PhD-FDEF-2018-07
The Faculty of Law, Economics and Finance

DISSERTATION

Defence held on 28/06/2018 in Luxembourg

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN ÉCONOMIE

by

Thorsten DOHERR
Born on 23 May 1968 in Ludwigshafen (Germany)

DISAMBIGUATION OF RESEARCHER CAREERS: SHIFTING THE PERSPECTIVE FROM DOCUMENTS TO AUTHORS

Dissertation defence committee

Dr Katrin Hussinger, dissertation supervisor
Professor, Université du Luxembourg

Dr Nicolas Jonard, Chairman
Professor, Université du Luxembourg

Dr Christoph Schommer, Vice-chairman
Professor, Université du Luxembourg

Dr Dirk Czarnitzki
Professor, Catholic University of Leuven, Belgium

Dr Michele Pezzoni
Professor, Université Nice Sophia Antipolis, France

Acknowledgements

This dissertation would not have been possible without all the people commuting between Germany and Belgium causing regular congestions. Almost five years ago, Dirk Czarnitzki, a friend, co-author, mentor and most importantly one of these commuters, became an involuntary participant of such a traffic jam, providing him with enough time to contemplate. As he called me, still stuck in the traffic, to persuade me into doing my PhD, everything necessary was already arranged. Katrin Hussinger, Professor at the University of Luxembourg, has already agreed to become my promotor and Georg Licht, my boss at the Center for European Economic Research (ZEW), approved the notion that a member of his technical staff transformed into a doctoral student. It is understood that I had to yield. It is also worth mentioning that this was the best decision somebody else had made for me and I am very grateful to that trio for their collusion.

Being way beyond the typical age of a PhD student, the challenging experience I made in the last five years was rejuvenating thanks to the wonderful community of researchers I had the luck to become a part of. I thank Katrin Hussinger and Dirk Czarnitzki not only for this opportunity but also for promoting and mentoring me, giving direction through the depths of economic research as my co-authors, and providing the examples every "young" researcher should pursuit. I also would like to thank Nicolas Jonard for donating his precious time as a member of my dissertation supervisory committee.

The story above is prove of the administrative sorcery Georg Licht is capable of, always supporting his people without compromising the overall goals of the department. Talking about my peer group at the ZEW, my fellow doctoral students and post-docs are such an inspiring, open-minded, supportive and clever bunch that subsuming them as "good working environment" would be a grave insult. I have to thank Paul Hünermund for being a valiant defender of causality and having a critical eye on my lenient ways and Maikel Pellens for always giving profound comments to any of my premature ideas. I also want to thank Heidi Halder and Heidrun Förster for smoothly organizing everything administrative or bureaucratic intruding my comfort zone. This thesis would still be a convoluted mess without the help of Thomas Eckert, my lifelong roommate and document wizard.

A special thank goes to Riccardo Cappelli who not only invited me to a very productive research stay at the University of Bologna but also to his wonderful hometown Montefiore dell'Aso, combining knowledge transfer with la dolce vita. These acknowledgements would not be complete without mentioning my gratitude to Andrew Toole, an inspirational force sporting flabbergasting creativity. I owe Fabio Montobbio and Paula Schliessler a big thank you, for being efficient, friendly and competent co-authors.

Life does not only consist of work and research. I am grateful to my family and friends for their love, tolerance and patience. In particular, I thank my mother for her incessant encouragement, my girlfriend Alex for her healthy skepticism towards economics and my Dungeon & Dragons crew: Stumbl, Tom, Anette and Alex. You are so much more than a counter-balance to my work life.

Contents

List of Tables

List of Figures

Introduction

"Patents do not promote progress of science and technology, Inventors do."

Kalyan C. Kankanala

Conducting a topic search on the Web of Science bibliometric platform for the term "patent" returns 3,260 hits in the science category "economics" (February 2018). A considerable number emphasizing the importance of patents in this field. Although patents are a fundamental topic for economic research, the ones conceiving them, the inventors, still do not get the attention they deserve. I retrieved only 304 documents putting a focus on the term "inventor". A first assessment of this literature revealed that the majority of the empirical studies rely on relative small samples of survey data or concentrate on manually cleaned subsamples. Although the information of the inventors is attached to every patent record in publicly available databases, researchers refrain from shifting the perspective from documents to inventors because of the ambiguity of inventor names. The risk of linking documents of different homonymous individuals is too high to be neglected. Mistakenly linked document pairs - false positives - introduce a whole phalanx of identification issues, while missed links - false negatives - may suffocate any research question. Early efforts to disambiguate inventor careers in patent data supplemented the name by taking additional criteria like technology classes, last name frequencies and co-inventors into account (Singh, 2003; Jones, 2005; Fleming and Marx, 2006). Trajtenberg et al. (2004 and 2006) extended the feature selection with meta information aggregated from the patent data. Further approaches refined the catalogue of criteria and improved on the optimization of the parameters (Pezzoni, Lissoni and Taraskoni, 2012; Schön, Heinisch and Bünsdorf, 2014). All algorithms balance precision and recall by comparing the calculated output with the verified document links of a training dataset. The generation of these training datasets is a laborious exercise, based on either survey data or diligent assessment. The public availability of training data promoted machine learning methods, i.e. neural networks (Petrie, Julius and Thomson, 2017) and decision trees (Kim, Kabsa and Giles, 2016). Of course, training data is only applicable for the database it represents by providing labeled document links. Whether the resulting parametrization can be transferred onto other bibliometric databases, for example using the

parameters based on EPO data for USPTO patents, depends on compatibility of the data structure.

The prospect of relying on external training data or even creating this data by myself drove me into conceiving a method to disambiguate researcher careers without the need of such. Future projects may require the disambiguation of data not sporting a set of conveniently pre-linked careers. The existing datasets are criticized for being not exhaustive and swaying researchers in false security about the precision of their disambiguation methods (Magerman, 2015). My approach abandons the necessity of training data by replacing it with a theory for the namesake conundrum. The first chapter describes this theory and gives guidelines for the implementation. It further explains why most of the existing benchmark datasets are afflicted by an inherent namesake bias leading to the aforementioned overstated optimism regarding the precision rate of the applying disambiguation algorithms. The "disambiguation by namesake risk assessment" provides an unbiased tool which is frictionless deployable on any patent or bibliometric database containing information about inventors respectively authors, opening new branches of research questions pertaining researcher careers. By using documents and associated information as milestones in the careers of these individuals, which, after all, are the instigators of human progress, we can observe their mobility, the networks they create and the impact of their presence or absence. As bibliometric information always provide panel data, we are able to control for their proficiency and investigate their reaction to policy changes. Researchers are the immediate channel of knowledge transfer (Rahko, 2016). On a more macroscopic scale, their behavior can influence whole economies through brain gain and drain (Nathan, 2014; Docquier and Rapoport, 2012). In the second paper of my thesis, we explore the effect of inventor mobility on the productivity of Italian regions.

Skilled labor mobility between countries is well studied. It affects innovative capacity (Trajtenberg, 2005; Hoisl, 2007; Hunt and Gauthier-Loiselle, 2010), productivity growth (Pieri, 2012) and is considered a key device for knowledge spillovers (Almeida and Kogut 1999; Rosenkopf and Almeida, 2003; Song et al., 2003; Kerr, 2008; Breschi and Lissoni, 2009; Trippl, 2011). Although mobility within a country has lower barriers than international migration, potentially providing stronger effects, the inter-regional inflows and outflows of high skilled labor is barely represented in the literature. Italy, a country well known for its international

Diasporas, also shows high rates of mobility of its high skilled work force within the country, born of disparate opportunities on the job market with a clear decline from the north to the southern and central regions (Alma Laurea, 2016). Riccardo Cappelli, one of my co-authors, provides the anecdotic evidence to this situation. Together with him, Fabio Montobio and Dirk Czarnitzki, I applied the inventor mobility index on Italian regional data to identify the movements of inventors during their careers. In our paper "Inventor Mobility and Productivity in Italian Regions", constituting chapter 2, we explore whether inventor mobility, as a proxy for high skilled labor, affects differential regional total factor productivity. Mobile inventors may find jobs in the destination region, better suiting their skill sets, leading to higher efficiency through task specialization. They stimulate innovation in these regions by generating absorptive capacity (Miguélez and Moreno, 2015) and fostering knowledge diffusion (Trajtenberg, 2005; Giuri et al., 2007; Hoisl, 2007). The receiving regions benefit from increased entrepreneurial activity and from the convenience of filling labor shortages to sustain their productivity. We find that inflow of inventors leads to a significant growth of the TFP of the receiving region while outflow has a significantly negative impact for the sending region. The effect in mainly driven by inventors not only changing the region but also their employer (patent applicant). The receiving regions enjoy all the benefits of brain grain while the sending regions are exposed to the inevitable, unwanted sibling, the brain drain. Taking the job matching theory of Jovanovic (1979) into account, this leads to a vicious circle. By not being able to provide appropriate jobs for the residing high skilled individuals, mobility to regions delivering these kind of jobs is spurred, further withdrawing the capabilities of creating just these jobs in the sending regions.

Mobility is a prime example of the applicability of output oriented career information on economic research questions. Further enriching the information by linking the affiliations to firm level data and intersecting the inventor career path with the publishing author career path facilitates even deeper insights into the activities and motivations of researchers. Together with my co-authors Dirk Czarnitzki, Katrin Hussinger, Paula Schliessler and Andrew Toole, I exploited an exogenous change in German federal law to observe the reactions of university researchers to policies directly pertaining their relationships to firms. The third chapter "Knowledge Creates Markets: The Influence of Entrepreneurial Support an Patents Rights on Academic Entrepreneurship" accompanies a reform in 2002 called "Knowledge

Creates Markets" which mimicked the Bayh-Dole Act the United States issued 1980 facilitating institutional ownership of inventions conceived by researchers supported by federal funds. In Germany, the law replaced the "professor's privilege" which granted the University researchers full control on how to exploit their inventions, be it by renunciation of property rights through publication, founding start-up companies, direct collaboration with private firms or licensing of patents. Under the new regime, the university has the priority claim on all inventions of their employees. Any commercialization of claimed patents is channeled through technology transfer offices (TTO) with a strong focus on academic start-ups fostered by financial and management support. The literature is criticizing TTOs for overemphasizing revenue maximization impairing their function as intermediaries for technology transfer (Litan et al., 2007), being too bureaucratic, informational limited and providing misaligned incentives (Kenny and Patton, 2009). Additionally, they fail their main task of spurring academics to start new companies (Clarysse et al., 2007). Patents, on the other hand, are uniformly considered a key driver of academic start-ups by increasing initial funds (Clarysse et al., 2007) and providing leverage as signaling device on the financial scope of the venture (Conti et al., 2013; Haeussler and Colyvas, 2011; Graham et al., 2009; Audretsch et al., 2013). To evaluate the introduction of TTOs on academic start-ups, we observe the careers of German university professors together with a control group of researchers of public research organizations, e.g. Max-Plank or Fraunhofer institutes, not affected by the law change. Our difference in differences framework confirms the importance of patents for academic start-ups and shows that TTOs, despite their bad reputation, are operational but do not improve upon the "professor's privilege". The overall patent output of the researchers experiencing the reform is reduced, implicitly curtailing the fuel for academic start-ups. In particular, the collaborative patents between faculties and firms suffered a strong decline. These patents constitute an informal and, from an institutional viewpoint, effortless channel for knowledge transfer, which is substituted by elaborate mediation of TTOs, leading us back to the criticism of their efficiency.

Our research emphasizes the fragility of the status quo pertaining researchers and inventors. They belong to a highly mobile group with excellent job market prospects contributing significantly to the economic development of their destination regions. Settled researchers react highly sensitive to changes to their working conditions especially if an increase of

bureaucracy is involved. Usually policymakers perceive this group of individuals mostly through statistics of their output as this is the prevailing form of representation in the literature neglecting the finer nuances a career centric vantage point could provide. This thesis illustrates the utility of a large-scale disambiguation effort for economic research. Although, there is research regarding the career of these protagonists, it is in most cases based on small, specialized pockets within the huge ecosphere of science. A generalized disambiguation approach allows tackling broader research questions without the limitations imposed by selection.

1 Disambiguation by Namesake Risk Assessment

Thorsten Doherr[a,b]

a) Centre for European Economic Research (ZEW), Mannheim, Germany
b) University of Luxembourg

Abstract Most bibliometric databases only provide names as the handle to their careers leading to the issue of namesakes. We introduce a universal method to assess the risk of linking documents of different individuals sharing the same name with the goal of collecting the documents into personalized clusters. A theoretical setup for the probability of drawing a namesake depending on the number of namesakes in the population and the size of the observed unit replaces the need for training datasets, thereby avoiding a namesake bias caused by the inherent underestimation of namesakes in training/benchmark data. A Poisson model based on a master sample of unambiguously identified individuals estimates the main component, the number of namesakes for any given name. To implement the algorithm, we reduce the complexity in the data by resolving similarity in properties through cascaded traversal. At the core of the implementation is a mechanism returning the unit size of the intersected mutual properties linking two documents. Because of the high computational demands of this mechanism, it is a necessity to discuss means to optimize the procedure.

Keywords: homonymy, namesakes, disambiguation, scientific careers, inventors, patents, publications

JEL: C18, C36

1.1 Introduction

Bibliographic and patent databases have comparable structures as they both are collections of documents cataloged by fixed retrieval criteria, like author or inventor name, title, abstract, academic discipline or international patent classifications, keywords, filing and publications dates, citations, affiliations respectively applicants and so on. Patent offices, public institutions and even private providers foster the access to this data to extend the knowledge about the treasures they hoard through academic research. The ubiquity of the data facilitates new research approaches especially in the fields of innovation economics and bibliometric analysis. It does not take long until researchers were not content by only exploiting the document, i.e. patent or publication, as observation unit. Linking the documents to other sources, for instance firm panels, extends the utility for researchers to deepen the insights into the mechanisms of innovation and research. However, the actual protagonists of these mechanisms, the authors and inventors, are in most cases not the focus of these efforts. The fuzziness of names as the main handle to their careers requires disproportionately complex identification strategies. Nevertheless, these individuals are the main driver of human progress and deserve close inspection.

The best solution to the issue would be the assignment of a unique author identification number (UAIN) to every author or inventor retrievable from every document or patent published (Fallgas, 2016). An implementation of a mandatory author identifier only exists for Brazil, the Netherlands and for some selected research fields (Fenner, 2010). As far as we now, no patent authority has introduced a mandatory identifier for inventors. Other efforts, like the ORCID (Open Researcher and Contributor ID), target specifically large publication data providers like Web of Science or Scopus, which are more open to the needs of their prime audience, the researchers. Patent authorities are less inclined to support researchers because their assignment is the administration of legal documents. Their key audience consists of lawyers, patent assignees, firms and other patent authorities. Even though some institutions already apply administrative methods to identify authors, this only covers documents filed under the new regime. Older documents remain unchanged and the associated author careers ambiguous. This situation forced researchers to implement their own disambiguation methods. An early effort by Singh (2003) relies on a combination of name and patent subcategory match, while Jones (2005) and Fleming and Marx (2006) concentrate mainly on

the names, latter taking the frequencies of the last names and mutual co-inventors into account. Trajtenberg, Shiff and Melamed (2004 and 2006) inaugurate the “Names Game” season, introducing a score based method with matching parameters fine-tuned by using a dataset of manually disambiguated Israeli inventors. The paper already acknowledges the effect of large assignees or cities in conjunction with common names on the probability of causing false positives. The size of an assignee or city and the commonness of an inventor name is measured by the number of patents that share this specific unit. A link between two documents by these criterions is regarded weaker for high patent counts. These frequencies are part of the parameters, determining the strength of a document link, to be weighted by the iterative fine-tuning process. Trajtenberg et al. (2006) also discusses the existence of an intransitivity “conundrum”: document A can be linked to document B with a high probability. The same is valid for B and C, but A and C do not match. The authors decide to impose transitivity in such cases stating this as the only plausible action, even though they consider this not an “innocent decision”.

In another approach, Torvik and Smalheiser (2009) apply the “Author-ity” model to 15.3 million articles in the MEDLINE database. It requires training data consisting of a match set of pairs of articles with a high probability being from the same author and a non-match set of document pairs of obviously different authors. For a given document pair a similarity profile is computed and compared with both training sets, returning the respective relative frequency of the profile within each set. The ratio of both values is the so-called r-value. The main formula to estimate the pairwise probability of a valid match incorporates, next to the r-value, the document count per name as a priori match probability. Although the authors criticize the inaccuracy of this proxy, they consider it a reasonable heuristic value for the following steps. Given the Bayesian approach, tolerating intransitivity is not an option and therefore solved by iterative smoothing of triplet violations. The final clustering of the comprising tuples relies on a maximum likelihood framework.

Pezzoni, Lissoni and Tarasconi (2012) construct a list of 17 matching criteria for their “Massacrator” algorithm. Some of these include meta information based on aggregated data. They classify an applicant as small, if less than 50 inventors are affiliated. The paper omits the method of the size estimation. We insinuate an aggregation on the inventor name level by applicant as the most obvious procedure. They identity a rare surname by counting its

occurrence by patents within the inventor's country. The criteria are weighted by a Monte Carlo simulation balancing recall and precision measured by a training dataset consisting of the "Noise Added French Academic" (NAFA) and the "Noise Added EPFL" (NAE) benchmark data promoted by the "Name Game" algorithm challenge of the APE-INV (Academic Patenting in Europe) initiative of the European Science Foundation (Lissoni, 2010). Schön, Heinisch and Bünsdorf (2014) apply a similar method based on less matching criteria, called classifiers. A classifier differentiates between matching, non-matching and missing patent properties implementing an additional degree of freedom. The classifier patterns returning the highest accuracy is determined by testing them against a set of manually matched document pairs, enjoying a high confidence of being from the same inventor, and a randomly matched negative set. The team obviously found a way to identify "common surnames" as a classifier, but did not elaborate on how that was conducted.

Other approaches are deeply rooted in the realm of machine learning. Therefore, they always require a training dataset of already classified documents. Kim, Kabsa and Giles (2016) use pairs of patents from the same inventor and pairs from different inventors to train a random forest classifier to produce a decision tree on the matching criteria. The output of this tree is the distance between documents. By applying a DBSCAN clustering algorithm, they resolve the resulting document network. Petrie, Julius and Thomson (2017) use a similar training set consisting of matching and non-matching document tuples to train a neural network called AlexNet (Krizhevsky, 2017) specialized in image classification. To feed the network they converted the document data into a graphical representation based on color coding and 2D mapping.

The literature so far does not provide deeper insights of the main culprit inciting all these efforts, the namesake. A namesake is "someone or something that has the same name as another person or thing" (Merriam-Webster's Learner's Dictionary). This issue is generally circumvented by methods optimizing parameters and thresholds for sets of matching criteria using training data. In the ideal case, that data is based on real word observation of authors or inventors like the surveyed samples of French and Swiss inventors of the NAFA and NAE benchmarks, or the self-assigned ORCHID. As the availability of those convenient datasets is the exception, researchers have to fall back on classification of sample documents based on intuitive assessment of whether a document is from the same individual or from two

namesakes. This article introduces a theoretical model for the probability of encountering namesakes simulating the intuitive namesake risk assessment and replacing the necessity of training data.

Some names are comparable with unique identifiers without any risk of encountering namesakes, while other, more common, names involve a high risk of linking documents from namesakes. One challenge is to find a way to differentiate between unique and common names to assess the risk of linking careers of namesakes. For any given name, this risk increases with the number of namesakes in the population, but it also depends on the size of the observed reference unit. Although a person may have a very common name, it has a low risk of encountering a namesake, if the reference is a small firm. A reference unit is not limited to physical stations like affiliations accrued during an author's or inventor's career, but can also be a research area, a technology field, a co-author network, special interests manifested by citations, keywords, repeating topics in titles or even a combination of multiple contexts. In this paper, we discuss the theory of namesake risk assessment, a method to estimate the number of namesakes, the identification of unit sizes and the implementation of a universal disambiguation algorithm based on this knowledge. We further discuss the inherent underrepresentation of namesakes in training/benchmark data inevitably resulting in a namesake bias. The paper concludes with application examples omitting benchmarks for reasons explained in the preceding chapter.

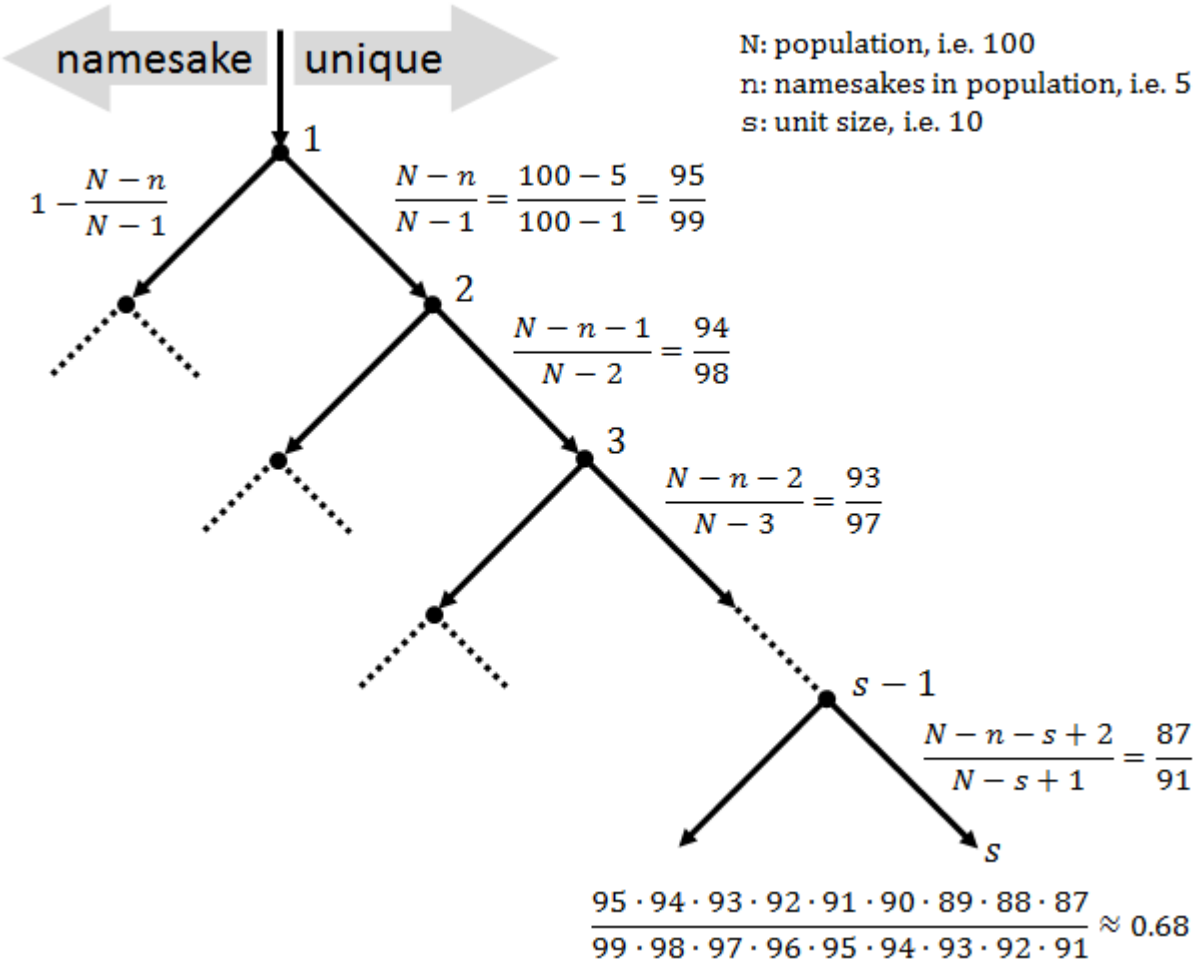
1.2 Namesakes

1.2.1 Likelihood of drawing a namesake

To explain the theory of namesakes, it is helpful to picture a concrete example for a reference unit: a firm. For our analogy, we assume that our firm has only one employee at the beginning, the founder. The firm employs more and more individuals from a finite pool, until it reaches its current size. With every new entrant, the risk for a namesake to the founder increases. The extreme case of employing/drawing the complete population dictates this interrelation. As we are only interested in the risk of drawing any namesakes, it is sufficient to handle the reverse case of drawing no namesakes. The probability of drawing a valid employee equals the remaining number of individuals in the population, which are no namesakes to the founder, divided by the remaining population size. With every new employee, the numerator

and the denominator of this relation decrease by one. Figure 1.1 illustrates the development of the probability of the founder staying unique.

Figure 1.1: Probability tree for drawing a namesake vs. staying unique



In our example, the population N consists of 100 individuals. The number of namesakes n in the population is 5, including the founder. The final unit size, respectively firm size, s is 10. The probability of drawing a namesake with the first employee is $1 - 95/99 \approx 0{,}0404$. The probability of drawing a namesake to the founder with the second employee is only slightly larger: $1 - 94/97 \approx 0{,}0408$. The parallel decrement of the numerator and the denominator has only a very small impact on the probabilities for larger populations, a circumstance we will exploit later on. The product of the probabilities of the unique branch is 68%, therefore is the likelihood for at least one other person in the firm with the same name as the founder 32%.

The most straightforward implementation of the probability of drawing a namesake for an individual with n namesakes within a population of N individuals for a unit with size s is 1 minus the product of all stepwise probabilities:

$$P(namesake) = 1 - \prod_{i=0}^{s-2} \frac{N-n-i}{N-1-i} \tag{1.1}$$

The number of operations to calculate the probability depends on the unit size s, making (1.1) an unwieldy proposition for large units. Figure 1.1 already hints a way for a formula that is independent from the unit size. The term at the bottom shows a division of product sequences, which can be constructed by using factorials:

$$P(namesake) = 1 - \frac{(N-n)! - (N-n-s+1)!}{(N-1)! - (N-s)!} \tag{1.2}$$

Of course, factorials are even more cumbersome as they grow very fast beyond the capabilities of contemporary computing systems. For example, the factorial of 171 is already too large to be properly represented by the numerical data type with the highest precision used by statistical software packages (double, 8 bytes). It is suggested to use the natural log of the factorials approximated by James Stirling's formula published 1730:

$$lnf(x) = \left(x + \frac{1}{2}\right) ln(x+1) - (x+1) + \frac{1}{2} ln(2\pi) \tag{1.3}$$

With the support of this handy approximation, it is possible to rewrite (1.3) avoiding factorials altogether:

$$P(namesakes) = 1 - exp\big(lnf(N-n) - lnf(N-n-s+1) - lnf(N-1) + lnf(N-s)\big) \tag{1.4}$$

Because the number of operations to calculate the probability stays always constant, we use the final form (1.4) in our implementation of the disambiguation algorithm.

To fill this theory with life, we need to determine the number of namesakes for any given name in a population. Of course, we do not have the luxury of name repositories for any occasion. The following section discusses a method to estimate the number of namesakes

based on a representative sample and a highly correlated indicator, which fulfills the requirement of being derivable for any name population.

