University-Industry Knowledge Transfer: An Empirical Analysis of Channels, Motives and Distances

Vom Promotionsausschuss der Technischen Universität
Hamburg-Harburg
zur Erlangung des akademischen Grades

Doktor der Wirtschafts- und Sozialwissenschaften (Dr. rer. pol.)

genehmigte Dissertation

von
Jan Willem Reerink

aus
Rio de Janeiro, Brasilien

2018

Gutachter:

- Prof. Dr. Jan Christoph Ihl
- Prof. Dr. Robin Kleer

Tag der mündlichen Prüfung:

09.02.2018

Acknowledgements

I would like to thank the many people who contributed to this work over the last years. First and foremost I extend my thanks to my supervisor, Prof. Ihl, for his dedicated support and mentorship (as well as occasional mountain rescue efforts). I would also like to thank Prof. Kleer for his feedback and contributions as well as Prof. Lüthje for his evaluation. Furthermore, thanks to my friends at the RWTH and TUHH universities for both the fun times and learning experience. Last but not least, thanks to my family, especially Tanya, Paul, Anna, Max and Greta.

Abstract

The present thesis investigates several technology transfer channels in the context of university-industry knowledge transfer. Both traditional channels such as collaborative research or academic patenting and innovative alternatives such as broadcast search are considered. Theoretical advances and implications for policy can be derived by implementing novel methods from the field of Data Science.

Zusammenfassung

Die vorliegende Arbeit untersucht mehrere Technologietranferkanäle im Kontext des Wissenstransfers von Universitäten in die Industrie. Sowohl traditionelle Transferkanäle wie zum Beispiel kollaborative Forschung oder akademisches Patentieren sowie innovative Alternativen wie broadcast search werden berücksichtigt. Beiträge zu Theorie und politische Implikationen können durch den Einsatz neuartiger Methoden aus dem Bereich Data Science gewonnen werden.

Contents

List of Figures

List of Tables

List of Abbreviations

DCA	**D**ocument **C**o-Citation **A**nalysis
DOCDB	Master documentation database of the European Patent Office
ECLA	**E**uropean **CLA**ssification
EPO	**E**uropean **P**Patent **O**ffice
GB	**G**iga**b**yte , 10^9 bytes
GPL	**G**eneral **P**urpose **T**echnology
GML	**G**eneralized **L**inear **M**odel
IPC	**I**nternational **P**atent **C**lassification
KPF	**K**nowledge **P**roduction **F**unction
LDA	**L**atent **D**irichlet **A**llocation
NLTK	**N**atural **L**anguage **T**ool**k**it
NUTS	**N**omenclature des **U**nités **T**erritoriales **S**tatistique
RFP	**R**equest **F**or **P**roposals
R&D	**R**esearch and **D**evelopment
SDT	**S**elf **D**etermination **T**heory
SVD	**S**ingular **V**alue **D**ecomposition
SQL	**S**tructured **Q**uery **L**anguage
TF-IDF	**T**erm **F**requency - **I**nverse **D**ocument **F**requency
TTO	**T**echnology **T**ransfer **O**ffice
OECD	**O**rganisation for **E**conomic **C**o-operation and **D**evelopment
PATSTAT	European **Pat**ent Office **Stat**istical Database
ZINB	**Z**ero-**I**nflated **N**egative **B**inomial

Chapter 1

Introduction

This chapter briefly discusses the motivation for this thesis and its research goals. An outline of the structure of the thesis is also given.

1.1 Motivation

The motivation of this work is the emergence of the field of academic entrepreneurship in conjunction with methodological advances in computer science as well as increasing availability of relevant data. The traditional role of universities, to educate students and to conduct basic research, has been expanded into the area of commercialization due to the increasing importance of information to knowledge-based economies (Abramovitz, 1956; Solow, 1957). The “entrepreneurial university” (Etzkowitz, 1988) is supposed to transfer knowledge via more direct channels to society, i.e. by commercializing knowledge with spin-offs, patents or licenses based on research findings. As the traditional university was perceived as inefficient in commercialization of knowledge, either due to university staff lacking the entrepreneurial skills to commercialize the mass of knowledge provided by researchers or due to researchers being too taken up with publishing to commercialize, policy instruments have been implemented to encourage academic entrepreneurship. These include changes in legislation such as the often-cited Bayh-Dole act (Mowery et al., 2001) or, in the German context, the similar “Arbeitnehmererfindergesetz”[1] (Kilger and Bartenbach, 2002). At the institutional level, technology transfer offices (TTOs) have been established at most universities to improve the disclosure of knowledge created by university scientists and a number of projects are underway to encourage young academics in

[1] For details on these policy instruments see Chapter 2.

commercializing their work in start-ups (Maritz, Koch, and Schmidt, 2016; Siegel, Waldman, and Link, 2003). Given the importance of innovation to a knowledge-based economy and the high potential for innovation inherent in organizations that are dedicated to teaching and research, the topic of knowledge transfer between academia and industry has received significant attention from researchers (Bozeman, 2000; Crespi et al., 2011; Macho-Stadler and Pérez-Castrillo, 2010; Markman et al., 2005; Nilsson, Rickne, and Bengtsson, 2010; Owen-Smith and Powell, 2001; Siegel, Waldman, and Link, 2003). In recent years new methods and data sources have become available that may enable researchers to shed new light on some of the important questions that are associated with the topic of these knowledge transfers: which transfer channels are most suitable? What motivates scientists to transfer knowledge? Which incentives exist and how are they perceived? What role does the academic culture play in all of this? To what degree is geographic distance a barrier to successful transfers?

An improved understanding of these relations is required to improve our ability to influence and steer the system into the desired direction. This work will explore new transfer channels such as privately owned innovation platforms that match industrial solution seekers to individuals able to solve complex problems. Using methods adopted from the field of computer science, untapped data sources will be used to extend findings on some of the traditional transfer channels.

