist and Pischke, 2008), if we admit that the likelihood that a professor collaborates with a student it is also a function of the student's quality and productivity. However, this should bias downwards our estimates. We restrict the sample to those students who have published at least one article, since the new control variable is only defined in these cases. The results are displayed in Model 5 and 6 of Table 3.7. Importantly, our variable of interest remains highly significant. Interestingly, the coefficient on *Share of pubs with prof* is negative, contrary to what expected, which suggests that students coauthoring a higher share of their papers with their supervisors are overall less productive. However, note that there is no causal claim in this latter result. What matters to our purpose is to verify that the intensity of collaboration with the supervisor does not explain the effect encountered on our variable of interest.

3.5.4 Pre-existing research collaborations

Finally, we observe that a certain number of students presents publications where both the supervisor and a researcher affiliated with the University of the master 's degree are indicated as co-authors, which we label as "collaborative publications". Focusing on students with at least one publication, the percentage of students with this type of publications is obviously particularly high for students that obtained their master's degree from the same university of the PhD (about 90%), but it is also not negligible for external students (10%). The number of such "collaborative publications" is positively correlated with our variable of interest. Therefore, we worry about the fact that similar cases (especially for external students) reveal instances where students hired from the professor network are initially involved in pre-existing collaborations earlier during the PhD or, even before, during their master's degree studies. This might be an alternative and rather trivial explanation of the positive effect we encountered on our variable of interest. In order to rule out this hypothesis, in Model 7 and 8, we estimate our model restricting the sample to students with at least one publication but not presenting any of the collaborative publications mentioned above. The coefficient remains highly significant and positive. The number of observations in both samples is substantially reduced, which might explain the slight decrease in the significance. Similarly to the discussion in the previous paragraph, these estimates might be downward biased if the likelihood of a student being involved in a "collaborative publication" is also a function of her quality and productivity. We conclude that collaborative publications do not explain the positive effect we found in our main analyses.

3.6 Conclusion

In this paper we examine the scientific productivity of a population of 4'666 PhD students in the two major engineering universities in Switzerland, ETH and EPFL. First, similar to previous studies, we distinguished between internal students (who obtained their master's degree from the same university of the PhD) and external students (who obtained their master's degree from a different university of the one of the PhD). In addition, we distinguish within the latter category, PhD students who obtained their master's degree from a university within the network of their supervisors (where their supervisors have coauthors).

As a main result, external PhD students from the network of the supervisor result to be about 12% more productive than internal students and 15% more productive than other external students. Also interestingly, other external students (not from the network of the supervisor) are not more productive than internal students. In our most complete model specification, we controlled for the heterogeneity of the supervisor, the heterogeneity of the university where the student obtained her master's degree and for proxies of the specific relevance of a particular university for a given supervisor. Therefore, we conclude that the presence of the supervisors' co-authors in different universities has a positive effect on the productivity of the students hired from these universities.

Such results provide evidence of positive effects of networks in hiring. Also, the results have a direct implication for the literature on the productivity of PhD students and researchers. While this literature has identified a general higher productivity of researchers from external environments (either from different institutions or countries), our results adds to the evidences of existing heterogeneity within this category (Hunt, 2011). In particular, in our case, the higher productivity of external PhD students (which notably largely overlaps in our sample with the category of foreign or foreign-educated students) is mainly (if not fully) explained by the higher productivity of external PhD students from the supervisors' network.

For universities and similar contexts, our study suggests that the higher productivity of external researchers cannot only be attributed to a simple process of indirect selection, but that informal connections are crucial in order to benefit from the inflow of students and researchers from external institutions and foreign countries. Nonetheless, normative implications are dependent on the possibility to disentangle the underlying mechanisms explaining the results. A series of additional analyses we performed suggest that the reduction of information asymmetries are likely the main explanation. Accordingly, our results would imply that institutions intending to attract productive researchers from other institutions might not succeed simply by "opening their doors". Conversely, a more comprehensive set of initiatives aiming at increasing the professional integration of the institution with external environments – including the faculty professional networks - would be necessary.

Our findings can be extended to a number of settings. For instance, they could be extended to research centers in public or private institutions in which knowledge production is a fundamental objective and in which the head of a research group is confronted with the problem of maximizing the research output of her members. The findings could also be extended to other universities as long as supervisors have some autonomy in choosing the PhD students whom they wish to admit to their group. Finally, our results are obtained from a European context, which, despite its heterogeneity in terms of a number of aspects, comprises fairly culturally homogeneous countries. Thus, we expect that if we were to extend this analysis to a broader context, the effects of supervisor's research networks on could be even larger.

However, further research is required in order to discern different possible mechanisms and to allow for more informed policy initiatives. This goes together with a series of limitations of our study. First, it would be interesting to assess the impact of network effects on the probability that candidates are selected for PhD student positions. In addition, although our tests suggest that the positive relationship between belonging to a supervisor's research network and student productivity are driven by hiring effects, it would be important to corroborate our results with models and data allowing to directly control for the unobserved idiosyncratic quality of students. Third, it would be interesting to disentangle the typologies of tie characterizing the student and the coauthor of the supervisor: for instance, we could not distinguish whether the coauthor actually knew and recommended the student or if the coauthor simply provided general information to the supervisor.

Finally, despite the generalizability of our results, extending the analysis to other universities in other countries would allow an understanding of how different institutional settings affect the role of supervisors' networks. In particular, we do not neglect the possibility that depending on the cultural and social background of an institution or country, negative effects of networks might prevail. Nonetheless, our results show that most of the direct benefit in terms of scientific productivity from hiring external students in two of the main research institutions in Switzerland is mediated by the presence of networks of their faculties. Evidence of the high productivity of external researchers and students is sometimes equated to the need of more objective and centralized hiring procedures that might indeed be necessary and complementary. However, our evidence suggests that a virtuous adoption of professional networks in order to reduce information asymmetries could be an indispensable element in order to benefit from the hiring of external and international students.

Chapter 4 Interregional knowledge integration and firms' innovative productivity

(With Guillaume Burghouwt)

4.1 Introduction

Knowledge diffusion is a determinant of technological progress, and consequently, of economic growth (Grossman and Helpman, 1993; Romer, 1990). Accordingly, firms' innovative performance depends on access to diverse knowledge sources, beyond their own efforts in internal Research and Development (R&D) expenditures (Cohen and Levinthal, 1990; Jaffe, 1986). Some notable literature has found geographic proximity to be an antecedent of knowledge spillovers among firms located in geographical clusters (Audretsch and Feldman, 1996; Jaffe, et al. 1993; Marshall, 1891; Porter, 1998; Singh and Marx, 2013). However, other authors have documented the existence of knowledge flows crossing regional and national boarders (Coe and Helpman, 1995; Keller, 2004; Mancusi, 2008; Maurseth and Verspagen, 2002), and have discussed the role of non-geographic proximity dimensions as drivers of knowledge diffusion (Boschma, 2005; Breschi and Lissoni, 2001; Crescenzi, 2014; Kerr, 2008; Singh, 2005). Moreover, there is some evidence that the distance to which knowledge diffuses is increasing over time due to the effects of transportation costs and information technology improvements (Keller, 2002). As a consequence, interregional knowledge integration, which we define as a region's degree of access to and adoption of knowledge developed in other geographically dispersed regions, is an increasing phenomenon often at the center of attention for firms and policy makers (Archibugi and Iammarino, 2002; Chessa et al., 2013).

While there is extensive empirical evidence on the effects of localized knowledge diffusion, only recently scholars have started to devote more attention to the role of interregional knowledge integration for firms' innovative performance (Breschi and Lenzi, 2012; Crescenzi, 2014). Some authors have pointed out the presence of a dichotomy between benefits emerging from the local diffusion of knowledge in geographical clusters and the need to access distant knowledge in order to trigger and sustain innovation (Arikan, 2009; Bathelt, et al., 2004; Boschma, 2005). Indeed, historical evidence demonstrates that the most innovative and

competitive regions often show higher levels of knowledge integration with other regions (Bresnahan et al., 2001; Saxenian, 1994, 2005; Kerr, 2008). However, empirical evidence on the relationship between interregional knowledge integration and innovative performance is still limited.

We adopt an unbalanced panel of 3,871 innovative companies in Germany between 1992 and 2010, for a total of 15,819 observations, and study their innovative productivity. We measure interregional knowledge integration as the geographic dispersion of patent backward citations of the region toward other regions worldwide. In fixed effects estimations, we find that interregional knowledge integration positively affects innovative productivity of local firms. To address concerns of endogeneity due to the possibility of reverse causality and omitted time-variant variables, we exploit airline liberalization realized in Europe as a source of exogenous shock to the interregional knowledge integration of German regions.

The main result of liberalization has been the entry of low cost carriers (LCCs) that introduced new direct connections and offered flights at extremely inferior prices as compared to previous traditional airlines (Calder and Laker, 2002; Dobruszkes, 2006). More generally, the entry of new airlines in an airport delimits the shift from mainly monopolistic markets toward more competitive markets. This airline liberalization in Europe was formally accomplished in 1992, but it became gradually effective only starting from 1997. Furthermore, the entry of LCCs in different European regions was not instantaneous, but distributed in a period of approximately five years. Delays in the entry of LCCs were usually determined by factors independent of the strategic timing of the new entrants, such as a lack of available slots in airports or the resistance of local administrations to effectively adopt the liberalization mandate (Calder and Laker, 2002).

We estimate the impact of the entry of an LCC in close airports on the level of interregional integration of a region. The entry of an LCC determines a significant increase in our indicators of interregional knowledge integration. We find that firms located in regions where airline liberalization induced a higher level of interregional knowledge integration significantly increased their innovative productivity. Finally, we investigate the heterogeneity of the effect of interregional knowledge integration across regions with different levels of R&D. When we do not use the entry of an LCC as an instrument, firms located in regions with higher levels of R&D show a stronger association between innovative productivity and interregional knowledge integration. Interestingly, we find the opposite sign when we use the entry of an LCC as instrument. However, in both cases, the differences across regions are small in magnitude.

4.2 Innovative productivity and interregional knowledge integration

The core research question of this paper is: what is the impact of the interregional knowledge integration of a region on the innovative productivity of local firms? A first series of studies have analyzed the relationship between firms' performance and characteristics of the region or cluster where firms are located

(Cruz and Teixeira, 2009; Frenken et al., 2014). Firms located in regions endowed with a certain critical mass of firms and institutions performing innovative activities are expected to benefit from agglomeration economies: 1. broader access to specialized labor; 2. access to specialized suppliers and collaborators; 3. access to localized knowledge spillovers (Marshall, 1891). Baptista and Swann (1998) found that firms located in regions with a higher concentration of labor in their own sector are more innovative. Similar studies have also found that geographic proximity is effective given a certain level of technological similarity that enables knowledge spillovers to be captured by recipient firms (Autant-Bernard, 2001).

Other studies have explored the extent to which knowledge diffuses at longer distances and across countries (Coe and Helpman, 1995; Bottazzi and Peri, 2003; Keller, 2004). Some authors have questioned the relative importance of local knowledge searches and exploitation, against access to knowledge external to the region. Boschma (2005) distinguishes five dimensions of proximity (cognitive, organizational, social, institutional, and geographical) and claims that while too much distance along these dimensions might impede communication and collaboration, too much proximity might deter innovation due to technological lock-in and lack of sources of novelty. Specifically, Bathelt et al. (2004) pointed out that the combination of "internal learning processes - local buzzes" and "communication channels" with other external environments - "pipelines" - is required to maintain and increase innovative performance.

Important previous studies have investigated the relationship between innovative performance and distant knowledge (Breschi and Lenzi, 2012; Crescenzi, 2014; Eisingerich et al., 2010; Frenz and Ietto-Gillies, 2009; Lecocq et al., 2012). Many of these have focused on the innovative performance of multinational companies and have found a positive correlation between the presence of firms in different regional contexts and their innovative performance. Phene et al. (2006) suggest that the types of external knowledge combinations, in terms of technology and geographical distant knowledge, determine the likelihood of breakthrough innovation. Breschi and Lenzi (2012) find that the coexistence of dense internal collaborations within a city and a certain number of external connections, measured through co-inventor network indicators, is positively associated with patenting productivity.

In line with these contributions, we expect interregional knowledge integration to have a positive impact on innovative productivity for several, non-exclusive, reasons. First, the inflow of new knowledge developed in another context might constitute a novel input to the knowledge production function of firms located within a region (Coe and Helpman, 1995). These knowledge externalities potentially overcome diminishing returns to the exploitation of knowledge internal to the firm and the region. Second, innovation requires the recombination of different technologies and approaches; different sources of knowledge might result in complementarity (Cassiman and Veugelers, 2006; Cohen and Levinthal, 1990). Firm-level analyses have demonstrated that internal R&D strategies and searches for external knowledge at a broad international level are complementary strategies (Cassiman and Veugelers, 2006). Finally, higher knowledge integration to other regions worldwide might allow for the timely identification of technological opportunities and of

potential areas of specialization in order to develop competitive advantages and investment in high productivity sectors, as compared to competitors (McCann and Ortega-Argilés, 2013). Accordingly, we formulate the hypothesis that *firms located in regions with a higher level of interregional knowledge integration have higher innovative productivity*.