1.2.2 The indicator

In most cultures, the last name is inherited from the parents and there are only few instances where it may be subject to change, i.e. marriage. Parents that bear a common last name may choose an exotic first name for their child for it to stand out more. They may choose to name their child according to their family tradition, i.e. first name of a grandparent and thus explicitly creating a namesake. The density of namesakes also depends on temporary trends that influence naming decisions. A common last name generates a high number of variants in terms of first names. For a common first name, we naturally observe a high amount of different last names in the population. If we pick an uncommon first or last name, we expect less variation within the other part of the name. A look in an old-fashioned telephone book reveals that the combination of common last names with common first names is responsible for most namesake occurrences. In fact, such an entry can only be saved from having a namesake by a less common first name. This trivial observation leads us to the assumption, that the number of namesakes is positively correlated with the occurrence of the part of a name that appears less often in the population.

To define the indicator, we aggregate a population on the name level by removing duplicate entries. This harmonizes every population containing names, regardless of the original context of an observation, to a name aggregate. For every name, a minimum occurrence is calculated by counting the frequencies of every name part in the name aggregate and choosing the respective minimum. After sorting the name aggregate by the minimum occurrences in ascending order, we apply a dense ranking ("1223444...") and a final normalization to the range]0,1]. The intention of the normalized minimum occurrence rank, further called *minocc*, is the harmonization of the distributions of different name aggregates by obfuscating the frequencies. The *minocc* of a common name is close to 1 whereas the minocc of a unique name is not far from zero.

1.2.3 The master sample

To estimate the number of namesakes, we need a representative master sample containing unambiguously defined individuals along with their names. Our master sample is the

stakeholder database of the German credit rating agency "creditreform". It contains 6731543 owners, managers and major shareholders of almost all German firms over a period of 15 years. It also includes a large proportion of German micro firms. Therefore, we do not expect a bias towards specific ethnical groups, which would be the case, if only larger firms were in the sample. From a demographic point of view, the master sample is not representative. For instance, only 27% of the stakeholders are female. However, we consider the data a healthy sample in regard of its representation of the variation of names.

The names are cleaned from additional clutter like academic titles and other not birth name related appendages, converted to upper case and special letters, like the German "ß" or French “âcçènts”, are replaced with the most common alphabetical letter representation. Potential target data needs to be prepared in a similar way. Because the name format greatly influences the distribution of namesakes over the minimum occurrence rank and access to the master sample, to reiterate the estimation, should not be a requirement for a disambiguation run, we conducted our analysis on the three most common name formats encountered in patent and bibliographic data:

- **Format A: last name, first name**
  Last and first name are or can be separated into two distinct fields.
  minocc is based on the minimum occurrence of the last or first name in population.

- **Format B: last name, initials**
  First names are represented by starting letters only.
  minocc is based on the number of initials per last name in the population.

- **Format C: unordered name**
  Last and first name are in one field without specific order.
  minocc is based on the word of a name with the lowest occurrence in the population.

In case of format C, we have to handle a methodical weakness of this specific name representation. As there is no way to distinct between last and first name, there exists a group of outliers with a high *minocc* and relatively low number of namesakes because they have a common first name as the last name, e.g. Maria Peter. We eliminate this group from the master sample by identifying the percentile of the most common first names and removing all observations with these first names as last name. We lose 145,587 names (345,576 persons) of the aggregated master sample. For these cases, the predicted number of namesakes will be consistently overestimated, but keeping them in the sample would lead to a general

underestimation. As format A and format C both provide the non-truncated name information, format A should always be the preference if applicable. The following table shows the effect of the different name formats on the name aggregation, minimum occurrence and the number of namesakes per name:

Table 1.1: Master sample by name formats

			Minimum occurrence				Namesakes per name			
F	People	Names	min	max	E	sd	min	max	E	sd
A	6,731,543	4,691,779	1	7,975	105.2	230.4	1	1,118	1.43	3.56
B	6,730,633	2,544,481	1	1,264	32.2	77.5	1	5,366	2.65	15.37
C	6,385,967	4,546,192	1	27,150	192.2	570.3	1	1,035	1.40	3.35
C*	6,731,543	4,691,779	1	98,363	279.3	1264.5	1	1,118	1.43	3.56

*including outliers

1.2.4 Predictive model

Our predictive model consists of a weighted Poisson regression of the number of namesakes on a polynomial of the 5th degree of *minocc*. The model takes the following form:

$$namesakes = exp\left(\beta_0 + \beta_1 minocc + \beta_2 minocc^2 + \beta_3 minocc^3 + \cdots + \beta_5 minocc^5\right) + \varepsilon \quad [pweight: namesakes] \tag{1.5}$$

The population weight is the number of namesakes itself, as every observation in the aggregated data is a name representing namesake persons in the master sample. Using an unweighted model would underestimate the namesakes, because the aggregation inflates the relation of unique names against namesake-afflicted names. The model is equivalent to an unweighted regression of the non-aggregated data, where every observation unit is an individual represented by its number of namesakes and a name with standard errors clustered by names. As we also want to show the difference to an unweighted model, we adhere to the name-based version. Nevertheless, we do expect a proper estimator for namesake predictions of individuals represented by their names. This estimator will overestimate the population size, if we accumulate all predicted namesakes for every name in the aggregated population. An issue we need to address before we can calculate unit sizes.

Because the interpretation of high-degree polynomials is not very intuitive and because of the univariate design of the model, we discuss the results of the regressions based on figures showing scatter plots of namesakes on *minocc* overlaid by predicted namesakes (weighted and unweighted) for all three formats. We present the actual regression results in the Appendix, Table 1.4.

Figure 1.2: Predicted namesakes (weighted vs. unweighted) on scatter plot: Format A

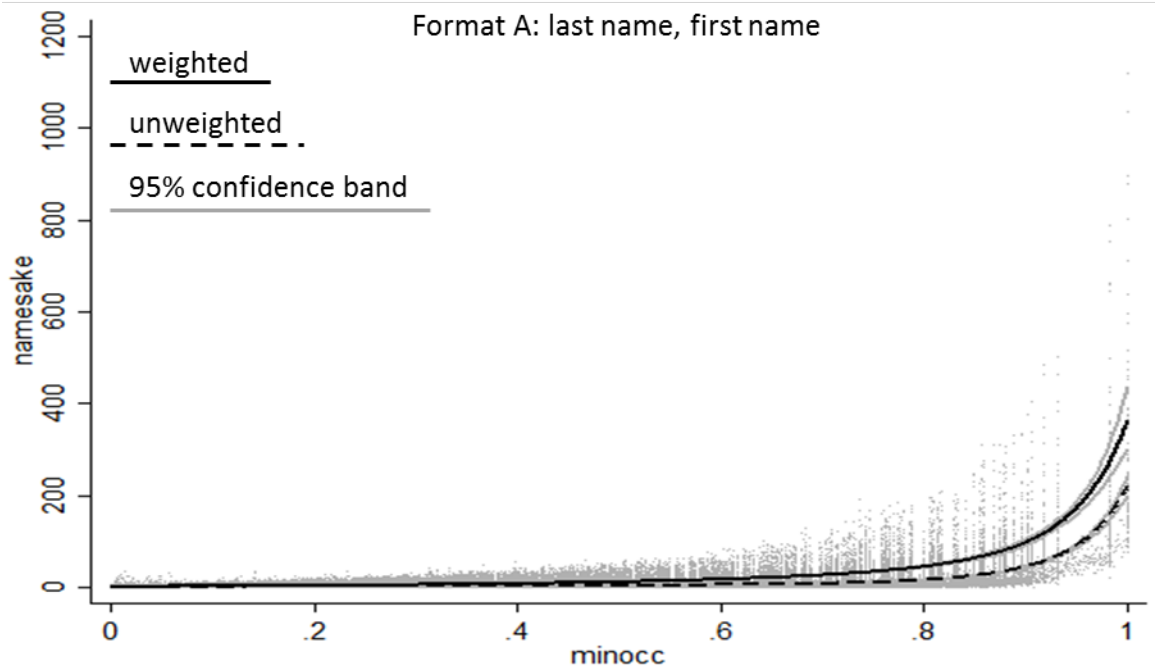


Figure 1.3: Predicted namesakes (weighted vs. unweighted) on scatter plot: Format B

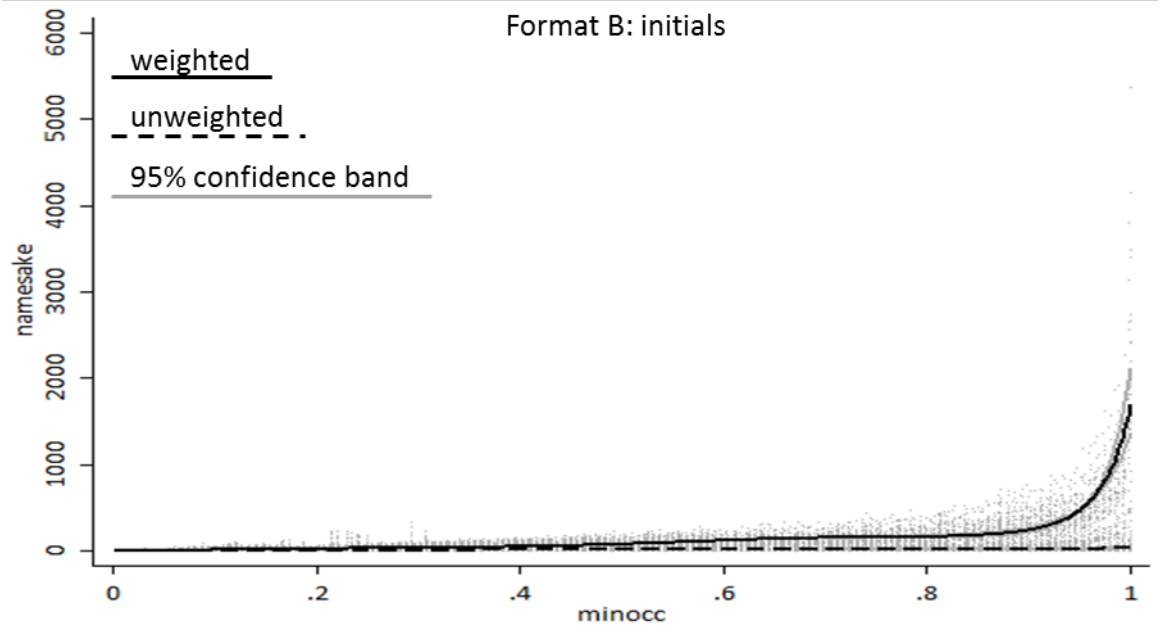
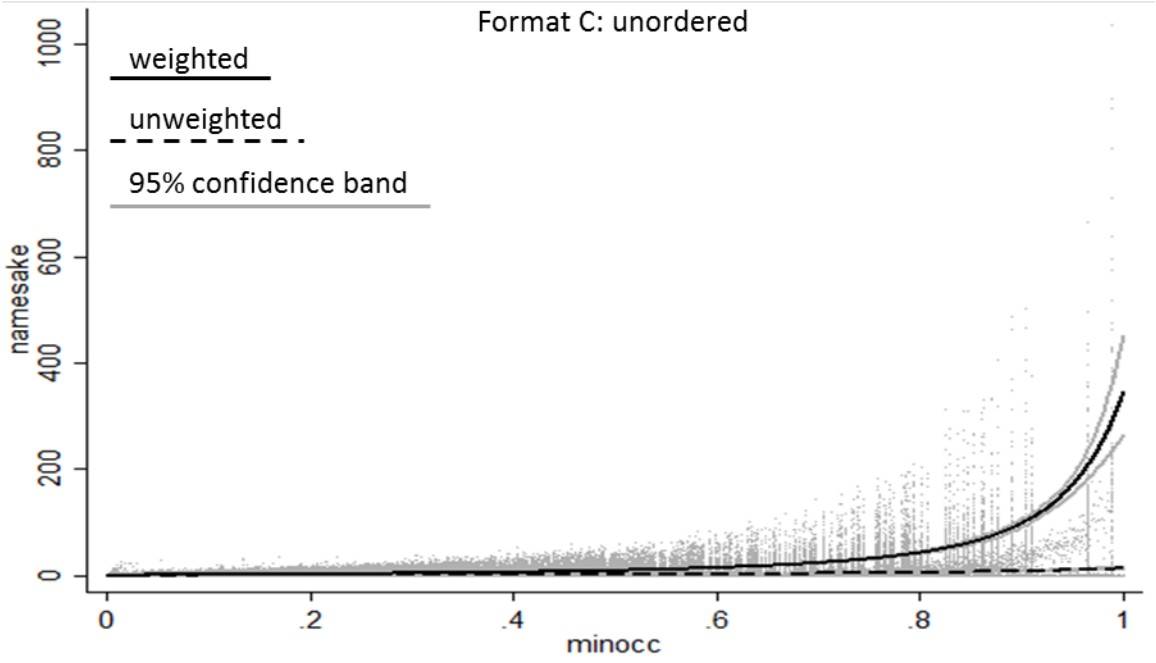


Figure 1.4: Predicted namesakes (weighted vs. unweighted) on scatter plot: Format C



The graph for format A already shows an unweighted curve that follows the rising of namesakes with higher values of *minocc*. Both variables are positively correlated. We observe only a negligible number of outliers with a high *minocc* but a low number of namesakes. The other formats display a much larger discrepancy between the two curves. The unweighted prediction is forced to the bottom on the right side of the graph by a relatively high number of names seemingly violating our key assumption. However, the weighted curve tells us, that rare names have a lower individual support, i.e. namesakes, than common names. Even though the R^2 for format B is the highest of all formats, we consider the predictive power being the lowest, because of the high information loss by using initials instead of first names. In any case, the risk of encountering a namesake-afflicted individual is larger for higher values of *minocc*. Figure 1.8 in the Appendix shows the graph for the unmodified format C (including outliers).

1.2.5 Adjustment to target populations

Our prediction of namesakes is based on the master sample and scaled in relation to this specific population. The target data may be considerably larger or smaller than the master sample, raising the question, if the size of the population N and the estimated number of namesake n has to be adjusted accordingly. Even if such an adjustment can easily be applied

by a scaling factor f on both parameters, we would need the population size of the target data, which is unknown. As a work around, we could use the ratio of the aggregated name counts of both populations as a proxy for f. The fact, that the scaled namesake numbers have to be truncated if they fall below 1, suggests that the rescaling is not properly defined by a constant factor. Simulations with samples from the master data show a statistically, but not practically, significant difference between the namesake distribution of the samples and their respective rescaled counterparts, so using a constant factor is still adequate. We also observe that the potential factor based on name aggregates consistently overestimates the actual ratio of the populations sizes used for the simulations. We explain this by the higher density of the name aggregate, which leads to a disproportionate draw of names. In the case of format B, with the highest name density of all formats, a sample may contain 25% of the individuals but already 36% of all names. Before we delve too deep into this issue, we prefer to prove that an adjustment to different population sizes in not required.

We start with the straightforward implementation of the namesake probability based on the products of the stepwise probabilities (1.1). The parallel decrement of the numerator and denominator has only a minimal impact on the result. It is only relevant for already small values, which only occur for s or n being of the same magnitude as N, a highly unlikely situation. Therefore, we approximate the namesake probability with

$$P(namesake) \approx 1 - \left(\frac{N-n}{N}\right)^{s-1} \tag{1.6}$$

by replacing the stepwise probabilities with a constant probability. When we adjust our approximation by plugging in a scaling factor f for N and n we get

$$P_f(namesake) \approx 1 - \left(\frac{fN-fn}{fN}\right)^{s-1} = 1 - \left(\frac{N-n}{N}\right)^{s-1} \tag{1.7}$$

and witness the elimination of the scaling factor. This relieves us from the issue of adjustment. We can always pretend that the target population is of the same size as the sample population without the apprehension of consequences.

1.2.6 Unit sizes

To assess the risk of a namesake, we need the number of namesakes n in the population N and the unit size s. The parameter n can be estimated using the normalized minimum occurrence based on the target data. The estimate is scaled in relation to the population size N of the master sample needing no adjustment as shown in the previous section. A preliminary proxy $\check{s}$ is defined by the size of the name aggregate of the unit, for instance, the number of distinct inventor names appearing for a specific applicant in the patent documents. We call this set of names U. The number of names in U is always underestimating the real unit size because of namesakes. We receive the augmented proxy $\hat{s}$ by accumulating the estimated number of namesakes in U in respect to a unit with size $\check{s}$:

$$\hat{s} = \check{s} + \delta(\check{s} - 1) \sum_{i \epsilon U} 1 - \frac{N - \hat{n}_i}{N - 1} \tag{1.8}$$

The probabilities of drawing a namesake on the first step in the probability tree (see Figure 1.1) for all involved names are summarized and multiplied by the number of steps required to reach unit size $\check{s}$. The size $\hat{s}$ is the number of expected additional individuals because of namesakes plus the number of names $\check{s}$, as every name in U already represents at least one individual. We already know, that by using the predicted number of namesakes $\hat{n}$, we will overestimate the population on the name level. This positive bias is challenged by the negative of using the name count $\check{s}$ as a lower-bound proxy for s. In addition, we use the probability of the first step although there is a very slight incremental shift in the probability of drawing a namesake by going further down the tree. Finally, we have to take the fragmentation of the population into account. For small units the proxy $\check{s}$ is closer to the real size than for large units, because the probability of encountering namesake-afflicted names increases with the size and therefore the difference to the name aggregate. To balance these contradictory effects, we introduce the parameter δ, which is retrieved from a Monte Carlo experiment simulating a fragmented population.

First, we separate the randomly sorted master sample into virtual units with the following equally distributed and randomly chosen sizes: 10, 20, 100, 500, 1000, 2500, 5000, 10000, 50000 and 200000. We aggregate the data on the name and unit level, keeping the actual size s as a reference. We repeat this process 15 times, appending the resulting data to get a

virtual, fragmented population. For every unit, we calculate the improved proxy for the unit size $\hat{s}$ without the balancing coefficient ($\delta = 1$) using equation (1.8). After aggregating the data on unit level, we regress the real number of additional individuals by namesakes on the estimated number:

$$s - \check{s} = \delta(\hat{s} - \check{s}) \tag{1.9}$$

We omit the intercept to force the slope through the origin. The level of fragmentation is an arbitrary choice mimicking the natural separation of a population into units. The actual fragmentation of a population depends on the context. The context "applicant" generates a different fragmentation than the context "technological classification". As we also have to consider combinations of contexts, which create additional layers of fragmentation, we decided to tackle this issue by a generous mixture of unit sizes.

Table 1.2 shows the improvement of the preliminary proxy $\check{s}$ to the balanced estimate $\hat{s}$ by applying equation (1.8) based on the virtual population derived from the master sample for format A and format B. The need to include the estimated namesakes increases with the density of the name aggregate. Format B has a higher density, meaning less name variation, whereby the degree of the bias for the unmodified proxy is exacerbated compared to format A. For the latter format, the additional effort shows a smaller improvement, but it is still justified by the robustness gain against outliers in regard of the unit composition, i.e. units with a high share of common or rare names. Further, the upper half of the table alludes to the issue that the share of namesakes within a unit is not a linear function of the unit size, a circumstance leading to the namesake bias introduced by unbalanced training respectively benchmarking data. In section 1.4, we explain why this bias is almost unavoidable but, fortunately, not affecting our disambiguation approach.

Table 1.2: Estimated unit sizes for format A and format B

	last name, first name					last name, initials				
s	**N**	**min**	**max**	$\boldsymbol{E(\check{s})}$	$\boldsymbol{\sigma}$	**N**	**min**	**max**	$\boldsymbol{E(\check{s})}$	$\boldsymbol{\sigma}$
20	285	20	20	20.00	0.00	271	19	20	19.99	0.09
100	263	99	100	100.00	0.06	248	99	100	99.93	0.26
1000	275	995	1000	999.32	0.87	288	986	1000	993.53	2.71
50000	304	48696	48896	48791	36.64	258	43816	44200	43999	74.20
200000	259	187175	187801	187485	124.0	282	153909	155012	154469	186.0
	$\boldsymbol{\delta = 0.487103\ (0.00031)}$					$\boldsymbol{\delta = 0.444869\ (0.00438)}$				
s	**N**	**min**	**max**	$\boldsymbol{E(\hat{s})}$	$\boldsymbol{\sigma}$	**N**	**min**	**max**	$\boldsymbol{E(\hat{s})}$	$\boldsymbol{\sigma}$
20	285	20	20	20.00	0.00	271	19	20	20.00	0.09
100	263	99	100	100.00	0.06	248	99	100	100.00	0.26
1000	275	996	1001	1000.00	0.87	288	991	1006	999.13	2.74
50000	304	49889	50090	49987	38.31	258	49176	49800	49423	105.9
200000	259	199631	200330	200007	131.3	282	199131	201966	200365	469.8

Note: cluster robust standard errors in parentheses

1.3 Implementation

The representation of documents like patents or scientific publications in bibliographic databases accessible to researchers usually does not include the full document itself. It concentrates mainly on bibliometric properties to support the retrieval of documents by their authors or inventors, affiliations, locations of the aforementioned, keywords and topics, titles, classifications, journals, date of publishing and so on. To identify the documents of a specific person, one would first search for the name of the person. If the name is exotic, the result of the search already documents the career of the person. For a common name, it is required to supplement the search with additional information about the person, for example the name of a co-author or an affiliation. Whether the found documents belong to the person of interest or are from a namesake depends on the commonness of the name and the identification potential of the additional information. If the co-author has an exotic name or the affiliation is only a small company, the likelihood of getting the wrong person is small. Obviously, the identification potential corresponds with the perceived size of the search criteria relating to

the peer group of authors or inventors. The search results of the first step reveal new document properties to be included in further-reaching queries, leading to new documents to be subject of namesake risk assessment and, again, the retrieval of new search criteria. A good depiction of this recursive procedure is a network analogy, where the documents are nodes connected by mutual properties. The searcher traverses along the edges from node to node, collecting all touched nodes into a cluster list. The accessibility of an edge depends on whether the risk of connecting documents of namesakes is below a general threshold. To assess the risk, the searcher has to estimate the number of namesakes for the given name and the unit size. The latter is determined by intersecting the peer groups of the connecting mutual properties. The disambiguation algorithm separates the network, spread out by mutual properties of documents sharing a specific name, into clusters with a low risk of containing the work of namesakes.

1.3.1 Reducing complexity

We separate document properties into two different kinds. *Hard properties* have no variation in regard of the entity they designate. Categorical memberships of documents like standardized technological classifications, research field categories or unique identifiers like cited patent numbers are hard properties. We consider properties whose variation can be eliminated by trivial cleaning procedures, like removing non-numerical characters or transferring all characters to upper case, as hard. *Soft properties* have no trivial to eliminate variation in regard of the entity they designate. They require the usage of the adjective “similar” to describe their relation, e.g. similar inventor name, similar affiliation, similar applicant, similar title and so on. There is a high variation in the portrayal of the same entity in bibliographic or patent data because the focus is the proper representation of documents but not the administration of specific databases to harmonize inventor names, affiliations, addresses and so on. Besides the identification of entities within the data, we are also interested in detecting similarity in descriptive properties like titles or keyword lists sharing a specific topic. For these cases, the topic is the entity. The goal is to identify cluster ids for variants of the same entity for all properties. We further call these cluster ids *traits* of a document.

First, we compress the data of all properties into respective versions without duplicates. For hard properties, the sequence number of the compressed data is already the cluster id. For

soft properties, the compressed table is the source and the target for a self-referential search algorithm to identify the similarities between the entries. For every entry, the algorithm selects potential candidate entries using meta information about the frequencies of words within the data, retrieved from the data source itself. Every word of the search entry is weighted by the inverse of its respective frequency retrieved from the meta data. The algorithm perceives anything separated by blanks as a word. Internal preparation routines guarantee a general harmonization level (upper case, replacement of special characters and so on) and optionally implement linguistic methods like Soundex, Metaphone or n-grams to improve the robustness against misspellings. Common words, like legal forms or frequent phrases, get low weights compared to more identifying words. The algorithm further separates the meta data according to the originating source field to avoid the blending of frequencies of different contexts. A common street name does not swamp the frequencies of the applicant name field where the same word appears less often. Superordinate weights on these contexts allow for extensive control over the search and the measurement, i.e. putting 70% on the applicant name and 30% on the address with a threshold of 90% enforces the requirement of partial similarity of the address even if the name matches perfectly. The share of the weights, the joint words accumulate, measures the quality of a candidate.

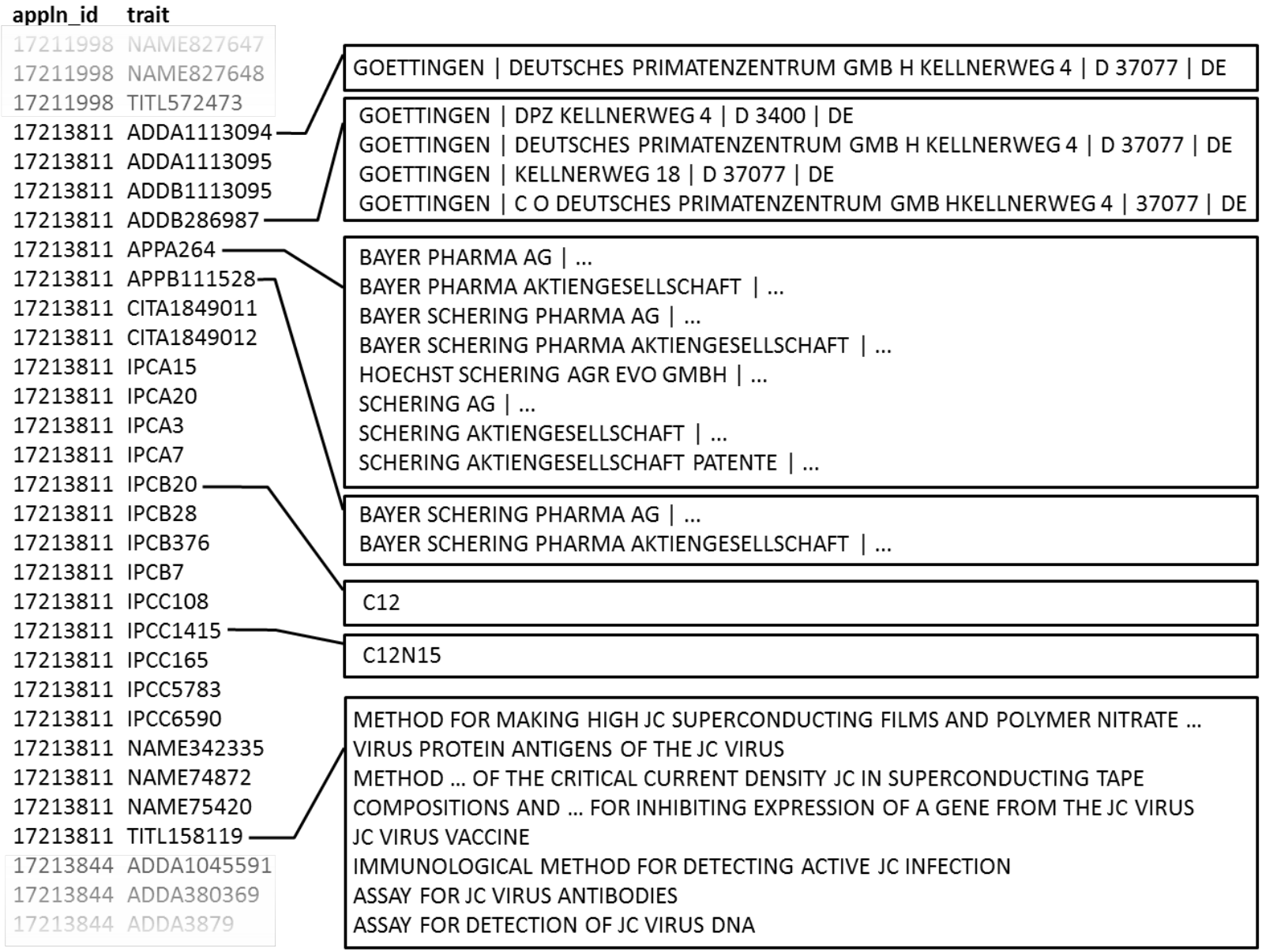
The result of the self-referential search is a list of matches consisting of all distinct property entries, the respective candidates and their similarity scores, which are greater equal a high threshold. This list is neither commutative nor transitive allowing the following cases: A matches B but B does not match A; A matches B and B matches C but C and A do not match. If the list would have been transitive, we could already designate the cluster ids by simply choosing the minimum or maximum entry id per candidate. This method is not applicable to our result, which needs recursive traversal following the intransitive links through an implicitly constructed network to let the inherent clusters emerge. On first sight, this seems to be a disadvantage to methods producing or enforcing transitive results, but the additional freedom creates flexibility. For example, a firm group name is matched with several subsidiaries containing the mother's name as part of a much longer specification. Because of the additional clutter in the names, the subsidiaries are in most cases not matched with the mother. Some subsidiaries may be joint ventures connecting to a different group of firms spreading out the network. Even links between different historical versions of the same firm name, including

mergers, can be detected, as long as there are still some overlaps. This is especially important as bibliographic and patent data notoriously contain historical information. On the other hand, this behavior also creates completely unrelated connections, usually if matches with a low identification potential are involved. Traversing these networks without further precautions will lead to unexpectedly large clusters containing mostly unrelated entities. We call the method to handle this issue *nested cascaded traversal*. In short, it introduces a sequence of arbitrary size limiters combined with incremental conditions on the match quality, both based on experience with the data and educated guesses. During traversal, every time the cluster size exceeds a cascade limit, the associated rule set activates, enforcing higher requirements on the quality of a match and reinitiating the traversal at the respective start node. The separation of the search process and the clustering facilitates a high level of flexibility. Different cascade definitions can be applied without repeating the time consuming search. The universal approach of our disambiguation routine allows for multiple differently granulated cluster formations of the same context. For technical details, see Doherr (2016).