1.2 Research goals

The point of this thesis is to extend existing research on knowledge transfer between academia and industry with the help of state of the art methods from the field of data science. Data science, which combines automated acquisition and preparation of data with machine learning algorithms and statistical analysis, enables us to gain insights from data sources that have, so far, not been exploited for the purpose of scientific analysis on knowledge transfer. These methods serve as tools to improve our understanding of some of the key concepts underlying knowledge transfer, in particular its drivers: what motivates scientists to actively engage in knowledge transfer? Which incentives are efficient? Which transfer channels are suitable? And not least of all, what is the economic impact of transfers? Even small improvements to our understanding of these questions may yield significant returns as knowledge will continue to be the most important resource for our economy.

1.3 Structure of thesis

The remainder of this thesis is structured in the following way: chapter 2 gives an overview of the theoretical context of academic entrepreneurship and derives research questions. Chapter 3 describes the data sources accessed for the subsequent research projects as well as some of the procedures that were necessary to obtain and process the data. Chapter 4 introduces the concept of Data Science, its importance to quantitative analysis in economics and presents some of the machine learning algorithms used. The following four chapters describe the projects outlined in chapter 2 in detail and present findings. Chapter 5 shows how to tackle increasing amounts of literature using machine learning methods to derive insights that are often required in early stages of scientific projects, as an applied example the chapter shows the development of academic entrepreneurship as a branch of entrepreneurship science. Chapter 6 discusses the motivational aspects of academic entrepreneurship by analyzing the traditional transfer channel of academic patenting. Chapter 7 encompasses research on the more innovative transfer channel of broadcast search. Chapter 8 and 9 measure the impact of academic entrepreneurship on knowledge production by expanding the traditional knowledge production framework with novel measures for collaborative research. The thesis concludes with chapter 10, which sums up the findings, highlights methodological advances, discusses limitations and implications for future research opportunities that can be derived from this work.

Chapter 2

Theoretical context

This chapter introduces the theoretical background to the core topics of this thesis. As described in chapter 1, this thesis is concerned with the analysis of knowledge transfer between academia and industry. After a brief overview of the various types of knowledge transfer and the importance of the topic, we will take a closer look at three transfer mechanisms which have been selected as they represent different perspectives on the issue. The first mechanism, academic patenting, explores the perspective of individual researchers with a focus on the motivational aspects of academic patent disclosure. As an innovative alternative transfer channel we consider broadcast search as a potentially more flexible tool for facilitating knowledge transfer. The third mechanism, collaborative scientific projects, allows for an analysis at the regional level. This mechanism facilitates an analysis of the impact of knowledge transfer on knowledge generation. The theoretical context is then used to derive research questions which lead to the projects described in later chapters. Since each project is concerned with a relatively specialized field, the chapters on the various projects contain additional information on the specific literature required to build hypotheses.

2.1 Knowledge transfer from academia to industry

As a preface to this section it is useful to delimit some of the terms used in this thesis. The term "university-industry technology transfer" is often used to describe the field of science that explores the potential of academic knowledge for application in the private sector. It encompasses the transfer of knowledge from the public sector (universities and other research institutions) to the private sector (industry). This transfer has usually the purpose of commercializing knowledge or making it relevant in an applied context. Technology transfer may also refer to the transfer of knowledge between companies (Zhao and Reisman, 1992) or, in the case of international technology transfer, between countries (Robinson, 1988). However, these perspectives are not explicitly regarded in this work. Whereas the definition of transfer is straightforward, the definition of technology in the context of technology transfer is slightly more involved. The definition of Sahal (1981) stresses that a technology has aspects of both a product and a process and appears suitable in the context of technology transfer, where a significant aspect is not the transfer of a physical ojbect but also the transfer of knowledge required to understand and apply the procedural aspects of technology (Bozeman, 2000). Often the terms "technology transfer" and "knowledge transfer" are used more or less interchangeably. Sometimes technology transfer is more closely associated with the transfer of knowledge (Zhao and Reisman, 1992), in other cases it is linked to innovation (Rogers, 2010). The latter perspective stresses the importance of transferring knowledge that creates value to the recipient primarily by virtue of being novel. Since "knowledge transfer" is the more abstract term compared to "technology transfer" and since it is more closely associated with the problems that occur in the transfer of either, we will use the term "knowledge transfer" throughout this thesis.

The generation of new information is of increasing importance as nations tend to adopt the concept of the knowledge-based economy (Abramovitz, 1956; Solow, 1957; Cooke and Leydesdorff, 2006). Innovation is widely regarded as main driver of economic growth, a claim that has been demonstrated in research by considering the inputs to and outputs from an economy (Abramovitz, 1956). If the inputs, such as labour available from a nation's population or capital from its economy, cannot explain all of the outputs, an omitted variable must exist that explains the unexplained growth. Thus a large residual, usually a cause for alarm among statisticians, contributed to important findings in economics. The importance of innovation is closely associated with the uncertainty that goes hand in hand with the processes that lead to innovation. The success or failure of an invention is difficult to predict, in large part due to a large potential for improvement in the early phases of a new product or a technological development (Rosenberg, 2004). Thus even (or especially) important innovations often come as a surprise, which can create a certain degree of upheaval in established markets (Christensen, 2013). This finding has spurred a new way of thinking about economies in general as primarily impacted by an economy's capability in implementing innovations, that is its ability to produce knowledge and the degree to which its policies empower entrepreneurs. This theory challenges established theories such as the neoclassical view under which the accumulation of capital is the main aspect defining the dynamics of economies (Atkinson and Ezell, 2012). In this context entrepreneurs can be individuals who discover and act upon opportunities (Schumpeter, 1934) or large organizations equipped with resources that enable them to shoulder the increased risk of entrepreneurial activities (Schumpeter, 2013).