A further point of interest is whether interregional knowledge integration differently affects firms located in regions with different characteristics. In particular, we address the question as to whether regional knowledge integration has a stronger impact on firms located in large regions with high levels of investments in R&D or, on the contrary, firms located in smaller regions benefit the most. In other words, we wonder if the phenomenon enhances (loosens) the effect of agglomeration economies increasing (decreasing) the attractiveness of large existing geographical clusters. From a theoretical stand point, there exist arguments in favor of both hypotheses. On the one hand, regions with a higher level of R&D investments might be endowed with the necessary absorptive capacity required to adopt external knowledge and translate it into successful innovative products (Cohen and Levinthal, 1990; Mancusi, 2008). To the extent that geographic proximity in large technological clusters can be complementary to other forms of proximity to geographically distant environments (Eisingerich et al., 2010), the attractiveness of specific locations might increase, and agglomeration economies might become strengthened (Sonn and Storper, 2008). Furthermore, larger innovative clusters might become more attractive toward external resources and sources of knowledge once a stronger connection with the external environment is established. As such, firms located in larger regions would benefit the most from interregional knowledge integration. On the other hand, small regions might be more flexible and capable in adapting to the changes required by the adoption of external knowledge (Menzel and Fornahl, 2010). Also, larger regions might already have a higher level of integration with other regions and, compared to them, smaller regions might benefit proportionally more from the phenomenon (Mancusi, 2008).

4.3 European airline liberalization

There is a general consensus on the fact that progress in information technology and the reduction of transportation costs are expected to allow for easier access to distant knowledge and to possibly reduce the relative importance of geographic proximity per se (Tranos, 2013). We explicitly take into account the latter of these two factors as a driver of a region's capacity to access external environments. Exploiting the exogenous shock to the transportation costs provided by European airline liberalization, we also attempt to explicitly address the endogeneity issue, which is likely to affect the relationship under study. First, firms and regions showing the ability to reach and connect with different sources of knowledge might also be endowed with other unobserved characteristics, such as better infrastructures, and organizational and managerial skills, which are likely to affect performance. Second, reverse causality can affect the results if

the most innovative regions and firms are more capable of reaching distant knowledge and of developing technologies with a broader geographic scope. Simply considering the level of a region's connectivity, for example, in terms of airport proximity or number of flights, might be subject to similar concerns to those previously discussed: the level of investment in transportation infrastructures and transportation costs from and towards a certain region can, themselves, be a function of the innovativeness and attractiveness of a region. Particularly for this reason, we exploit airline liberalization in Europe as a source of an exogenous shock to transportation costs, therefore affecting interregional knowledge integration.

Before this regulation change, European aviation markets were mainly regulated by bilateral agreements and were dominated by monopolistic markets. Airline liberalization was a deregulation process started in 1986 and accomplished in 1992 with the Third Aviation Liberalization Package (Calder and Laker, 2002). However, effective implementation of the deregulation was considerably delayed in many countries and, even afterwards, the entry of new airlines in several European airports was constrained by a lack of available slots or the resistance of local administrators. Calder (2002) summarizes the process: "Europe's skies... have officially been open since 1997. But leading airports remain effectively closed to newcomers because of the shortage of available slots…And in parts of Europe, obstructive governments act...to constrict the freedom of the skies." As a consequence, the effects of liberalization propagated gradually in European regions for reasons mostly independent of the strategic planning of new entrants and of the time-variant characteristics of the regions. Most importantly for our purposes, the consequences of airline liberalization were likely diffused independently of time-specific shocks to the demand of flight connections in a region.

In each airport, the effects of liberalization materialized with the entry of new airlines operating at substantially lower prices and toward destinations previously not reachable with direct connections: low-cost carriers (LCCs). More generally, the entry of an LCC determined the shift from monopolistic to competitive markets. LCC prices have been from one-half to eight times lower than the average of previous traditional carrier prices. Traditional flag carriers reacted to the higher competition by also reducing prices, offering, for temporary periods, prices at the levels of LCCs and entering new markets. Therefore, prices decreased substantially while the number of direct destinations and the frequency of connections increased.

Overall, generalized travel costs decreased towards most of the main destinations. While enthusiastic, the following words by Calder (2002) provide a feeling about the perceived strength of the impact of LCC entry on connectivity: "Thanks to low-cost airlines, second home ownership abroad has rocketed. Lifestyles have been transformed, and long-distance relationships formed, thanks to a newly affordable Europe." We hypothesize that airline liberalization had a considerable positive impact on the interregional integration of European regions, and we consider the entry of an LCC to be a potential instrument of our variables of interest.

4.4 Data and methods

4.4.1 Data and variables

In order to test our hypothesis, we combine data from different sources. We use the Mannheim Innovation Panel (MIP) as a source of information on the innovative firms in Germany. The MIP combines survey information on innovation activities of German firms from 1992 to 2010 and constitute a representative sample of innovative firms in Germany. Firms participating in the survey report on several indicators related to their economic and innovation activities. By considering firms observed for at least 2 periods and with a positive amount of innovative sales for at least one period, we obtained an unbalanced panel of 3,871 firms for a total of 15,819 observations within the period just mentioned. We assigned firms based on the postal code of their addresses to regions in Germany at level 3 of the Eurostat nomenclature of territorial units for statistics (NUTS3 level of the NUTS classification). While the total of the NUTS3 regions in Germany is 428, only 405 are represented because not all regions host innovative firms in our sample.

We constructed region and firm patent based indicators combining information from the European Patent Office (EPO) Worldwide Patent Statistical Database 2013 (PATSTAT) and the REGPAT 2013 Database provided by the Organization for Economic Co-operation and Development (OECD). REGPAT 2013 contains information on EPO and PCT (Patent Cooperation Treaty) patents and the geographic location of inventors and applicants at the level of the NUTS3 regions. Finally, information regarding the entry of LCCs in airports relevant to German regions was obtained by the Official Airline Guide (OAG) database on historical flight status. The data have been used to obtain yearly information on European airports relevant to German regions and on the airlines operating in these airports. Whenever possible, we categorized airlines as LCCs or traditional airlines based on the categorization proposed by the literature in transportation economics (e.g., Dobruszkes, 2006). Few airlines not found in this literature have been categorized based on complementary search on the Internet, and specifically, airline web-sites.

We assigned airports in Germany and those close to the German borders based on their relative distance. Therefore, both airports and the NUTS3 regions have been localized according to their longitude and latitude (considering the NUTS3 regions for their geographic center). For each region and airport pair, we estimated their average travel time distance through a query in the Google API geocoding database. Finally, for each region, we kept only airports at a maximum of 3 hours driving distance, since this is approximately the minimum travel time at which all regions considered could reach at least one airport. As alternatives, we considered airports at a maximum of 2 hours and 1 and-a-half hours of driving distance, matching regions with no accessible airports in this travel time with their closest airport. Table 4.4 reports the list of variables considered in the analyses, their description and descriptive statistics.

Table 4.1: Variable description and descriptive statistics

Variable	Description	Data source	Obs	Mean	Std.Dev	Min	Max
Dependent variable:							
Innovative productivity	Innovative sales per employee	MIP 1992-2010	15,819	0.0325212	0.2226614	0	22.08855
Interregional knowledge integration:							
Region citations' invH	Inverse Herfindahl Index of the distribution of citations of a NUTS3 german region across worldwide NUTS2 regions (excluding self inventor citations)	REGPAT 2013 and PATSTAT 2013	15,819	14.33042	8.277598	1	40.33333
Region copatents' invH	Inverse Herfindahl Index of the distribution of copatenting activities of a NUTS3 german region across worldwide NUTS2 regions abroad	REGPAT 2013 and PATSTAT 2013	15,819	9.353444	10.99868	0	63.77087
Firm controls:							
N of employees*	Number of employees	MIP 1992-2010	15,819	867.852	7,006.66	1	282,758
R&D_employee	R&D expenditure per employee	MIP 1992-2010	15,819	0.0233492	0.2809853	0	11.88121
Export_employee	Export per employee	MIP 1992-2010	15,819	0.3147266	3.626616	0	165.2895
Patent stock	Firm patent stock (discounted at the 15% discount rate)	PATSTAT 2013	15,819	4.799182	59.41618	0	3,642.41
Region controls:							
Region R&D*	Aggregated value of firms' R&D expenditure in the region	MIP 1992-2010	15,819	323.5177	1,952.69	0	39,720.49
Region N of employees*	Total number of employees in the region	MIP 1992-2010	15,819	25,711.26	88,732.38	4	953,992
Region export*	Aggregated value of firms export of the region	MIP 1992-2010	15,819	3,576.41	20,046.92	0	305,108.80
Region patents**	Number of patents of the region	REGPAT 2013	15,819	106.081	307.4564	1	2,629
Firm R&D collaboration:							
R&D coop in DE	Each year equals 1 if the focal firm has engaged in R&D collaborations with firms or institutions within Germany in the first of the previous years where the information is available	MIP 1992-2010	15,819	0.250648	0.4333999	0	1
R&D coop abroad	Each year equals 1 if the focal firm has engaged in R&D collaborations with firms or institutions abroad in the first of the previous years where the information is available	MIP 1992-2010	15,819	0.1340793	0.3407482	0	1
Instrument:							
LCC entry	Equal to 1 if a low cost carrier is operating in an airport close to the region (max. 3 hours driving distance)	AOG	15,819	0.4717744	0.4992185	0	1
Interaction:							
Pre-entry region R&D*	Aggregated average value of firms' R&D expenditure in the region over the period 1992-1995	MIP 1992-2010	15,819	347.0625	1,500.89	0	13,310.70

Note: * the variable is rescaled by 1,000 in the analyses. ** The variable is rescaled by 100 in the analyses.

Among the variables reported, we focus primarily on their innovative productivity as a dependent variable, defined as the amount of innovative sales per employee. Innovative sales are defined as the amount of sales that respondents consider attributable to innovative products (new to the market). The main variables of interest are the indicators of interregional knowledge integration which we measure through indicators of the geographic dispersion of the knowledge sources of technologies developed within a region. As an indicator of dispersion we adopt the inverse Herfindahl index *(invH*). *Region citations' invH* corresponds to the inverse Herfindahl index of the distribution of citations from patents belonging to a NUTS3 German region across other worldwide regions at the NUTS2 level. *Region copatents' invH* corresponds to the inverse Herfindahl index of the distribution of co-patenting activities of NUTS3 German region inventors with inventors across other worldwide regions at the NUTS2 level abroad. The inverse Herfindahl index is defined by:

$$invH = \frac{1}{H} = \frac{1}{\sum_{i=1}^{N} s_i^2}$$

where *H* is the Herfindahl index, *i,* which goes from 1 to *N,* represents the regions cited (or where at least one coinventor is located), and *s* is the share of citations to (or co-patenting activities with inventors in) region *i,* in a given year. Here, the inverse Herfindahl index is a measure of the geographic dispersion of the citations (or co-patenting activities), which has a direct intuitive interpretation as the effective number of

regions cited (or involved in co-patenting activities). In other words, the backward citations (co-patenting activities) of a region are distributed across other regions in such a way that they are as concentrated as they would be if they were divided evenly across a number of regions corresponding to the value of the variable *Region citations' invH* (*Region copatents' invH*).

We further consider three sets of variables. First, we consider basic controls at the firm level: the number of employees, R&D expenditure per employee, exports per employee and the patent stock. The first three are survey-based variables. The patent stock is calculated as the cumulated number of patents in PATSTAT reporting the firm as an applicant up to the year before the focal year, discounting previous years with a 15% discount rate. Second, we consider a set of controls at the regional level: R&D expenditure, exports, the number of employees and the number of patents filed in the focal year. The first three variables are obtained by aggregating the amount of R&D expenditure, exports and employees at regional level, as reported in the MIP survey by all firms in the region (also including firms not considered in our analyses that appear for only one period or those that do not have innovative sales). The number of patents corresponds to the number of EPO patents reporting inventors located in the region, as reported in REGPAT. Importantly, these are the same patents adopted to construct the citation-based indicators of interregional integration. Therefore, this variable also controls for a possible omitted variable problem caused by the positive correlation of our citation based indicators and the amount of patenting activity in the region.

Third, we consider two variables related with the R&D collaboration activities of the firm. These are two dummy variables that indicate respectively if the firm collaborated in R&D with partners within Germany or abroad. These variables are only available every four years in the MIP. In order not to lose observations, a company (not) reporting a collaboration in a given year is considered (not) to have collaborated in the following years where the information was not available. Finally, we consider a variable indicating the level of R&D expenditure within a region before the entry of LCCs in European airports across all periods, which we adopt as interaction variable with our variable of interest. In particular, we consider the average R&D expenditure within the region in the period 1992-1995 (*Pre-entry region R&D*). Considering shorter periods does not change the results.