After creating the clusters of the soft properties and the recoding of the hard poperies, we consolidate all cluster ids into a single table. The *trait vector* associates the document IDs with the respective traits of the documents. A trait is a key composed of a prefix designating the context and a cluster id. A complex relational database becomes a simple vector of tuples.

For a better understanding, the following paragraph describes in full detail an excerpt of the trait vector we constructed for the EPO patents, shown in Figure 1.5. The traits with the prefix NAME refer to three different inventors. The patent has two citations designated by the prefix CITA. The choice of the prefixes indicates that we conducted multiple cluster formations for the different soft properties and for the hard property IPC (International Patent Classification). We create three different aggregation levels for the IPC by truncation at coherent positions prefixed by IPCA, IPCB and IPCC. The inventor address prefixes ADDA and ADDB and applicant prefixes APPA and APPB are based on different cluster building cascades. The trait ADDA1113094 refers to only one address, but cluster ADDB286987 reveals that there are actually four different variants in the data describing this specific location. Because the cascade definitions for the inventor addresses are complementary, it is not necessarily the case that ADDB always returns a stricter defined cluster than ADDA. The applicant cluster

Figure 1.5: Excerpt from the trait vector for the EPO data



APPA264 is a typical representation of name changes during mergers. The pharmaceutical company “Schering” was bought by “Bayer” to become a part of the “Bayer Pharma” group. Before that merger, “Schering” and “Höchst” had a joint venture, called “Höchst Schering AGR EVO Gmbh”, to join their crop protection divisions. The whole process is traceable because the partial overlapping of the names creates intransitive links between these applicants. For reasons of clarity, addresses are not listed. The stricter APPB cluster returns only the name variants close to the time of invention. Finally, the title cluster TITL158119 shows some incoherent entries. Obviously, the term “JC” plays a defining role for two completely different technologies, once as a virus and as a component for superconducting tapes. These incoherent clusters do not pose a high risk, as the probability of joining namesakes by a cluster representing such a small peer group, aka unit size, is marginal at worst. On the other hand, a cluster that completely got out of bounds is hedged by the fact that the inflated unit size automatically increases the calculated probability for a namesake, therefore mitigating false positives.

1.3.2 Mutual traits

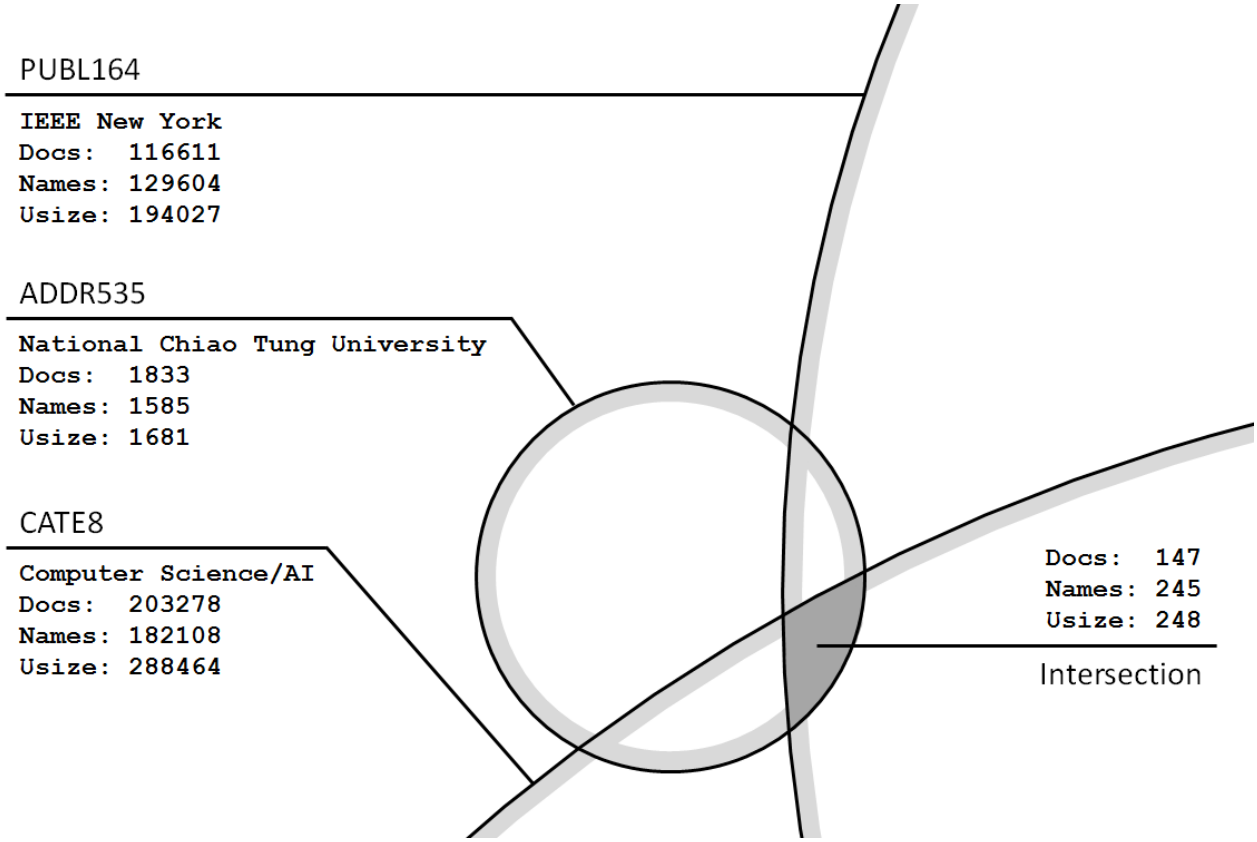
The trait vector reduces the complexity of documents as bags of words, where similarities are obfuscated by noisy variation into a simple collection of traits. The otherwise complex task of identifying common properties between documents transforms into a single SQL statement. Another prerequisite for the disambiguation algorithm is the calculation of the estimated number of namesakes for every author respectively inventor name encountered in the target data. A name is defined as the cluster designating the name in the trait vector, i.e. NAME74872 (see Figure 1.5). As a name cluster may contain several variants because of misspellings, positional variation and so on, only the maximum of the minimum occurrences within every name cluster is ranked and normalized into the *minocc*, ready to be plugged into the estimated equation (1.5) according the target data format.

The algorithm sequentially creates enclosed networks of linked documents for every name cluster. We call such an enclosed network a *namespace*. A link between two documents is defined by the mutual traits of these documents. Of course, the respective name trait is excluded from the mutual traits as this would lead to a completely connected network. The exclusion has to be exerted for all contextually related traits, like author, inventor or applicant,

bearing the risk of tautological links. Independent careers appear already as separated sub-graphs. However, all sub-graphs need to be scrutinized for potential breaking lines. With the number of namesakes n and the population size N already in place, we only need the unit size s to assess the risk of a link connecting documents of two different persons sharing the same name. For every mutual trait of a link, we collect the document IDs in the trait vector sharing this trait. We intersect the resulting documents by means of simple SQL joins. The two originating documents are always in the center of the final intersection. The next step is the identification and aggregation of the involved names to get the preliminary proxy $\check{s}$ by selecting all name traits of these documents in the trait vector. Now we have all parameters required to resolve equation (1.8). With the approximated parameter $\hat{s}$, replacing unit size s in equation (1.5), we can assess the risk of having a namesake for the given name in a peer group of authors or inventors whose profiles match the mutual traits of the two documents. If the risk is above an arbitrarily defined threshold, the link will be destroyed. We repeat this process for every link in the network. Finally, we just need to traverse the remaining links of the namespace to identify the document clusters describing specific author or inventor careers.

Figure 1.6 shows an intersection of mutual traits from a trait vector based on all documents of the research area “Computer Science” in the “Web of Science” database. The areas of the circles (or circular segments) representing the three traits ADDR535, PUBL164 and CATE8 are in proportion to the respective unit sizes. The example illustrates the extremely high computational effort to determine the unit size for a single combination of mutual traits. Intersecting trait PUBL164 with trait CATE8 requires 116611 comparison operations. Starting the sequence with trait ADDR535, the first intersection costs only 1833 operations with a decreasing amount for subsequent intersections. Being the central part of the algorithm, requiring most of the computational resources, improving the performance of the intersection procedure is a worthwhile endeavor.

Figure 1.6: Example of an intersection of mutual traits



1.3.3 Optimization

The magnitude of the document count weak traits yield calls for a more sensible approach than intersecting mutual traits in a random order. Starting with a trait appearing in hundreds of thousand documents significantly slows down the whole process. For that reason, we introduce an additional vector table called *meta vector*. It contains all traits of the trait vector with already calculated unit sizes. Joining mutual traits with this vector allows for efficient ordering of the intersection sequence. Having direct access to the actual unit sizes of the mutual traits may even lead to skipping the intersection effort altogether, if the unit size of the smallest trait already returns a namesake probability equal or below the threshold.

Parallel runs of different approaches have shown that calculating the namesake risk after every intersection, requiring the aggregation of the intersected documents on the name level, is still faster than always conducting the complete intersection sequence. The linkage between the names and their corresponding number of namesakes to calculate the unit size $\hat{s}$ by

equation (1.8) is only required if the provisional unit size *š*, defined by the size of the name aggregate, already returns a preliminary namesake risk equal or below the threshold.

Furthermore, there is a high level of repetition among the links of a namespace. Combinations of mutual traits repeat themselves within the network because of the state dependency found in most inventor or author careers and because of weak links based on combinations of common categories. Identifying the different mutual trait combinations within a namespace reduces the frequency of mutual trait evaluations to the number of combinations. For instance, on average a namespace within the EPO data consists of 123 links based on only 17 combinations.

We also apply two derivations, reasoned by simple set theory, on the intersection procedure: First, if we have completely intersected a combination, but the resulting unit size is still too large in regard of the namesake risk threshold, all remaining combinations that are a subset of the unsuccessful combination are also invalid. Second, if a unit size emerges during the intersection sequence that is small enough to satisfy the namesake risk threshold, all remaining combinations that are a superset of the successful combination are also valid. In both cases, we skip the evaluation of the indirectly rated combinations.

All these optimizations only affect the actual namespace. To convey already made efforts beyond the actual namespace, we introduce a *shortcut table* containing all already assessed combinations and the associated unit sizes. Every combination is represented by a string of traits concatenated in a fixed order. Only new combinations have to be evaluated to end up as another shortcut record in this table.

Finally, the separation into namespaces is a textbook example for applicability of parallelization. A CPU process cycle consists of looking for a free namespace in a namespace registry and reserving it for disambiguation. Contemporary computer systems allow for multiple parallel processes. The only bottleneck is the collective access on key tables like the trait and the meta vector. Every process has its own shortcut table to prevent further accessibility conflicts. The following list summarizes all implemented optimizations:

- Preparatory sorting of mutual traits by unit size using the meta vector.
  → Intersections start already small.

- Intersecting stops early when risk of namesakes is equal or below threshold.
  → Not all mutual traits have to be intersected.

- Identification of mutual trait combinations defining the links in the namespace
  → Number of combinations is much smaller than the number of links.

- If a completely intersected unit size is still too large, all remaining combinations that are a subset of the unsuccessful combination are also invalid.

- If a valid unit size is intersected, all remaining combinations that are a superset of the successful combination are also valid.

- Saving of evaluated combinations and associated unit sizes in a shortcut table prevents repetition of already made efforts beyond the actual name space.

- Separation into name spaces allows for a simple separation of the workload for a multiprocessing approach.

1.4 Namesake bias

The need for reliable benchmark respectively training datasets always accompanies the development of the various disambiguation efforts. These datasets are not only used to compare the performance of different approaches, but also to tune the parameters of the algorithms to produce the desired outcome: improving precision while maintaining a high recall rate. These goals cannot be maximized independently as this would lead to conflicting solutions, i.e. deeming all names unique minimizes the number of false negatives while maximizing the number of false positives. Researchers use training dataset to adjust the weights of matching criteria and algorithm specific parameters in a multitude of ways to balance both goals. There exists several benchmark datasets like the Benchmark Israeli Inventors Set (BIIS) (Trajtenberg et. Al., 2008), the Noise Added French Academe (NAFA) and Noise Added EPFL datasets (NAE) (Lissoni et. Al., 2010) which are publicly available. These datasets trace the careers of individual inventors by their output. As the personal inquiry of this information is an expensive process, the samples are often not randomly drawn but chosen by ease of access. This by itself can already introduce a bias caused by clustering of similar career profiles. Although, the more concerning bias is systematically inherent in the fact that the samples are based on individuals and not on names.

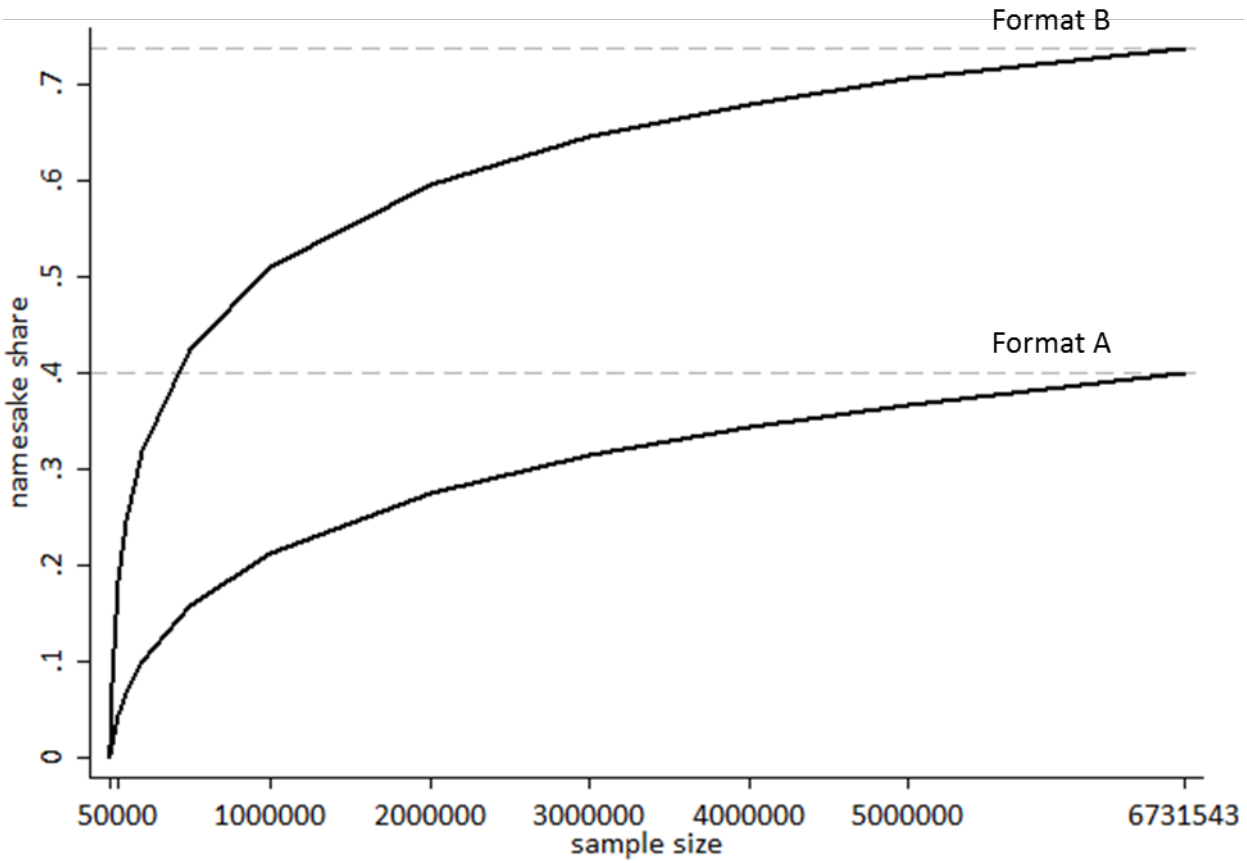
Magerman (2015) criticized benchmark datasets in general for severely underrepresenting careers of homonymous researchers and therefore not being exhaustive. Even one of the largest datasets, the "E&S" labeled dataset (Chunmian, Ke-Wei, Ping, 2016), linking 96,104 patents to 14,293 inventors, contains only 10 homonymous cases, a circumstance implying the deliberate selection of uncommon inventor names to reduce workload. Unfortunately, these datasets are not suited as training data because they concentrate on identifying the careers of individual authors or inventors and not on namespaces. The Monte Carlo experiment, outlined in Table 1.2, simulates this drawing process. In the top half, we can see that the risk of encountering a namesake is usually very low for small units and not representative to the whole population. Even a very common name may appear unique in a relatively small sample, not reflecting the need of disentangling multiple individuals sharing that name in the whole population.

Observing the lower bound name proxy for unit size $\check{s}$ in Table 1.2 culminating in the final name aggregate of the population, as seen in Table 1.1, clearly shows the non-linearity of the relation between a sample size and the encountered namesakes. This systematic underrepresentation of namesakes in a sample stems from the fact that the property "namesake" is only defined in the name aggregate and not on the individual level, requiring at least two randomly drawn individuals with the same name in the sample to be identified as such. At the beginning of a sequential drawing procedure, this conditional probability is much smaller than the probability of drawing an individual with a fresh or unique name but increases with the exhaustion of the name population. Hence, any sample size will lead to an underestimation of the real namesake distribution. Of course, this is also true for the master sample, but, given the size, we are confident that it represents the larger share of the targeted name population.

Figure 1.7 depicts the results of a second Monte Carlo experiment. For a selection of sample sizes between 100 and 5 Million individuals, we repeated a random draw without replacement from the master sample 200 times for every sample size. In contrast to the first Monte Carlo experiment, the master sample is replenished after every sample draw. We record the number of namesake-afflicted individuals for every draw to calculate the average share of namesakes per sample size. For the name format A (last name, first name), we observe 40% of namesake-afflicted individuals in the master sample. This number rises to 74% for the denser aggregate

on initials and last name (format B). Given these high shares and under the assumption that a random draw is representative in terms of namesakes, the observed namesake shares should be close to the dotted lines even for relative small samples. The graph reveals that this is not the case. The curves show a steep catchup of the namesake share consolidating in an almost linear progression as the name population is exhausted with larger sample sizes. Representativeness is not achieved until the population is completely depleted.

Figure 1.7: Share of namesakes in relation to sample size



Note: Confidence intervals are too narrow to be visible on this scale.

All algorithms exploiting training data based on random or selective draws of individuals suffer the namesake bias. They will persistently underestimate the risk of encountering namesakes up to the point where other matching criteria beyond the name itself become pointless, because names are perceived as reliable unique keys. This is especially true for training data deliberately constructed from uncommon names to save the effort of validating document links. The namesake bias does not affect training data based on the semi-random draw of

document tuples as long as both documents belong to the same namespace and there is no selection preferring uncommon names to minimize the effort of validation.

A training dataset providing a robust framework for algorithms should be based on a two- step procedure: First, a representative selection of names has to be determined by the name aggregation of a reasonably sized draw of individuals or, if that is not possible, documents. Second, the complete disambiguation of all careers manifested by documents bearing the drawn names. Unfortunately, the verification by questioning all the authors or inventors sharing a specific name is an impossible task for obvious reasons, i.e. deceased authors, language barriers, obsolete or insufficient addresses. One could argue that identification on the personal level is not required, as this would include information based on confounding parameters having no representation in the data. Even then, the creation of such a dataset is an enormous task requiring coordinated action of several teams to install an overlapping monitoring system to prevent biases.

Even a perfectly balanced training dataset is only valid for the parametrization for one specific bibliometric database. The transferability of these parameters onto other databases requires a high level of compatibility. The method of assessing the risk of encountering a namesake for every single document link does not require a training dataset and is therefore by definition whether bound to a specific database nor affected by the namesake bias. It rather embraces the concept of namesakes by dynamically adjusting its parameters instead of relying on a predetermined statistical average. We will see in the final chapter discussing applications, that the algorithm is still not free from arbitrary decisions, typically permeating most heuristic approaches. Nevertheless, this is a small price to pay given the advantages of not requiring training data, which, as a side effect, allows rapid deployment of the method on any person related bibliometric database.

1.5 Discussion and applications

The key advantage of this approach is the independency of training data. No sample of disambiguated document sets has to be created, be it by common sense assessment, surveys on authors or inventors or by exploiting existing identification keys like ORCHID. Of course, having such a reference group can provide a guideline to improve the settings of the

algorithm, which have emerged during the development phase. Until now, we have only discussed the threshold for the namesake risk as the only parameter. A threshold of 10% seems acceptable for any link between two documents, but by the very nature of the intransitive networks defining the namespaces, the risk of falsely linking separate individuals accumulates, leading to the intransitivity "conundrum" mentioned by Trajtenberg et.al. (2006). A lower threshold alleviates this issue at the cost of increasing the amount of false career splits especially within namespaces of common names. Inspection of these large namespaces has shown that the culprit are in most cases rarely used classifications issued by external authorities and not by the creators of the document. Hence, it is possible to designate specific trait prefixes as supplemental only, if the estimated number of namesakes exceeds an arbitrary limit defining the upper bound of a "small" namespace, e.g. 10. Supplemental traits are used to intersect the unit size, but never define a link exclusively. Besides downgrading specific types of traits, it is also possible to declare trustworthy trait prefixes. Links between documents also based on trustworthy traits enjoy a relaxed namesake risk threshold, if the temporal difference between the documents filing respectively publishing date is below a limit, e.g. 4 years. Finally, it is possible to set a lower bound for the number of namesakes to prevent that namespaces with a very low namesake estimate always are bundled into one cluster regardless of unit sizes. All these additional parameters support the basic idea of the algorithm to simulate the intuitive namesake risk assessment of a manual web search.

Of course, having a proper training dataset to tune the parameters is much more convenient than relying on intuition. Unfortunately, as shown in chapter 1.5, this sentiment may lead to an involuntarily introduced namesake bias. A simple test based on the combined Noise Added French Academe and the Noise Added EPFL datasets illustrates this conflict. We disambiguate the EPO data using our elaborate approach to achieve a recall of 94% and a precision of 99%. However, if we pretend, that every name is unique and namesakes does not exist, we still end up with a recall of 97% and a precision of 97%. Apparently, a small sample of 517 individuals, already containing noise in the form of random namesakes to the properly identified inventors, is not sufficient to represent the population in regard of namesakes. However, even significantly larger benchmark datasets may produce similar results due to the inherent namesake bias.

We applied the algorithm to the data of three major patent offices: EPO, USPTO and JPO. For the EPO and the USPTO we can rely on original data sources provided by the offices. The Japanese data is obtained from the Patstat, a worldwide patent data repository maintained by the EPO. All data sources were released in 2015. We define traits based on inventor names, inventor addresses, applicant names and addresses, title topic clusters, forward and backward citations and international patent classifications (IPC). All soft properties have two context prefixes representing a strict and a more lenient clustering. The trait vector contains three aggregation levels of the IPC: class level (length 3), sub class level (length 4) and group level (delimited by slash). Table 1.3 shows the results and settings of our disambiguation efforts.

Table 1.3: Disambiguation results for three major patent offices

Office	**EPO**	**USPTO**	**JPO**
Source	EPO 2015	USPTO 2015	Patstat 2015
Patents	2,796,553	5,282,235	10,625,369
Names	1,872,103	2,603,181	1,963,483
Inventors	2,382,035	3,295,523	4,004,029
Inventors/Names	1.27	1.27	2.04
Patents/Inventors	1.17	1.60	2.65
Crossing Borders	46,033	67,635	low data quality
Threshold	2.5%|Δt >3y 10%|Δt ≤3y	2.5%|Δt >3y 10%|Δt ≤3y	1%|Δt >2y 5%|Δt ≤2y
Lower Bound	5	5	5

For the EPO and USPTO the main settings are equal. We declare IPC traits as supplemental if the estimated number of namesakes exceeds 10. For every namespace, we enforce at least 5 namesakes as the lower bound. The default threshold is 2.5% respectively 10% if we consider the mutual trait combination as trustworthy. That is the case, if it contains a trait other than an IPC and the filing dates are not more than 3 years apart. Given the expected higher density of namesakes in the Japanese data, we have to adjust the settings accordingly by reducing the limit of the validity of IPC exclusivity to 5 namesakes and demanding a lower namesake risk

threshold of 1% in the default case and 5% in the trustworthy case, which additionally has a shorter time window of 2 years.

The line “Names” designates the name aggregate of the respective population. For example, there are around 1.8 Million distinct inventor names in the EPO data. A slight name clustering to handle misspellings already curtails this number. It would also be the total number of inventor careers given the naïve assumption of name uniqueness. The “Inventors” line shows the count of disambiguated individual inventors. Both, the USPTO and the EPO have a similar ratio in regard of average inventor careers spawned per name. As expected, the higher namesake density in the Japanese population leads to more inventors hiding in a namespace. The average Japanese inventor has also more patents than her European or US counterpart, reflecting the fact that the Japanese patent system requires a patent for every claim of an invention. The USPTO and the EPO have roughly the same share of 2% of inventor careers yielding patents in different countries (see “Crossing Borders”). We were not able to retrieve this information from the Japanese data because the source database Patstat is notorious for insufficient data quality in terms of addresses.