Innovations are the result of knowledge creation, which can roughly be divided into basic and applied research. Generally, applied research is understood as efforts for finding a solution to a specific problem, while basic research is aimed at improving the general understanding in a scientific discipline. Applied research promises greater economic viability in the short term, while basic research involves higher risks but also higher rewards (Kleer, 2010). Due to the required investments and long-term ramifications of basic research, it has traditionally been the objective of academic research by governmental institutions. Applied research is traditionally associated with industry. However, there are cases of industry engaging in basic research and public institutions delivering valuable applications. Due to an acceleration in scientific development, the interval between major technological breakthroughs grow shorter and the resources that were intended to improve either basic or applied research prove useful in both areas (Sandmo, 2011; Rosenberg and Nelson, 1994). Hence there is an increasing overlap in the roles of industry, academia and government with regard to their roles in the production of knowledge (Owen-Smith, 2003; Etzkowitz and Leydesdorff, 2000).

Traditionally, universities' main task was to educate and to conduct research, with the focus shifting from teaching to research during the 20th century (Hounshell, 1996). This view was reflected in researcher's conceptualization of universities as distinct from commercial enterprises in order to warrant efficient advancement of basic scientific research that might be negatively impacted by industry's more short-term perspective on commercial applications (Merton, 1973; Dasgupta and David, 1987). This negative view on interactions of university and industry has been labeled "corporate manipulation" thesis (Florida, 1999), which cautions against sacrificing long-term scientific insights for short-term commercial interests (Noble, 1979). When universities receive most of their funding from industry, they may be subjected to restrictions (such as increased secrecy) that compromises their ability to excel at basic research (Brooks, 1993). A contrasting theory highlights the potential of universities as providers of opportunities for industry based on university's stocks of knowledge (Etzkowitz, 1988).

Regardless of which side one favors in this argument, universities are involved in a type of knowledge transfer as per their original mission: teaching students who may later be employed by industry is one form of dispersing the knowledge that is created at universities. However, in practice there is a divide between the two missions. The majority of students are taught to a degree level that incorporates most of the status quo of existing knowledge in a field. The information generated by universities, for example in the form of research projects, is not usually taught outside of master or Ph.D level educational programs. Since university research is likely to be important to industry (Etzkowitz and Leydesdorff, 2000), with the time frame between scientific breakthrough and application shortening, there exists a demand for a more direct transfer of knowledge from academia to industry. Given the importance of knowledge to modern economies and the available knowledge stocks in universities as potential source for innovation, a third role has been added to the two traditional roles of universities. According to this triple helix model (Etzkowitz, 2008) of the entrepreneurial university (Slaughter and Leslie, 1997), academia is also tasked with contributing to the commercialization of knowledge. As a result, collaboration between industry and academia (for example at university-industry research centers) attracts large amounts of R&D (Research and Development) funding. Other indicators, such as university revenues from patents licensed to industry, also show the importance of university-industry relations (Florida, 1999). It should be noted that, while the topic of university-industry collaboration is very popular in recent years, relations between industry and academia played a role before the advent of large policy changes, which are often regarded as starting point of modern university-industry collaboration (Mowery et al., 2001).

Why are universities interesting for industry and vice versa? Academia has access to significant resources useful for the generation of knowledge. Funding of projects with high risk, equipment and skilled employees offer opportunities for the creation of information that are immediately useful in applied contexts or of interest to firms active in domains where basic scientific knowledge is likely to offer competitive advantages in the relative short term. From the perspective of university researchers, collaboration with industry is attractive due to potential access to funds and the prospect of employment (Roach and Sauermann, 2010). Collaboration with industry may be a means to the end of advancing a scientist's research (D'este and Perkmann, 2011). However, for either side to benefit from collaboration, the knowledge created at universities needs to be transferred. This process is not straightforward due to some characteristics of knowledge and several additional barriers: the codification of knowledge is expensive and often insufficient to capture all relevant aspects. Written texts do not convey the same amount of information as an interpersonal exchange. Some knowledge cannot easily be transmitted through personal communication either, for example if that transfer requires a mutual understanding of the subject area (Polanyi, 1966). The choice of a suitable transfer channel can mitigate barriers. Several transfer mechanisms are frequently mentioned in research as channels through which knowledge can flow from academia to industry (Bekkers and Freitas, 2008, OECD, 2013):

1. Training of graduates: universities train students to a degree level in the scientific status-quo. Graduates then seek employment in industry, where they can apply the skills they learned,

effectively transferring knowledge from academia into the rest of the economy.

2. Academic patents: scientists may file a patent based on their research findings. In the long term the publication of that patent enables the adoption of the invention, while in the short term licenses can be used to make the invention available to industry.

3. University spin-offs: closely associated with academic patents, universities may attempt to commercialize inventions with high potential through start-ups.

4. Scientific publications: associated with relatively low costs, this transfer channel makes research findings available to a broad audience.

5. Informal communication: personal communication is usually regarded as highly effective means of transferring knowledge, even though the audience is limited. Scientific staff may use informal communication at conferences or within their professional network to transfer knowledge.

6. Collaborative research: formal collaboration between industry and academia on research projects comprises some aspects of the other transfer channels such as efficient personal communication, publication and patenting. However, it also involves high costs associated with the selection of suitable collaboration partners and the investment of resources.

7. Broadcast Search: internet platforms that enable companies to describe technical problems. Potential solvers can submit solution proposals with the winning submissions usually being rewarded. This is an a-typical transfer channel that bears some similarities to formal collaboration.

As mentioned above, this work will focus on the three channels of broadcast search, academic patenting as well as collaborative research. This selection enables to study university-industry knowledge transfers from several perspectives. It also enables us to understand in which case a channel is a good choice for knowledge transfer: for channels to be used effectively one needs to be aware of their advantages and shortcomings as well as general obstacles to knowledge transfer. Regardless of the chosen channel, the recipient must be able to process the transmitted information. Depending on the channel, some barriers to the transfer exist and are more or less pronounced, as shown in the next section.

2.2 The role of distance in knowledge transfers

One of the more important barriers to the successful use of a transfer channel is the distance over which information is to be transferred. For the context of this work, we differentiate between three types of distance: institutional distance (sometimes also referred to as organizational distance), knowledge distance (also known as technological distance) and geographic distance.