4.4.2 Model

In our main model, we relate innovative productivity with the indicators of interregional knowledge integration. Given the characteristics of the dependent variable, non-negative and with potential zero-inflation, we adopt a Quasi Maximum Likelihood Estimation (QMLE) method with fixed effects. The QMLE method has desirable properties for the analyses under study, such as consistency of the estimates independently from the variance functional form and robustness to zero-inflation (Wooldridge, 1997, 2010). The model specification includes firm fixed effects and year fixed effects. Year fixed effects control for any

shock to productivity over time, common to all firms in Germany. Firm fixed effects control for time-invariant characteristics of the firms.

$$InnProd_{irt} = \exp(\beta_0 + \beta_1 RegIntegration_{rt} + \beta_2 FirmCtrls_{irt} + \beta_3 RegCtrls_{irt} + \beta_4 FirmFE_{ir} + \beta_5 YearFE_t) + \varepsilon_{irt} \quad (1)$$

Note that these two levels of fixed effects control for several potentially omitted variables such as policy changes at the country level, the diffusion of ICT technologies, geographic position, sector-specific characteristics, etc., as far as these are not correlated with the firm-year specific idiosyncratic error term. Also, region fixed effects would be quasi-perfectly collinear with respect to firm fixed effects: controlling for firm fixed effects automatically implies that regional time-invariant characteristics are controlled for, with the exception of a few firms that appeared over time in different regions (6%). Removing these firms does not change the results. In order to further control for the most relevant time-variant variables, the three sets of controls (firm level controls, region controls, and firm R&D collaborations) are considered and added incrementally.

Unobservable variables correlated with interregional knowledge integration and affecting innovative productivity might be time variant so that controlling for fixed effects at the relative level of analysis might not solve the problem. Also, reverse causality is not solved by fixed effects estimation. Taking into account lagged independent variables might partially address this issue. Nonetheless, innovative performance based on patent indicators are already measured with delays determined by the patent application process, which raises concerns on the right timing to consider. Finally, the indicators used to measure knowledge flows, especially patent-based indicators, are subject to considerable measurement errors leading to a downward bias of the estimations (Alcacer and Gittelman, 2006; Breschi and Lissoni, 2005; Criscuolo and Verspagen, 2008; Jaffe et al., 2000). We attempt to address these issues by exploiting the entry of LCCs as an exogenous shock to the level of interregional knowledge integration. We adopt a two-stage IV model. In the first-stage equation, we estimate the *Region citations' invH* (or alternatively, the *Region copatents' invH*) as a function of firm fixed effects, year fixed effects and the set of firm and region controls.

$$RegIntegration_{rt} = \gamma_0 + \gamma_1 LccEntry_{rt-3} + \gamma_2 FirmCtrls_{irt} + \gamma_3 RegCtrls_{rt} + \gamma_4 FirmFE_i + \gamma_5 YearFE_t + \epsilon_{irt} \quad (2)$$

We use a linear model with fixed effects and robust errors to estimate this equation. Accordingly to the analyses presented in the following paragraph, *LCC entry* is considered with a lag of 3 years in order to

optimize the power of the estimation. In a second-stage equation equivalent to equation (1), we include the fitted values of the indicator of interregional knowledge integration from the first-stage equation instead of the original variable. In the second-stage equation, errors are bootstrapped to take into account the non-normality of the residuals in the two-step IV estimation. We consider a variant of the model discussed where we interact the indicator of interregional integration with the level of R&D expenditure in the region before the period of entry of an LCC, *Pre-entry region R&D*, in European airports. In this case, when adopting the two-stage IV model, we will have two first-stage equations with dependent variables the indicator of interregional integration and the interaction between interregional integration and *Pre-entry region R&D,* respectively. In both equations, two instruments are included: *LCC entry* and the interaction between *LCC entry* and *Pre-entry region R&D*. The second-stage equation includes the fitted values from the two first-stage equations (Wooldridge, 2010).

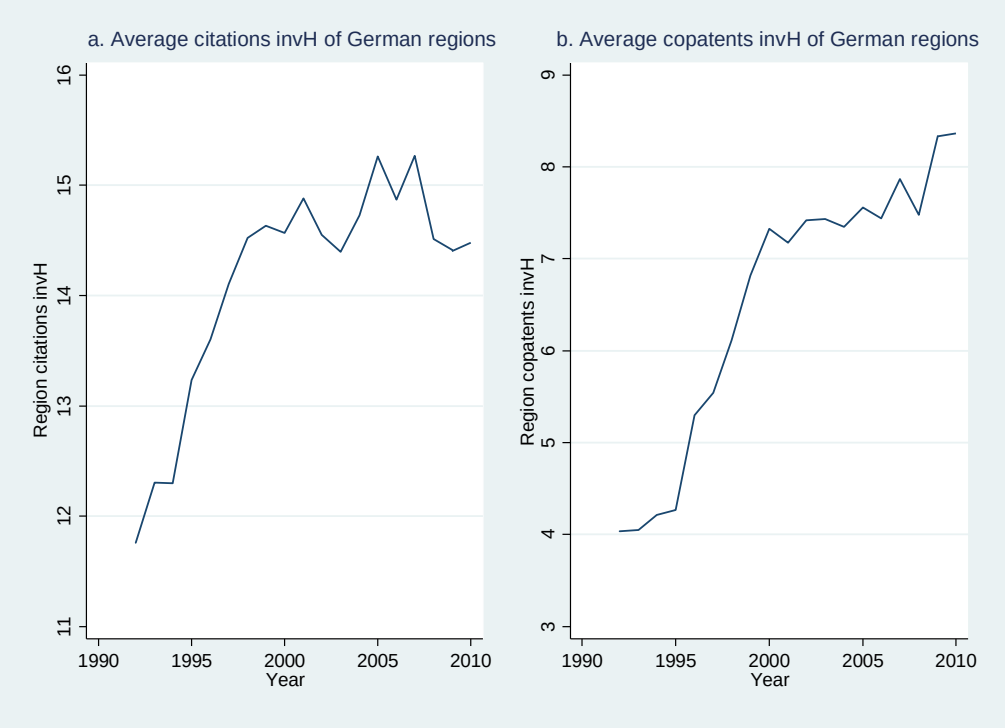
Our two-stage IV model estimation strategy is based on the assumption that the entry of an LCC is uncorrelated with the firm-year specific idiosyncratic error term: that is, it is not directly correlated with firms' innovative productivity having controlled for firm and region controls and, in particular, year and firm-region specific characteristics. In other words, our estimation strategy is valid as far as the entry of an LCC is uncorrelated with year-firm specific factors (exogeneity) and if it affects innovative productivity solely through the higher interregional knowledge integration of regions (exclusion restriction). Our confidence in the exogeneity assumption is based on the modalities of European airline liberalization, as described in the previous paragraph. The use of important control variables and the robustness of the results to the introduction of these controls provide some evidence regarding the validity of the exclusion restriction. However, the exogeneity and exclusion restriction cannot be formally tested, and we discuss potential challenges in the last section of the chapter.

4.5 Descriptive statistics

4.5.1 Interregional knowledge integration indicators

We report descriptive graphs relative to our variables of interest. The graphs reported in Figure 4.1 show the effective number of regions cited by inventors in German regions - *Region citations' invH* – (Figure 4.1a) and the effective number of regions involved in co-patenting activities - *Region copatents' invH* - (Figure 4.1b). Both measures increased significantly over the period considered. Interestingly, Breschi and Lenzi (2012) found a similar trend for the indicator of geographic distance connection of US cities. This evidence indicates an increasing average interregional integration of regions with other regions worldwide.

Figure 4.1: Citations and co-patenting activities inverse Herfindahl index over time



In addition, we show in Figure 4.2 the trend of citations abroad, citations to other German regions and citations to the same NUTS2. Similarly, in Figure 4.3, we show the trend of co-patenting activities abroad, co-patenting activities with other German regions and co-patenting activities in the same NUTS2. It is interesting to note that the number of citations abroad is consistently higher than the number of citations within Germany and within the same region. The number of citations abroad also appears to have sharply increased over time, although proportionally to the increase in the total number of citations. Similarly, the total number of co-patenting activities abroad increased from an average of 4 in 1992 to approximately 20 in 2009. However, this increase is proportional to the increase of co-patenting activities in general, and, in this case, co-patenting activities within the region and in Germany remain largely prevalent.

Figure 4.2: Average number of citations of German regions to other regions

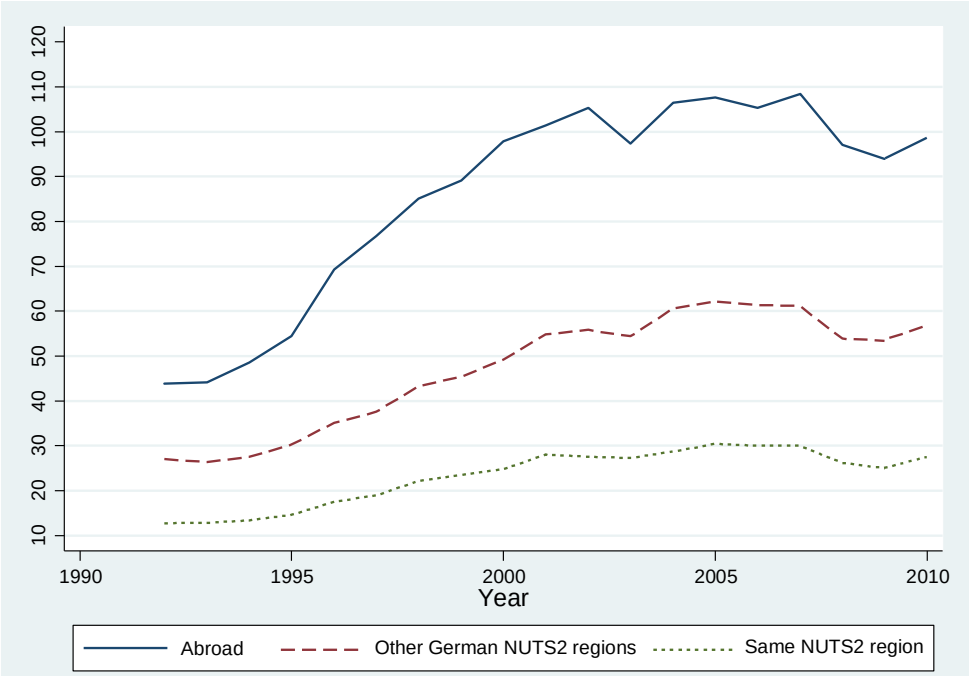
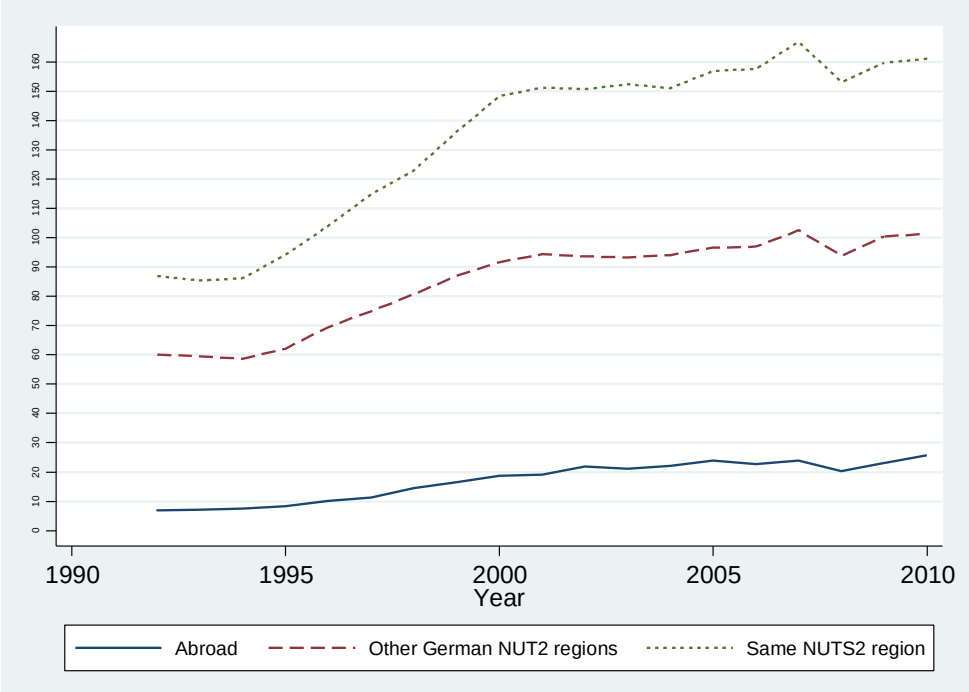