We have shown that it is possible to disambiguate large bibliometric databases without the requirement of training datasets, which, given the precarious representation of namesakes in samples based on individuals, are of questionable value. The algorithm, at its core, relies also on a training dataset of mostly German individuals to aggregate a name population for the namesake estimator. We are aware of the fact that this estimator may under-perform for especially homogenous name populations. Applying a more restrictive threshold strategy can alleviate this shortcoming at the cost of uncomfortably arbitrary decisions. Nevertheless, the intuitive nature of the namesake risk assessment is well suited to monitor the impact of different settings on handpicked namespaces allowing a relatively quick and hassle free disambiguation of any bibliometric data source.

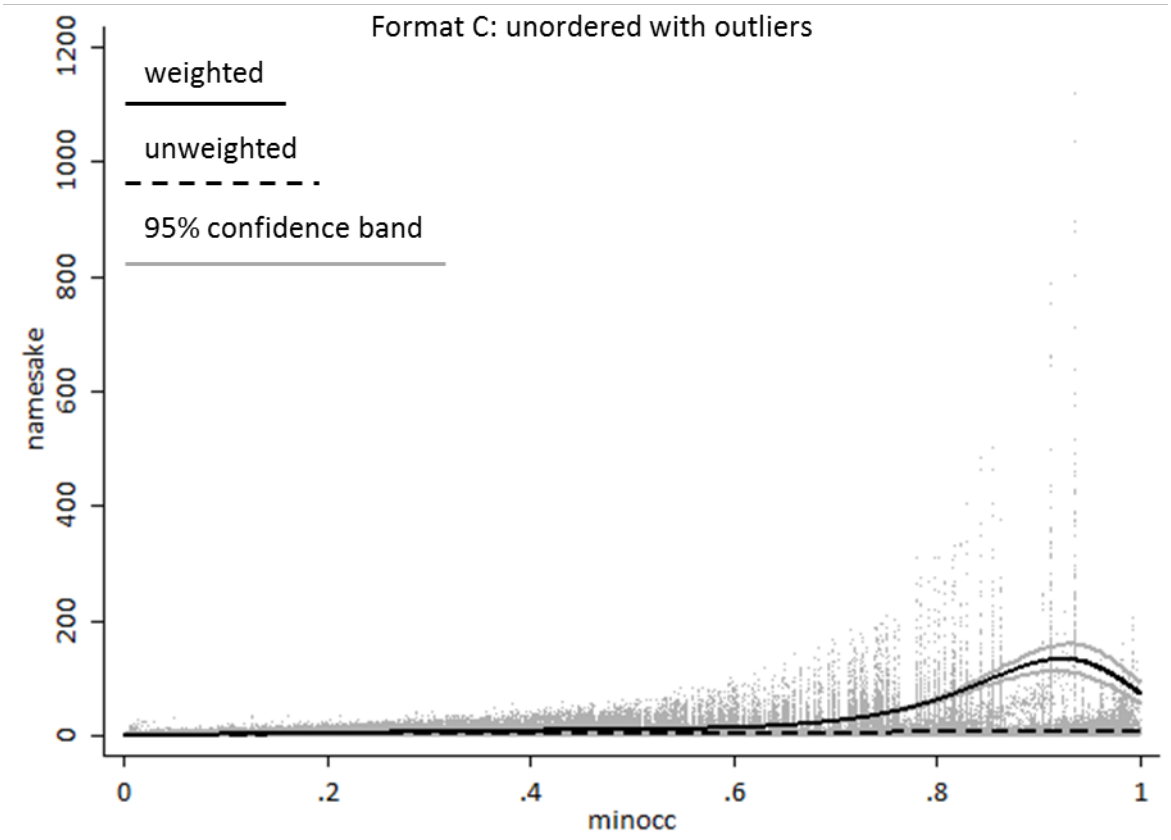
1.A Appendix: Tables and figures

Table 1.4: Weighted Poisson regression of namesakes on *minocc* (5th degree polynomial).

namesakes	Format A first name, last name	Format B initials	Format C unordered
minocc	9.393937*** (1.0015)	45.904456*** (3.6871)	9.050922*** (1.1975)
$minocc^2$	−25.628787*** (8.6151)	−210.223650*** (24.4674)	−24.064323** (10.3669)
$minocc^3$	58.247685** (25.4380)	483.059392*** (64.1039)	51.376449* (31.0283)
$minocc^4$	−64.434441** (30.4690)	−511.296700*** (71.6261)	−52.977835 (37.6794)
$minocc^5$	28.335309** (12.7382)	200.781658*** (28.5667)	22.484273 (15.9620)
const	−0.0215839 (0.01635)	−0.791724*** (0.1542)	−0.025849 (0.0217)
Pseudo R^2	0.7733	0.8411	0.7179
Observations	4,691,779	2,544,481	4,546,192

Notes: Equivalent to the non-aggregated model (one observation designates a person instead of a name) with standard errors clustered on names.

Figure 1.8: Predicted namesakes (weighted vs. unweighted) on scatter plot: Format C, including outliers



2 Inventor Mobility and Productivity in Italian Regions

Riccardo Cappelli [a], Dirk Czarnitzki [b], Thorsten Doherr [c,d], Fabio Montobbio[e]

c) Department of Management, University of Bologna
d) Department of Managerial Economics, Strategy and Innovation, KU Leuven
e) Centre for European Economic Research (ZEW)
f) Centre for Research in Economics and Management (CREA), University of Luxembourg
g) Department of Economics "S. Cognetti de Martiis", University of Turin

Abstract This paper describes the inter-regional and international mobility of inventors in Italy and estimates its impact on total factor productivity (TFP) at the regional level for the period 1996-2011. A new database of mobile inventors is constructed and, using a set of geography based instruments to address endogeneity, the paper shows that inventor inflows and outflows affect regional TFP growth. Moreover, the positive effects of the inventors' mobility (inflow) between different applicants take more time to materialize (relative to movements within the same company). Finally, the negative effects of inventor outflows are mainly driven by mobility between applicants.

Keywords: inventor disambiguation; inventor mobility; migration; economic growth, regional TFP, Italy

JEL: O47; J61; R23; O3

2.1 Introduction

Understanding the economic impact of diaspora, skilled labor mobility, and migration, on regional economic development is a critical issue for both sending and receiving regions. For the receiving regions, skilled migrants contribute to economic growth and generate significant externalities, promoting creativity, innovation and entrepreneurship (Nathan, 2014). For the sending regions, there is potential brain drain and a loss of human capital (Docquier and Rapoport 2012). In short, "communities on the move" influence the welfare and the economic conditions in both the region of origin and destination.

Several papers have studied how skilled labor mobility and immigration have affected innovative capacity (e.g. Trajtenberg, 2005; Hoisl, 2007; Hunt and Gauthier-Loiselle, 2010) and productivity growth (e.g. Peri, 2012). In addition, skilled labor mobility and diaspora networks are increasingly considered a key vehicle of knowledge spillovers (Almeida and Kogut 1999; Rosenkopf and Almeida, 2003; Song et al., 2003; Kerr, 2008; Breschi and Lissoni, 2009; Trippl, 2011).

Despite the growing interest in the diffusion of knowledge through labour mobility of high skilled workers, few studies analyse the relationship between skilled labour mobility (both outflows and inflows) within countries in Europe and regional economic performance. In fact, diaspora effects can arise especially within a country as inter-regional migrants face lower barriers to migration than international migrants (Faggian et al, 2017). To the best of our knowledge, no study directly explores whether inter-regional outflows and inflows of skilled workers within a country affects differential regional economic productivity, as measured by total factor productivity (TFP).

This paper tries to fill this gap by analysing the effects of inventor mobility on TFP growth of Italian regions for the period 1996-2011. Italy is a relevant case because it has experienced very high rates of inter-regional and international mobility of the work force in recent years. In particular, movements of graduated and skilled labour have attracted the attention of policy makers, raising important issues of brain drain and brain gain at the regional level (Becker et al. 2004). Actually, the central and southern regions lose their human capital in an almost systematically manner. A recent report shows that five years after graduation (in 2010), 13%

of graduates in Italian central regions and 26% of the graduates in the South moved mainly to the North (Alma Laurea, 2016).

This paper faces three main empirical challenges: the construction of a skilled labour mobility index, the estimation of the TFP for Italian regions and the identification of the effect of mobility on TFP. First, this paper builds a skilled labour mobility index using a novel database on Italian inventors of patents at the European Patent Office (Czarnitzky et al., 2015) to measure the regional rate of inflow and outflow (and, consequently, the net flow) of inventors between regions. Second, to measure TFP at the regional level, it adopts a growth accounting approach, estimating the regional capital stock (Maffezzoli, 2006). Finally, the last empirical challenge is to identify the effects of labor mobility and TFP growth in the presence of endogeneity, simultaneity and omitted variable biases. This paper identifies some regional characteristics that are likely to be related to skilled mobility and much less to other determinants of productivity at the regional level. These are the geographical distance between the origin and the destination, a common border effect and regional fixed effects (see Miguelez and Moreno, 2015 and Frenkel and Romer, 1999; Peri, 2012). For each different calendar year, these geographic variables are good predictors of the inflow and outflow of inventors and, at the same time, being essentially based on geography, are a priori not correlated with shocks on TFP. The empirical estimates are performed using both OLS and IV 2SLS fixed-effects techniques. In the instrumented regression of productivity on the inventor mobility indexes, this paper uses the following proxies for other relevant determinants of regional productivity growth: R&D expenditure per capita, the ratio of patents to R&D expenditure and the population density.

This paper suggests that inventor mobility has a positive a significant impact on TFP. In particular, it shows that the inflows of inventor has a positive impact and the outflow of inventor has a negative impact supporting the idea of a substantial economic adverse effect of the loss of human capital. In addition, this paper analyses heterogeneous effects for “within applicant”, i.e. an inventor moves between subsidiaries of the same employer across regions, and “between applicants” movements, i.e. between different firms. It shows that the effects of “between applicants” inventor inflows on TFP take more time to materialize than the “within applicant” inventor inflows. Moreover, the negative effects on TFP by inventor outflows are driven by “between applicants” movements.

The rest of the paper is as follows: Section 2 analyses the background literature and discusses why labour mobility and migration affects TFP. Section 3 describes the methodology and the estimation strategy. Section 4 explains the data. Section 5 shows the OLS and IV 2SLS estimates of the effect of labour mobility on productivity. Section 6 provides some concluding remarks.

2.2 Background

Why do skilled labor mobility and, in particular, mobility of inventors affect TFP at the regional level? First, skilled labor mobility produces a better match of jobs and task specialization. Secondly, it stimulates innovation in regions, filling labor shortages in specific sectors, generating absorptive capacity and fostering significant knowledge diffusion. Regional TFP can be generated by entrepreneurial and innovation efforts involving migrants between different regions. For the same set of reasons, the authors of this paper expect that an outflow of skilled labor force induce a decrease in regional TFP.

The first set of explanations builds upon the job matching theory (Jovanovic, 1979); the idea is that inventors stay in jobs in which their productivity appears to be relatively high and that they leave those jobs in which their productivity is perceived to be low. Therefore, mobility is likely to increase the match quality between inventors and employers with a consequent increase in inventors' productivity (Hensen et al., 2009; Marinelli, 2013; Topel and Ward, 1992).

A better match between complementary skills induces also task specialization that, in turn, involves all the employees. Mobile inventors may contribute to make the colleagues more inventive. Complementary skills could involve management and entrepreneurship. At the same time, the inventor further benefits from the knowledge of his/her new colleagues.[1] It is worthwhile noting that this view implies that productivity increases after inventor mobility took place.

[1] Related to this there is the idea that productivity increases because of improved working conditions. For example, Clark et al. (1998) using the German data of the Socio-Economic Panel show that workers are more likely to leave when are not satisfied with their jobs.

A number of papers (Peri and Sparber 2011; Peri 2012; Peri et al. 2013) has analyzed migration in US cities. For example, Peri (2012) finds that immigration has a positive effect on state-level TFP. He also finds that a substantial portion of this effect depends upon increased task specialization of native workers. Relatedly cross-regional labor mobility generates a more culturally diverse workforce.[2] Individuals coming from different regions have different, complementary skills in respect to workers in the receiving region, which can lead to the production of new ideas.

These mechanisms also take place within a country. A good example is Italy providing substantial labor mobility, which is observed by Fratesi and Percoco (2014). Using data on internal mobility for the period from 1980 to 2001, they find a positive relationship between the net inter-regional inflows of skilled people, measured using by educational level, and the GDP per-capita growth of regions. The authors highlight the negative effects of the loss of human capital for the Southern regions of Italy.

A second set of explanations underlines that skilled labor mobility increases productivity because it stimulates innovation activities. Inventors play a crucial role in the literature, which tends to support the idea that international labor movements in science and engineering have a positive effect on innovation, in most of the cases measured by patents. Peri (2007) shows that foreign PHD affects state-level patenting. Using a 1940-2000 state panel in US, Hunt and Gauthier-Loiselle (2010) find as well that an increase in the share of tertiary educated migrants increases the number of patent applications per capita. Kerr (2010) shows that there is localized patent growth in US cities after breakthrough inventions. The spatial reallocation of patenting activities across US cities is faster if the technology has a more mobile workforce. In addition, using data on US H1-B visa program, Kerr and Lincoln (2010) find a positive effect of high skilled immigrants in science and engineering on innovation performance of American firms and cities (see also Stephan and Levin, 2001; No and Walsh, 2010; Ortega and Peri, 2014).

Concerning Europe, Bosetti et al. (2015), using a panel of twenty European countries, observe that skilled migrants contribute positively to the number of patents and citations of scientific

[2] There is a large literature discussing the impact of cultural diversity on productivity. See for example Ottaviano and Peri (2006) and Alesina et al. (2014).

publications. Fassio et al. (2015) show that highly skilled migration has a positive effect on innovation in Germany, France and UK. Finally, tracking inventor mobility for 274 EU NUTS2 regions over 8 years, Miguélez and Moreno (2015) show that inventor mobility and co-patenting have a positive effect on regional patents per capita. They interpret their results in terms of absorptive capacity of the regions that benefit from the knowledge and information brought in by mobile inventors and cooperation networks.

Mobile inventors could increase innovation and productivity at the regional level if there are barriers to mobility and self-selection leads them to be more educated, more entrepreneurial or of higher unobserved inventive ability. However, there is substantial evidence that labor mobility of high skilled workers is a key mechanism of knowledge diffusion that overcomes geographic barriers and other constraints. When an employee changes jobs, he/she transfers from the old to the new firm detailed information on the technologies used in the previous employment and the knowledge, skills and experience embedded in the mobile worker. Previous research shows a positive relationship between mobility and productivity of inventors (Trajtenberg, 2005; Giuri et al., 2007; Hoisl, 2007).[3]

Inventor mobility brings about "learning by hiring" (Almeida and Kogut 1999; Rosenkopf and Almeida, 2003; Song et al., 2003; Breschi and Lissoni, 2009). Rosenkopf and Almeida (2003) show that inventor mobility increases the likelihood of knowledge flows between two firms (measured by patent citations) irrespective of the geographic location of the two firms, i.e. in the same or different regions.[4] These results are confirmed by Song et al. (2003) using data for all the firms in the semiconductor industry for the period 1980-1999. At an aggregate level, Almeida and Kogut (1999) show the close link between knowledge flows and labor mobility. Analyzing the determinants of knowledge flows (measured by patent citations) between regions in the US semiconductor industry, they show that the mobility of engineers is an

[3] Using data on USPTO patent inventors, Trajtenberg (2005) shows that mobile inventors are more productive, as measured by patent citations, than non-mover inventors. Hoisl (2007), using a sample of German inventors with EPO patents, confirms these results. The author also shows that inventors that are more productive are less likely to move.

[4] They empirically test the effectiveness of inventor mobility as mechanism of interfirm knowledge flows for a sample of 74 entrant firms founded between 1980 and 1989 in the semiconductor industry. The authors also find that the effectiveness of inventor mobility as mechanism of knowledge diffusion increases when the involved firms are technologically distant.

important factor explaining the localized diffusion of knowledge within regions. This stems from the fact that an important part of inventions is represented by the tacit knowledge embedded in engineers.

Beyond these direct effects, inventor mobility positively affects firms and regions performance through knowledge externalities (Griliches, 1995; Breschi and Lissoni, 2009). The mobility of workers creates links between firms through social ties, which involve the worker that moves and the workers in his or her previous firm. These ties favor the diffusion of knowledge among firms and regions (Breschi and Lissoni, 2009; Miguélez and Moreno, 2015 Cappelli and Montobbio, 2016). Agrawal et al. (2006), analyze the diffusion of knowledge (measured by patent citations) between US regions, generated by the mobility of inventors. They show that an inventor who moves from one region to another is more likely to cite inventors in the previous region, compared to those who have never lived in that region. The social networks between inventors reduce the frictions in knowledge flows exerted by geographical factors such as physical distance.

2.3 Methodology

2.3.1 Empirical specification

To conduct the empirical analysis, this paper represents a region by its production function to calculate the TFP, which in turn depends upon knowledge and technological variables.[5] Knowledge generated in a region is mainly measured using technological input measures like R&D expenditure and number of high skilled people and/or technological output measures like the number of patents and patent forward citations. On the other side, knowledge flows between regions are measured using indirect measures like the stock of foreign R&D or considering explicitly a channel of knowledge flows like inventor mobility and citations between inventors or scientists. In this work, the dynamics of TFP growth in Italian regions are modelled using the following equation:

[5] The empirical model of this paper is also in line with the technology gap approach (Nelson and Winter, 1982; Fagerberg and Verspagen, 2002). The traditional technology gap model (Fagerberg, 1987, 1988) considers regional economic (or productivity) growth as driven primarily by innovation and takes the distinction between the development of new knowledge in a region and the diffusion of knowledge between regions.

$$ln\left(\frac{TFP_{i,t}}{TFP_{i,t-1}}\right) = \beta_0 + \beta_1 Mobility_{i,t-1} + \beta_2 ln\left(TFP_{i,t-1}\right) + \beta_3 R\&Dpc_{i,t-1} + \beta_4 PATrd_{i,t-1} + \beta_5 Density_{i,t-1} + \alpha_i + \varepsilon_{i,t-1} \quad (2.1)$$

where $ln\left(TFP_{i,t}/\ TFP_{i,\ t-1}\right)$ is the TFP growth rate between year *t-1* and *t* of a given Italian region *i* and ($TFP_{i,t-1}$) represents the lagged regional level of TFP. Following a standard growth accounting approach (Solow, 1957), these variables are constructed using a Cobb-Douglas production function with two input factors, i.e. labor and capital, and constant return to scale. The innovative efforts of regions are measured using R&D expenditure per capita ($R\&Dpc_{i,t-1}$). The ratio of patents and R&D expenditure ($PATrd_{i,t-1}$) is also included to check for an additional effect exerted by successful R&D. The placeholder $Mobility_{i,t-1}$ takes account of the interregional diffusion of knowledge when inventors move between regions. It will be substituted by three types of inventor mobility indexes: inflow of inventors from other regions ($Inflow\_rate_{i,t-1}$); outflow of inventors to other regions ($Outflow\_rate_{i,t-1}$) and net flows of inventors ($Netflow\_rate_{i,t-1}$), i.e. the difference between inflow and outflow of inventors. These indexes are expressed as the ratio of the number of the respective mobile inventors in a period and the regional stock of inventors in the previous period.[6] In addition, population density ($Density_{i,t-1}$), measured by the number of thousand inhabitants per square kilometer, is included to control for agglomeration effects (Glaeser, 2010). Finally, ε_{it} represents the error term.

2.3.2 Identification strategy

In order to estimate equation (2.1) this paper needs to address some econometric issues that affect the correct identification of the effects of our variables on TFP growth regarding the inventor mobility indexes. The literature on labor mobility (see e.g.: Ortega and Peri, 2013) clearly demonstrates that economic factors like the expected earnings in the destination area are important in explaining the observed migration patterns. According to the job matching theory (Jovanovic, 1979), it could be the case that highly efficient regions attract inventors more than less efficient regions. Moreover, TFP shocks might affect the relative degree of

[6] To give an example, the inventor inflow index of Lombardy in 2000 is calculated as the ratio of the number of inventors that move to Lombardy in 2000 to the stock of inventors in Lombardy in 1999.

attractiveness of regions. This means that the relationship between TFP and inventor mobility might be bi-directional which introduces an endogeneity problem resulting in biased estimates of the coefficients of the inventor mobility indexes.

To solve the endogeneity issue, this paper adopts several strategies. Firstly, as visible from equation (2.1), the mobility indexes and all the other independent variables are lagged by one year. Moreover, a fixed-effects OLS estimator is applied to take all the unobservable factors related to region's attractiveness to inventors into account. However, these expedients still do not completely solve the potential omitted variables and endogeneity biases. To address these issues further, we implement a 2SLS fixed-effects technique.

Following Frankel and Romer (1999), the instruments used are constructed using a gravity model where bilateral inventor flows are explained by regional geographic characteristics. Migration costs related to physical distance and national border clearly affect inventor mobility between regions (Ortega and Peri, 2014; Migueléz and Moreno, 2015), while, on the other side, it can be safely assumed that these geographic factors are not correlated with regional TFP.

As a first step, we conduct cross-sectional poisson pseudo maximum likelihood (PPML) estimations for inventor inflow respectively outflow rates for every of the 16 years covered by our data.[7] The gravity model takes to following form:

$$M_{ij} = exp\left[\beta_0 + \beta_1 ln(Dist_{ij}) + \beta_2 Border_{ij} + \beta_3 Italy_{ij} + \sum_{k=1}^{IT} \beta_{4k}\, r_{ik} + \sum_{k=1}^{IT+C} \beta_{5k}\, r_{jk}\right] + e_{ij} \quad (2.2)$$

Where M_{ij} captures the inventor mobility rate between region i and j in a given year. The index i always designates a region in Italy (IT) while j also includes involved countries (C), implying that inflow and outflow ratios are based on the inventor stock of Italian region i. As geographic variables, we use the logarithm of the geographical distance between the two areas ($Dist_{ij}$), a dummy indicating if the two regions are neighbors ($Border_{ij}$) and a dummy

[7] Santos Silva and Tenreyro (2006) demonstrate the appropriateness of the PPML estimator to address econometric issues inherent to gravity model like heteroscedasticity and observations without bilateral flows.

designating if the two regions belong to Italy ($Italy_{ij}$). Two dummy sets control for region-specific effects and correct for cross-sectional bias (Anderson and van Wincoop, 2003; Baldwin and Taglioni, 2006). One dummy set controls exclusively for the Italian regions ($IT = 20$) identified by i, while the other, controlling region j, also includes countries which are involved in a relocation of an inventor from/to Italy ($C = 52$).[8]

The seconds step consists of the aggregation of the predicted mobility rates per region i:

$$\widehat{M}_i = \sum_{j \neq i} \widehat{M}_{ij} \tag{2.3}$$

As a result we receive the predicted inflow respectively outflow rates of inventors for every region and every year in our data. The net inflow rate is constructed by the difference between the two instrumental variables.[9]

2.4 Data

For the empirical analysis, this paper constructs a set of variables for the period 1995-2011 for the 20 Italian regions, i.e. the first level of administrative divisions in the Italian state.[10] A first group of variables is constructed using data from the Italian National Institute of Statistics (ISTAT), i.e. total R&D expenditure (used to construct the variable $R\&Dpc_{i,t-1}$), population and area in square km (used to construct the variable $Density_{i,t-1}$). In addition we rely on other data sources: PASTAT data to construct the number of patents (used to build the variable $PATrd_{i,t-1}$); EUROSTAT data on the coordinates of regional centroids (used to build the variable

[8] Each of the 16 annual cross-sections consists of 1420 observations (20*(19+52)). Intra-regional mobility is excluded.

[9] Gravity model estimates are not performed for interregional net-flows because of inappropriateness of the gravity-type variables. Variables like geographical distance and territorial borders represent migration costs, which affect bilateral inventor inflows and inventor outflows in the same way. Thus, these variables are not useful in explaining the bilateral net effects resulting from inflows and outflows

[10] This paper uses a balanced panel dataset. TFP growth data refers to period 1996-2011, while data on the lagged independent variables refer to the period 1995-2010. The period of analysis cannot be extended because of data constraints.

$Dist_{ij}$).[11,12] As mentioned above, to develop the empirical analysis, we have to address the challenge of measuring both geographical inventor mobility and regional TFP.

2.4.1 TFP of Italian regions

We measure the TFP of Italian regions, both in level and growth rate, as Solow's residual to GDP once the contribution of two input factors, i.e. labor and capital, are taken into account. This paper relies on ISTAT data to construct the variables on regional TFP: GDP at constant prices[13]; number of full time equivalents[14] as measure of labor input; ratio between compensation of employees and GDP as measure of GDP elasticity to labor. Data on capital stock is not available at regional level, but ISTAT provides data on regional fixed investment for the period 1995-2011. These short time series data on regional investments allow one to obtain, through a perpetual inventory method, only a partial approximation of the capital stock of regions. Moreover, the quality of the approximation worsens for the first part of the period, as the length of the series on fixed investments is getting shorter. In order to reduce this shortcoming of the simple perpetual inventory method, we use the procedure developed by Maffezzoli (2006). The basic idea is to integrate the regional fixed investments data with the time series data on national capital stock (available from ISTAT) in order to construct a measure of regional capital stock using as much as information as possible. For further details, the authors of this paper refer to the appendix.

2.4.2 Mobility of Italian inventors

The large patent datasets supplied by patent offices make it possible to construct various measures based on patents (number of patents, patent citations, etc.) at country or regional

[11] Geographical distance, measured in km, is calculated as great circle distance between regional centroids.
[12] This paper considers also the potential effect of the human capital. ISTAT provides data on the number of graduates in Science & Technology (S&T) for the period 1998-2011. Using these data, a variable measuring human capital (high skilled) is computed as the ratio between the number of S&T graduates and the total population. *High skilled* is highly correlated with other control variables already included in the model, i.e. R&D per capita (0.64) and TFP level (0.48). Thus, the authors of this paper argue that R&D per capita and TFP level capture most of the regional differences in human capital. Moreover, the results of 2SLSE FE estimates that include *high skilled* as additional control variable (available from the authors upon request) are similar to those reported in the paper, and also the coefficient values of *high skilled* are not statistically significant.
[13] All the variable measured in Euros are expressed at constant prices (reference year: 2005).
[14] Data on total hours worked is not available at regional level.

level.[15] However, for the construction of measures related to the mobility of inventors, the data suffers some important limitations because of the "who is who" and the "John Smith" problems (Trajtenberg et al., 2006). The former refers to the fact that the name of an inventor with two or more patents may be spelled differently on these documents. The latter refers to the same name sometimes referring to different inventors. To overcome these limitations, this paper builds a separate dataset using a procedure belonging to the class of the so called "name game" methods (Trajtenberg et al., 2006; Raffo and Lhuillery, 2009) conducted on PATSTAT data of EPO patent applications. To tackle the common name issue, an inventor career in patent data is represented by documents that not only share the name of the inventor but also additional characteristics like the assignee, addresses, co-inventors, citations and so on. Whether a set of mutual characteristic between two documents is sufficient for a valid connection depends on a heuristic plausibility check. The algorithm is based on a hierarchical order of the characteristics starting with the inventor address. All patents sharing a similar address for a given inventor name are considered to be from the same person. These patent clusters by them self are not able to identify mobility, but in association with the next entry in the hierarchy, the assignees, we are able to create an intransitive network between these non-mobile clusters. Traversal of this network leads to an onion like structure of layered clusters, already containing mobility. This process is repeated for the remaining characteristics like citations or co-inventors ending with the international patent classification. To avoid huge clusters of documents perceived as unrelated, a circumstance explained by the intransitivity of the connections, a system of plausibility checks, based on aggregated meta information within a cluster and exogenous assessment of the resulting mobility, has to be passed. If a cluster is not able to pass the test, the contained network is successively traversed with increasingly restrictive rule sets until it is separated into plausible sub-clusters. For more information on this topic, please consult Doherr, 2017.