Geographic distance is the most obvious form, which still has a strong effect on knowledge transfer. Increasing the distance between the actors tends to decrease the effectiveness of knowledge transfer, in the same way as spatial dependence affects most relations that are spatially distributed (Tobler, 1970). Some transfer channels are less affected by geographic distance than others: texts such as scientific publications or patent abstracts are available regardless of distance to the author. Any transfer that is related to personal communication, such as informal communication or collaborative research, are more strongly affected by geographic distance, as the likelihood of interpersonal communication decreases with distance (Allen, 1984). This relationship also applies when taking into account the decreasing costs of digital communication (Allen and Henn, 2007).

Institutional or organizational distance refers to the distance between actors from different backgrounds. In the context of university-industry knowledge transfers, employees in industry and academia have adapted to their respective organizational cultures and associated norms. The difference between the two may complicate the transfer of knowledge or influence the motivation for an actor to engage in a transfer (Cummings and Teng, 2003). Actors with similar organizational background are usually more efficient at transmitting knowledge (Uzzi, 1996). As university-industry transfer necessarily spans organizational boundaries, this distance is relevant to all transfer channels. Since trust is facilitated by lower organizational distance, those channels where trust is more of an issue may be more affected by organizational distance. An example is broadcast search, where the solver needs to trust the seeker to fairly reward proposals.

Knowledge distance refers to the degree to which two actors possess similar knowledge (Cummings and Teng, 2003). For a successful transfer, which includes that the recipient is able to absorb the knowledge, the knowledge distance may not be too great (Hamel, 1991). Knowledge from a completely different context may turn out to be useless to the receiver as understanding it would involve the acquisition of additional (usually more basic) knowledge required to fully appreciate the state of the art. Hence knowledge distance is also closely associated with the receiver's experience in absorbing unfamiliar knowledge (Cohen and Levinthal, 1990). As with organizational distance, knowledge distance affects all attempts to transfer knowledge from academia to industry. An important aspect that differentiates knowledge distance from other types of distance is that an increase in distance does not necessarily come with the negative effects that physical distance is often associated with (admittedly, physical distance may in some cases also prove beneficial for collaboration, for example when it is precondition for linking of suitable partners). With an increase in knowledge distance usually comes an increase in the novelty of the transmitted knowledge. Recombining existing knowledge with novel knowledge is likely to produce more successful innovations. Hence a greater knowledge distance can improve the receiving actor's innovation performance (Nooteboom et al., 2007).

This thesis focuses on three transfer channels to highlight the import of the three distance types outlined above. Academic patenting is closely associated with organizational distance due to the relevance of scientific norms for the motivational aspects of academic patenting. Broadcast search is a suitable transfer mechanisms to study the effects of knowledge distance as the promise of innovation intermediaries is to boost innovation performance by making available novel information. Finally, geographic distance is a factor in collaborative efforts that are distributed over regions.

The following sub-chapters give an overview of three transfer mechanisms and how they are affected by distance.

2.2.1 Academic patenting

A patent is a legal title for protecting an invention that provides several rights to the title's owner. The owner has the right to prevent others from making, selling or importing the protected product or from using or selling the protected process. Patents are transferable, allowing the owner to sell or license the right (*OECD Patent Statistics Manual*; 2009). In return for having one's invention protected (typically for a time of up to 20 years) the patent's owner agrees to the publication of the patent upon its grant. Since the patent needs to contain a technical description of the invention, the disclosure makes knowledge available to others, enabling downstream innovation. When the patent expires, the technology becomes essentially a public good, which enables more widespread adoption. Hence the patent system provides incentives to engage in inventive work and improves the dissemination of knowledge in the long term. For a patent to be granted, the invention must fulfill some requirements:

(i) it must be novel, i.e. the invention may not be already patented or published in another fashion.

(ii) The patent must include an inventive step, i.e. it must be non-obvious, and the invention must generally require the use of state-of-the-art methods.

(iii) the patent has to be susceptible to industrial application.

Aside from a precise description of the patented invention, a list of prior art needs to be included with the application. This serves to show links of the patent to prior work and to differentiate the patent from similar existing patents. These citations are a popular metric for scientific analysis (Zuniga et al., 2009), even though their value is somewhat questionable as there is a tendency to inflate the number of included citations in order to improve the likelihood of being granted the patent (missing citations can be regarded by patent examiners as an attempt to disguise the similarity of the application to prior art) (Zuniga et al., 2009).

Since academia creates large quantities of new knowledge, an increasing number of patents are assigned to scientists working at universities or public research institutions (Lissoni et al., 2008). However, since the traditional role of research institutions is to conduct basic research, there exists a conflict of interest. Academics are traditionally trained in an environment that stresses the Mertonian norms of science (Merton, 1973), i.e. scientists may regard a disinterestedness in commercial application of their work as part of their professional ethos. Patents, intended for the commercialization of inventions, are thus not an ideal incentive for scientists. When scientists do disclose inventions by filing for a patent with their university, that can lead to a different type of problem: university technology transfer offices (TTOs) may amass large numbers of patents without being able to commercialize the underlying inventions. The inventors are typically more interested in their academic careers, for which patents (and in particular the application for patents) have been largely insignificant in the past. The TTO, in turn, may lack the technical know-how, the resources and the entrepreneurial spirit to commercialize an invention. In the United States patents resulting from federally sponsored research used to be owned by the federal government. In the hope of easing the commercialization of research results, various U.S. government agencies negotiated deals with universities that would enable universities instead of government to own the intellectual property resulting from publicly funded research (Mowery et al., 2001). Eventually the individual arrangements were replaced by the Patent and Trademark Law Amendments Act of 1980 (more commonly known as the Bayh-Dohle, a reference to the two senators who sponsored the act). Subsequent policy changes further strengthened the concept of private ownership of publicly funded research results with the goal of more efficient commercialization (Eisenberg, 1996). The policy in the United States before and after the Bayh-Dole act represents two viewpoints on the relation of governmental research fundings: prior to the act, the opinion that publicly funded research should be a public good was dominant: granting ownership of the intellectual property to private organizations would basically force the public to pay twice for the same invention — once through taxes and the second time due to the monopoly created by the patent protection. Proponents of the opposing viewpoint realized that the creation of knowledge and the subsequent transfer into an applied context were separate. Private organizations were regarded as more efficient in commercializing inventions. Hence granting patent ownership to those with an incentive and the abilities to use them was hoped to prove beneficial for the economy (Eisenberg, 1996). In essence, paying twice for a useful innovation was seen as preferable to paying once for an invention that would be discarded. As a result of the Bayh-Dole act, patenting and licensing at universities has increased, although it has been noted that a significant proportion of the observed increase in patent applications and licenses can be better explained by the higher relevance of some scientific disciplines for commercial endeavors (Mowery et al., 2001). For example, the emerging field of biomedicine in the 20th century led to many patents that were easier to commercialize compared to patents from other disciplines.