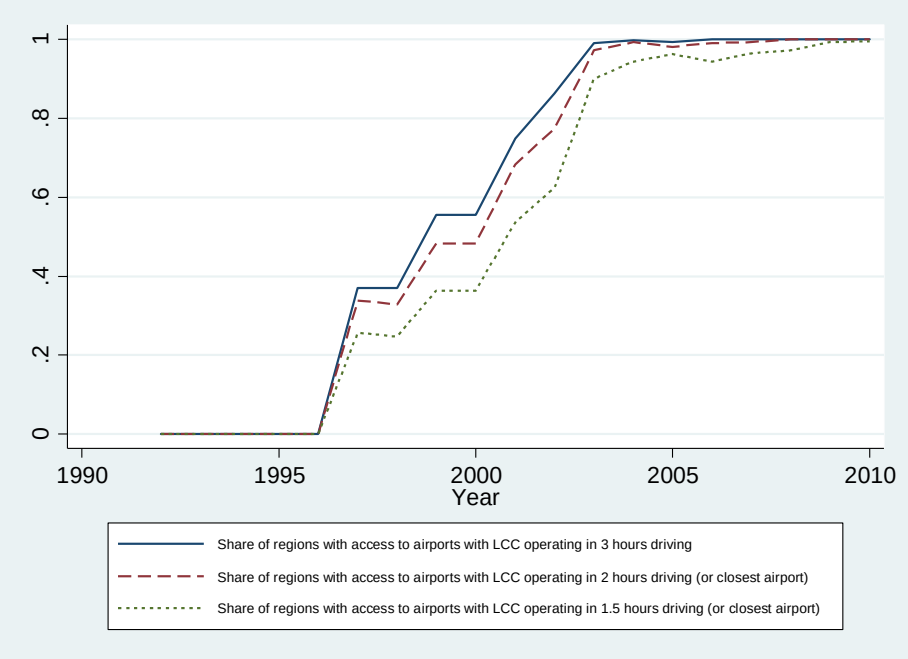
Figure 4.3: Average number of German regions' co-patents with other regions



4.5.2 LCC entry

The graph reported in Figure 4.4 shows the entry pattern of LCCs in German regions. Each line corresponds respectively to the share of regions with access to at least one airport with an LCC operating in 3 hours driving distance, 2 hours driving distance (or closest airport) and 1 and-a-half hours driving distance (or closest airport). Importantly to note, the first entries are registered in 1997, and the number of regions with access to LCC flights sharply increase in a period of approximately 5 years. After 2003, the totality of regions in Germany had access to at least one airport with LCCs operating at a reasonable travel distance.

Figure 4.4: LCC entry in German regions



4.5.3 Effect of LCC entry on interregional knowledge integration

In the following graphs (Figure 4.5 and Figure 4.6), we present a graphical semi-parametric analysis to explore the effect of LCC entry on interregional integration. We estimate the relative averages of *Region citations' invH* and *Region copatents' invH* at different periods before, after and at the moment of LCC entry (time 0) having controlled for region fixed effects and year fixed effects. Averages are plotted in reference to the average in the period minus 1, one year before LCC entry.

The analyses allow us to observe that the averages of these indicators, especially *Region citations' invH*, once controlling for year and region fixed effects, are approximately constant up to the year of LCC entry and significantly increase after three years from LCC entry. While this indicates that the effect of LCC entry is not sharp at the moment of entry, it is consistent with a gradual penetration of LCCs in the airport and other airport destinations. Furthermore, observed citations and co-patenting activities are the results of

innovation activities likely performed in the previous few years, such that is natural to expect a delay in the observable consequences of LCC entry. Most importantly, the lack of a pre-trend - an increase of the indicators previous to entry - provides descriptive support to the exogeneity of the entry decision that does not seem to be anticipated by a demand shock.

Figure 4.5: Effect of *LCC entry* on *Region citations' invH*

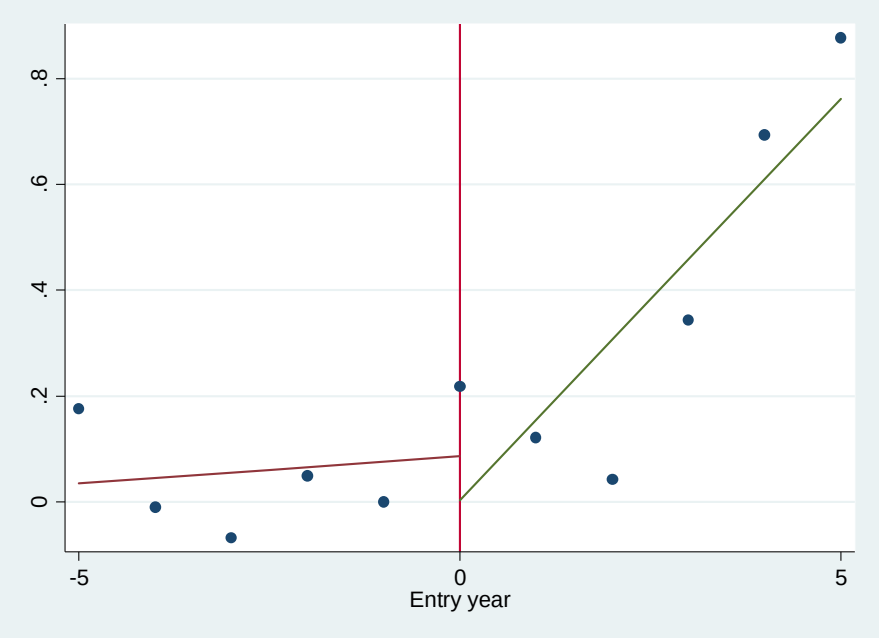
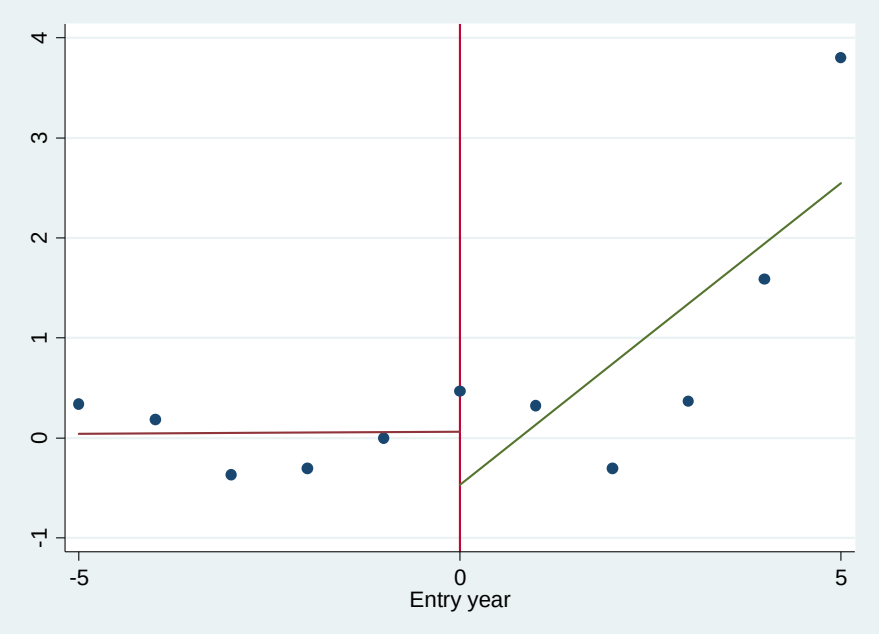


Figure 4.6: Effect of *LCC entry* on *Regions copatents invH*



4.6 Results

In this section, we present the results for the analyses relative to the variable *Region citations' invH*. Table 4.2 presents the results for the productivity equation without instrumenting the variable of interest. Table 4.3 presents the first stage of the two-step IV estimation, while Table 4.4 presents results for the second-stage equation where *Region citations' invH* is replaced by its predicted value from the first-stage equation. In both tables, we include the control variables gradually from Model 2 to Model 4. Model 1 does not include controls. Model 2 includes controls at the firm level. In Model 3, region-level controls are considered. Finally Model 4 includes additional controls at the firm level to control whether the firm is directly involved in R&D cooperation activities in Germany or abroad. Beyond being standard controls, the firm-level controls limit the concern that the effect of LCC entry on the innovative productivity of firms can be mediated by factors other than the level of interregional knowledge integration. A reduction in generalized transportation costs might determine higher growth of the firm (number of employees), free resources that could be dedicated to innovation (R&D expenditure), ease exports with an effect on incentives to innovate and on knowledge acquisition from customers abroad (Salomon and Shaver, 2005), or it might directly affect the patenting activities of the company (patent stock). Similarly, a reduction of transportation costs might have an impact at region level on the total amount of employees, R&D performed, exports and inventions (patents), consequently to the entry (or exit) of new firms in the region, as well. These variables can have a spillover effect on the innovative productivity of the company. Finally, R&D cooperation in Germany and abroad are included to test whether the effect of interregional integration is fully mediated by direct R&D collaborations of the firm.

From Table 4.2, we observe the results when the variable of interest is treated as exogenous. Interregional integration, as measured by the *Region citations' invH* variable, has a positive effect on innovative productivity, and the result is robust to the inclusion of the listed controls. The magnitude of the coefficient is not high, indicating that one additional region effectively cited by a region increases innovative productivity of firms by approximately 2%[35]. However, this corresponds to an effect of 16% higher innovative productivity for a standard deviation of the variable. Controls have the sign that might be expected, especially regarding R&D expenditure per employee, export per employee and patent stock. R&D expenditure per employee and patent stock in particular show significant and positive coefficients. On the contrary, innovative productivity decreases with the number of employees, but is weakly significant only in Model 2. Among the region control variables, it is interesting to note that the amount of R&D performed in the region negatively affects the outcome variable, although the coefficient is only weakly significant. Other regional characteristics do not show a significant impact. Finally, among the R&D collaboration variables,

[35] QMLE estimates are interpreted as $(e^{\beta} - 1) * 100$ percentage change.

R&D collaborations abroad positively affect innovative productivity, but the inclusion of these variables does not affect the result on the variable of interest.

Table 4.2: QMLE on innovative productivity (Region citations' invH)

	Model 1	Model 2	Model 3	Model 4
Region citations' invH	0.018**	0.018**	0.018**	0.018***
	(0.008)	(0.008)	(0.007)	(0.007)
N employees		-.009*	-0.008	-0.008
		(0.004)	(0.005)	(0.005)
R&D_employee		4.961**	5.080**	4.743**
		(2.269)	(2.328)	(2.081)
Export_employee		0.312	0.305	0.305
		(0.281)	(0.284)	(0.285)
Patent stock		0.002**	0.002**	0.002**
		(0.001)	(0.001)	(0.001)
R&D coop in DE				-0.190
				(0.152)
R&D coop abroad				0.283*
				(0.146)
Region R&D			-0.024*	-0.024*
			(0.012)	(0.013)
Region export			0.003*	0.003*
			(0.002)	(0.002)
Region N employees			0.001	0.001
			(0.001)	(0.001)
Region patents			-0.065	-0.067
			(0.055)	(0.056)
Year dummies	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
Observations	15,819	15,819	15,819	15,819
Number of firms	3,871	3,871	3,871	3,871
Chi-squared test	162.4	192.9	204.5	237.7

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

The first-stage results (Table 4.3) show that the entry of an LCC significantly affects *Region citations' invH*. The indicator increases approximately by 1 unit, and the F-test on the omitted instrument has a value higher than 30, largely beyond the indicative threshold of strong instruments. It is interesting to note that firm-level variables have no significant impact on the *Region citations' invH* variable. On the contrary, region-level variables affect the indicator, as it might be expected. In particular, *Region R&D* and *Region Patents* have the expected positive sign. More surprisingly *Region export* is negatively associated with *Region citations' invH*[36].

[36] However, note that the significance of the coefficient of *Region Patents* and *Region export* is not robust to cluster robust standard errors at the NUTS3 regional level. We discuss in paragraph 4.7 how the results on the main variable of interest are affected by considering cluster robust standard errors.

Table 4.3: First-stage regression on Region citations' invH

	Model 1	Model 2	Model 3	Model 4
LCC entry (lag 3 years)	1.184***	1.185***	1.156***	1.157***
	(0.185)	(0.185)	(0.185)	(0.185)
N employees		0.002	0.001	0.000
		(0.014)	(0.014)	(0.014)
R&D_employee		1.396	1.127	1.074
		(2.130)	(2.116)	(2.123)
Export_employee		-0.408	-0.428	-0.430
		(0.341)	(0.342)	(0.342)
Patent stock		-0.002	-0.003	-0.003
		(0.003)	(0.003)	(0.003)
R&D coop in DE				0.148
				(0.173)
R&D coop abroad				-0.044
				(0.214)
Region R&D			0.101***	0.101***
			(0.013)	(0.013)
Region export			-0.004***	-0.004***
			(0.001)	(0.001)
Region N employees			0.001	0.001
			(0.001)	(0.001)
Region patents			0.098*	0.098*
			(0.052)	(0.052)
Year dummies	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
Constant	10.802***	10.798***	10.717***	10.687***
	(0.189)	(0.190)	(0.194)	(0.196)
Observations	15,819	15,819	15,819	15,819
Number of firms	3,871	3,871	3,871	3,871
F-test on ommitted instrument	41.08	41.21	39.06	39.18
Chi-squared test	49.12	40.67	36.61	34.30

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

The second-stage results (Table 4.4) again report a significant effect of *Region citations' invH* on innovative productivity, importantly, robust across different specifications. The estimation coefficient increases considerably, passing from 0.02 to about 0.33. This coefficient would imply, after correction for the logarithm approximation, an effect of about a 40% higher innovative productivity per unit increase of *Region citations' invH.* This sharp increase is typical, to some extent, of IV estimators: in this sense, it might be due to the lower efficiency of the two-step IV estimation method and to a considerable measurement error of the endogenous variable, as mentioned in section 4.4. Also, the possibility must be acknowledged that unobserved variables determine a downward bias of the estimation if interregional integration is treated as exogenous. First, a higher number of citations toward other regions might be pushed by a negative shock to the marginal productivity of internal knowledge exploitation, such that the search and use of external novel inputs might be simultaneous to lower innovation performance. Second, an increase in the number of citations toward other regions might be driven by the emergence of new firms and regions operating in similar sectors, as such representing a sign of higher external competition from other regions negatively

affecting innovative sales[37]. Indeed, it is important to note that the coefficient resulting from the QMLE model in Table 4.2 and the coefficient from the two-step IV estimation in Table 4.4 are significantly different, revealing, under our assumptions, endogeneity of the variable of interest.