To verify the overall quality of this procedure in regard of precision and recall rate[16], we conducted the benchmark Lissoni et al. (2010) proposed for the algorithm challenge of the APE-INV (Academic Patenting in Europe - Inventors) initiative of the European Science

[15] Patent documents provide a variety of information regarding the invention (e.g. description of the invention and its technological class), applicants (names and addresses) and inventors (names and addresses).
[16] Precision = true positive / (true positive + false positive); Recall = true positive / (true positive + false negative).

Foundation. It is based on individually verified EPO patent links of 424 French and 121 Swiss researchers, enriched with false positives as noise. Because we disambiguated not only the Italian inventors but also the entirety of EPO patents, we were able to follow the guidelines of the APE-INV for these benchmark datasets. Our method achieved a recall rate of 90.98% with a precision of almost 100% (99.9903%). Given these numbers, we are confident to identify inventor mobility in Italy without having concerns related to the disambiguation procedure.

This dataset of Italian inventors allows us to identify movement of inventors between regions by observing the patents they developed over time. We consider inventors with at least two EPO patent applications and look at the inventors' addresses of these patents. If an inventor, in a given period, has an EPO patent with a region address and the same inventor, in a later period, appears on an EPO patent with a different region address, this paper assumes that this inventor moved from one region to another during the two periods. Since the exact date of an inventor movement cannot be tracked from the patent documents, the inventor flows are computed assuming that the mobile inventors move in the priority year of the patent of the destination region.

2.5 The empirical analysis

2.5.1 Descriptive evidence

This section displays the main characteristics of our data on TFP and inventor mobility in the twenty Italian regions. Table 2.1 shows descriptive statistics of the TFP annual growth rates (in percentage values) for the period 1996-2011. The TFP growth rates range from -5.95% (Umbria) to 4.86% (Calabria). The region with the highest average value of TFP growth rates is Basilicata (0.53%); the region with the lowest average value of TFP growth is Molise (-0.46%).

Table 2.1: Descriptive statistics of TFP annual growth rate (percentage change) - 1996-2011

Region (North to South)	Mean	Std. Dev.	Min	Max	Obs.
Trentino Alto Adige	–0.19	1.55	–3.52	2.19	16
Aosta Valley	0.03	2.23	–4.06	3.11	16
Friuli-Venezia Giulia	0.04	2.00	–4.60	3.04	16
Lombardy	–0.11	1.83	–4.71	4.29	16
Venetia	–0.04	1.79	–4.13	2.44	16
Piedmont	0.02	1.97	–5.65	3.82	16
Emilia-Romagna	0.23	2.00	–5.03	3.22	16
Liguria	0.09	1.72	–4.48	2.94	16
Tuscany	0.10	1.39	–3.44	2.08	16
Marche	0.07	1.71	–3.78	1.74	16
Umbria	–0.27	1.80	–5.95	1.94	16
Abruzzi	0.10	1.55	–2.71	2.75	16
Latium	–0.07	1.27	–2.77	1.92	16
Molise	–0.46	1.69	–4.85	2.24	16
Campania	0.53	1.24	–2.49	2.39	16
Apulia	0.06	1.48	–2.78	2.54	16
Basilicata	0.53	1.95	–3.39	3.20	16
Calabria	0.28	2.29	–3.99	4.86	16
Sardinia	–0.15	1.20	–2.95	1.73	16
Sicily	–0.02	1.31	–3.74	1.99	16
Italy	0.04	1.43	–3.97	2.20	16

Note: the TFP growth rates for Italy are calculated as weighted average of the TFP growth rates of the 20 Italian regions.

Table 2.2 displays the stock of inventors in the year 1995 and the total number of inventor flows for the Italian regions in the period 1995-2010.[17] Column 1 shows the geographical distribution of Italian inventors in 1995, i.e. the stock of Italian inventors with at least one EPO patent application. The total number of Italian inventors in 1995 is 7143 with the highest number of inventors (2570) in Lombardy and the lowest (5) in Aosta Valley. The other columns show the interregional inventor inflows, outflows and net inflows. For each of these three categories of inventor mobility, the total number of flows (Total) are also separated in inventor flows between Italian regions (National) and inventor flows between Italian regions

[17] Our database obtained using the disambiguation algorithm contains 52696 inventors with at least one patent during the period covered by our analysis. The number of inventors with only one patent is 33459 (63.49% of the total). Since inventor mobility is observed only if an inventor has 2 or more patents, the inventor mobility measures adopted in this paper do not capture potential movements of inventors with only one patent. In general, we recognize that the computed inventor mobility indexes underestimate the real inventor flows, and overall the inter-regional flows of high skilled people.

and non-Italian regions (International). All of these values are constructed aggregating the annual data on inventor mobility observed during the period 1995-2010. The region with the highest value for total inventor inflows (720) is Lombardy; the region with the lowest value for total inventor inflows is Molise (2). The region with the highest value for total inventor outflow is Lombardy with 695 cases; Molise and Aosta Valley are the regions with the lowest value of total inventor outflows with 6 cases each. The region with the highest value for total net inflows is Emilia Romagna with 28 cases; the region with the lowest value of total net inflows is Piedmont with a value of -69. Regarding the distinction between national and international flows, it emerges that (on average) two thirds of inventor flows, both inflows and outflows, are represented by inventor mobility within Italy. The relatively low or even detrimental net flows imply that the significant catch-up of some regions, like Campania, Apulia or Tuscany, observed in 2010 stems largely from interregional efforts than from shifts explained by mobility. Considering the international mobility, Table 2.3 shows the top 10 countries of origin (destination) of inventor inflows (outflows). USA ranks first in both categories of inventor flows and, with the exception of China ranking tenth as destination country of inventor outflows, the other top 10 countries are European countries.

Table 2.2: Stock of inventors and total number of inventor flows during the period 1995-2010

Regions North to South	Stock of Inventors 1995	Inflow			Outfow			Netflow			Stock of Inventors 2010
		Total	National	Intern.	Total	National	Intern.	Total	National	Intern.	
Trentino Alto Adige	81	40	24	16	38	19	19	2	5	–3	176
Aosta Valley	5	14	14	0	6	6	0	8	8	0	19
Friuli-Venezia Giulia	226	67	51	16	51	42	9	16	9	7	346
Lombardy	2570	720	420	300	695	402	293	25	18	7	2637
Venetia	743	199	141	58	200	131	69	–1	10	–11	948
Piedmont	992	232	169	63	301	212	89	–69	–43	–26	1059
Emilia-Romagna	929	248	168	80	220	137	83	28	31	–3	1267
Liguria	213	61	44	17	70	59	11	–9	–15	6	262
Tuscany	376	173	120	53	155	111	44	18	9	9	720
Marche	141	46	34	12	40	39	1	6	–5	11	260
Umbria	49	29	21	8	31	26	5	–2	–5	3	81
Abruzzi	81	53	39	14	60	39	21	–7	0	–7	82
Latium	458	210	139	71	176	116	60	34	23	11	550
Molise	6	2	1	1	6	5	1	–4	–4	0	4
Campania	79	46	42	4	68	58	10	–22	–16	–6	226
Apulia	43	29	23	6	47	39	8	–18	–16	–2	166
Basilicata	8	5	3	2	8	6	2	–3	–3	0	13
Calabria	14	10	9	1	9	9	0	1	0	1	28
Sardinia	34	19	16	3	24	20	4	–5	–4	–1	40
Sicily	95	35	26	9	39	28	11	–4	–2	–2	130
Italy	7143	2238	1504	734	2244	1504	740	–6	0	–6	9014

Table 2.3: Inventor inflows and outflows (period 1995-2010): top 10 countries per country of origin and destination

Inventor inflows			
		Percentage on total	
Country of origin	Number	Including Italy	Excluding Italy
USA	198	8.85	26.98
Germany	118	5.27	16.08
France	95	4.24	12.94
United Kingdom	81	3.62	11.04
Switzerland	61	2.73	8.31
Sweden	30	1.34	4.09
Netherlands	29	1.30	3.95
Belgium	22	0.98	3.00
Spain	19	0.85	2.59
Austria	7	0.31	0.95
Inventor outflows			
USA	209	9.31	28.24
Germany	111	4.95	15.00
Switzerland	85	3.79	11.49
France	80	3.57	10.81
United Kingdom	57	2.54	7.70
Netherlands	30	1.34	4.05
Belgium	26	1.16	3.51
Sweden	25	1.11	3.38
Spain	21	0.94	2.84
China	10	0.45	1.35

Note: The percentage values under the column Including Italy (Excluding Italy) are calculated including (excluding) inventor mobility within Italy.

2.5.2 Results

Table 2.4 shows the descriptive statistics and correlations of the variables used to estimate the determinants of TFP growth rates of Italian regions according to equation (2.1).

Table 2.4: Descriptive statistics

Variable	Description	Mean	Std.Dev.	Min	Max
$\ln(TFP_{i,t} / TFP_{i,t-1})$	Regional TFP growth rate	0.0004	0.017	−0.06	0.05
$\ln(TFP_{i,t-1})$	Regional TFP level	1.895	0.085	1.663	2.08
$Inflow\_rate_{i,t-1}$	Interregional inventor inflow rate	0.020	0.037	0.0	0.50
$Outflow\_rate_{i,t-1}$	Interregional inventor outflow rate	0.022	0.038	0.0	0.33
$Netflow\_rate_{i,t-1}$	Interregional net inflow rate	−0.002	0.041	−0.333	0.25
$R\&Dpc_{i,t-1}$	R&D expenditure (in 1K €) per capita	0.217	0.133	0.018	0.55
$PATrd_{i,t-1}$	Patents per 1000K € R&D expenditure	0.236	0.105	0.0	1.39
$Density_{i,t-1}$	Population density (people per km^2)	0.176	0.105	0.035	0.42

Table 2.5 shows the results of equation (2.1) obtained using OLS FE (models with suffixes **a**) and 2SLS FE estimates (models with suffixes **b**). These analyses are performed by alternating through the three categories of inventor flows. As mentioned in Section 3, to instrument the observed inflow or outflow rates in the 2SLS analysis, variables are constructed containing the predicted values from cross-sectional gravity model estimates using the respective observed mobility rate as dependent variable (see equations (2.2) and (2.3)). The instrumental variable for the observed net-flow rate is calculated as the difference between the predicted inflow rate and the predicted outflow rate.

For each of the three 2SLS FE models a weak identification test is performed by computing the Kleibergen-Paap Wald rk F statistic. The Kleibergen-Paap test coincides with the Angrist and Pischke (2009) test since only one endogenous variable is used in the 2SLS analysis. The values of these tests are well above both to the traditional rule of thumb of 10 and to the highest critical value (16.38) reported by Stock and Yogo (2005), and, thus, support the relevance of the selected instruments. In general, the results of OLS and 2SLS FE are very similar. The sign and the significant level associated to the coefficient of these three inventor flow variables are the same in both OLS and 2SLS FE estimates. Only slight differences are observed in the coefficient and standard error values.

As expected, the inflow of inventors (*Inflow_rate*) has a positive effect on regional TFP growth (0.025 in Model 1b). As outlined above, several reasons can explain this result. First, immigrating inventors positively affect innovation capacity of the destination regions increasing the stock of inventors. Second, incoming inventors help destination regions to gain access to different and complementary knowledge. Third, even though the direct beneficiaries of inventor mobility are the local hiring firms, knowledge externalities allow other local actors to benefit from the knowledge embodied in the incoming inventors. In general, this result is in line with the existing literature that support the effectiveness of inventor mobility as channel of knowledge diffusion (Migueléz and Moreno, 2015), but extend this literature by providing a first empirical evidence of the direct contribution of inventor inflows in explaining the changes in the regional TFP.

Table 2.5 also shows a negative coefficient for outflow of inventors (*Outflow_rate*) (-0.037 in Model 2b). This results suggests that the negative effects associated with inventor outflows, i.e. the reduction in the inventor stock and the weakening in the network ties of inventors within region (brain drain effect), prevails over the positive effects represented by the facilitated access to the knowledge generated outside the region (brain gain effect).

The results also show a positive coefficient for the net inflow of inventors (*Netflow_rate*) (0.048 in Model 1c), which overall confirms the benefits on regional TFP growth associated with the immigration of inventors.

Finally, the effects of the controls are worth mentioning. Lagged TFP level ($ln(TFP)$) is significant and has a negative sign. Thus, during the period examined there was a process of catching-up, which means that regions with lower levels of TFP per capita, ceteris paribus, show higher TFP growth rates. The results do not show any significant effect of R&D activities (*R&Dpc*) on TFP growth rates. However, there is a positive and significant effect of the variable for patent intensity (*PATrd*). This means that successful R&D activity, i.e. which results in a patent application to the EPO, contributes positively to regional growth. Lastly, population density is positive but not significant (*Density*).

Table 2.5: Determinants of TFP growth rates - OLS FE and 2SLS FE estimates

$ln\left(\frac{TFP_{i,t}}{TFP_{i,t-1}}\right)$	Inflow		Outflow		Netflow	
	Model 1a OLS FE	Model 1b 2SLS FE	Model 2a OLS FE	Model 2b 2SLS FE	Model 3a OLS FE	Model 3b 2SLS FE
$\ln(TFP_{i,t-1})$	−0.275***	−0.274***	−0.316***	−0.316***	−0.298***	−0.298***
	(0.075)	(0.070)	(0.077)	(0.072)	(0.069)	(0.065)
$R\&Dpc_{i,t-1}$	−0.001	−0.002	0.004	0.004	−0.004	−0.004
	(0.018)	(0.017)	(0.018)	(0.017)	(0.018)	(0.017)
$PATrd_{i,t-1}$	0.014**	0.014**	0.017**	0.017**	0.014*	0.014**
	(0.007)	(0.006)	(0.007)	(0.007)	(0.007)	(0.006)
$Density_{i,t-1}$	0.179	0.179	0.170	0.170	0.168	0.168
	(0.210)	(0.198)	(0.220)	(0.207)	(0.202)	(0.190)
$Inflow\_rate_{i,t-1}$	0.024*	0.025**				
	(0.013)	(0.012)				
$Outflow\_rate_{i,t-1}$			−0.038**	−0.037***		
			(0.014)	(0.014)		
$Netflow\_rate_{i,t-1}$					0.048***	0.048***
					(0.010)	(0.009)
Year dummy	Yes	Yes	Yes	Yes	Yes	Yes
Region FE	Yes	Yes	Yes	Yes	Yes	Yes
Instrument: predicted…		Inflow rate		Outflow rate		Netflow rate
Observations	320	320	320	320	320	320
Number of regions	20	20	20	20	20	20
R-squared	0.682	0.682	0.685	0.685	0.691	0.691
Log Likelihood	1038.42	1038.42	1039.95	1039.95	1043.13	1043.13

Note: clustered standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

2.5.3 "Within applicant" and "between applicants" inventor flows

Are these results driven by "within applicant" or "between applicants" inventor flows? To address this question the full sample of inventor movements is separated into two groups. If the applicant of the last patent in the originating region equals the assignee of the first patent in the destination region, the inventor movement is considered as "within applicant", while a change constitutes a "between applicants" movement. In case of patents with multiple applicants, inventor movements are considered as "between applicant" when the originating region and destination region patents do not show any similarity in the applicants. The

observed percentage of inventors that move “between applicants” at least once is 43.5% (973/2238).

A preliminary set of 2SLS FE estimates (not shown here but available from the authors upon request) are performed substituting the aggregate mobility indexes with the two mobility indexes computed distinguishing between mobility “within applicant” and mobility “between applicants”. We find that the effect of inventor inflows resulting from “within applicant” movements is positive and significant at 10% level, while the effect of inventor inflows resulting from “between applicants” is not significant. Inventor outflows are negative for both types of mobility, but (more interestingly) only the effect of “between applicants” movements is significant. Inventor netflows are positive and significant for both types of movements (at 10% level for “between applicant” movements).

We suspect that the effects of “between applicants” inventor movements will occur only after few years (e.g. because of different organizational routines between the origin and destination firms). Thus, new 2SLS FE estimates are performed including also the 2 year lagged values of the inventor flows variables. The estimates results (see Table 2.6) show that, as expected, the coefficient of the 2 year lagged values of inventor inflows resulting from “between applicants” is significantly positive (see Model 1b).

Table 2.6: Determinants of TFP growth rates and distinction between "within applicant" and "between applicant" mobility

$ln\left(\frac{TFP_{i,t}}{TFP_{i,t-1}}\right)$	Inflow		Outflow		Netflow	
	Model 4a	Model 4b	Model 5a	Model 5b	Model 6a	Model 6b
$\ln(TFP_{i,t-1})$	−0.275*** (0.071)	−0.282*** (0.072)	−0.291*** (0.072)	−0.317*** (0.067)	−0.288*** (0.071)	−0.304*** (0.063)
$R\&Dpc_{i,t-1}$	−0.001 (0.015)	−0.002 (0.019)	0.004 (0.016)	0.003 (0.017)	0.001 (0.016)	−0.006 (0.021)
$PATrd_{i,t-1}$	0.015*** (0.005)	0.015** (0.007)	0.015** (0.007)	0.015*** (0.006)	0.016*** (0.006)	0.014** (0.007)
$Density_{i,t-1}$	0.175 (0.200)	0.169 (0.191)	0.171 (0.207)	0.149 (0.185)	0.183 (0.205)	0.176 (0.188)
$Inflow\_rate\ within_{i,t-1}$	0.041* (0.023)					
$Inflow\_rate\ within_{i,t-2}$	0.032 (0.022)					
$Inflow\_rate\ between_{i,t-1}$		0.015 (0.043)				
$Inflow\_rate\ between_{i,t-2}$		0.046*** (0.010)				
$Outflow\_rate\ within_{i,t-1}$			−0.022 (0.021)			
$Outflow\_rate\ within_{i,t-2}$			0.017 (0.011)			
$Outflow\_rate\ between_{i,t-1}$				−0.076*** (0.024)		
$Outflow\_rate\ between_{i,t-2}$				0.031 (0.063)		
$Netflow\_rate\ within_{i,t-1}$					0.062*** (0.011)	
$Netflow\_rate\ within_{i,t-2}$					0.002 (0.012)	
$Netflow\_rate\ between_{i,t-1}$						0.047* (0.025)
$Netflow\_rate\ between_{i,t-2}$						0.039** (0.020)
R-squared	0.68	0.68	0.68	0.68	0.69	0.69
Log Likelihood	1038.99	1038.37	1038.39	1036.42	1040.58	1039.95
Kleibergen-Paap F stat.	11334.80	6264.77	37225.75	150.26	41876.41	881.50
All Models: 2SLS FE; 320 obs.; year dummies; region fixed effects; instrumented mobility rates						

Note: clustered standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

2.5.4 Robustness checks

We conducted various checks to validate the robustness of the main results of this paper. The results of these robustness checks, not reported here, are available upon request from the authors.

The data used in the analysis include the recent financial crisis period. The shocks caused by the crisis might affect both the inventor flows and the TFP of Italian regions. To exclude the possibility that the relationships between inventor flows and TFP are driven by these shocks, new estimates are performed excluding the period 2009-2011. The obtained results are very similar to those discussed before.

To measure the regional TFP growth rates, we rely on an estimated measure of the regional capital stocks. To control for potential biases due to possible errors in the measurement of the capital stocks and, thus, of the TFP growth rates, new estimates are performed using as dependent variable the labor productivity growth rates.[18] The traditional high correlation between TFP and labor productivity is in support of this strategy. The estimates results are very similar to those discussed before.

In order to check for heterogeneous effects related to inventors' inventive performance, additional estimates are performed based on a recalculated mobility index using the past inventive productivity of each inventor as a weight. In particular, this paper adopts two alternative measures of inventor productivity, i.e. the inventor's patent depreciated stock and the average number of forward citations associated to the inventor's patents. The former are calculated using the perpetual inventory method with a depreciation rate of 15%. The latter calculated using temporal windows of 3 years to control for the well-known truncation bias problem (Hall et al., 2005). The results are similar to those discussed before in terms of significance levels, but the coefficient values of all the mobility indexes are lower. These results can be explained by a demographic selection. Mobile inventors are, in general, younger than non-mobile inventors and, consequently, the past productivity of the mobile inventors is lower than the past productivity of non-mobile inventors.

[18] Labor productivity is measured as the ration between the GDP and the number of full time equivalent workers

Another issue is whether the uncertainty about the exact date of inventor movement affects our estimate results. Our data shows that the mean value of the temporal lags between the origin region patent and the destination region patent is 2.4 years and that in 53% of the cases the observed temporal lag is 1 year or less. Moreover, the percentage of inventor movements increases to 77% if we consider a temporal lag of 3 years or less. These figures show that in most cases the move date is reasonably well observable. In addition, as a robustness check, we perform new estimates excluding cases of inventor movements for which the temporal lag between the origin and destination patents is bigger than a temporal threshold value. Using both a temporal value of 5 years and 3 years the obtained results are very similar to those discussed before, except that inventor inflows are significant at 10% level instead of 5% level when we use a temporal threshold value of 3 years.[19] Temporary movements, i.e. cases where the inventor moves from region i to region j and then back to region i are observed for 207 inventors (8.93% of the 2318 mobile inventors). Our inventor flow indexes are constructed considering only the last movement. Therefore, for example, if an inventor moves from Marche to Lombardy both in 1995 and in 2002, it is assumed that this inventor moves from Marche to Lombardy only in 2002. However, since the bias of the effect of temporary movements is only partially mitigated by the correction adopted to control for double movements, 2SLS FE estimates are performed excluding all cases of temporary movements. The estimates results are similar to those reported in the main text, except for the inventor inflows that are significant only for the 2 year lagged values. Moreover, inventor outflows have a stronger effect (-0.086 vs. -0.037).

2.6 Conclusion

This paper combines a new database on geographical mobility of patent inventors with estimates of the regional TFP in Italy in the 1996–2011 period. It uses an aggregate production function at the regional level to test the relationship between the geographical mobility if

[19] A question related to the uncertainty about the exact move date is that the mobility variable might capture different lags of the R&D or patent variables. As a further robustness check, two separate set of 2SLS FE estimates are performed including additional control variables. The first one includes the lagged values from t-4 to t-2 of both R&D per capita and patent per R&D expenditure. The second one includes the lagged values from t-3 to t-1 of the annual rates of growth of both R&D per capita and patent per R&D expenditure. The results (available from the authors upon request) are similar to those reported in the paper.

inventors and TFP, instrumenting mobility with the geographical distance between the origin and the destination, a common border effect and regional fixed effects. Even if caution is needed in the causal interpretation of the performed estimates because the considered aggregate framework does not provide a genuinely random variation of the mobility, the findings of this paper are new in the literature.

We find that inventor mobility across regions in Italy is significantly associated with TFP growth. In particular, a 1% increase in the inflow of inventors in a region increases TFP by 2.5%. In parallel, a 1% increase in the outflow of inventors in a region decreases TFP by 3.7%. These correlations are robust to including several control variables (such as R&D expenditure per capita, the ratio of patents to R&D expenditure and the population density). The coefficients from the 2SLS estimates imply that the net flow of inventors is correlated with significant productivity gains for the receiving regions and a significant productivity loss for the sending regions.

In addition, the results of this paper show heterogeneous effects for “within applicant” and “between applicants” movements. The effect of “between applicants” inventor inflows on TFP take more time to materialize than the “within applicant” inventor inflows. Moreover, the negative effects on TFP by inventor outflows are driven by “between applicants” movements.

These results contribute to shed light on the economic effects of the movement of the regional communities in Italy that transfer their own tacit knowledge and social capital to receiving regions and contribute to task specialization, innovation and, possibly, entrepreneurship. It also raises a warning flag for the sending regions bearing the ramifications of the loss in human capital. In doing so, the results of this paper add evidence to the effect of brain drain in Italy where a substantial and increasing portion of geographically mobile workers are tertiary educated (Becker et al. 2004).

Finally, more generally, the results of this paper highlight that diaspora effects arise also within a country. Comparing to international migrations, inter-regional migrations might be of greater magnitude because of the lower barriers faced and, thus, resulting in a major geographical reallocation of the human capital. This spatial heterogeneous redeployment of human capital might generate different regional growth patterns within the same country depending on the ability of regions to maintain and attract high-skilled people. Cumulative

processes might be engendered where the richer regions benefit from the inflow of high skilled people attracted by the higher wages of these regions, which in turn entails a greater demand of high-skilled people and a greater investment in knowledge-intensive activities. Conversely, the poorer regions might enter a vicious circle characterised by outflow of high-skilled people, decrease in demand of high-skilled people and in knowledge-intensive activities (Faggian and McCann, 2009). Consequently, inter-regional mobility might engender regional divergence.

2.A Appendix: TFP of Italian regions

To calculate the TFP of Italian regions for the period 1996-2011, we follow a standard growth accounting approach (Jorgenson, 1995; OECD, 2011; Solow, 1957). The starting point is a Cobb-Douglas production function with constant return to scale

$$GDP_{i,t} = TFP_{i,t} Capital_{i,t}^{1-\beta} Labor_{i,t}^{\beta} \tag{2.4}$$

where $GDP_{i.t}$ is the GDP of region *i* at time *t*. Capital and labor are the two input factors considered, and $1-\beta$ and β are, respectively, the GDP elasticity of capital and the GDP elasticity of labor. We can calculate TFP growth rates ($ln\left(TFP_{i,t}/\ TFP_{i,t-1}\right)$) and the annual TFP levels ($TFP_{i,t}$) of regions by deriving the following equations:

$$ln\left(\frac{TFP_{i,t}}{TFP_{i,t-1}}\right) = ln\left(\frac{GDP_{i,t}}{GDP_{i,t-1}}\right) - \left(1-\bar{\beta}\right) ln\left(\frac{Capital_{i,t}}{Capital_{i,t-1}}\right) - \bar{\beta} ln\left(\frac{Labor_{i,t}}{Labor_{i,t-1}}\right) \tag{2.5}$$

$$TFP_{i,t} = \frac{GDP_{i,t}}{\left(Capital_{i,t}^{1-\beta} Labor_{i,t}^{\beta}\right)} \tag{2.6}$$

A basic problem on the calculation of equations (2.5) and (2.6) arises because of the lack of data on regional stock capital (Capital). At regional level, ISTAT provides only data on regional fixed investments for the period 1995-2011. Unfortunately, the length of these data on regional investments does not allow us to construct an accurate estimate of the gross and net capital stocks independently. Hence, we use these short regional time series on fixed investment to build a partial approximation of the regional capital stock using the perpetual inventory method. Secondly, we use this partial approximation to build regional shares. Third, we apply the regional shares to the complete data available at the national level (see Maffezzoli, 2006; Quatraro, 2009).

The first step is to calculate the partial initial stock of capital at the regional level:

$$PC_{i,1995} = \frac{I_{i,1995}}{g_i + \delta_{1995}} \tag{2.7}$$

where $I_{i,1995}$ is the real investment in 1995, g_i is the average growth rate of fixed investments in Italian regions and δ_{1995} is the national depreciation for capital stock in 1995. We assume a constant depreciation rate across all regions equal to 0.0048, as regional rates are not available.