Irregardless, the perception of the Bayh-Dole act as effective policy tool influenced policy in other countries. Germany introduced a policy that was similar in intent to the Bayh-Dole act but differed in some details. German policy on patenting for most of the 20th century was based on decrees introduced during the second World War. The Göring-Speer decree forced inventors to disclose inventions to their employer, while providing that the inventor would retain rights to compensation. The decree was intended to provide incentives for the development of technologies that would be valuable to the military (Koblank, 2012). With regard to inventions at universities, German law had one peculiarity, it permitted university employees full rights to their inventions. Even though this provision ostensibly favored the commercialization of academic knowledge, it caused similar issues to those of the United States provisions that would reserve ownership of publicly funded research invention to the federal government: the inventions were rarely successfully commercialized. A probable cause was the focus of academics on their career and thus on publications. However, publishing an invention in an academic journal could conflict with patent law, which requires inventions to be novel and thus not previously published for a patent to be granted. The so-called professor's privilege was abolished with a reform of the "Arbeiternehmererfindergesetz" (employee inventor law) in 2002: the ownership of inventions by university staff was divided between university and inventor. This resulted in a situation that is largely similar to that established by the Bayh-Dole act, as German universities were thus able to attempt to commercialize the inventions of their employees (Kilger and Bartenbach, 2002). Other countries have followed suit or shown interest in emulating the Bayh-Dole act (Mowery and Sampat, 2005). However, there are also exceptions to the rule: Italy has taken a policy step that is close to the opposite of recent German policy changes (Breschi, Lissoni, and Montobbio, 2007). While Bayh-Dole-like policy is widely regarded as working as intended (Grimaldi et al., 2011) and not detrimental to the basic-research mission of universities (Thursby and Thursby, 2011), there is also criticism of the concept of universities as institutions for the commercialization of inventions as suffering from unnecessary delays and largely inefficient (Kenney and Patton, 2009). Researchers also warn that emulating U.S. policy in other countries without taking into account the specifics of the national innovation system may be precipitous (So et al., 2008).

The wide range of available policy options and the relative lack in consensus which option is preferable paired with the large interest in the topic indicates that the potential of knowledge transfers from university to industry has been widely realized. It also indicates that further research is required to understand how to unlock this potential. Prior research has determined that the best way to accomplish this task is to improve understanding of the academic patenting process shifting policy focus to the individual who may engage in academic patenting (Clarysse, Tartari, and Salter, 2011). The key issue with regard to this question is what motivates scientists to engage in academic patenting. Once the motivational aspects of academic patenting are understood, factors that influence this motivation may be added to arrive at a better understanding of the whole process. The role of incentives and barriers in the academic patenting process and how these interact with motivation should explain whether scientists engage in invention disclosure with their university. While incentives play an important role and represent useful policy tools for encouraging commercialization of inventions, they are not sufficient to explain the motivation of scientists in either participating or refraining from academic patenting. The motivation is more likely shaped by the scientist's environment. In this respect the organizational distance described in previous sections may affect the effectiveness of academic patenting as knowledge transfer mechanism. Even though scientists are part of the same organization that has an interest in commercializing the result of scientific work, a distance may be perceived between the traditional norms of scientists (Merton, 1973) and the aspirations of university staff intent to commercialize (e.g. staff of TTOs). Understanding the interaction of scientist's normative distance to the concept of commercialization and the role of incentives should improve our understanding of the motivation scientists may have to commercialize. This, in turn, enables an understanding of which policy instruments are effective at eliciting academic patenting.

An overview of existing literature on this question is given in chapter 6 along with a proposal how to advance existing research in this field.

2.2.2 Broadcast search

The previous section described two opposing views on the commercialization of academic knowledge. This section demonstrates an alternative. To briefly aggregate the previously mentioned views: one side regards governments and, in extension, universities as inefficient when it comes to commercialization and cautions against policy that makes consumers pay twice for an invention — once through taxes used for publicly funded research and a second time for the acquisition of the resulting products or services from private organizations that commercialized the knowledge. The other side claims that universities should play a role in the commercialization of knowledge that would otherwise be neglected by the government or by individual researchers who are not particularly interested in commercialization. While the latter view appears to be dominating, research on the effectiveness of policy comes up with mixed results. Critics of current policy recommend to turn back to previous policy. However, there exist alternative transfer channels that may differ in important aspects to traditional means of commercialization of academic knowledge. One such channel is Broadcast Search. Prior to defining the method of Broadcast Search it is useful to briefly illustrate the concept of Open Innovation, as Broadcast Search can be considered as a method that is based on insights from research on open innovation in the area of innovation management.