Table 4.4: Second-stage regression on innovative productivity (Region citations' invH)

	Model 1	Model 2	Model 3	Model 4
Region citations' invH	0.297***	0.309***	0.330***	0.327***
	(0.102)	(0.115)	(0.101)	(0.095)
N employees		0.009	0.007	0.007
		(0.014)	(0.014)	(0.009)
R&D_employee		4.495**	4.665**	4.385**
		(2.130)	(2.033)	(2.048)
Export_employee		0.446	0.456	0.455
		(0.425)	(0.405)	(0.475)
Patent stock		0.002**	0.003***	0.003**
		(0.001)	(0.001)	(0.001)
R&D coop in DE				-0.228
				(0.152)
R&D coop abroad				0.283*
				(0.150)
Region R&D			-0.058***	-0.057**
			(0.021)	(0.024)
Region export			0.004*	0.004*
			(0.002)	(0.002)
Region N employees			0.001	0.001
			(0.001)	(0.001)
Region patents			0.098	0.099
			(0.063)	(0.063)
Year dummies	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
Observations	15,819	15,819	15,819	15,819
Number of firms	3,871	3,871	3,871	3,871
Chi-squared test	273.5	327.9	323.0	245.7

Bootstrap standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

In Table 4.5, we interact the variable *Region citations' invH* with the variable *Pre-entry region R&D*. In Model 1, all variables are treated as exogenous. We find a positive effect of the interaction term, implying that firms located in regions with higher levels of R&D expenditure at the beginning of the period considered increase their innovative productivity more when *Region citations' invH* increases. However, the effect is small in magnitude, and the difference is weakly significant. The coefficient of *Region citations' invH* varies, from a mean of 0.013 for firm in regions with *Pre-entry region R&D* equal to the 25th percentile, to a mean 0.014 for firm in regions with *Pre-entry region R&D* equal to the 75th percentile, and the difference is not significant. The difference is only appreciable for firms in regions beyond the 95th percentile, for which the coefficient value is 0.032. Still, the difference is significant only at the 10% confidence level. In the

[37] In the paragraph 4.7. we discuss how considering cluster robust standard errors decreases the F-test on the omitted instruments which might imply the presence of a weak instrument bias. This could be a further explanation of the large difference in the coefficients.

following models, we adopt the two-stage IV method. As a first result, we find that the entry of an LCC has a stronger impact for firms in smaller regions with lower levels of R&D expenditures (Model 2). Interestingly, in the second-stage equation, we observe that the effect of the interaction effect is inverted, in this case, implying that the effect of *Region citations' invH* is stronger for smaller regions. However, again, the variation is not strong in magnitude and is not significant for reasonable values of the interacting variable.

Table 4.5: Regression results for models with interaction with Pre-entry region R&D

		Two-stage IV model		
	Model 1	Model 2	Model 3	Model 4
	QMLE	First stage	First stage (interaction)	Second stage
Region citations' invH	0.013**			0.333***
	(0.006)			(0.111)
Region citations' invH * Pre-entry region R&D	0.017**			-0.041**
	(0.007)			(0.017)
LCC entry (lag 3 years)		1.276***	15.222***	
		(0.186)	(2.410)	
LCC entry (lag 3 years) * Pre-entry region R&D		-0.337***	-34.79***	
		(0.072)	(7.95)	
N employees	0.008	0.001	0.409	0.005
	(0.006)	(0.014)	(0.741)	(0.005)
R&D_employee	4.772**	1.149	-1.191	4.348**
	(2.108)	(2.126)	(20.171)	(2.054)
Export_employee	0.304	-0.426	-4.389	0.438
	(0.272)	(0.342)	(4.467)	(0.508)
Patent stock	0.002**	-0.002	0.052	0.003**
	(0.001)	(0.003)	(0.071)	(0.001)
R&D coop in DE	-0.162	0.148	-3.494**	-0.242
	(0.154)	(0.174)	(1.616)	(0.162)
R&D coop abroad	0.259*	-0.033	2.063	0.289**
	(0.149)	(0.214)	(2.168)	(0.137)
Region R&D	-0.026**	0.103***	0.063***	-0.032*
	(0.013)	(0.013)	(0.006)	(0.018)
Region export	-0.003*	-0.003*	-0.059*	-0.003
	(0.001)	(0.002)	(0.072)	(0.002)
Region N employees	-0.001	-0.001	-0.012	-0.001
	(0.002)	(0.001)	(0.023)	(0.001)
Region patents	-0.075	0.125**	9.44***	-0.072
	(0.057)	(0.053)	(2.077)	(0.076)
Year dummies	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
Observations	15,819	15,819	15,819	15,819
Number of firms	3,871	3,871	3,871	3,871
F-test on omitted instruments		*31.09*	*20.51*	
F-test		33.98	11.10	
Chi-squared test	254.2			945.8

Model 1-3: Robust standard errors in parentheses; Model 4: Bootstrap standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 4.6 reports the main models, considering the entire list of controls, of the analyses using the variable *Region copatents' invH* as an indicator of interregional integration. Model 1 reports the results for the QMLE model on innovative productivity, where *Region copatents' invH* is treated as exogenous. Model 2 reports the first-stage analysis on the same variable. Model 3 shows the second-stage equations. The results are significant and equivalent to those obtained with the citation-based indicator. However, the instrument does not pass the threshold of 10 of the F-test; and the instrument cannot be considered strong in this case. Also importantly, the interaction effect with the level of R&D expenditure in the region is not significant. In the two-stage IV model the instruments are not significant (therefore, the model is not presented).

Table 4.6: Regression results on innovative productivity with copatents' indicator (Region copatents' invH)

		Two-stage IV model		
	Model 1	Model 2	Model 3	Model 4
	QMLE	First stage	Second stage	Interaction
Region copatents' invH	0.024***		0.767***	0.019***
	(0.007)		(0.220)	(0.007)
Region copatents' invH * Pre-entry region R&D				0.003
				(0.003)
LCC entry (lag 3 years)		0.493**		
		(0.193)		
N employees	-0.009*	0.051	-0.041***	-0.009*
	(0.004)	(031)	(0.016)	(0.005)
R&D_employee	4.881**	-1.349	5.770***	4.850**
	(2.124)	(2.470)	(2.051)	(2.125)
Export_employee	0.302	0.231	0.137	0.296
	(0.289)	(0.311)	(0.456)	(0.289)
Patent stock	0.002**	0.002	0.001	0.002**
	(0.001)	(0.003)	(0.001)	(0.001)
R&D coop in DE	-0.175	-0.228	-0.005	-0.165
	(0.155)	(0.155)	(0.125)	(0.155)
R&D coop abroad	0.261*	-0.025	0.288**	0.252*
	(0.144)	(0.184)	(0.122)	(0.145)
Region R&D	-0.022*	0.024	-0.043**	-0.020*
	(0.012)	(0.027)	(0.017)	(0.011)
Region export	0.003*	0.003	0.001	0.003
	(0.002)	(0.005)	(0.002)	(0.002)
Region N employees	0.001	0.001	-0.002**	0.001
	(0.001)	(0.002)	(0.001)	(0.001)
Region patents	-0.084	1.026***	-0.854***	-0.092
	(0.059)	(0.089)	(0.219)	(0.062)
Year dummies	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
Observations	15,819	15,819	15,819	15,819
Number of firms	3,871	3,871	3,871	3,871
F-test on ommitted instrument	-	*6.50*	-	-
F-test	-	60.30	-	-
Chi-squared test	240.4	-	555.9	240.9

Model 1, 2, 4: Robust standard errors in parentheses; Model 3: Bootstrap standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

4.7 Robustness

In this section, we discuss two main possible concerns on the robustness of our results. For the sake of brevity, we only discuss the main conclusions without reporting the relative results. First of all, in our models, we assumed independence of the error terms across observations. However, this assumption is likely not valid, especially due to the fact that some of the variables varied at the regional level. To address this issue, we estimated our models with cluster robust standard errors and with bootstrapped cluster robust standard error estimations, where clusters are considered at the level of the NUTS3 regions. The coefficient on the effect of *Region citations' invH* on innovative productivity remains significant at the 1% confidence level in all model specifications. The effect of LCC entry on *Region citations' invH* also remains significant,

however, the associated F-test is lower. Furthermore the indicative threshold of the F-test equal or higher than 10 is not valid if the errors are not identically or independently distributed (i.i.d). In order to further explore this issue, we estimated a linear regression IV model with fixed effects and clustered robust standard errors. In this model, the Anderson-Rubin Wald test and the Stock-Wright LM statistic (weak instruments robust statistics) both reject the null hypotheses that the coefficient of the endogenous repressor is null, providing confidence at least on the significance and the sign of the results. The same conclusion applies for the model where we included the interaction with *Pre-entry region R&D*. In the first stage, we do not find a significant interaction term any more in the first of the two first-stage equations; however, the excluded instruments are jointly significant in both equations. On the contrary, the two-stage IV models where we adopt *Region copatents' invH* as an indicator of interregional knowledge integration, are not robust to the cluster robust standard errors estimation methods and yield insignificant estimates.

A second aspect to take into account is that the results we obtained might be driven by the entry choices of new firms in a region. More (or less) innovative firms might show a preference to settle in regions showing higher levels of interregional knowledge integration and with lower transportation costs toward other destinations. Fixed effect estimation is supposed to partly take this issue into account. Nevertheless, it is still possible that new entrant firms have different growth trends of innovative productivity, which would imply that also fixed effect estimations might be biased. In order to address this issue, we estimated our models in a sample of firms that always appeared in the same region and with a foundation year lower than 1996 (earlier than the entry of an LCC). Results are mostly unchanged.

4.8 Conclusion

The geography of innovation is gradually evolving due to the increasing level of connectivity among regions, driven by communication technology improvement, the reduction of transportation costs, as well as political initiatives (Chessa et al., 2013; Tranos, 2013). Firms, regions and policy makers have invested increasingly in the search for distant knowledge in order to be connected with different locations. Overall interregional integration has been increasing. However, only recently have scholars started to explore the dynamics and effects of the phenomenon on innovative activities.

In this chapter, we study the innovative productivity of 3,871 innovative firms in Germany between 1992 and 2010, for a total of 15,819 observations. We find that higher levels of a region's interregional knowledge integration - measured as the geographic dispersion of the knowledge sources of inventions developed in the region and of co-patenting activities - lead to higher innovative productivity of local firms. We exploit airline liberalization in Germany and find that the entry of new operators (LCCs) in airports accessible to a region affects its interregional knowledge integration. The effect of interregional integration on innovative productivity, when measured as the geographic dispersion of backward citations in the patents of the region,

is robust and increases in magnitude when we adopt the entry of an LCC in a close airport as an instrument of interregional knowledge integration of the region. Results were similar when adopting a measure of interregional knowledge integration based on copatenting activities. However, the analyses showed a problem of weak instruments in this case, when adopting a two-stage IV model.

Different theories would predict that the effect of interregional knowledge integration on innovative performance might vary accordingly with the size of the region where companies are located (here measured as the average amount of R&D expenditure in the region at the beginning of the period considered). When we adopt the citation-based measure of interregional knowledge integration, we find a positive interaction effect in models where we do not use LCC entry as instrument. On the contrary, the effect is negative in the two-stage IV model. However, in both cases, the moderating effect of the regional R&D expenditure is not strong and is only weakly significant for extreme values of the interacting variable. Furthermore, we do not find any significant interaction effect when adopting the measure based on co-patenting activities and the analyses. Therefore, we conclude that there is no evidence of strong differences of the effect of interregional knowledge integration across firms located in regions with low or high levels of R&D investment.

This evidence supports policies oriented to improve interregional integration and has implications for the location and investment decision of companies. To the extent that access to distant knowledge is important for firms and regions' innovative productivity, a location in a smaller, but well-integrated region might be convenient, compared to a location in a relatively bigger geographical cluster with a lower level of interregional integration. Further research is needed to confirm these results and to disentangle the different mechanisms underlying the relationships observed here. A series of limitations of the analyses presented and considerations for future research are discussed in the following.