Then, we calculate the regional partial capital stock for the period 1996-2011:

$$PC_{i,t} = PC_{i,t-1}(1 - \delta_t) + I_{i,t} \tag{2.8}$$

Thus, for each region i, we calculate the share of the partial capital stock in 2011:

$$S_{i,2011} = \frac{PC_{i,2011}}{\sum_j PC_{j,2011}} \tag{2.9}$$

We use these shares to redistribute the net national capital stock NC, provided by ISTAT for 2011, on the 20 regions:

$$C_{i,2011} = S_{i,2011} NC_{2011} \tag{2.10}$$

Finally, we follow the procedure outlined in the literature (Maffezzoli, 2006; Quatraro, 2009) to extend the series before 2011:

$$C_{i,t-1} = \frac{C_{i,t} - I_{i,t}}{1 - \delta_t} \tag{2.11}$$

3 Knowledge Creates Markets: The Influence of Entrepreneurial Support and Patent Rights on Academic Entrepreneurship

Dirk Czarnitzki [a,b], Thorsten Doherr [b,c], Katrin Hussinger [c,b,a], Paula Schliessler [a,b] and Andrew A. Toole [d,b]

a) KU Leuven, Dept. of Managerial Economics, Strategy and Innovation, Leuven, Belgium
b) Centre for European Economic Research (ZEW), Mannheim, Germany
c) University of Luxembourg, Luxembourg
d) US Patent and Trademark Office, Washington D.C., United States

Abstract We use an exogenous change in German Federal law to examine how entrepreneurial support and the ownership of patent rights influence academic entrepreneurship. In 2002, the German Federal Government enacted a major reform called Knowledge Creates Markets that set up new infrastructure to facilitate university-industry technology transfer and shifted the ownership of patent rights from university researchers to their universities. Based on a novel researcher-level panel database that includes a control group not affected by the policy change, we find no evidence that the new infrastructure resulted in an increase in start-up companies by university researchers. The shift in patent rights may have strengthened the relationship between patents on university-discovered inventions and university start-ups; however, it substantially decreased the volume of patents with the largest decrease taking place in faculty-firm patenting relationships.

Keywords: Intellectual property, patents, technology transfer, policy evaluation

JEL: O34, O38

Acknowledgements We are grateful for the funding of this research project by the Centre for European Economic Research (ZEW) within the research program "Strengthening Efficiency and Competitiveness in the European Knowledge Economies" (SEEK). The views expressed in this article are the authors' and do not necessarily represent the views of the United States Patent and Trademark Office.

3.1 Introduction

Based on the belief that academic research is an important driver of economic growth and the perception that academic institutions should have an entrepreneurial mission beyond teaching and research, policymakers are increasingly interested in stimulating entrepreneurial behaviors among academic researchers. The idea is to change the incentives researchers face so that entrepreneurial choices are more attractive. Numerous policy levers are available including tax policies, employment policies, subsidies, entrepreneurial education, and intellectual property (IP) policies.

In the area of IP policies, the United States has become the de facto leader. In 1980, the Bayh-Dole Act facilitated institutional ownership of inventions discovered by researchers who were supported by federal funds. Many observers credit the Bayh-Dole Act with spurring university patenting and licensing that, in turn, stimulated innovation and entrepreneurship (The Economist, 2002; OECD, 2003; Stevens, 2004). With this success, the Bayh-Dole Act has become a model of university IP policy that is being debated and emulated in many countries around the world including Germany, Denmark, Japan, China, and others (OECD, 2003; Mowery and Sampat, 2005; So et al., 2008).

But how do intellectual property rights (IPRs) influence the incentives for university researchers to form start-up companies? Perhaps surprisingly, this question has not received much attention in either the theoretical or empirical literatures. From a theoretical point of view, Damsgaard and Thursby (2013) examined the mode and success of commercialization under an individual ownership system (i.e. the academic inventor keeps the patent rights) and a university ownership system. In a number of cases, their model shows less faculty entrepreneurship (i.e. fewer faculty start-ups) under university ownership. Using survey and case study evidence, Litan et al. (2007) and Kenney and Patton (2009) argued that conflicting objectives and excessive bureaucracy make university ownership ineffective and suggest an individual ownership system may be superior. In a follow-on study looking at technology-

based university spin-offs, Kenney and Patton (2011) found suggestive evidence that an individual ownership system is more efficient for generating spin-offs.[20]

In this paper, we use an exogenous change in German Federal law to examine how entrepreneurial support and the ownership of patent rights influence academic entrepreneurship.[21] The new German policy strengthened the institutional and financial support for academic start-ups and fundamentally changed who owns the patent rights to university-discovered inventions. Prior to 2002, university professors and researchers had exclusive intellectual property rights to their inventions. This "Professor's Privilege" allowed university researchers to decide whether to patent or not and how to commercialize their discoveries. After 2002, universities were granted the intellectual property rights to all inventions made by their employees and this shifted the decision to patent from the researchers to the universities.

Based on a novel researcher-level panel database that includes a control group not affected by the IP policy change, we find no evidence that the new infrastructure resulted in an increase in start-up companies by university researchers. The shift of patent rights to the universities not only changed the ownership distribution, but also affected the volume of patents on university-discovered inventions. The policy reform may have strengthened the relationship between patents on university-discovered inventions and university start-ups (i.e. increased the marginal impact of university-owned patents on university start-ups); however, it substantially decreased the volume of patents with the largest decrease taking place in faculty-firm patenting relationships. By displacing so many faculty-firm relationships, our evidence suggests the policy reform probably decreased overall university technology transfer.

---

[20] In a recent working paper, Astebro et al. (2016) compare entrepreneurship between the Bayh-Dole system in the U.S. and Sweden's faculty ownership system. Their analysis finds that Swedish academics are twice as likely to enter entrepreneurship, but average earnings deteriorate for academic entrepreneurs in both countries after founding a new company.

[21] Academic entrepreneurship is defined as the formation of a new company in which the university researcher is part of the founding team. This includes *all* university researcher start-ups – those that license university technologies and those that do not license (Toole and Czarnitzki, 2007; Kenney and Patton, 2011; Czarnitzki et al., 2015).

The remainder of the paper is as follows: the next section reviews the German policy reform, develops our conceptual background using the literature and states the hypotheses to be tested. The third section describes the empirical identification strategy and introduces the data. Section 4 discusses the econometric results and the fifth section concludes.

3.2 Background and hypotheses

In 2002, the German Federal Government introduced a major reform called Knowledge Creates Markets to stimulate technology transfer from universities and other public research organizations to private industry for innovation and economic growth. The program was largely a reaction to the "European paradox" (European Commission, 1995). At that time, policymakers believed that Germany had one of the world's leading scientific research enterprises, but was lagging the United States in terms of technology transfer and commercialization. The new program addressed four broad areas of science-industry interactions including the processes and guidelines governing knowledge transfer, science-based new firms, collaboration, and the exploitation of scientific knowledge in the private sector.

One part of the Knowledge Creates Markets reform created new institutions with new financing to facilitate the movement of university research to the private sector. Unlike most of Germany's public research organizations (PROs)[22], German universities had little experience undertaking technology transfer activities, and only a few universities maintained professionally managed technology transfer offices (TTOs) (Schmoch et al., 2000). The government established regional patent valorization agencies (PVAs) that were supported with a budget of 46.2 million EURO (Kilger und Bartenbach, 2002). Universities were free to choose whether to use the PVAs' services or not. To date, 29 PVAs serve different regional university networks and employ experts specialized in these universities' research areas. The

[22] In addition to universities, Germany's research enterprise includes other public research institutions that have many branches in a variety of different scientific disciplines. For instance, the Fraunhofer Society has 59 institutes in Germany with about 17,000 employees, the Max Planck Society has 76 institutes with about 12,000 employees. The Leibniz Association employs 16,100 people in 86 research centers. The Helmholtz Association has about 30,000 employees in 16 research centers.

PVAs support the entire process from screening inventions, finding industry partners, and determining fruitful commercialization paths, including the formation of faculty start-up companies.

While the PVAs were intended to fill a void in the institutional structure supporting commercialization of university research, the reform also called for the expansion of Federal subsidies to university-specific TTOs. Among other initiatives, the legislation included vocational training for university and PRO administrative staff on intellectual property and innovation management, financial assistance to offset the costs of university patent applications (application and counselling fees), and subsidies for early stage entrepreneurial activity such as business plan development.

The idea that more support services through the PVAs and subsidies to university TTOs could stimulate more technology transfer and academic entrepreneurship finds mixed support in the scholarly literature. One strand of the literature investigates how the presence of a TTO, its resources and its capabilities influence technology transfer indicators such as licenses and spin-off companies. For instance, Siegel et al. (2003) found that the number of TTO staff was positively associated with the number of licensing agreements based on a sample of US universities. For university spin-offs created through licensing, Di Gregorio and Shane (2003) found that specific TTO policies such as inventor royalty rates and the willingness to make equity investments were important. Lockett and Wright (2005) added TTO business development capabilities as a further factor. The development of these capabilities depends on the experience and skill level of the TTO staff (see Grimaldi et al., 2011 and the literature reviews by Rothaermel et al., 2007; O'Shea et al., 2008; Bradley et al., 2013; Kochenkova et al., 2015).

On the other side, several studies identify problems with TTOs as intermediaries, which suggests additional infrastructure and financing may not spur entrepreneurship. Litan et al. (2007) suggest TTOs are misguided due to an overemphasis on revenue maximization and centralization. Kenney and Patton (2009) believe TTOs are ineffective due to bureaucratic problems, informational limitations and misaligned incentives. Using survey data, Siegel et al. (2004) found that 80% of managers and 70% of scientists at US research universities cited

bureaucracy and inflexibility as barriers. Based on European data, Clarysse et al. (2007) found TTOs play only a marginal, often indirect role, in spurring academics to start new companies.

Although the results in the scholarly literature are mixed, the following hypothesis is based on what policymakers expected:

H1: Infrastructure and financing support provided through the Knowledge Creates Markets reform stimulated university start-up companies

Beyond the infrastructure and financing, the Knowledge Creates Markets reform included one of the most significant changes from both a legal and cultural perspective: the abolishment of Professor's Privilege. Professor's Privilege originated from Article 5 of the German constitution that protects the freedom of science and research. The new program repealed Clause 42 of the German employee invention law that had granted university researchers - as the only occupational group in Germany - the privilege to retain the ownership rights to their inventions that otherwise rest with the employer.

During the Professor's Privilege era most of the responsibility for university technology transfer was in the hands of German professors and patents played an important role.[23] Patenting provided the legal means for negotiating and partnering with private firms to pursue development and commercialization, especially as most academic discoveries are early-stage or "embryonic" (Colyvas et al., 2002; Jensen and Thursby, 2001). Through this process, most German professors gave up their IP to firms, but they also established relationships that involved the exchange of technology with some sort of compensation (pecuniary and/or non-pecuniary). In other words, university-industry technology transfer in Germany had evolved over time into a fairly extensive network of faculty-firm interactions. Presumably, most of these relationships were bilateral in the sense that the universities were not legal partners and did not receive any financial compensation.

[23] University patents are one mechanism for transferring academic research results to the market. Other mechanisms include collaborative and contract research, licensing, networking, publications and so forth (Grimaldi et al., 2011).

Also, by owning the patent rights, university researchers could leverage the advantages of patents for creating start-up companies. Hsu and Ziedonis (2013) suggest patents have a "dual function." Beyond knowledge protection, patents may be an important device for reducing asymmetric information and signaling the "quality" of the venture and thus expected returns of the business idea to potential lenders, which provides easier access to finance (Conti et al., 2013; Haeussler et al., 2011; Graham et al., 2009; Audretsch et al., 2013). Similarly, Shane (2001) argues that patents are disproportionately important to independent entrepreneurs who lack complementary assets. Clarysse et al. (2007) confirm that patents increase the initial funding that university start-ups raise. Levin et al. (1987) state, that "[…] for small, start-up ventures, patents may be a relatively effective means of appropriating R&D returns, in part because some other means, such as investment in complementary sales and service efforts may not be feasible. The patents held by a small, technologically oriented firm may be its most marketable asset" (Levin et al., 1987, p. 797).[24]

In the current era without Professor's Privilege, German university researchers are required to cull their research findings for inventions and report any inventions to the university – unless the researcher decides to keep his or her inventions secret by not publishing or patenting. The university has four months to consider any submitted inventions for patenting. If the university does not claim the invention, the rights to pursue patenting and commercialization are returned to the researcher. If the university does claim the invention, the inventor receives at least 30% of the revenues from successful commercialization, but nothing otherwise. Furthermore, the university handles the patenting process and pays all related expenses such as processing fees, translation costs and legal expenses. University researchers retain the right to disclose the invention through publication two months after submitting the invention to the university. Prior contractual agreements with third parties also remained valid during a prescribed transition period.

[24] Graham and Sichelmann (2008) and Graham et al. (2009) conduct a comprehensive investigation of the various expected benefits of patents for technology foundations. They conclude that protection against imitation and easier access to finance are the main reasons for start-ups to patent (Graham et al., 2009). Other functions of patents of almost an equal importance include an improved likelihood and value of an IPO or acquisition, a stronger reputation, a better negotiation position with other companies, the prevention of IP suits and licensing revenues (Graham et al., 2009).

The abolishment of Professor's Privilege created a complex situation regarding the incentives to form start-up companies. It took the initial patenting and commercialization decisions away from the researchers and gave them to the universities. The researcher became secondary to the university TTO in the search, negotiation, partnering with private firms, and forming start-ups. Individual researchers, however, remained the primary decision makers regarding the formation of start-up companies. The critical issue is how the loss of patent rights changed the researchers' costs and benefits associated with the decision to found a start-up company.[25]

University ownership of the patent rights could strengthen the relationship between patents and the formation of start-ups if, for patented technologies, university ownership lowered "entry" costs for starting a company and/or increased expected returns. This seems to be the outcome German policymakers had in mind. They argued that academic researchers were so resource constrained that the costs of patenting and the market uncertainty surrounding the potential value of discoveries were limiting commercialization. Prior to the reform private firms owned most patents on university-discovered inventions. Researchers gave up their patent rights to industry partners as part of a *quid-pro-quo*, but this meant they lost the opportunity to form start-up companies based on those discoveries. With the university as the primary patent owner, a researcher could regain patent rights if the university does not claim the invention or if the university decides to license the discovery back to the researcher, making it easier for faculty members to found new companies.[26] Moreover, the university TTOs and regional PVAs perform various kinds of services such as market value assessment before patenting (Debackere and Veugelers, 2005). These services may increase the expected return on a discovery by decreasing the uncertainty about its potential value and thereby stimulate more start-up companies.

[25] This only applies to start-ups that are based on patented technologies. For those that do not rely on patents, the abolishment of Professor's Privilege is irrelevant and any effect of the reform on non-patent start-ups is captured in hypothesis #1.

[26] Hellman (2007) found this will happen in cases where the researcher is more efficient than the TTO at searching for an industry partner. In his model, a spin-off is an alternative mechanism for organizing the search for an industry partner.

H2: The relationship between university start-ups and university owned patents became stronger following the Knowledge Creates Markets reform (i.e. increased the marginal effect of patents on the number of start-ups).

Even if the strength of the relationship between patents and start-ups increased, the effect on the total number of start-ups depends indirectly on the level of patenting in the post-reform era. Prior work has found the Knowledge Creates Markets reform decreased the volume of patents in university-discovered inventions (Czarnitzki et al., 2015; Von Proff et al., 2012). This effect was primarily due to heterogeneity among university researchers in the costs of patenting, which was reflected in the patent ownership distribution. For instance, under Professor's Privilege, academic researchers who maintained a well-functioning network with industry partners had relatively low costs of patenting by assigning the IPRs directly to industrial partners, but had to forego starting a company on those inventions. After the reform, patenting costs increased as the new university-ownership of the IP disrupted the existing ties between academic inventors and industry (Czarnitzki et al., 2015), but start-ups became a new possibility. Those academic researchers without industry partners had relatively high patenting costs before the reform. Afterward, both the costs of patenting and the costs of starting a company may be lower for these researchers. Overall, the impact of the reform on the formation of researcher start-ups will reflect these two effects.[27]

H3: The net effect of the Knowledge Creates Markets reform on the number of start-ups is determined indirectly by the change in the volume of patents.

[27] Three recent studies use a different framework than we present above, but suggest the net impact of the reform will be fewer spin-offs. Damsgaard and Thursby (2013) consider both regimes using a theoretical model that incorporates the need for continued inventor effort in development. They found the university ownership leads to less entrepreneurship if established firms have some advantage in commercialization. Kenney and Patton (2011) compared inventor versus university ownership using data on technology-based spin-offs from six universities. The University of Waterloo, which was the only university with inventor ownership, matched University of Wisconsin Madison and exceeded the other US universities even though it had less research and development support and fewer faculty members. The authors point to ineffective incentives, information asymmetries, and contradictory goals as the primary reasons university ownership produces fewer spin-offs. Hvide and Jones (2016) found a 50% decline in faculty start-ups and patenting after the abolishment of Professor's Privilege in Norway.

3.3 Empirical model and data

3.3.1 Identification strategy and estimation approach

The Knowledge Creates Markets reform provides a unique opportunity to analyze how policy initiatives influence academic entrepreneurship. The changes in technology transfer support and the new IP ownership rules outlined above were targeted primarily at university-discovered inventions. To identify the policy effects, we use a difference-in-difference (DiD) research design with university inventors as the treatment group and PRO researchers as the control group. Like university professors, PRO researchers conduct academic research at publicly funded institutions in Germany. They work in similar academic fields and experience similar changes in research opportunities that affect the discovery of new knowledge. But unlike university professors, PRO institutions already had a strong technology transfer infrastructure and the patent rights to the inventions by PRO researchers were always owned by the institution. Our researcher-level DiD setup accounts for common macroeconomic trends and individual-specific unobserved effects that capture an academic inventor's "taste" for patenting and entrepreneurship.

Academic entrepreneurship is measured as the number of firm foundations by academic inventors per year. Note that we deliberately label the dependent variable as start-ups as we will measure all firm foundations by academic inventors in the empirical study and not only those that went through the university (or PRO) TTOs, which are commonly labeled as spin-offs.

$$Startups_{it} = f\left[\begin{array}{c}\beta_o + \beta_1 Prof_i NewPol_t + \sum_{j=1}^{J} \beta_{2j} Prof_i NewPol_t Pat_{ijt} + \\ \sum_{j=1}^{J} \beta_{3j} Pat_{ijt} + \beta_4 3yAvgPubs_{i,t-1} + \delta_i + \gamma_t\end{array}\right] + \varepsilon_{it} \qquad (3.1)$$

The direct impact of the reform is captured by the coefficient β_1 of the interaction term $(Prof \cdot NewPolicy)$. $Prof$ is a dummy variable that takes the value of 1 when the inventor is a university professor and 0 when the inventor is a PRO researcher. $NewPolicy_t$ is a dummy variable that takes the value of 1 following the policy change, 2002 onward, and 0 otherwise. We use a three year moving average of past research publications, $(3yrAvgPubs)_{i,t-1}$, to

capture the arrival of new knowledge. δ_i is a researcher-level fixed effect and γ_t is a full set of annual time dummy variables.[28] Note that the professor dummy variable gets absorbed into the researcher fixed effects. Similarly, the annual time dummy variables absorb the new policy dummy variable.

In addition to the direct impact of the reform, we are interested in how the abolishment of Professor's Privilege changed the relationship between university start-ups and patents on university-discovered inventions (hypothesis #2). To test this, we include the variable *PAT* and its interaction with $(Prof \cdot NewPolicy)$. As the coefficient on *PAT* shows the strength of the relationship before the reform, a positive and significant coefficient on $(Prof \cdot NewPolicy \cdot PAT)$ would indicate the relationship became stronger.

Notice that equation (3.1) includes summation operators over the index j on the explanatory variable *PAT*. This index captures ownership types for patented academic inventions. We classified patents on university and PRO-discovered inventions into three ownership types (J=3): industry, employer institution (university/PRO), and personal (i.e. held by the individual). This was accomplished by manually reviewing the list of applicants and coding the records. Also, for notational simplicity, we are using the variable *PAT* to represent patent counts and citation-weighted patents. As will be clear in the discussion of the results, we use citation-weighted patents in some specifications.[29]

In the results section, we present two versions of equation (3.1) in separate tables. First, we look at the overall effect indicated by aggregating all patents and ignoring the variation by ownership type. This will test whether the relationship between start-ups and patents became stronger overall. In a separate set of regressions, we implement a more flexible specification that estimates separate coefficients for employer-owned (e.g. university) and

[28] Note that the literature on life cycle models of researcher productivity often includes career age of the researchers and the square of career age in regression specifications (Diamond, 1986; Levin and Stephan, 1991; Turner and Mairesse, 2005; Hall et al. 2007). As we estimate fixed effects regressions, the model would be fully saturated with the fixed effects, the full set of time dummies and career age. Thus, we do not include the variables career age and its square as regressors; career age is included implicitly by the time dummies in combination with the fixed effects.

[29] We weight patents by the number of citations received over a four year window following application. To avoid dropping patents with zero citations, the citation-weighted patents are constructed as (patents + citations).

personal-owned patents in the post-reform period. This allows us to investigate whether the strength of the relationship increased for these ownership types.

Intuitively one might expect that the start-up equation (equation 1) would be modeled as a binary choice. However, a few researchers are involved in multiple firm foundations in some years. Therefore, the variable *startup* becomes a count variable and not a dummy variable. Consequently, we estimate eq. (3.1) using a fixed effects Poisson quasi-maximum likelihood estimator (QMLE). As a member of the linear exponential family of distributions, the Poisson QMLE produces consistent estimates of the population parameters as long as the conditional mean is correctly specified (Gourieroux et al. 1984; Wooldridge 1999). Consequently, the function *f* is chosen to be the exponential function in the Poisson regression. We use robust standard errors clustered at the researcher-level. As a robustness test, we also estimate conditional fixed effects logit regressions where the link function is logistic instead of exponential.

It is possible that the number of patents is endogenous in the firm foundation equation. For instance, unobserved market opportunities could influence the decision to found a new firm and be correlated with the decision to seek patent protection. We would like an instrumental variable that influences patent protection, but is unrelated to the market opportunities facing the academic founder. Aggregate patent trends in the United States (US) are attractive instruments because they are arguably exogenous to the firm foundation decision by German academic entrepreneurs, but correlated through broader technology trends. We decided to use the growth rate of US patents by technology class. Higher growth in US patents within a technology area indicates the technology area is increasingly crowded. As more patents crowd a given technology space, costs of patenting exogenously increase. As long as the growth in US patents within technology areas is not related to the error term in the start-up equation for German professors, the IV is exogenous.

We implemented the robust endogeneity test recommended by Wooldridge (2010, p. 742) for count data models. The instrument was constructed using the 35 technology fields according to the Fraunhofer technology classification and linked to each researcher according to his/her main field of activity. The growth rate was defined over the past three years as: [(USPAT(t) – USPAT(t-3)) / USPAT(t-3)]. For the first-stage regression, which is a linear model

with fixed effects, the F-statistic on the growth of US patents was 13.86, p-value < 0.001. In the second-stage explaining start-ups, the residuals were insignificant with a z-statistic of 0.25 and a p-value = 0.803. Based on these results, we do not consider patent as endogenous in our subsequent models.[30]

As outlined in Section 2, the overall effect of the policy on entrepreneurship also depends on how the reform influenced the volume of patents on university-discovered inventions. To investigate this indirect impact, we follow prior work by Czarnitzki et al. (2015) and use a DiD setup for the volume of academic patents. These DiD models take the for

$$Pat_{ijt} = \mathrm{g}\left[\beta_o + \beta_1(Prof_i NewPol_t) + \beta_2 3yAvgPubs_{i,t-1} + \beta_3 z_{i,t-1} + \delta_i + \gamma_t\right] + \varepsilon_{it} \qquad (3.2)$$

where the notation is as above in eq. (3.1) and *z* stands for the vector of instrumental variables as described above.

As patent counts take only nonnegative integer values, we use the fixed effects Poisson quasi-maximum likelihood estimator (QMLE) again, i.e. the function *g* is chosen to be the exponential function. We use robust standard errors clustered at the researcher-level.

3.3.2 Data and descriptive statistics

The relevant population of researchers includes all academic inventors affiliated with a university or PRO and appeared as an inventor on at least one patent submitted to the German or European Patent Offices between 1978 and 2008. Academic inventors are a subpopulation of all academic researchers in Germany. The broader population includes academic researchers who only published. The core of the Knowledge Creates Markets reform, however, was the abolishment of Professor's Privilege and this did not affect researchers who never participated in the intellectual property system over the entire period.[31]

[30] Note that we experimented also with specifications where we additionally used patent applications at the Japanese Patent Office (JPO) as instrumental variables in addition to the US variable. However, these specifications did not improve or change any result.

[31] As noted by a referee, the population of academic researchers who patent is not representative of all academic researchers. The policy reform may have had indirect effects that are not fully captured with our data. One should keep this limitation in mind when interpreting the results.

We constructed a researcher-level panel dataset of academic inventors following a multistep procedure. In addition, we searched for all of these inventors in the "Mannheim Enterprise Panel," a database containing all German firm foundations and detailed information on the founding persons. Appendix A summarizes the data compilation. This process yielded a sample with 17,417 university and 35,353 PRO researcher-year observations.[32] We defined the study period to extend from 1995 through 2008 so that we observed enough years before and after the policy change. For each inventor, our data contain the individual's history of patenting between 1978 and 2008 and the individual's history of publications between 1990 and 2008. Each researcher enters the panel when we observe either the first patent application or the first publication. The researcher stays in the panel for a maximum of 35 career years after which we assume the researcher retires. To account for earlier exit, we adopted a 5-year rule that has a researcher leaving the panel if he or she had no patenting or publishing activity for five consecutive years. The estimation sample contains 52,770 researcher-year observations corresponding to 1,946 different university researchers and 4,551 PRO researchers.[33]

In total, the sample contains 1,030 start-ups founded between 1995 and 2008 by the researchers in the sample. Thus, most of the 52,770 researcher-year observations in the sample have a value of zero (98.4%). In some cases, researchers formed more than one start-up in a given year. In the sample, we have 674 observations (1.3%) where a single start-up was founded by a researcher in a given year; 127 cases (0.2%) where two start-ups were formed, and in about 0.05% of the cases, more than two start-ups were formed (with the maximum being five).[34]

[32] This sample excludes those researchers employed at both, PRO and university, as it is not clear which patent regime applied to these researchers. Furthermore, we had to drop persons with very common German names to ensure clean matches across the patent, publication and firm foundation databases. See Appendix A for more details.

[33] Note that our sample is smaller than the one used by Czarnitzki et al. (2015). This is because we had to drop some common inventor names when linking the inventors to the firm foundation data.

[34] We checked the right tail of the distribution manually and the data are correct. Some exceptional researchers apparently build a small portfolio of different start-up companies at a certain point in their careers.

Table 3.1: Academic entrepreneurship and patents before and after the 2002 policy reform (annual mean values, 1995-2008)

		Start-ups per year	Patents per year
University researcher	Before 2002	46.43	755.86
	After 2002	42.57	397.43
PRO researcher	Before 2002	29.43	1230.00
	After 2002	28.71	1132.43

Note: The sample of patenting university researchers comprises 1,946 different inventors and the sample of PRO researchers amounts to 4,551 people. The 1,946 university researchers were, on average, involved in about 46 start-ups and 1,312 patents per year before 2002 and these numbers dropped to about 43 start-ups and 1,177 patents per year after the law change. The numbers for PRO researchers read equivalently.