The basic premise of open innovation is that organizations are more efficient at innovating when they consider knowledge from outside their organizational boundaries in addition to knowledge obtained by internal R&D efforts (Chesbrough, Vanhaverbeke, and West, 2006). Embracing openness as norm has enabled the phenomenon of open source software, which led to the production of high quality software tools as well as their adaption and subsequent modification by users. Hence open source software is an example how switching from a focus on financial rewards to an open sharing of information can be just as effective, if not more so, in the creation and diffusion of knowledge (Feller, 2005). This is, in part, possible as proponents of open source software consider participating in the open development as a reward in itself (Cristina and Rossi, 2004). A related mechanism for applying open innovation principles to overcome organizational boundaries when accessing external knowledge is crowdsourcing. Crowdsourcing can be regarded as a more abstract form of Broadcast Search: in crowdsourcing web platforms are used to distribute tasks, such as solving technical problems but also tasks involving product design or relatively simple manual labor, to a large number of participants (Howe, 2006). The exact definition of crowdsourcing is given by Howe as:

> *"Simply defined, crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call. This can take the form of peer-production (when the job is performed collaboratively), but is also often undertaken by sole individuals. The crucial prerequisite is the use of the open call format and the large network of potential laborers." (Howe, 2016)*

Division of labor is applied when the complexity of a task surpasses what an individual can reasonably accomplish in a given time frame.

Crowdsourcing extends this logic using web-based technologies for the organization of a large network of participants. However, crowdsourcing goes beyond a more extensive version of division of labor as it enables solutions that are beyond the scope of what the organizing entity is capable of creating (Brabham, 2008). This is due to a special characteristic of problem solving by crowds: the aggregation of knowledge leads to an increase in performance solution with the size of the crowd (Surowiecki, 2004). This feature is also called "collective intelligence" and is likely to have an impact on society that goes beyond the design of commodities (Pierre, 1997). Some negative aspects of crowdsourcing are closely related to its commercial success: the crowd is usually paid very little. Even the providers of winning solutions or successful product designs receive a compensation that is rather small compared to the revenue that the solution provides (Postigo, 2003). However, the opportunity to participate in open projects allows participants to acquire new skills which may be of use in their career or intrinsically motivating and thus sufficient to compensate the relative lack of financial rewards (Lakhani et al., 2007).

A popular example for crowdsourcing is Threadless, a company that allows users to design T-shirts. The shirts are manufactured and sold by Threadless with some of the proceeds going to the designer. The result is a web-shop that leverages customer's insights into market demands directly. The use of solution information provided by users, i.e. those who are most exposed to the problem, enables companies to be more successful by providing products that are aligned with market demands that may not have been fully realized by either the company nor, indeed, the market (Von Hippel, 2005).

Broadcast Search is a problem solving methodology that enables organizations to access knowledge relevant to finding a solution to their problems by broadcasting the problem to potential solvers. Broadcast Search is a mechanism that enables companies to open up their innovation process. Problems, usually related to technical issues that a company's internal R&D department cannot resolve given time and budget restraints, are published by the company (in this context referred to as seeker) to individuals who may have access to knowledge external to the company and suitable for solving the problem (hence these individuals are referred to as solvers). This contrasts with a closed innovation approach of relying on internal capabilities alone: internal R&D will be limited by the search method usually applied in problem solving. Specialists rely on past experiences and learning and therefore are more likely to rely on a "local search" of the solution space (Simon, 1991), i.e. only that part of the solution space is considered that mostly overlaps with the solvers existing knowledge. Hence the term bears some resemblance to the equally named computer science algorithm for finding the optimum of a function by iteratively comparing adjacent function values to the current value. While this approach is effective, there is a risk of identifying only local maxima and ignoring the global maximum, which may differ significantly from the local solution. Since the information on the problem and the information required to solve the problem must exist in one location (or locus) and since it is often more difficult to move the "sticky" solution information (Von Hippel, 1994), broadcast search is a promising method for dealing with challenging problems. The value of the concept as knowledge transfer mechanism for academic knowledge has been recognized by Lakhani et al. (2007). They find that broadcast search not only serves to solve problems that internal R&D departments cannot solve but also that the knowledge provided by solvers was often based on existing solutions for other problems, which indicates that broadcast search efficiently transmits existing knowledge for application in new contexts. Jeppensen and Lakhani (2010) use the example of the longitude problem to illustrate some key characteristics of broadcast search: in the 18th century naval navigation was crucially limited by the lack of a reliable method to determine a vessel's longitude. An innovation contest with financial reward was organized for a method that would reliably determine the longitudinal position of a ship at sea with greater accuracy than estimates based on a starting position and extrapolation with course and speed (which would quickly lead to an accumulation of errors because of difficulties in measuring speed at sea). The winning solution, a reliable chronometer that enabled the comparison of local time to a reference time, was based on a technology that existed but was deemed unsuitable due to being not robust enough for use at sea (the pitching, yawing and rolling motions of a ship in strong seas would interfere with the mechanisms of existing chronometers). The example illustrates that specialists may overlook viable solutions that are outside of their field of expertise and that the solution may have been implemented in similar form in other contexts. Solutions submitted by "the crowd" have been found to perform better when compared to solutions offered by internal staff (Poetz and Schreier, 2012). The innovation contest underlying broadcast search differs from other types of contests, such as salesforce contests, that have been studied in the past: participation is voluntary and the goal is to obtain one high performing rather than many satisfactory solutions (Terwiesch and Xu, 2008).

In broadcast search information on the problem to be solved is usually broadcasted by an innovation intermediary. This can be a web-based platform that publishes the problem to the general public or a company internal platform that shares problem details only to employees. In that regard broadcast search is similar to the more general concept of crowdsourcing.