First, LCC entry is likely not the only factor affecting interregional integration in the period analyzed. Importantly, Information and Communication Technologies (ICT) are diffused in the same period. In this respect, year fixed effects and region fixed effects are expected to control for any factor affecting interregional integration over time and across regions, as far as this is not correlated with the entry of an LCC in a specific airport in a specific year. Since the timing of entry of an LCC in different airports has been mainly determined by exogenous factors or by the time invariant-characteristics of a region, we consider it unlikely that LCC entry is correlated with region-year specific characteristics determining the level of interregional integration, including the diffusion of ICT technologies.

Second, the reduction of generalized transportation costs, determined by the entry of an LCC, may affect innovative productivity via factors other than a higher level of interregional knowledge integration. We tried to limit this concern by showing the robustness of the results to the inclusion of several variables at the firm and regional levels. However, future research might further investigate the effects of transportation costs on

regional economic systems and other possible mechanisms through which these effects might lead to innovative performance.

Third, the scope of our results is limited by a lack of firm-level indicators equivalent to the indicators of interregional integration. These indicators cannot be computed based on the patent data in our sample, given that most of the firms did not apply for a patent or did it very seldom over time. However, it would be important to study whether it is correct to conceptualize interregional knowledge integration as a property of the region, generating externalities within the region, or, alternatively, the effect encountered emerges as an average effect of firms actually investing in access to distant knowledge.

Chapter 5 Conclusion

Proximity affects the possibility of individuals sharing information and knowledge, thereby influencing the process of knowledge production and innovation. Three broad areas of research have been the object of recent debates. First, various forms of proximity, beyond geographic proximity, interact with and differently affect innovation processes. Second, different mechanisms, beyond a simple learning process, might explain the role of proximity, both positively and negatively affecting the level and direction of technological progress. Third, increased attention has been devoted to the importance of combining the exploitation of proximate knowledge with access to distant knowledge sources. This thesis proposes three studies that, while being heterogeneous for the contexts of their analyses, provide novel empirical evidence on how different types of proximity can affect science and innovation activities through various mechanisms. In addition, particular attention is given to the importance of accessing relatively distant knowledge and information. In this final chapter, I summarize the main conclusions and insights for future research.

Previous literature has primarily conceptualized knowledge spillovers and knowledge flows as the antecedents of cumulative innovation. By sharing and accessing existing knowledge, economic agents gain an awareness and understanding of the existing state of the art in one field, which enables them to develop novel technologies. The second chapter of this thesis further suggests that the diffusion of knowledge determines the extent to which innovative efforts lead to duplicate inventions. Accordingly, the phenomenon of duplication is not randomly distributed geographically. Proximity allows inventors to avoid duplication, while inventors run the risk of duplicating inventions already discovered in distant environments, even several years before. However, proximity can also potentially increase the rate of duplication if economic agents, sharing the same pool of knowledge, end up competing on the same technological path. Indeed, for recent inventions, we find that duplication is more likely at short distances. This evidence suggests that failing to access relevant knowledge not only might impede innovation, but can also cause duplication. The extent to which the mechanisms of knowledge diffusion at higher distances (such as, for instance, ICT and the patent system itself) can or can better address this issues remain a relevant policy issue. At the same time, whether proximity and knowledge diffusion lead to cumulative innovation also depends on the strategic decisions of economic agents.

The third chapter of the thesis analyses how social proximity to distant environments might affect the possibility to attract human capital. The role of networks and referrals in hiring processes has been widely acknowledged in the economics and sociology literature. Accordingly, the results have shown that most of the benefit from hiring external students for PhD positions in two of the main research institutions in

Switzerland is mediated by the presence of their faculty networks. As such, the results suggest that social proximity with geographically distant environments helps attract external human capital with higher productivity, beyond the possible indirect benefits from network constructions and knowledge diversity. Conversely, institutions, and, more in general, innovation systems lacking in connections with external environments might face excessive information asymmetries and screening costs required to successfully and profitably attract external resources and knowledge. However, further research is needed in order to assess the extent to which our results are generalizable to other cultural and institutional contexts. Also, social proximity might be rather heterogeneous and networks might take different forms (such as professional or personal networks) that we could only explore partially.

Finally, the fourth chapter looks at the effects of interregional knowledge, defined as a region's degree of access to and adoption of knowledge developed in other geographically dispersed regions, on innovative firm performance. We empirically test the hypothesis that interregional knowledge integration positively affects the innovative productivity of local firms. We find evidence for this hypothesis, which is robust to the adoption of the LCC entry in European airports as an exogenous shock to the level of interregional knowledge integration of regions in Germany. As such, access to external sources of knowledge appears, indeed, as a determinant of local innovative performance. Therefore, we suggest that the phenomenon of interregional integration, which descriptive evidence shows to be increasing, has the potential to positively increase innovation. Based on our framework, competition in the upstream sectors determining the costs to access external environments is emerging as an important policy target. However, further research is needed to understand how the reduction of transportation costs and other determinants of the phenomenon (above all ICT) interact to determine knowledge diffusion and innovative performance.

The geography of innovation has been a flourishing area of research that has thoroughly documented the relationship between the geographic distribution of economic activities, knowledge diffusion and innovation. Decreasing transportation costs, ICT improvements and the increasing size and geographic extension of professional networks and human capital mobility promise to further shape the relationship between proximity and innovative and scientific performance. This thesis discussed few selected aspects of these phenomena which are likely to open new important areas of investigation for future research.

References

Adams, J. D., Black, G. C., Clemmons, J. R., and Stephan, P. E. 2005. Scientific teams and institutional collaborations: Evidence from U.S. universities, 1981–1999. *Research Policy*, 34(3): 259–285.

Agrawal, a., Cockburn, I., and McHale, J. 2006. Gone but not forgotten: knowledge flows, labor mobility, and enduring social relationships. *Journal of Economic Geography*, 6(5): 571–591.

Agrawal, A., Kapur, D., and McHale, J. 2008. How do spatial and social proximity influence knowledge flows? Evidence from patent data. *Journal of Urban Economics*, 64(2): 258–269.

Alcacer, J., and Gittelman, M. 2004. *How do I know what you know? Patent examiners and the generation of patent citations*. Working Paper.

Alcacer, J., and Gittelman, M. 2006. Patent citations as a measure of knowledge flows: The influence of examiner citations. *Review of Economics and Statistics*, 88(4): 774–779.

Alcácer, J., Gittelman, M., and Sampat, B. 2009. Applicant and examiner citations in U.S. patents: An overview and analysis. *Research Policy*, 38(2): 415–427.

Almeida, P., and Kogut, B. 1999. Localization of Knowledge and the Mobility of Engineers in Regional Networks. *Management Science*, 45(7): 905–917.

Amin, M., and Mabe, M. 2000. Impact factors: use and abuse. *Perspectives in publishing*, 1(2): 1–6.

Angrist, J. D., and Pischke, J.-S. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton university press.

Antoninis, M. 2006. The wage effects from the use of personal contacts as hiring channels. *Journal of Economic Behavior & Organization*, 59(1): 133–146.

Archibugi, D. 1992. Specialization and size of technological activities in industrial countries: The analysis of patent data. *Research Policy*, 21(1): 79–93.

Archibugi, D., and Iammarino, S. 2002. The globalization of technological innovation: definition and evidence. *Review of International Political Economy*, 9(1): 98–122.

Arikan, A. T. 2009. Interfirm knowledge exchanges and the knowledge creation capability of clusters. *Academy of Management Review*, 34(4): 658–676.

Arora, A., and Gambardella, A. 2005. The Impact of NSF Support on Basic Research in Economics. *Annales d'Economie et de Statistique*, Special Is: 79–80.

Arrow, K. 1972. Models of job discrimination. In A. H. Pascal and R. Corporation (Eds.), *Racial discrimination in economic life*. Lexington Books.

Arundel, A., and Kabla, I. 1998. What percentage of innovations are patented? empirical estimates for European firms. *Research Policy*, 27(2): 127–141.

Atal, V., and Bar, T. 2010. Prior art: To search or not to search. *International Journal of Industrial Organization*, 28(5): 507–521.

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

Audretsch, D. B., and Feldman, M. P. 1996. Innovative clusters and the industry life cycle. *Review of Industrial Organization*, 11(2): 253–273.

Autant-Bernard, C. 2001. The Geography Of Knowledge Spillovers And Technological Proximity. *Economics of Innovation and New Technology*, 10(4): 237–254.

Baptista, R., and Swann, P. 1998. Do firms in clusters innovate more? *Research Policy*, 27(5): 525–540.

Baruffaldi, S. H., and Landoni, P. 2012. Return Mobility and Scientific Productivity of Researchers Working Abroad: The Role of Home Country Linkages. *Research Policy*, 41(9): 1655 – 1665.

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

Bessen, J., and Meurer, M. J. 2008. *Patent failure: How judges, bureaucrats, and lawyers put innovators at risk*. Princeton Univ Pr.

Bikard, M. 2012. *Simultaneous Discoveries as a Research Tool: Method and Promise*. SSRN Electronic Journal.

Black, G. C., and Stephan, P. E. 2010. The Economics of University Science and the Role of Foreign Graduate Students and Postdoctoral Scholars. *American Universities in a Global Market*: 129–161. University of Chicago Press.

Bloom, N., Schankerman, M., and Van Reenen, J. 2013. Identifying Technology Spillovers and Product Market Rivalry. *Econometrica*, 81(4): 1347–1393.

Bonaccorsi, A., Čenys, A., Chorafakis, G., Cooke, P., Foray, D., Giannitsis, A., et al. 2009. *The Question of R&D Specialisation: Perspectives and policy implications*. (D. Pontikakis, D. Kyriakou, & R. van Bavel, Eds.). Luxembourg: European Commission, Joint Research Centre - Institute for Prospective Technological Studies.

Boschma, R. A. 2005. Proximity and innovation: a critical assessment. *Regional studies*, 39(1): 61–74.

Boschma, R. A., and Frenken, K. 2005. Why is economic geography not an evolutionary science? Towards an evolutionary economic geography. *Journal of Economic Geography*, 6(3): 273–302.

Boschma, R. A., and Frenken, K. 2011. The emerging empirics of evolutionary economic geography. *Journal of Economic Geography*, 11(2): 295–307.

Bottazzi, L., and Peri, G. 2003. Innovation and spillovers in regions: Evidence from European patent data. *European Economic Review*, 47(4): 687–710.

Brannigan, A., and Wanner, R. A. 1983. Multiple Discoveries in Science: A Test of the Communication Theory. *The Canadian Journal of Sociology / Cahiers canadiens de sociologie*, 8(2): 135–151.

Breschi, S., and Lenzi, C. 2012. Net city: how co-invention networks shape inventive productivity in us cities. *American Association of Geography 2012 Conference*.

Breschi, S., and Lissoni, F. 2001. Knowledge spillovers and local innovation systems: a critical survey. *Industrial and Corporate Change*, 10(4): 975.

Breschi, S., and Lissoni, F. 2005. Knowledge Networks from Patent Data. In H. Moed, W. Glänzel, and U. Schmoch (Eds.), *Handbook of Quantitative Science and Technology Research*: 613–643. Springer Netherlands.

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

Bresnahan, T., Gambardella, A., and Saxenian, A. L. 2001. "Old Economy" Inputs for "New Economy" Outcomes: Cluster Formation in the New Silicon Valleys. *Industrial and Corporate Change*, 10(4): 835.

Brogaard, J., Engelberg, J., and Parsons, C. A. 2014. Networks and productivity: Causal evidence from editor rotations. *Journal of Financial Economics*, 111(1): 251–270.

Calder, S., and Laker, F. 2002. *No frills: The truth behind the low-cost revolution in the skies*. *Virgin Books*. Virgin London.

Cappelli, R., and Montobbio, F. 2013. *European Integration and Knowledge Flows across European Regions*. Department of Economics and Statistics Cognetti de Martiis - Working Papers.

Cassiman, B., and Veugelers, R. 2006. In Search of Complementarity in Innovation Strategy: Internal R&D and External Knowledge Acquisition. *Management Science*, 52(1): 68–82.

Castilla, E. J. 2005. Social networks and employee performance in a call center. *American Journal of Sociology*, 110(5): 1243–1283.

Catalini, C. 2012. *Microgeography and the Direction of Inventive Activity*. Working Paper SSRN 2126890.

Chessa, A., Morescalchi, A., Pammolli, F., Penner, O., Petersen, A. M., and Riccaboni, M. 2013. Is Europe Evolving Toward an Integrated Research Area? *Science*, 339(6120): 650–651.

Coe, D. T., and Helpman, E. 1995. International R&D spillovers. *European Economic Review*, 39(5): 859–887.

Cohen, W. M., Goto, A., Nagata, A., Nelson, R. R., and Walsh, J. P. 2002. R&D spillovers, patents and the incentives to innovate in Japan and the United States. *Research Policy*, 31(8-9): 1349–1367.