Table 3.1 gives the first indication of how the policy reform influenced academic start-ups and patenting. It shows the number of researcher-founded start-ups, university-discovered and PRO-discovered patented inventions before and after the reform. Looking at the third column, the number of start-ups decreased for university and for PRO researchers after the reform. This is the opposite of what policymakers expected and casts doubt on hypothesis #1. Column four shows the average annual number of patents decreased for both university discoveries and PRO discoveries. However, the patenting activity of university researchers fell much more dramatically following the reform. This suggests that the abolishment of Professor's Privilege did not stimulate university patenting, however, the strength of the relationship between patenting and start-ups may have increased. This will be investigated in the subsequent econometric models.

Recall, the reform was a fundamental change in the ownership structure for university-discovered inventions. Its impact on university start-ups will depend in part on how the ownership distribution on patented university discoveries changed. For instance, when private firms hold the patent rights, researchers have limited opportunities to use these inventions for start-up companies. Industry firms are unlikely to support new companies that may be competitors in their technology space.

Table 3.2 shows the average number of patents on university-discovered inventions by ownership type before and after the reform. In line prior results reported in Czarnitzki et al. (2015), we see the overall decrease in patented university inventions. Before the policy change, the university inventors filed on average 0.58 patents per researcher per year, and

this number drops to 0.34 patents after the policy change (see bottom row labeled "Total" in Table 3.2). For the pre-reform period, the first row shows the extent of faculty-firm bilateral interactions before the reform. Industry applicants owned an average of 0.45 patents per researcher per year. After the shift to university ownership, industry ownership was cut in half to 0.23 on average. This decrease may reflect higher transaction costs after the reform as university TTOs interrupted these bilateral relationships. Even at this much lower level, faculty-firm relationships still accounted for the majority of university-invented patents after the policy change (62%). Personal-owned patents also fell from 0.14 to 0.04 per researcher per year after the abolishment of Professor's Privilege. In contrast, university-owned patents increased from 0.02 to 0.10 per researcher per year and accounted for 27% of all patents afterward. The econometric models will show how these ownership changes affected university start-ups.

Table 3.2: University-discovered patented inventions by applicant type before and after the 2002 policy reform (mean values, 1995-2008)

	Before 2002		After 2002	
Industry applicant	0.45	74%	0.23	62%
Personal applicant	0.14	23%	0.04	11%
University applicant	0.02	3%	0.10	27%
Sum	0.61	100%	0.37	100%
Total	0.58		0.34	

Note: An applicant is equivalent to a US patent assignee. The total row is not the sum of the cells of the columns because some patents are co-applications of different owner types (e.g. industry and personal). In these cases, we counted the patent for all owners (instead of applying fractional counting) as each of them maintains unrestricted disposal rights (unless specific contracts over-rule the default rights). The control group of PRO researchers is omitted as the law change in 2002 did not apply to them. See the descriptive statistics in Appendix 3.B for more information.

More detailed descriptive statistics of the sample employed in the following regressions are presented in Appendix 3.B.

3.4 Econometric results

Using the scientist-level DiD research design, we begin with a baseline evaluation of the Knowledge Creates Markets reform. Table 3 shows the regression results explaining the number of university/PRO start-ups using fixed effects Poisson QMLE as well as conditional

fixed effects logit regressions. Models 1 and 2 use a count of total patents on academic discoveries while models 3 and 4 use patents weighted by forward citations (a form of quality adjustment).

In the recent applied econometric literature, scholars have raised some doubts about the validity of standard errors in common DiD regressions that estimate treatment effects of policy reforms. Typically, relatively long panels are used and the policy reform variable is just a dummy that switches from 0 to 1 for the treatment group and then remains at the value 1. As this is a regressor not varying a lot across the sample observations, scholars have been concerned about biased standard errors, particularly referring to the Moulton bias and to serial correlation problems (Moulton 1990). Therefore, we conducted a number of robustness tests where we follow the discussions in Bertrand et al. (2004) and Angrist and Pischke (2009). First, we tested for autocorrelation by estimating a linear fixed effects within regression model with AR(1) disturbances and calculated the Bhagarva et al. (1982) Durbin-Watson tests as well as the Baltagi-Wu (1999) tests. All test were always close to the value 2, indicating that no auto correlation is present. This is in line with our expectations as start-up creation at the level of the individual researcher is an intermittent activity rather than a persistent one. Furthermore, we calculated cluster-bootstrapped standard errors using 400 bootstrap replications. These were always close to the analytical cluster-robust standard errors. Because of space limitation, we do not show all these numbers. In what follows, the results of the fixed effects Poisson model using analytical cluster-robust standard errors are reported, and for the conditional fixed effects logit models bootstrapped standard errors are shown.

The difference-in-difference regressions in the context of treatment effects estimation are based on the assumption that before the intervention the treatment and the control groups show similar trends in the dependent variable. See Appendix C for a discussion of the common trend assumption. Statistical tests do not reject the hypothesis that the start-up variable shows a common trend in the pre-treatment period for university researchers and PRO researchers.

Turning back to the results in Table 3.3, policymakers expected the reform to increase the number of start-ups by university researchers due to infrastructure and financing support as stated in hypothesis #1. Looking across all four models, the variable (*Prof*Newpolicy*) is not

statistically significant in any model. The new PVAs and the additional support for university TTOs did not produce an increase in the number of university researcher start-ups above PRO researcher start-ups. In fact, from the descriptive statistics in the last section, we saw that start-ups among both groups declined following the reform.

Table 3.3: Regression on academic entrepreneurship (aggregate patents)

Startup	Model 1	Model 2	Model 3	Model 4
QMLE Poisson FE				
Prof•NewPolicy	−0.036	−0.006	−0.032	−0.012
	(0.174)	(0.180)	(0.175)	(0.180)
Patents	0.155***	0.123***		
	(0.030)	(0.032)		
Prof•NewPolicy•Patents		−0.057		
		(0.066)		
Patents cited			0.042***	0.045***
			(0.012)	(0.013)
Prof•NewPolicy•PatentsCit				−0.025
				(0.029)
Avg. Pubs	0.018	0.019	0.020	0.021
	(0.013)	(0.013)	(0.013)	(0.013)
Conditional FE logit				
Prof•NewPolicy	0.041	0.043	0.044	0.053
	(0.171)	(0.181)	(0.176)	(0.181)
Patents	0.097***	0.098***		
	(0.030)	(0.032)		
Prof•NewPolicy•Patents		−0.005		
		(0.074)		
Patents cited			0.033**	0.035**
			(0.013)	(0.014)
Prof•NewPolicy•PatentsCit				−0.013
				(0.039)
Avg. Pubs	0.024*	0.024*	0.025*	0.025*
	(0.013)	(0.013)	(0.013)	(0.014)
Time dummies (1995-2008)	Yes	Yes	Yes	Yes
Observations	6,035	6,035	6,035	6,035

Note: In the case of the Poisson regression, we use cluster-robust standard errors, and for the logit models, we computed cluster-bootstrapped standard errors using 400 replications. Standard errors in parentheses. Significance: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Regarding the abolishment of Professor's Privilege, hypothesis #2 stated that university ownership could have strengthened the relationship between patents on university-discovered inventions and university start-ups. At least in principle, with university ownership, more patented university-discoveries could be available for start-ups and value-added services by the TTOs could increase the expected returns to forming a start-up. For the models in Table 3.3, the variables *patents* and *patents-cited* capture the marginal effect of patented university discoveries before the Knowledge Creates Markets reform. In all four models (and across both estimation methods), the effect is positive and significant at the 5% level indicating a strong relationship between patents and start-ups. Looking at Model 1, the marginal effect suggests an additional patent leads to about a 12% [exp(.115)-1] increase in the number of university start-ups before the reform. For citation weighted patents, the results in Model 3 are smaller in magnitude, about a 4.4% increase in start-ups, on average. The marginal effects obtained from the conditional fixed effects logit models are similar in size, yet slightly smaller. The post-reform explanatory variables (*Prof*Newpolicy*Patents*) and (*Prof*Newpolicy*Patents-cited*) are not statistically significant in any model. Contrary to the prediction in hypothesis #2, this indicates that the strength of the relationship between patents and start-ups did not get stronger following the reform.

In Table 3.4, we disaggregate patents into the three ownership types and re-evaluate how the reform changed the strength of the relationship between patents on university-discovered inventions and university start-ups. Looking at the pre-reform relationships in Models 5 and 7, the results are consistent with prior expectations. Patents owned by private firms are not related to university start-ups. Patents held by the researchers' employers (university or PRO) are related to start-ups. This suggests that universities and PROs were somewhat successful at connecting patents to start-ups before the reform. Patents held by the individual researchers (i.e. personal patents) are positive and highly statistically significant. Each additional personal patent in the pre-reform period is associated with about a 34% [exp(.290)-1] increase in the number of start-ups (for citation-weighted patents in Model 7 the marginal effect is about 10%).

Table 3.4: Regressions on academic entrepreneurship (patent ownership type)

Startup	Model 5		Model 6		Model 7		Model 8	
QMLE Poisson FE								
Prof•NewPolicy	−0.026	(0.174)	−0.054	(0.176)	−0.052	(0.176)	−0.072	(0.178)
Firm Patents	0.059	(0.039)	0.057	(0.039)				
Employer Patents	0.086*	(0.049)	0.047	(0.055)				
Personal Patents	0.285***	(0.083)	0.290***	(0.089)				
Firm Patents cited					0.008	(0.018)	0.006	(0.018)
Employer Patents cited					0.051**	(0.022)	0.041*	(0.024)
Personal Patents cited					0.100***	(0.036)	0.098**	(0.039)
Prof•NewPolicy•EmplPat			0.254**	(0.105)				
Prof•NewPolicy•PersPat			0.067	(0.203)				
Prof•NewPolicy•EmplPatCit							0.086	(0.054)
Prof•NewPolicy•PersPatCit							0.045	(0.089)
Avg. Pubs	0.020	(0.013)	0.019	(0.013)	0.022	(0.013)	0.022	(0.013)
Conditional FE logit								
Prof•NewPolicy	0.034	(0.174)	0.002	(0.185)	0.005	(0.191)	−0.018	(0.178)
Firm Patents	0.026	(0.041)	0.025	(0.038)				
Employer Patents	0.098**	(0.049)	0.061	(0.059)				
Personal Patents	0.291***	(0.089)	0.281***	(0.098)				
Firm Patents cited					−0.011	(0.021)	−0.013	(0.021)
Employer Patents cited					0.062**	(0.026)	0.054**	(0.027)
Personal Patents cited					0.096***	(0.047)	0.090*	(0.048)
Prof•NewPolicy•EmplPat			0.266**	(0.125)				
Prof•NewPolicy•PersPat			0.105	(0.243)				
Prof•NewPolicy•EmplPatCit							0.091	(0.070)
Prof•NewPolicy•PersPatCit							0.060	(0.148)
Avg. Pubs	0.025	(0.013)	0.024	(0.014)	0.026	(0.014)	0.026	(0.014)
Time dummies (1995-2008)	Yes		Yes		Yes		Yes	
Observations	6,035		6,035		6,035		6,035	

Note: In the case of the Poisson regression, we use cluster-robust standard errors, and for the logit models, we computed cluster-bootstrapped standard errors using 400 replications. Standard errors in parentheses. Significance: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

But did the Knowledge Creates Markets reform increase the strength of the relationship between patents and start-up activities? Based on the findings in Models 6 and 8, the answer is somewhat mixed. For simple patent counts, Model 6 shows that the reform increased the strength of the relationship for university-owned patents. The coefficient is highly significant at the 1% level and the suggests each additional patent on university-discovered inventions increases the number of researcher start-ups by about 29% [exp(.254)-1], on average. When using citation-weighted patents, however, the coefficient on (*Prof*NewPolicy*Employer Patents-cited*) is much smaller in magnitude and statistically insignificant. As can be seen in the bottom panel of Table 3.4, the results for the conditional fixed effects logit models are very similar. This casts some doubt on the robustness of the finding that the reform increased the linkages between patents and start-ups. Citations are intended to be a correction for the quality of the inventions under the idea that a "high quality" invention should attract more follow-on patenting. While standard, this assumption about citation-weighting is quite strong and may actually be correlated with different factors than the market value of the invention or its potential to earn private returns.

It is clear from the results in Table 3.3 and Table 3.4 that patents on university-discoveries are strongly related to the number of university start-ups, although one may question whether the reform strengthened this relationship. The net impact of the reform on academic entrepreneurship, however, also depends on how the reform affected the volume of patents. To investigate this we start by replicating the main result of Czarnitzki et al. (2015) using equation (3.2) and the smaller sample available for this analysis. Table 5 presents the parameter estimates based on Poisson QMLE with cluster-robust standard errors for patent counts as well as the citation-weighted patent counts. Looking at the upper panel, i.e. at the regressions using patent counts, we find that the overall treatment effect, which is revealed by the coefficient on $(Prof \cdot NewPolicy)$ in Model 9, is negative and statistically significant at the 1% level. This indicates that the overall effect of abolishing Professor's Privilege was to decrease the volume of patents obtained on university-discovered inventions in Germany. It is economically significant as well. Holding the arrival of new knowledge and other exogenous trend factors constant, the coefficient estimate shows the volume of university patents decreased by about 14% [exp(-.153)-1], on average.

Table 3.5: Poisson models of academic patents (aggregated and by ownership type)

Dep. var.	Covar	Model 9 overall patents	Model 10 firm patents	Model 11 personal patents	Model 12 employer patents
Patent counts					
	Prof•NewPolicy	–0.153**	–0.754***	–0.128	1.746***
		(0.078)	(0.111)	(0.194)	(0.135)
	Avg. Pubs	0.039***	0.036***	0.012	0.054***
		(0.009)	(0.014)	(0.010)	(0.011)
	Growth US Patents	–0.377***	–0.229	–0.751***	–0.377**
		(0.122)	(0.171)	(0.240)	(0.157)
4y forward citation-weighted patent counts					
	Prof•NewPolicy	–0.153*	–0.711***	–0.017	1.685***
		(0.090)	(0.122)	(0.233)	(0.159)
	Avg. Pubs	0.029***	0.028*	0.002	0.052***
		(0.009)	(0.015)	(0.009)	(0.011)
	Growth US Patents	–0.198	–0.001	–0.622**	–0.248
		(0.147)	(0.206)	(0.299)	(0.188)
Time dummies (1995-2008)		Yes	Yes	Yes	Yes
Observations		52,770	24,312	9,607	39,406

Note: Standard errors in parentheses. Significance: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Models 10-12 in the upper panel of Table 3.5 rerun the regression specification in equation (3.2) using the patents by ownership type as alternative dependent variables. In Model 10, firm patents decrease dramatically after the shift to university ownership. The roughly 53% [exp(-.754)-1] decrease in the number of patents represents an economically significant decline in technology transfer through faculty-firm relationships. The decrease in personal patents is not significant, but the increase in employer patents due to the shift to university ownership is very large and significant. The point estimate reveals a 473% [exp(1.746)-1] increase, albeit from a small starting base. The results suggest that we can expect a lower university start-up rate after the policy change because patents have been shown to be essential for technology start-ups (Graham et al. 2009). The regressions using citation-weighted patent counts in the lower panel of Table 5 show similar results.

3.5 Conclusion

Following the US Bayh-Dole Act in 1980, university ownership became the leading intellectual property model for stimulating university-industry technology transfer for many countries around the world including Germany. In 2001, Germany introduced the Knowledge Creates Markets policy that not only set up new infrastructure and subsidies to support technology transfer, but more fundamentally, it shifted the ownership rights of university-discovered inventions from individual researchers to the university. Policymakers hoped to stimulate more patents on university-discoveries with the expectation that increased patenting would allow more licensing and the formation of new start-up companies.

The German policy experiment provides a unique opportunity to learn how academic entrepreneurship responds to policy changes, specifically to greater resources (i.e. infrastructure and subsidies) and to university ownership of IP rights. To identify these effects, we use a difference-in-difference research design with the university researchers as the treatment group and researchers at German public research organizations (PROs) as the control group. Unlike university researchers, PRO researchers already had well established TTOs and the rights to their inventions were already owned by their employing institutions.

The empirical analysis found no impact of the new infrastructure or its associated financing on the number of university start-ups. University start-ups followed the same trends as PRO start-ups after the policy: researchers in both groups of institutions formed fewer start-up companies. Our analysis focuses on the six year period right after the policy change when German TTOs where new at most universities and PVAs were completely new. Some might argue that these institutions lacked the necessary capabilities and routines that are important for fostering academic entrepreneurship (Lockett and Wright, 2005). However, recent studies, including an evaluation report of German PVAs, suggest inefficiencies may be a better explanation for our finding (Cuntz et al., 2012). This is consistent with a growing literature suggesting that intermediaries such as PVAs and TTOs are subject to numerous inefficiencies (e.g. Chapple et al., 2005; Markman et al., 2005; Anderson et al., 2007; Siegel et al., 2004; Kenney and Patton, 2009; Hertzfeld et al., 2006).

We found that the strength of the relationship between patents and start-ups increased (i.e. the marginal effect of patents on start-ups), but only for university-owned patents following

the reform and not for citation-weighted patents. As expected, firm-owned patents were not significantly related to faculty start-ups, but personally-owned patent were strongly related to start-ups in the pre-reform period. The post-reform coefficient for personally-owned patents was insignificant, which indicates the relationship did not change due to the policy. This evidence suggests university ownership increases the dependence of academic entrepreneurship on patent protection, but the resulting incentive effects on faculty start-ups remain mixed. On the one hand, patent protection confers advantages to new companies such as signally for financing and the ability to prevent imitation. In principle, this helps spur academic entrepreneurship. On the other hand, the time and money required to obtain patent protection is costly and, at least for some technologies and markets, this may not be necessary. In these cases, a requirement for patent protection could be a bureaucratic barrier that impedes academic entrepreneurship. How these benefits and costs balance out will depend on the specific circumstances facing the academic entrepreneur.

But even if the relationship between patents and start-up activities got stronger, the impact on the number of start-ups still depends on how the number of patents changed as a result of the policy. Consistent with prior work, we found significant decreases in the volume of firm-owned patents, an increase in the volume of university-owned patents, and no change for personally-owned patents. This suggests a trade-off emerged in the modes of technology transfer due to the abolishment of Professor's Privilege. Faculty-firm exchanges decreased dramatically and faculty start-ups increased to some degree. By displacing so many faculty-firm relationships, our evidence suggests the Knowledge Creates Markets reform likely decreased overall university technology transfer, although a final conclusion will need to wait until more research is completed.

For policymakers, our findings highlight the need for careful consideration of the institutional and cultural context before implementing reforms on IP ownership. Too often, the university ownership model is assumed the most effective IP policy for spurring academic entrepreneurship and/or other forms of technology transfer. It is important to remember that the Bayh-Dole Act was negotiated to clarify IP ownership for non-governmental US institutions within the US cultural environment. For Germany, in the era of Professor's Privilege, IP ownership rights were clearly delineated and privately held. Our evidence suggests the network of faculty-firm relationships in place prior to the Knowledge Creates

Markets reform was disrupted without compensating benefits. It appears the value and extent of this network was poorly investigated at the time of the reform. One clear lesson is for policymakers to require more background research and information before adopting IP ownership reforms.

Our study is not free of limitations. It will be important in future research to examine the performance of university start-up companies to better understand how these policies affected the economic impact of academic entrepreneurship. Entrepreneurial support and the ownership of patent rights might change the economic contribution of university start-ups by altering the "quality" distribution of these new companies, which may be observable using firm sales or employment data. Furthermore, our inferences about technology transfer are based on patents and start-ups. A more inclusive analysis would add indicators such as licensing, contracting agreements, material transfers, and other less formal arrangements. For this, researchers will need to develop new databases. Overall, these limitations point to new opportunities for research, as policymakers need information on how to structure IP ownership rules for greater innovation and growth.

3.A Appendix: Data collection procedure

Our data process starts with all patent applications filed at the German Patent and Trademark Office (DPMA) and the European Patent Office (EPO) involving at least one German inventor since 1978 using the PATSTAT database. We collapse the list of relevant patent documents to the number of inventions to account for patent families. Between 1995 and 2008 (our sample period) the total number of patent families is 624,041. Based on our data process, German professors and PRO researchers appear as inventors on 58,252 patent families (9.3% of all patent families). Among those, 18,253 refer to professor-invented patent families.

Searching patents invented by university faculty

As no comprehensive list of German university faculty members exists, we followed an alternative strategy that has been used in prior research to identify patents of university professors (see e.g. Czarnitzki et al. 2007, 2009). In Germany, the award of a doctorate and holding a professorial position are considered great honors. The "Dr." becomes an official part of one's name and, for example, is even mentioned in the national IDs and passports. The professor title is protected by the German criminal code (article 132a) against misuse by unauthorized persons. Accordingly, this title is used as a name affix not only in academic environment, but also in daily life. Based on this, we search the inventor records in the database for the title "Prof. Dr." and a large number of variations of this.[35] After having obtained an initial list of patent documents, we also searched for these inventors again in order to see whether they also patented without the "Prof. Dr." title. Note that we do not claim to have identified all university-invented patents, but it is certainly a large share of this population. Our numbers are close to those reported in policy documents circulated during the debate on Professor's Privilege in the late 1990s. Those documents said that university-invented patents accounted for about 4% of all German invented patents. This is an

[35] One may be concerned that the Professor Doctor title is also given as an honorary title to individuals who are not employed at universities. While the granting of honorary titles seems to be relatively rare, some of these highly qualified individuals may be labeled as professors in our data process. We believe any misclassification error would work against finding a significant policy effect, as these individuals are not affected by the policy change.

intermediate data preparation step. The list of patent documents will be disambiguated in a subsequent step to identify the number of patenting professors.

Identifying patents by PRO researchers

The identification of patents by PRO scientists is more straightforward because they can be searched by institution (i.e. applicant) names. The intellectual property rights to inventions made by their researchers were always owned by the institutional. We obtained a list of about 500 PRO institutes existing in Germany from the "Bundesbericht Forschung und Innovation 2012" published by the federal government. These institutional were searched as applicants in the patent documents. In order to create a list of unique PRO inventors, we select all patents on this list that have the PRO as only applicant. These are 70% of all PRO patents. This was necessary to avoid including industry researchers is our data. Next, we searched for all patents by these inventors again, in order to come up with a comprehensive list of patents filed by PRO inventors.

Disambiguation routine

The two lists of retained patent documents were pooled. This merged list may include too many patents, because of name homonyms. In addition, some inventors may switch between the two groups of institutions and thus appear in both lists. Therefore, we then implemented a disambiguation routine leading to a list of unique inventors.

The disambiguation algorithm is based on a relation network analysis. Every node within this network is a patent connected to other patents by layers of relations defined by shared applicants, co-inventors, citations and joint sets of IPC codes. The analysis uses a hierarchical approach by first traversing connections of high reliability to define sub-clusters that function as new nodes for the next iterative step. By aggregating information within these 'hypernodes' new connections emerge that will also be traversed and so on. As every sub-cluster describes a part of an inventor career, suspiciously large sub-clusters can easily be identified, rejected and re-traversed with more restrictive requirements for the connections. This method implicitly solves the common name problem. The resulting list of unique individuals and their corresponding patents has been checked manually to the largest extent possible.

Some of the professors also appear as PRO researchers at some point in time. We exclude those researchers associated with both institutions from the regressions reported in the main body of the text. By doing so, we omit those researchers for whom we do not know which IP policy is binding, the policy of the university or the policy of the PRO.

Collecting publication data from the Web of Science

The list of inventors is used to perform name searches in the Thomson Reuters Web of Science publication database, 1990 – 2008. We first retrieve all publications from Web of Science that match with respect to the names in our inventor list and have at least one German affiliation. This amounts to 572,936 publications. Second, we disambiguate these authors from Web of Science using cross-referencing information on journals, coauthors, citations and affiliations. Out of the almost 600,000 possible publications, 296,320 are identified as being authored by the inventors in our sample from 1995 to 2008 (the publication data from 1990 to 1994 was only taken into account in order to improve the name disambiguation routine and are not part of our final sample).

Compiling the panel database

The final step of the database construction involves generating a panel of unique academic inventors that includes information on their patents, citation-weighted patents and publications for each year. We count patents at the family level to ensure that patents in different jurisdictions for the same invention are not counted more than once. The unit of observation is a researcher-year. We restrict the regression sample period to run from 1995 through 2008. However, we keep those researchers who patented before 1995 in the sample. This implies that a researcher does not need to have a patent in the 1995 to 2008 period to be in the sample. We define entry into research as the year the researcher first appeared as an inventor on a patent or as an author on a journal publication. The final database is an unbalanced panel.

Adding the firm foundation data to the panel

In order to add firm foundation data to the panel we matched the names and associated cities of the researchers (professors and PRO researchers) to the owners, founder and major stakeholders of firms located in Germany. We use the Mannheim Enterprise panel database

for this exercise. It is a panel data set of firms located in Germany. It is maintained at ZEW in cooperation with Creditreform, the largest business information service in Germany. Creditreform sends its firm data in six-month intervals to ZEW, where the data is cleaned and prepared as to panel database, the Mannheim Enterprise Panel ("MUP"). The MUP enables the analysis of, for example, market entrances and exits (start-ups and shut-downs), changes in numbers of economically active firms in specific sectors and regions, the development of firms over time or the dynamics of job creation in firms. Among other information, it includes the names of all founders and other shareholders. We use this information to match start-ups to our academic inventor data.

The match is based on name and associated cities of the researchers. We exclude those researchers that have matches in the firm database based on their name, but not on city as we cannot be certain that they are involved in a firm or not. We keep those with matches based on name and city and those for which no firm foundation entry is found. Note that this essentially means that researchers with very common German names are dropped. This reduces the number of observations in the database for this paper to 52,770 researcher-year observations (1995 to 2008). Of these, 830 researcher-year observations are associated with one or more start-ups per year. The 52,770 researcher-year observations are based on 1,946 different university researchers and 4,551 different PRO researchers.

3.B Appendix: Regression descriptive statistics

The following table presents the descriptive statistics of the variables used in the regression models for both university researchers and PRO researchers before and after the policy change in 2002. The variable PAT denotes all patents. This is subsequently split to the different ownership types, i.e. FIRM indicates firm ownership; EMPL stands for the employer of the scientist owning the patent which could be either the university of the public research organization for the control group; and PERS denotes patents that are owned by persons. The term CIT then denotes the patent counts weighted by citations these patents received in the 4 years following the patent application.

The variable US_PAT denotes the total number of patent applications at the US Patent and Trademark Office in the technology field of the corresponding researcher. We experimented with several specifications in the regression model and finally do not use the level of US patents but their three-year growth rate, i.e. $GR\_US\_PAT = (US\_PAT_t - US\_PAT_{t-3})/ US\_PAT_{t-3}$.