Broadcast search can be regarded as one form of crowdsourcing which is limited to solving problems that have been defined by the seeking company. Hence it is also sometimes referred to as innovation crowdsourcing. A key aspect in both broadcast search and crowdsourcing is providing suitable incentives for submissions. In broadcast search the best solution is usually rewarded with a fixed sum determined by the seeking organization. An obvious issue of this setup is the effect of the incentive in a public setting. In order to attract specialists, the incentive needs to be large enough. Yet a large incentive is likely attract larger numbers of submissions (Che and Gale, 2003). As is the case with academic patenting, research is required to identify the effects of incentives and interactions with the drivers of participation in innovation contests. In this context knowledge distance, as defined in previous sections, is likely to determine the effectiveness of broadcast search as a transfer channel. Knowledge distance is a key feature of broadcast search in that it mitigates the issues of local search encountered in a closed innovation setup. However, the effects of knowledge distance on participation in innovation contests remains unexplored.

Chapter 7 will give an overview of existing research on broadcast search and contribute to prior literature with an experiment that explicitly models the effects of knowledge distance on innovation contest participation.

2.2.3 Collaborative research

This section provides an overview on the topic of collaborative academic research and its potential as transfer channel. Cooperation can take place in a multitude of settings, from inter-firm alliances (Grant and Baden-Fuller, 1995) to division of labor in small startup teams (Forbes et al., 2006). Collaborative research is one such field, which involves cooperative efforts between several scientists in scientific projects. A formal definition is given by Katz and Martin (1997) who describe collaborators of scientific projects as:

> *"(a) those who work together on the research project throughout its duration or for a large part of it, or who make frequent or substantial contribution; (b) those whose names or posts appear in the original research proposal [...]; (c) those responsible for one or more of the main elements of the research (e.g. the experimental design, construction of research equipment, execution of the experiment, analysis and interpretation of the data [...] writing up the results in a paper) [...]; (d) those responsible for a key step (e.g., the original idea or hypothesis, the theoretical interpretation); (e) the original project proposer and/or fund raiser, even if his or her main contribution subsequently is to the management of the research (e.g., as team leader) rather than research per se." -Katz and Martin (1997)*

The definition excludes those who make only a small contribution or are not regarded as research staff such as technicians or assistants (Katz and Martin, 1997). Note that not all of these requirements need to apply at the same time and that Katz and Martin acknowledge the existence of many exceptions which make it difficult to exactly define research collaborations.

Scientific collaboration is usually measured using mulit-author research publications, also referred to as co-authored papers (Smith, 1958). Even though the measure is lacking in some regards — it does not provide information on the type and degree of collaboration, nor does it always indicate who contributed to the research as important participants may be left out or individuals who did not contribute may be included among the authors of a publication (LaFollette, 1992). Still, as a measure it is readily available through publication databases and thus frequently used in analyses on collaborative research (Solla Price et al., 1986). The large number of results obtainable from publication databases compared to the relatively low expenses makes bibliographic information on co-authorship attractive when compared to surveys, which may offer more detailed information on a much smaller selection of publications (Katz and Martin, 1997).

Scientific publications are increasingly the result of collaboration between multiple authors, although the degree to which co-authorship increases varies with discipline (Katz and Martin, 1997). Several reasons exist for this trend: with the increasing complexity of research projects, owing to the maturity of many fields, specialization has been increasingly important for scientific advancement. Cooperation between scientists increases the knowledge available to a team and therewith the chances of advancing the respective scientific field (Bush and Hattery, 1956; Goffman and Warren, 1980). Researchers with diverse backgrounds may benefit from cooperation due to cross-fertilization between disciplines (Braun et al., 1992). Additionally, costs for travel and communication have fallen in the 20th century, enabling cooperation over longer distances. The costs for fundamental research have increased in the same time, representing difficulties for funding bodies that need to rely increasingly on resource pooling with international teams. For example, the construction of experimental fusion reactors can take decades and cost billions of US Dollars, and are therefore funded by international consortia (Ikeda, 2009).

A further driver for research collaboration is that aspiring researchers are often taught by cooperating with their more experienced supervisors (Beaver and Rosen, 1978). As the relation between supervisor and aspirant often lasts past the time required for education, "invisible colleges" are established in the long term, which influence the patterns of collaboration in science (Solla Price and Beaver, 1966). For a more exhaustive review of drivers of multi-authorship see Katz and Martin (1997).

The interest in scientific collaboration can be explained by its impact of researcher's performance: highly productive scientists are rare (Lotka, 1926), which prompted research into what determines scientific productivity. As a result, a correlation between scientific performance and inclination to collaborate has been found (Solla Price and Beaver, 1966). As researchers are more likely to turn to other researchers when searching for information instead of using impersonal sources, collaboration occurs frequently, further adding to its impact (Allen, 1984). Co-authored papers are also more likely to be accepted for publication (Gordon, 1980) and generally have higher impact (Lawani, 1986).

The importance of collaboration for scientist's productivity highlights its use as a transfer channel for the diffusion of knowledge: the extent to which collaboration is useful to the individual scientist depends on the degree to which knowledge can be transmitted in collaborative projects. When studied as a transfer channel, scientific co-publications are associated with positive effects on scientific productivity (Beaver, 2004; Lee and Bozeman, 2005) and the economy in general (Perkmann and Walsh, 2009), although evidence for the latter is sometimes mixed (Dietz and Bozeman, 2005).