Cohen, W. M., and Levinthal, D. A. 1990. Absorptive Capacity: A New Perspective on Learning and Innovation. *Administrative Science Quarterly*, 35(1): 128–152.

Collins, H. 1992. *Changing Order: Replication and Induction in Scientific Practice*. University of Chicago Press.

Constant, E. W. 1978. On the Diversity and Co-Evolution of Technological Multiples: Steam Turbines and Pelton Water Wheels. *Social Studies of Science*, 8(2): 183–210.

Cornell, B., and Welch, I. 1996. Culture, Information, and Screening Discrimination. *Journal of Political Economy*, 104(3): 542–571.

Cowan, R., David, P. A., and Foray, D. 2000. The explicit economics of knowledge codification and tacitness. *Industrial and Corporate Change*, 9(2): 211–253.

Crescenzi, R. 2014. Changes in Economic Geography Theory and the Dynamics of Technological Change. *Handbook of Regional Science*, 649–666.

Crescenzi, R., Rodriguez-Pose, A., and Storper, M. 2007. The territorial dynamics of innovation: a Europe United States comparative analysis. *Journal of Economic Geography*, 7(6): 673–709.

Criscuolo, P., and Verspagen, B. 2008. Does it matter where patent citations come from? Inventor vs. examiner citations in European patents. *Research Policy*, 37(10): 1892–1908.

Cruz, S. C. S., and Teixeira, A. A. C. 2009. The Evolution of the Cluster Literature: Shedding Light on the Regional Studies-Regional Science Debate. *Regional Studies*, 9(1): 1–26.

Dasgupta, P., and David, P. A. 1994. Toward a new economics of science. *Research Policy*, 23(5): 487–521.

Dasgupta, P., and Maskin, E. 1987. The simple economics of research portfolios. *The Economic Journal*, 97(387): 581–595.

De Rassenfosse, G., Schoen, A., and Wastyn, A. 2014. Selection bias in innovation studies: A simple test. *Technological Forecasting and Social Change*, 81: 287–299.

Delgado, M., Porter, M. E., and Stern, S. 2010. Clusters and entrepreneurship. *Journal of Economic Geography*, 10(4): 495–518.

Denicolo, V., and Franzoni, L. A. 2003. The contract theory of patents. *International Review of Law and Economics*, 23(4): 365–380.

Ding, W., Levin, S., Stephan, P. E., and Winkler, A. 2010. The impact of Information Technology on Academic Scientists' Productivity and Collaboration Patterns. *Management Science*, 56(9): 1439–1461.

Ding, W., M., F., and Stuart, T. E. 2006. Gender Differences in Patenting in the Academic Life Sciences. *Science*, 4: 665–667.

Dobruszkes, F. 2006. An analysis of European low-cost airlines and their networks. *Journal of Transport Geography*, 14(4): 249–264.

Eisingerich, A. B., Bell, S. J., and Tracey, P. 2010. How can clusters sustain performance? The role of network strength, network openness, and environmental uncertainty. *Research Policy*, 39(2): 239–253.

Elkana, Y. 1971. The Conservation of Energy: a Case of Simultaneous Discovery?'. *Archives internationales d'histoire des sciences*, 24: 31–60.

Encaoua, D., and Ulph, D. 2005. *Catching-up or Leapfrogging: The effects of competition on innovation and growth*. Working paper.

Etherington, D., and Jones, M. 2009. City-regions: new geographies of uneven development and inequality. *Regional Studies*, 43(2): 247–265.

Feldman, M. P., and Kogler, D. F. 2010. Stylized Facts in the Geography of Innovation. *Handbook of the Economics of Innovation*, 1: 381–410.

Fernandez, R. M., Castilla, E. J., and Moore, P. 2000. Social capital at work: Networks and employment at a phone center. *American journal of sociology*, 1288–1356.

Foray, D. 2004. *The Economics of knowledge*. the MIT Press.

Forman, C., and van Zeebroeck, N. 2012. From Wires to Partners: How the Internet Has Fostered R&D Collaborations Within Firms. *Management Science*, 1–20.

Frenken, K., Cefis, E., and Stam, E. 2014. Industrial Dynamics and Clusters: A Survey. *Regional Studies*, 1–18.

Frenz, M., and Ietto-Gillies, G. 2009. The impact on innovation performance of different sources of knowledge: Evidence from the UK Community Innovation Survey. *Research Policy*, 38(7): 1125–1135.

Ganguli, I. 2010. Saving Soviet Science: The Impact of Grants when Government R&D Funding Disappears. *American Economic Journal: Applied Economics*.

Gaulé, P., and Piacentini, M. 2013. Immigration and Innovation: Chinese Graduate Students in US Universities. *Review of Economics and Statistics. Forthcoming*, 95(2): 698–701.

Gertler, M. S. 2003. Tacit knowledge and the economic geography of context, or The undefinable tacitness of being (there). *Journal of Economic Geography*, 3(1): 75–99.

Giuri, P., Mariani, M., Brusoni, S., Crespi, G., Francoz, D., Gambardella, A., et al. 2007. Inventors and invention processes in Europe: Results from the PatVal-EU survey. *Research Policy*, 36(8): 1107–1127.

Gómez, M. A. 2011. Duplication externalities in an endogenous growth model with physical capital, human capital, and R&D. *Economic Modelling*, 28(1-2): 181–187.

Graham, S., Merges, R., Samuelson, P., and Sichelman, T. 2009. High technology entrepreneurs and the patent system: Results of the 2008 Berkeley patent survey. *Berkeley Technology Law Journal*, 24(4): 255–327.

Granovetter, M. S. 1973. The strength of weak ties. *American Journal of Sociology*, 1360–1380.

Granovetter, M. S. 1995. *Getting a job: A study of contacts and careers*. University of Chicago Press.

Griliches, Z. 1979. Issues in assessing the contribution of research and development to productivity growth. *The Bell Journal of Economics*, 92–116.

Grossman, G. M., and Helpman, E. 1993. *Innovation and Growth in the Global Economy*. The MIT Press.

Guellec, D., Martinez, C., and Zuniga, P. 2012. Pre-emptive patenting: securing market exclusion and freedom of operation. *Economics of Innovation and New Technology*, 21(1): 1–29.

Guellec, D., and van Pottelsberghe de la Potterie, B. 2000. Applications, grants and the value of patent. *Economics Letters*, 69(1): 109–114.

Hall, B. H., and Ziedonis, R. H. 2001. The patent paradox revisited: an empirical study of patenting in the US semiconductor industry, 1979-1995. *RAND Journal of Economics*, 101–128.

Hausman, J. A., Hall, B. H., and Griliches, Z. 1986. Econometric Models for Count Data with an Application to the Patents-R&D Relationship. *Econometrica*, 52: 909–938.

Hegde, D., and Tumlinson, J. 2011. *Can Birds of a Feather Fly Together? Evidence for the Economic Payoffs of Ethnic Homophily*. Working paper.

Hoekman, J., Frenken, K., and Tijssen, R. J. W. 2010. Research collaboration at a distance: Changing spatial patterns of scientific collaboration within Europe. *Research Policy*, 39(5): 662–673.

Horta, H., Veloso, F. M., and Grediaga, R. 2010. Navel Gazing: Academic Inbreeding and Scientific Productivity. *Management Science*, 56(3): 414–429.

Hunt, J. 2011. Which immigrants are most innovative and entrepreneurial? Distinctions by entry visa. *Journal of Labor Economics*, 29(3): 417–457.

Jaffe, A. B. 1986. Technological Opportunity and Spillovers of R & D: Evidence from Firms' Patents, Profits, and Market Value. *American Economic Review*, 76(5): 984–1001.

Jaffe, A. B. 1989. Real effects of academic research. *American Economic Review*, 957–970.

Jaffe, A. B., Trajtenberg, M., and Fogarty, M. S. 2000. *The Meaning of Patent Citations: Report on the NBER/Case-Western Reserve Survey of Patentees*. No. w7631. NBER Working Paper.

Jaffe, A. B., Trajtenberg, M., and Henderson, R. 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics*, 108(3): 577–598.

Jones, B. F. 2009. The Burden of Knowledge and the Death of the Renaissance Man: Is Innovation Getting Harder? *Review of Economic Studies*, 76(1): 283–317.

Jones, B. F., Wuchty, S., and Uzzi, B. 2008. Multi-University Research Teams: Shifting Impact, Geography, and Stratification in Science. *Science*, 322(5905): 1259–1262.

Jones, C. I. 1995. R & D-Based Models of Economic Growth. *Journal of Political Economy*, 103(4): 759–784.

Jones, C. I., and Williams, J. C. 2000. Too Much of a Good Thing? The Economics of Investment in R&D. *Journal of Economic Growth*, 5(1): 65–85.

Jorde, T. M., and Teece, D. J. 1990. Innovation and Cooperation: Implications for Competition and Antitrust. *The Journal of Economic Perspectives*, 4(3): 75–96.

Keller, W. 2002. Geographic localization of international technology diffusion. *American Economic Review*, 92(1): 120–142.

Keller, W. 2004. International technology diffusion. *Journal of Economic Literature*, 42(3): 752–782.

Kerr, W. R. 2008. Ethnic scientific communities and international technology diffusion. *Review of Economics and Statistics*, 90(3): 518–537.

Kitch, E. W. 1977. The Nature and Function of the Patent System. *Journal of Law and Economics*, 20(2): 265–290.

Kortum, S. 1993. Equilibrium R&D and the Patent-R&D Ratio: U.S. Evidence. *American Economic Review*, 83(2): 450–457.

Krugman, P. 1998. What's new about the new economic geography? *Oxford Review of Economic Policy*, 14(2): 7–17.

Krugman, P. R. 1991. *Geography and trade*. the MIT Press.

Kuhn, T. S. 1996. *The structure of scientific revolutions*. University of Chicago press.

Laband, D. N., and Piette, M. J. 1994. Favoritism versus Search for Good Papers: Empirical Evidence Regarding the Behavior of Journal Editors. *Journal of Political Economy*, 102(1): 194–203.

Lamb, D., and Easton, S. M. 1984. *Multiple Discovery: The pattern of scientific progress*. Avebury.

Lampe, R. 2007. Strategic citation. *Review of Economics and Statistics*, 94(1).

Lecocq, C., Leten, B., Kusters, J., and Van Looy, B. 2012. Do firms benefit from being present in multiple technology clusters? An assessment of the technological performance of biopharmaceutical firms. *Regional Studies*, 46(9): 1107–1119.

Lee, C. Y. 2009. Do firms in clusters invest in R&D more intensively? Theory and evidence from multi-country data. *Research Policy*, 38(7): 1159–1171.

Leten, B., Landoni, P., and Van Looy, B. 2014. Science or graduates: How do firms benefit from the proximity of universities? *Research Policy*, 43(8): 1398–1412.

Levin, S. G., and Stephan, P. E. 1991. Research Productivity Over the Life Cycle: Evidence for Academic Scientists. *American Economic Review*, 81(1): 114–132.

Levin, S. G., and Stephan, P. E. 1999. Are the Foreign Born a Source of Strength for U.S. Science? *Science*, 285(5431): 1213–1214.

Li, D. 2012. *Information, bias, and efficiency in expert evaluation: evidence from the NIH*: 1–57.

Libaers, D. P. 2007. Role and Contribution of Foreign-Born Scientists and Engineers to the Public U.S. Nanoscience and Technology Research Enterprise. *Engineering Management, IEEE Transactions on*, 54(3): 423–432.

Lissoni, F., Mairesse, J., Montobbio, F., and Pezzoni, M. 2011. Scientific productivity and academic promotion: a study on French and Italian physicists. *Industrial and Corporate Change*, 20(1): 253–294.

Long, S. J. 1990. The Origins of Sex Differences in Science. *Social Forces*, 68: 1297–1315.

Mancusi, M. L. 2008. International spillovers and absorptive capacity: A cross-country cross-sector analysis based on patents and citations. *Journal of International Economics*, 76(2): 155–165.

Marshall, A. 1891. *Principles of Economics (1920)*. *McMillan, London*.

Martin, R., and Sunley, P. 2006. Path dependence and regional economic evolution. *Journal of Economic Geography*, 6(4): 395–437.

Maurseth, P. B., and Verspagen, B. 2002. Knowledge spillovers in Europe: a patent citations analysis. *The Scandinavian journal of economics*, 104(4): 531–545.

McCann, B. T., and Folta, T. B. 2008. Location matters: Where we have been and where we might go in agglomeration research. *Journal of Management*, 34(3): 532.

McCann, P., and Ortega-Argilés, R. 2013. Smart Specialization, Regional Growth and Applications to European Union Cohesion Policy. *Regional Studies*, (ahead-of-print): 1–12.

McPherson, M., Smith-Lovin, L., and Cook, J. M. 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27: 415–444.

Menzel, M.-P., and Fornahl, D. 2010. Cluster life cycles-dimensions and rationales of cluster evolution. *Industrial and Corporate Change*, 19(1): 205–238.