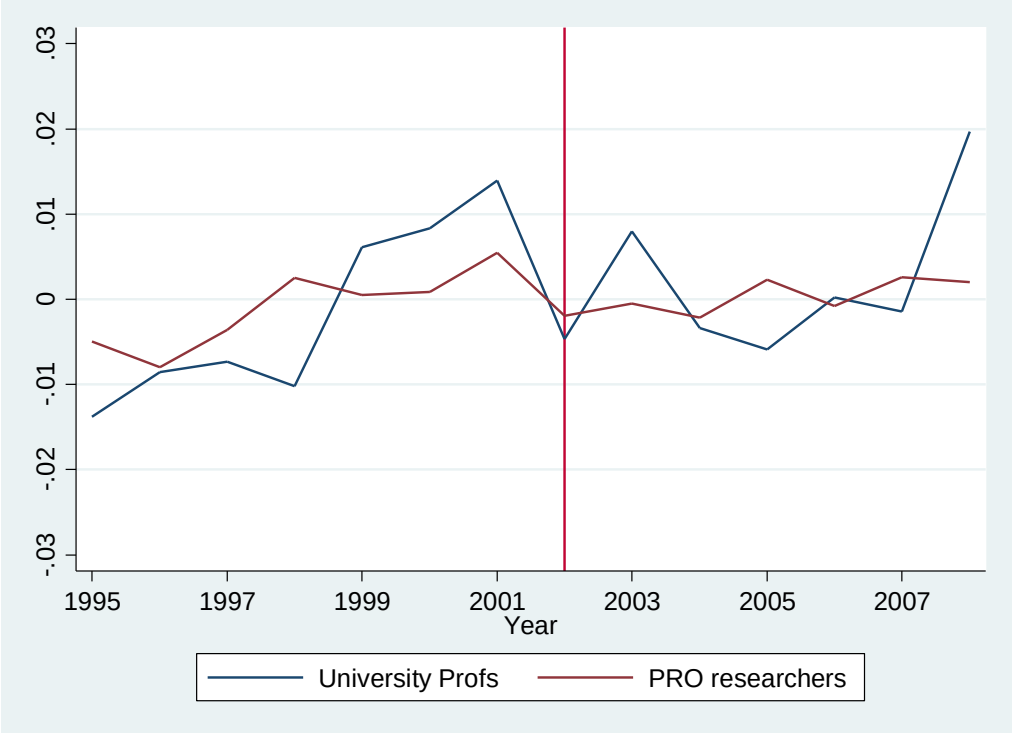
Table 3.6: Descriptive statistics

University researchers								
			Before 2002 reform (N = 9,180)				After 2002 reform (N = 8,237)	
Variable	**Mean**	**Std. Dev.**	**Min**	**Max**	**Mean**	**Std. Dev.**	**Min**	**Max**
STARTUP	0.04	0.23	0	5.00	0.04	0.23	0	4.00
PAT	0.58	1.41	0	24.00	0.34	1.03	0	28.00
PAT_FIRM	0.45	1.34	0	24.00	0.23	0.96	0	28.00
PAT_EMPL	0.02	0.19	0	4.00	0.10	0.39	0	6.00
PAT_PERS	0.14	0.51	0	10.00	0.04	0.24	0	5.00
PAT_CIT	0.97	2.79	0	62.00	0.52	1.85	0	56.00
PAT_CIT_FIRM	0.76	2.65	0	62.00	0.36	1.74	0	56.00
PAT_CIT_EMPL	0.04	0.38	0	17.00	0.13	0.63	0	15.00
PAT_CIT_PERS	0.24	1.05	0	26.00	0.06	0.45	0	13.00
3yr avg. pubs	2.38	4.87	0	67.33	3.22	6.22	0	73.33
GR_US_PAT	0.28	0.22	−0.19	1.00	0.29	0.17	−0.21	0.85
PRO researchers								
			Before 2002 reform (N = 15,507)				After 2002 reform (N = 19,846)	
STARTUP	0.01	0.14	0	4.00	0.01	0.12	0	4.00
PAT	0.56	1.27	0	29.00	0.40	1.07	0	26.00
PAT_FIRM	0.21	0.98	0	29.00	0.16	0.90	0	26.00
PAT_EMPL	0.39	0.91	0	16.00	0.28	0.71	0	17.00
PAT_PERS	0.02	0.21	0	9.00	0.01	0.09	0	4.00
PAT_CIT	0.97	2.58	0	61.00	0.62	1.91	0	51.00
PAT_CIT_FIRM	0.37	1.96	0	61.00	0.26	1.58	0	51.00
PAT_CIT_EMPL	0.67	1.85	0	42.00	0.43	1.32	0	28.00
PAT_CIT_PERS	0.05	0.48	0	22.00	0.01	0.16	0	11.00
3yr avg. pubs	0.87	2.11	0	44.00	1.12	2.46	0	63.67
GR_US_PAT	0.27	0.20	−0.19	1.00	0.28	0.18	−0.21	0.85

3.C Appendix: Trend graphs

Figure 3.1 shows the pre-treatment and post-treatment trends of the start-up variable. Note that the depicted variables are the averages of the within-demeaned dependent variable of the regressions presented in Table 3.3 and Table 3.4. A visual inspection may suggest that the pre-treatment trends in the period 1998/1999 differ between treatment and control group. Note, however, the scale of the vertical axis: the numerical differences are tiny. When implementing a formal test on whether the pre-treatment trends differ among the groups, the common trend assumption was never rejected at the 5% significance level. We implemented the test by annual t-tests on significant differences in the change of start-ups in first differences, and conducted a joint test in a regression on first differences of the start-up variable. The F-statistic on the test whether the annual slopes are jointly different only amounts to F = 0.99 with a p-value of 0.43.

Figure 3.1: Average trends of start-up activity



References

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

Alesina, A., Harnoss, J., and Rapoport, H., 2013. Birthplace Diversity and Economic Prosperity. National Bureau of Economic Research, NBER Working Papers No. 18699

Alma Laurea, 2016. XVIII Rapporto Alma Laurea sul Profilo e la Condizione occupazionale dei laureati. https://www.almalaurea.it/sites/almalaurea.it/files/comunicati/2016/cs_almalaurea_indagini_27_4_def.pdf

Almeida, P., and Kogut, B., 1999. Localization of knowledge and the mobility of engineers in regional networks. Management Science, 45(7), 905-917

Anderson, J.E., and van Wincoop, E., 2003. Gravity with gravitas: a solution to the border puzzle. American Economic Review, 93(1), 170-192

Anderson, T.R., Tugrul, U.D., Lavoie, F.F., 2007. Measuring the efficiency of university technology transfer. Technovation 27(5), 306-318

Angrist, J., and Pischke, S., 2009. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press

Angrist, J.D., Pischke, J.S., 2009. Mostly harmless econometrics: An empiricist's companion. Princeton: Princeton University Press

Astebro, T., Braguinsky, S., Braunerhjelm, P., Brostrom, A., 2016. Bayh-Dole versus the Professor's Privilege. Mimeo, HEC Paris

Audretsch, D.B., Leyden, D.P., Link, A.N., 2013. Regional appropriation of university-based knowledge and technology for economic development. Economic Development Quarterly 27(1), 56-61

Baldwin, R.E., and Taglioni, D., 2006. Gravity for dummies and dummies for gravity equations. National Bureau of Economic Research. NBER Working Paper 12516

Baltagi, B. H., Wu, P. X., 1999. Unequally spaced panel data regressions with AR(1) disturbances. Econometric Theory 15, 814-823

Becker, O, Ichino, A., and Peri, G., 2004. How large is the "brain drain" from Italy?. Giornale degli Economisti e Annali di Economia. 63 (Anno 117), No. 1 , 1-32

Bertrand, M., Duflo, E., Mullainathan, S., 2004. How much should we trust differences-in-differences estimates? Quarterly Journal of Economics 119(1), 249-275

Bhargava, A., Franzini, L., Narendranathan, W., 1982. Serial correlation and the fixed effects model. Review of Economic Studies 49, 533-549

Bosetti, V., Cattaneo, C., and Verdolini, E., 2015. Migration of skilled workers and innovation: A European Perspective. Journal of International Economics, 96(2), 311-322

Bradley, S.R., Hayter, C.S., Link A.N., 2013. Models and methods of university technology transfer. Foundations and Trends in Entrepreneurship 9(6). 571-650

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.

Cappelli, R., and Montobbio, F., 2016. European integration and knowledge flows across European regions. Regional Studies, 50(4), 709-727

Chapple, W., Lockett, A., Siegel, D., Wright, M., 2005. Assessing the relative performance of U.K. university technology transfer offices: parametric and non-parametric evidence. Research Policy 34(3). 369-384

Chunmian, Ge, Ke-Wei Huang, Ivan Png, 2016. Mobility of Scientists and Engineers: LinkedIn vis-a-vis patent records. Strategic Management Journal, Vol. 37 (1), 232-253

Clark, A., Georgellis, Y., and Sanfey, P., 1998. Job satisfaction, wage changes and quits: Evidence from Germany. Research in Labor Economics, 17, 95-121

Clarysse, B., Wright, M., Lockett, A., Mustar, P., Knockaert, M., 2007. Academic spin-offs, formal technology transfer and capital raising. Industrial and Corporate Change 16(4). 609-640

Colyvas, J., Crow, M., Gelijns, A., Mazzoleni, R., Nelson, R.R., Rosenberg, N., Sampat, B.N., 2002. How do university inventions get into practice? Management Science 48(1). 61-72

Conti, A., Thursby, J.C., Thursby, M.C., 2013. Patents as signals for startup financing. NBER Working Paper 19191. NBER: Cambridge, MA

Cuntz, A., Dauchert, H., Meurer, P., Philipps, A., 2012. Hochschulpatente zehn Jahre nach Abschaffung des Hochschullehrerprivilegs. Studien zum deutschen Innovationssystem 13-2012, Berlin

Czarnitzki, D., Doherr, T., Hussinger, K., Schliessler, P., Toole, A., 2015. Individual versus university ownership of university-discovered inventions. ZEW Discussion Paper 15-007. Mannheim

Czarnitzki, D., Glänzel, W., Hussinger, K., 2007. Patent and publication activities of German professors: an empirical assessment of their co-activity. Research Evaluation 16(4), 311-319

Czarnitzki, D., Glänzel, W., Hussinger, K., 2009. Heterogeneity of patenting activity and its implications for scientific research. Research Policy 38, 26-34

Damsgaard. E.F., Thursby, M.C., 2013. University entrepreneurship and professor

Debackere, K., Veugelers, R., 2005. The role of academic technology transfer organizations in improving industry science links. Research Policy 34(3), 321-342

Di Gregorio, D., Shane, S., 2003. Why do some universities generate more start-ups than others?. Research Policy 32(2), 209-227

Diamond, A.M., 1986. The life-cycle research productivity of mathematicians and scientists. Journal of Gerontology 41(4), 520–525

Docquier, F., and Rapoport, H., 2012. Globalization, brain drain, and development. Journal of Economic Literature, 50(3), 681-730

Doherr, Thorsten, 2017. Inventor Mobility Index: A Method to Disambiguate Inventor Careers. ZEW Discussion Paper No. 17-018, Mannheim

European Commission, 1995. Green paper on innovation. Brussels

Fagerberg, J., 1987. A technology gap approach to why growth rates differ. Research Policy, 16(2-4), 87-99

Fagerberg, J., 1988. Why growth rates differ. In: Dosi, G., Freeman, C., Nelson, R., Silverberg, G., Soete, L. (Eds.), Technical change and economic theory, Pinter Publishers, London

Fagerberg, J., 1994. Technology and international differences in growth rates. Journal of Economic Literature, 32(3), 1147-1175

Faggian, A., and McCann, P., 2009. Human capital and regional development. In R. Capello and P. Nijkamp (ed), Handbook of regional growth and development theories. Cheltenham: Edward Elgar

Faggian, A., Rajbhandari, I., and Dotzel K.R., 2017. The interregional migration of human capital and its regional consequences: a review. Regional Studies, 51(1), 128-143

Falagas, M.E., 2006. Unique author identification number in scientific databases: a suggestion. PLoS Medicine 3(5)

Fassio, C., Montobbio, F., and Venturini, A., 2015. How do native and migrant workers contribute to innovation? A study on France, Germany and the UK. IZA, Institute for the Study of Labor, Discussion paper no. 9062, http://ftp.iza.org/dp7922.pdf

Fenner, Martin, 2010. ORCID or how to build a unique identifier for scientists in 10 easy steps. Gobbledygook Blog, http://blogs.plos.org/mfenner

Frankel, J.A., and Romer, D., 1999. Does trade cause growth?. American Economic Review, 89 (3), 379-399

Fratesi, U., and Percoco, M., 2014. Selective Migration, Regional Growth and Convergence: Evidence from Italy. Regional Studies, 48(10), 1650-1668, DOI:10.1080/00343404.2013.843162

Geuna, A., Nesta, L.J.J., 2006. University patenting and its effects on academic research: The emerging European evidence. Research Policy 35, 790-807

Giuri, P., Mariani, M., Brusoni, S., Crespi, G., Francoz, D., Gambardella, A., Garcia-Fontes, W., Geuna, A., Gonzales, R., Harhoff, D., Hoisl, K., 2007. Inventors and invention processes in Europe: Results from the PatVal-EU survey, Research Policy, Elsevier, vol. 36(8), 1107-1127

Glaeser, E., 2010. Agglomeration Economics, University of Chicago Press, Chicago

Gouriéroux, C., Montfort, A., Trognon, A., 1984. Pseudo maximum likelihood methods: application to poisson models. Econometrica 52(3), 701-721

Graham, J.H.S., Merges, R., Samuelson, P., 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

Graham, J.H.S., Sichelman, T., 2008. Why do start-ups patent? Berkeley Technology Law Journal 23(3), 1063-1097

Griliches, Z., 1995. R&D and productivity: Econometric results and measurement issues, In P. Stoneman (ed.), Handbook of the Economics of Innovation and Technological Change, Basil Blackwell, Oxford

Grimaldi, R. Kenney, D., Siegel D.S., Wright, M., 2011. 30 years after Bayh–Dole: Reassessing academic entrepreneurship. Research Policy 40(8), 1045-1057

Haeussler, C., Colyvas, J.A., 2011. Breaking the ivory tower: Academic entrepreneurship in the life sciences in UK and Germany. Research Policy 40(1), 41-54

Hall, B.H., Jaffe, A., and Trajtenberg, M., 2005. Market Value And Patent Citations. Rand Journal of Economics, 2005, 36, 16-38

Hall, B.H., Mairesse, J., Turner, L., 2007. Identifying age, cohort and period effects in scientific research productivity: Discussion and illustration using simulated and actual data on French physicists. Economics of Innovation and New Technology 16(2), 159–177

Hellmann, T., 2007. The role of patents for bridging the science to market gap. Journal of Economic Behavior & Organization 63, 624-647

Hensen, M.M., de Vries, M.R., and Cörvers, F., 2009. The role of geographic mobility in reducing education–job mismatches in the Netherlands. Papers in Regional Science, 88(3), 667–682. doi:10.1111/j.1435-5957.2008.00189.x

Hertzfeld, H.R., Link, A.N., Vonortas, N.S., 2006. Intellectual property protection mechanisms in research partnerships. Research Policy 35(6), 825-838

Hoisl, K., 2007. Tracing mobile inventors-the causality between inventor mobility and inventor productivity. Research Policy, 36, 619-636.

Hsu, D.H., Ziedonis, R.H., 2013. Resources as dual sources of advantage: Implications for valuing entrepreneurial-firm patents. Strategic Management Journal 34(7), 761-781

Hunt, J., and Gauthier-Loiselle, M., 2010. How Much Does Immigration Boost Innovation?. American Economic Journal: Macroeconomics, 2(2), 31-56

Hvide, H., Jones, B., 2016. University Innovation and the Professor's Privilege. University of Bergen Discussion Paper in Economics No 16-1

Jensen, R., Thursby, M., 2001. Proofs and prototypes for sale: The licensing of university inventions. American Economic Review 91(1), 240-259

Jones, Benjamin, 2005. The Burden of Knowledge and the "Death of the Renaissance Man": Is Innovation getting harder? National Bureau of Economic Research Working Paper, No. 11360

Jorgenson D. W., 1995. Productivity Volume 1: Post-war US Economic Growth. Cambridge, MA, MIT Press.

Jovanovic, B., 1979. Job matching and the theory of turnover. Journal of Political Economy, 87 (5), 972-990.

Kenney, M., Patton, D., 2009. Reconsidering the Bayh-Dole act and the current university invention ownership model. Research Policy 38, 1407-1422

Kenney, M., Patton, D., 2011. Does inventor ownership encourage university research-derived entrepreneurship? A six university comparison, Research Policy 40(8), 1100-1112

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

Kerr, W.R., 2010. Breakthrough inventions and migrating clusters of innovation. Journal of Urban Economics, 67(1), 46-60.

Kerr, W.R., and Lincoln, W.F., 2010. The Supply Side of Innovation: H-1B Visa Reforms and U.S. Ethnic Invention. Journal of Labor Economics, 28(3), 473-508.

Kilger, C., Bartenbach, K., 2002. New rules for German professors. Science 298 (5596), 1173-1175

Kim, Kunho, Madjan Kabsa, Lee C. Giles, 2016. Inventor Name Disambiguation for a Patent Database using a Random Forest and DBSCAN. IEEE/ACM Joint Conference on Digital Libraries (JCDL)

Kochenkova, A., Grimaldi, R., Munari, N., forthcoming. Public policy measures in support of knowledge transfer activities: a review of academic literature. Journal of Technology Transfer

Krizhevsky, Alex, Ilya Sutskever, Geoffrey E. Hinton, 2017. ImageNet Classification with Deep Convolutional Neural Networks. Communications of the ACM, Volume 60, Issue 6, 84-90

Levin, R.C., Klevorick, A.K., Nelson, R.R., Winter, S.G., Gilbert, R., Griliches, Z., 1987. Appropriating the returns from industrial research and development. Brookings Papers on Economic Activity 1987(3), 783-831

Levin, S.G., Stephan, P.E., 1991. Research productivity over the life cycle: Evidence for academic scientists. American Economic Review 81(1), 114–132

Lissoni, Francesco, Andrea Maurino, Michele Pezzoni, Gianluca Tarasconi, 2010. APE-INV's "Name Game" Algorithm Challenge: A guideline for benchmark data analysis & reporting. European Science Foundation, http://www.esf-ape-inv.eu/download/Benchmark_document.pdf

Litan, R., Mitchell, L., Reedy, R., 2007. Commercializing university innovations: Alternative approaches. In: Jaffe, A., Lerner, J., Stern, S. (Eds.), Innovation policy and the economy, vol. 8, University of Chicago Press, Chicago, IL, 31-58

Lockett, A., Wright, M., 2005. Resources, capabilities, risk capital and the creation of university spin-out companies. Research Policy 34(7), 1043-1057

Maffezzoli, M., 2006. Convergence across Italian regions and the role of technological catch up. Topics in Macroeconomics, 6, Article 15

Magerman, Tom, 2015. PatentsView Disambiguation Inventor Workshop. Presentation at the USPTO, https://livestream.com/uspto/PatentsViewInventorWorkshop/videos/100138868

Marinelli, E., 2013. Sub-national graduate mobility and knowledge flows: An exploratory analysis of onward- and return-migrants in Italy. Regional Studies, 47(10), 1618–1633. doi:10.1080/00343404.2012.709608

Markman, G.D., Phan, P.H., Balkin, D.B. Gianiodis, P.T., 2005. Entrepreneurship and university-based technology transfer. Journal of Business Venturing 20(2), 241-263

Miguélez, E. and Moreno, R., 2015. Knowledge flows and the absorptive capacity of regions. Research Policy, vol. 44(4), 833-848

Moulton, B.R., 1990. An illustration of a pitfall in estimating the effects of aggregate variables in micro units. Review of Economics and Statistics 72(2), 334-338

Mowery, D.C., Sampat, B.N., 2005. The Bayh-Dole Act of 1980 and university-industry technology transfer: A model for other OECD governments? Journal of Technology Transfer 30 (1/2), 115-127

Nathan, M., 2014. The wider economic impacts of high-skilled migrants: a survey of the literature for receiving countries. IZA Journal of Migration, 3:4

Nelson, R.R., and Winter, S.G., 1982. An evolutionary theory of economic change, Harvard University Press, Cambridge, MA

No, Y., and Walsh, J., 2010. The importance of foreign-born talent for US innovation. Nature Biotechnology, 28, 289-291

O'Shea, R.P., Chugh, H., Allen, T.J., 2008. Determinants and consequences of university spinoff activity: a conceptual framework. Journal of Technology Transfer 33(6), 653-666

OECD, 2001. Measuring Productivity. Measurement of Aggregate and Industry-level Productivity Growth. Paris, OECD

OECD, 2003. Turning science into business: patenting and licensing at public research organizations, Paris: OECD

Ortega, F., and Peri, G., 2014. Openness and income: The roles of trade and migration. Journal of International Economics, 92(2), 231-251

Ottaviano, G., and Peri, G., 2006. The economic value of cultural diversity: evidence from US cities. Journal of Economic Geography, 6, 9-44

Peri, G., 2007. Higher Education, Innovation and Growth. In Education and Training in Europe. Edited by: Brunello G, Garibaldi P, Wasmer E. Oxford: Oxford University Press

Peri, G., 2012. The effect of immigration on productivity: evidence from U.S. States. Review of Economics and Statistics, 94(1), 48-358

Peri, G., and Sparber, C., 2009. Task Specialization, Immigration, and Wages. American Economic Journal: Applied Economics, 1(3), 135-169

Peri, G., and Sparber, C., 2011. Highly educated immigrants and native occupational choice. Industrial Relations, 50(3), 385-411

Peri, G., Shih, K., and Sparber, C., 2013. STEM Workers, H1B Visas and Productivity in US Cities. Norface Discussion Paper Series 2013009, UCL, London.

Petrie, Steve, T'Mir Julius, Russel Thomson, 2017. Author name disambiguation using a neural network-based algorithm. Presentation at the Web of Science data workshop at EPFL, Lausanne

Pezzoni, Michele, Francesco Lissoni, Gianluca Tarasconi, 2012. How to kill Inventors. Testing the Massacrator Algorithm for Inventor Disambiguation. Cahiers du GREThA, No. 2012-29

Privilege. Industrial and Corporate Change 22(1). 183-218

Quatraro F., 2009. Innovation, Structural Change and Productivity Growth: Evidence from Italian Regions 1980-2003. Cambridge Journal of Economics, 33(5), 1001-1022

Raako, Jaana, 2016. Knowledge spillovers through inventor mobility: the effect on firm-level patenting. Journal of Technology Transfer, Volume 42, Issue 3, pp 585–614

Raffo, J., and Lhuillery, S., 2009 How to play the "Names Game": patent retrieval comparing different heuristics, Research Policy, 38(10), 1617-1627

Rosenkopf, L., and Almeida, P., 2003. Overcoming local search through alliances and mobility. Management Science, 49(6), 751-766

Rothameral, F.T., Agung, S.D., Jiang, L., 2007. University entrepreneurship: a taxonomy of the literature. Industrial and Corporate Change 16(4), 691-791

Santos Silva, J.M.C, and Tenreyro, S., 2006. The log of gravity. The Review of Economics and Statistics, 88(4),641-658

Schmoch, U. Licht, G., Reinhard, M., 2000. Wissens- und Technologietransfer in Deutschland. Stuttgart: Fraunhofer IRB Verlag

Schmoch, U., 2007. Patentanmeldungen aus deutschen Hochschulen - Analysen im Rahmen der Berichterstattung zur Technologischen Leistungsfähigkeit Deutschlands. Studien zum deutschen Innovationssystem Nr. 10-2007, Berlin

Schoen, Anja, Dominik Heinisch, Guido Buenstorf, 2014. Playing the “Name Game” to identify academic patents in Germany. Scientometrics, Volume 101, Issue 1, 527–545.

Shane, S., 2001. Technology regimes in new firm formation, Management Science 47(9), 1173-1190

Siegel, D.S., Waldman, D., Link, A., 2003. Assessing the impact of organizational practices on the relative productivity of university technology transfer offices: an exploratory study, Research Policy 32, 27-38

Siegel, D.S., Waldman, D.A., Atwater, L.E., Link, A.N., 2004. Toward a model of the effective transfer of scientific knowledge from academicians to practitioners: qualitative evidence from the commercialization of university technologies. Journal of Engineering Technology Management 21, 115-142

Singh, Jasjit, 2003. Inventor Mobility and Social Networks as Drivers of Knowledge Diffusion. Havard University Working Paper Series

So, A.D., Sampat, B.N., Rai, A.K., Cook-Deegan, R., Reichman, J.H., Weissman, R. and Kapczynski, A., 2008. Is Bayh-Dole good for developing countries? Lessons from the US Experience. PLoS Biology 6 (10), 2078-2084

Solow, R. M., 1957. Technical change and the aggregate production function. The Review of Economics and Statistics, 39, 312-20

Song, J., Almeida, P., and Wu, G., 2003. Learning-by-Hiring: When Is Mobility More Likely to Facilitate Inter-firm Knowledge Transfer?. Management Science, 49, 351-365

Stephan, P.E., and Levin, S.G., 2001. Exceptional contributions to US science by the foreign-born and foreign-educated. Population Research and Policy Review, 20, 59-79

Stevens, A., 2004. The enactment of Bayh-Dole. Journal of Technology Transfer 29, 93-99.

Stirling, James, 1730. Methodus differentialis, sive tractatus de summatione et interpolatione serierum infinitarum. Royal Society, London

Stock J.H. and Yogo M., 2005. Testing for weak instruments in linear IV regression, In: Stock J.H. and Andrews D.W.K. (Eds.), Identification and Inference for Econometric Models:Essays in Honor of Thomas J. Rothenberg. Cambridge University Press

The Economist, 2002. Innovation's golden goose - The reforms that unleashed American innovation in the 1980s, and were emulated widely around the world, are under attack at hom. Technology Quarterly, Q4 2002

Toole, A.A., Czarnitzki, D., 2007. Biomedical academic entrepreneurship through the SBIR program. Journal of Economic Behavior & Organization 63(4), 716-738

Topel, R.H., and Ward, P.W., 1992. Job mobility and the careers of young men. The Quarterly Journal of Economics, 107 (2), 439-479

Torvik, Vetle I., Neil R. Smalheiser, 2009. Author Name Disambiguation in MEDLINE. National Institutes of Health Public Access, Author Manuscript

Trajtenberg, M., 2005. Recombinant ideas: the mobility of inventors and the productivity of research. In: Proceedings of the CEPR- Conference, Munich, May 26-28

Trajtenberg, Manuel, 2004. The “Names Game”: Using Inventors Patent Data in Economic Research. Presentation: http://www.tau.ac.il/~manuel/

Trajtenberg, Manuel, Shiff Gil, Melamed Ran, 2009. The “Names Game”: Harnessing Inventors’ Patent Data for Economic Research. NBER Working Paper 12479

Trippl, M., 2013. Scientific mobility and knowledge transfer at the interregional and intraregional level. Regional Studies, 47 (10), 1653-1667. DOI:10.1080/00343404.2010.549119

Turner, L., Mairesse, J., 2005. Individual productivity differences in public research: How important are non-individual determinants? An econometric study of French physicists’ publications and citations (1986–1997). Retrieved November 2, 2015, http://piketty.pse.ens.fr/fichiers/Turner2005.pdf

Von Proff, S. , Buenstorf, G., Hummel, M., 2012. University patenting in Germany before and after 2002: What role did the professors' privilege play? Industry & Innovation 19(1), 23-44.

Wooldridge, J., 1999. Distribution-free estimation of some nonlinear panel data models. Journal of Econometrics 90(1), 77-97

Wooldridge, J.M., 2010. Econometric Analysis of Cross Section and Panel Data. Cambridge, MA: The MIT Press