Merton, R. K. 1961. Singletons and Multiples in Scientific Discovery: A Chapter in the Sociology of Science. *Proceedings of the American Philosophical Society*, 105(5): 470–486.

Merton, R. K. 1979. *The sociology of science: Theoretical and empirical investigations*. University of Chicago Press.

Michel, J., and Bettels, B. 2001. Patent citation analysis. A closer look at the basic input data from patent search reports. *Scientometrics*, 51(1): 185–201.

Montgomery, J. D. 1991. Social Networks and Labor-Market Outcomes: Toward an Economic Analysis. *American Economic Review*, 81: 1408–1418.

Mowery, D. C., and Ziedonis, A. A. 2014. Markets versus Spillovers in Outflows of University Research. *Research Policy*, Forthcoming.

Mundlak, Y. 1978. On the pooling of time series and cross section data. *Econometrica: journal of the Econometric Society*, 69–85.

Munshi, K. 2003. Networks in the Modern Economy: Mexican Migrants in the US Labor Market. *Quarterly Journal of Economics*, 118(2): 549–599.

Murray, F., and O'Mahony, S. 2007. Exploring the foundations of cumulative innovation: Implications for organization science. *Organization Science*, 18(6): 1006–1021.

Niehans, J. 1995. Multiple discoveries in economic theory. *The European Journal of the History of Economic Thought*, 2(1): 1–28.

Ogburn, W. F., and Thomas, D. 1922. Are inventions inevitable? A note on social evolution. *Political Science Quarterly*, 37(1): 83–98.

Peri, G. 2005. Determinants of knowledge flows and their effect on innovation. *Review of Economics and Statistics*, 87(2): 308–322.

Pezzoni, M., Sterzi, V., and Lissoni, F. 2012. Career progress in centralized academic systems: Social capital and institutions in France and Italy. *Research Policy*, 41(4): 704–719.

Phene, A., Fladmoe-Lindquist, K., and Marsh, L. 2006. Breakthrough innovations in the US biotechnology industry: the effects of technological space and geographic origin. *Strategic Management Journal*, 27(4): 369–388.

Ponzo, M., and Scoppa, V. 2010. The use of informal networks in Italy: Efficiency or favoritism? *The Journal of Socio-Economics*, 39(1): 89–99.

Porter, M. E. 1998. *Clusters and the new economics of competition*. Harvard Business Review.

Prendergast, C., and Topel, R. H. 1996. Favoritism in Organizations. *Journal of Political Economy*, 104(5): 958–78.

Redding, S. J., and Sturm, D. M. 2008. The Costs of Remoteness: Evidence from German Division and Reunification. *American Economic Review*, 98(5): 1766–1797.

Romer, P. M. 1986. Increasing returns and long-run growth. *Journal of Political Economy*, 1002–1037.

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

Salomon, R. M., and Shaver, J. M. 2005. Learning by exporting: new insights from examining firm innovation. *Journal of Economics & Management Strategy*, 14(2): 431–460.

Saloner, G. 1985. The Old Boys' Networks as a Screening Mechanism. *Journal of Labor Economics*, 3: 255–267.

Salt, J. 1997. International movements of the highly skilled. *OECD Social, Employment and Migration Working Papers*.

Saxenian, A. 1994. *Regional advantage: culture and competition in Silicon Valley and Route 128 (1994) Cambridge*. MA: Harvard University Press.

Saxenian, A. 2005. From Brain Drain to Brain Circulation: Transnational Communities and Regional Upgrading in India and China. *Studies in Comparative International Development*, 40(2): 35–61.

Saxenian, A. L. 2007. *The new Argonauts: Regional advantage in a global economy*. Harvard Univ Pr.

Scotchmer, S. 1991. Standing on the Shoulders of Giants: Cumulative Research and the Patent Law. *The Journal of Economic Perspectives*, 5(1): 29–41.

Simonton, D. K. 1979. Multiple discovery and invention: Zeitgeist, genius, or chance? *Journal of Personality and Social Psychology 37*, 37(9): 1603–1616.

Singh, J. 2005. Collaborative networks as determinants of knowledge diffusion patterns. *Management Science*, 51(5): 756–770.

Singh, J., and Marx, M. 2013. Geographic Constraints on Knowledge Spillovers: Political Borders vs. Spatial Proximity. *Management Science*.

Song, J., Almeida, P., and Wu, G. 2003. Learning-by-Hiring: When Is Mobility More Likely to Facilitate Interfirm Knowledge Transfer? *Management Science*, 49(4): 351–365.

Sonn, J. W., and Storper, M. 2008. The increasing importance of geographical proximity in knowledge production: an analysis of US patent citations, 1975-1997. *Environment and Planning A*, 40(5): 1020.

Stephan, P. E. 1996. The economics of science. *Journal of Economic literature*, 34(3): 1199–1235.

Stephan, P. E. 2012. *How Economics Shapes Science*. Harvard University Press.

Sternitzke, C. 2009. Reducing uncertainty in the patent application procedure-Insights from invalidating prior art in European patent applications. *World Patent Information*, 31(1): 48–53.

Stuen, E. T., Mobarak, A. M., and Maskus, K. E. 2007. *Foreign Graduate Students and Knowledge Creation at US Universities: Evidence from Enrollment Fluctuations*. *Working Paper*.

Sylos Labini, M. 2005. *Social networks and wages: It is all about connections!* Working Paper.

Tan, D., and Roberts, P. W. 2010. Categorical coherence, classification volatility and examiner-added citations. *Research Policy*, 39(1): 89–102.

Thompson, P. 2006. Patent citations and the geography of knowledge spillovers: evidence from inventor-and examiner-added citations. *Review of Economics and Statistics*, 88(2): 383–388.

Thompson, P., and Fox-Kean, M. 2005. Patent Citations and the Geography of Knowledge Spillovers: A Reassessment. *American Economic Review*, 95(1): 450–460.

Tranos, E. 2013. The Geography of the Internet: Cities, Regions and Internet Infrastructure in Europe.

Van Looy, B., Landoni, P., Callaert, J., van Pottelsberghe, B., Sapsalis, E., and Debackere, K. 2011. Entrepreneurial effectiveness of European universities: An empirical assessment of antecedents and trade-offs. *Research Policy*, 40(4): 553–564.

Venturini, F. 2012. Looking into the black box of Schumpeterian growth theories: An empirical assessment of R&D races. *European Economic Review*, 56(8): 1530–1545.

Von Graevenitz, G., Wagner, S., and Harhoff, D. 2013. Incidence and Growth of Patent Thickets: The Impact of Technological Opportunities and Complexity. *The Journal of Industrial Economics*, 61(3): 521–563.

Walsh, J. P., Cohen, W. M., and Cho, C. 2007. Where excludability matters: Material versus intellectual property in academic biomedical research. *Research Policy*, 36(8): 1184–1203.

Wooldridge, J. M. 1997. Quasi-likelihood methods for count data. *Handbook of applied econometrics*, 2: 352–406.

Wooldridge, J. M. 2010. *Econometric Analysis of Cross Section and Panel Data*. The MIT press.

Yakubovich, V., and Lup, D. 2006. Stages of the recruitment process and the referrer's performance effect. *Organization science*, 17(6): 710–723.

Ziedonis, R. H. 2004. Don't Fence Me In: Fragmented Markets for Technology and the Patent Acquisition Strategies of Firms. *Management Science*, 50(6): 804–820.

Zinovyeva, N., and Bagues, M. F. 2012. *The role of connections in academic promotions*. INDEM Working Paper Business Economic.

Curriculum Vitae

STEFANO HORST BARUFFALDI

École Polytechnique Fédérale de Lausanne (EPFL)
Chair in Economics and Management of Innovation (CEMI)
ODY 4.16 Station 5, CH-1015 Lausanne | +41 21 6930037 | stefano.baruffaldi@epfl.ch

EDUCATION

École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland — 2010 to present
PhD in economics and management of innovation
Supervisor: Dominique Foray

Gerzensee Study Center, Gerzensee, Switzerland — 2012
Swiss Program for Beginning Doctoral Students in Economics

Politecnico di Milano, Milano, Italy — 2008 - 2010
Master in Management, Economics and Industrial Engineering
Thesis: "Mobilità scientifica internazionale: Una analisi sui ricercatori stranieri in Italia e Portogallo"

Politecnico di Milano, Milano, Italy — 2004 - 2008
Bachelor of Science in Management and Production Engineering

Liceo Classico Giosuè Carducci, Milano, Italy — 1999 - 2004
Diploma di Liceo Classico

RESEARCH EXPERIENCE

École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland — 2010 to present
Research assistant
Supervisor: Dominique Foray

University of Coimbra, Coimbra, Portugal — 2007
Research assistant
FCT financed project on Skilled Immigration in Portugal
Supervisor: Tiago Santos Pereira
Survey design and management, data construction, data analysis

PUBLICATIONS AND PAPERS

PUBLICATIONS IN REFEREED JOURNALS

With Paolo Landoni, *"Return mobility and scientific productivity of researchers working abroad: The role of home country linkages."*
Research Policy (2012), 41(9), 1655 – 1665

WORKING PAPERS

With Guillaume Burghouwt, *"Fly to learn: Interregional integration and firms' innovative productivity"*

With Julio Raffo, *"The geography of duplicated inventions: an analysis from patent citations" (under review for Management Science)*

With Annamaria Conti, and Fabiana Visentin, *"On the productivity of PhD students from the supervisors' networks."*

With Paolo Landoni, *"Motivations of scientific international mobility: The role of non-economic factors."*

WORK IN PROGRESS

With Marianna Marino, and Fabiana Visentin. *"Mobility research fellowships: A policy evaluation"*.

With Markus Simeth. *"The impact of patent disclosure on the diffusion of knowledge"*.

TEACHING EXPERIENCE

École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland 2014
Teaching assistant of Professor Stephane Lhuillery in "**Econometrics: Data Analysis & Empirical Method**" master course
Teaching exercise sessions, students' assistance

École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland 2012
Teaching assistant of Professor Julio Raffo in "**Econometrics: Data Analysis & Empirical Method**" master course
Teaching exercise sessions, teaching theory on endogeneity and IV estimators, students' assistance

École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland 2011 - 2013
Teaching assistant of Professor Thomas Weber in "**Information: Strategy and Economics**" master course
Students assistance and administered grades

École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland 2010
Teaching assistant – Professor Anu Wadhwa in "**Corporate strategy**"
Students assistance and administered grades

Politecnico di Milano, Milano, Italy 2009
Seminar on "**Bibliographic search and bibliometric indicators**" within the master course of Professor Paolo Landoni "Research and education systems management"

ACADEMIC PRESENTATIONS AND CONFERENCES

"Effects and determinants of the scientific international mobility: the cases of foreign researchers in Italy and Portugal"

- Triple Helix VIII conference, Madrid, Spain, October 2010

"*Network effects on knowledge production: Evidence for PhD students"*

- The 12th Roundtable for Engineering Entrepreneurship Research, Georgia Institute of Technology, Atlanta GA, November 2012
- The BRICK Workshop - The Organization, Economics and Policy of Scientific Research, Turin, Italy, March 2013
- Department of Economics, University of Lausanne, Lausanne, Switzerland, April 2013
- Department of Economics, Roma Tre University, Rome, Italy, April 2013
- Department of Economics, University of Lugano, Lugano Switzerland, March 2013

- Department of Economics, Politecnico di Milano, Milan Italy, March 2013

"The geography of duplicated inventions: An analysis from patent citations"

- Chair of Economics and Management of Innovation Workshop, Ovronnaz, Switzerland, January 2014
- Workshop on Regional and Urban Economics, Barcelona Spain, October 2012
- Annual Meeting of the Academy of Management, Orlando, FL, USA, August 2013
- European Policy for Intellectual Property Conference (EPIP), Paris, France, 2013

"Fly to learn: Interregional integration and firms' innovative productivity"

- Chair of Economics and Management of Innovation Workshop, St. Luc, Switzerland, January 2014
- The DRUID Society Conference, Copenhagen Business School, Copenhagen, Denmark, June 2014
- The international conference Governance of a Complex World, Turin, Italy, June 2014
- Annual Meeting of the Academy of Management, Philadelphia, PA, USA, August 2014

"The impact of patent disclosure on the diffusion of knowledge".

- European Policy for Intellectual Property Conference (EPIP), Brussels, Belgium, 2014

AWARDS

Ecole Polytechnique Fédérale de Lausanne 2012 – 2013
CDM best teaching assistant award

Ecole Polytechnique Fédérale de Lausanne 2012 – 2014
Swiss National Science Foundation Research Grant no. 100010_149931/1

RELATED EXPERIENCE

Reviewer for refereed journals: Research Policy, Small Business Economics, European Management Review, Higher Education

Reviewer for conference proceedings papers: Academy of Management Annual Meeting

LANGUAGES

Italian	Native language
English	Speaking, reading fluently and writing with high proficiency
Portuguese	Speaking, reading fluently and writing with high proficiency
French	Speaking, reading fluently and write with basic competence
Spanish	Speaking, reading intermediate and write with basic competence