Given the effectiveness of collaborative research as a transfer channel in the scientific domain, it is not surprising that the same channel may also be used for the transfer of knowledge from academia to industry (Caloghirou, Tsakanikas, and Vonortas, 2001). In this case collaborative projects may be indicated by multi-author publications with authors from both industry and academia. For industry, access to academic knowledge is helpful in developing products or services based on the scientific state of the art, enabling competitive advantages (Hanel and St-Pierre, 2006). For scientists, access to industrial resources, the opportunity to learn or to commercialize their research constitute incentives for collaboration with industry (D'este and Perkmann, 2011). In general, prior research has reached the consensus that this type of transfer is beneficial for the economy at large (Bozeman, Fay, and Slade, 2013), but also points out some challenges that influence the channel's effectiveness: collaboration is, as other transfer channels, affected by different types of distance. Social and spatial proximity are regarded as antecedents of collaboration: social proximity leads to informal communication, which can then result in formal collaboration. Spatial proximity eases communication and thus enables social proximity to act on the probability that collaboration ensues (Hagstrom, 1965; Katz, 1994). In the special case of university-industry knowledge, transfer through collaborative projects, spatial proximity plays a significant role as some spatial distance is usually implicit when collaboration spans institutional boundaries (Abramo et al., 2011). A framework that has proven useful for estimating the effect of collaboration is the extended knowledge production framework (Jaffe, 1989; Griliches, 1979). The knowledge production function (KPF) estimates the knowledge output of geographical areas (e.g. nations or regions) as function of industrial inputs with a special focus on the interdependence of geographical units. Specifically, spillovers between actors in spatially adjacent units can explain unintended effects of knowledge aggregations. These spillovers are referred to as Marshall-Arrow-Romer spillovers (Glaeser et al., 1991), named after the three economists who developed the concept (Marshall, 1898; Arrow, 1971; Romer, 1989), when they are assumed to effect industry positively due to a pooling of specialized knowledge. Jacobian spillovers provide positive effects by close proximity of diverse actors whose knowledge may lead to cross-fertilization (Jacobs, 1970). Recently, measures derived from the network structure of actors in collaborative projects have proven helpful in determining which collaborative efforts are relatively more successful (Liao, 2011)

Knowledge distance is also likely to be relevant for collaboration as a transfer channel since the type of research in the public and private areas can differ with respect to discipline or domain, leading to different types of benefits and different degrees of importance of public research for industry (Cohen, Nelson, and Walsh, 2002).

Chapters 8 and 9 will set out to test the effects of various types of distance on the effectiveness of university-industry collaboration as knowledge transfer channel using an extended knowledge production framework.

2.3 Research questions

In the previous sections we defined university-industry knowledge transfer and selected three transfer mechanisms for further study. The selection enables to capture a broad range of the aspects relevant to knowledge transfer that have been identified in prior literature. They are differently affected by various types of distance, as described in the last sections. This section condenses the theoretical background into three research questions that transform the general goals of this thesis described in Chapter 1 into more precise proposals for advances in the respective scientific fields. A research showing the three transfer channels between industry and academia (with technology transfer offices and crowdsourcing platforms as intermediaries) and the direction of knowledge flow is illustrated in Figure 2.1.

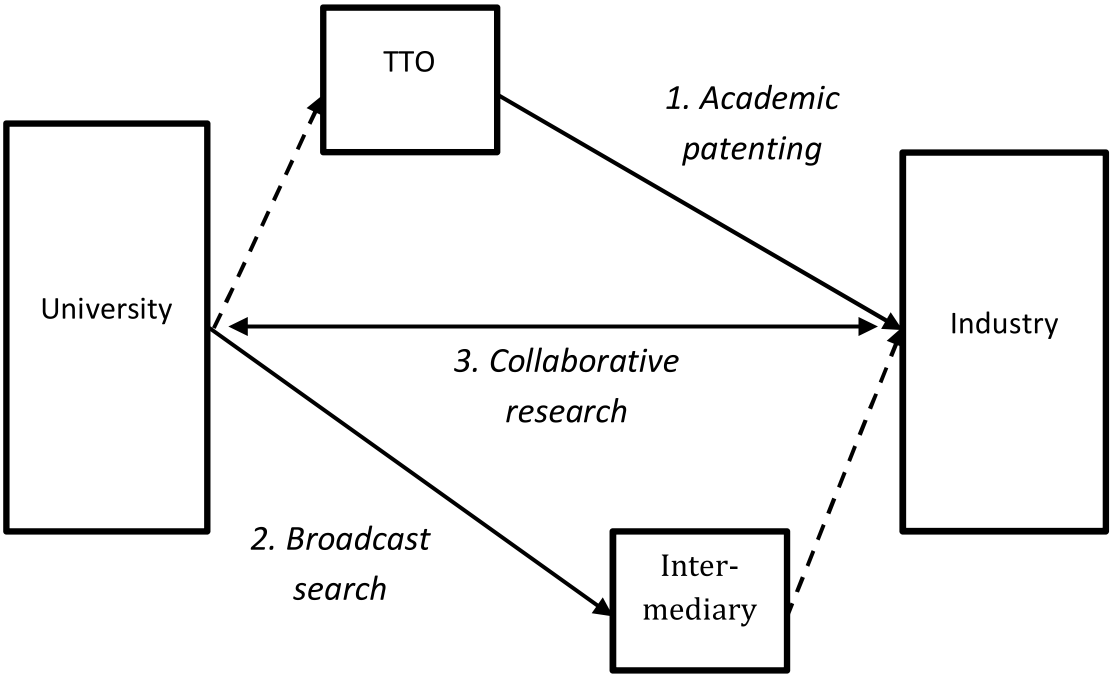


FIGURE 2.1: Research Agenda

The following chapters will deduce and attempt to contribute to research on the following topics: **Automatic Literature Reviews:** Chapter 5 illustrates how Data Science methods (see Chapter 4 for details) can be used to complement traditional research methods. The chapter also shows that such methods enable results that go beyond the scope of traditional literature reviews.

Motivational Aspects of Academic Patenting: In order to provide effective policy for the commercialization of academic knowledge, it is imperative to research academic patenting at the individual level. This provides us with an opportunity to understand scientists' motivation in engaging or ignoring academic patents as mechanism to transfer knowledge. The normative aspects influencing scientists' motivation in combination with incentives are likely to explain under which conditions scientists are willing to participate in commercialization of their inventions. Tailored policy based on such findings is likely to be more efficient at eliciting commercialization, thus enabling a more efficient transfer of knowledge from academia. Chapter 6 describes a discrete choice experiment used to study the motivational aspects of academic patenting.

Broadcast Search and Knowledge Distance: As an alternative to traditional transfer channels, web-based intermediaries provide a novel solution to some of the problems associated with knowledge transfer. By specifying the problem in advance and exploiting researcher's interest in explorative work broadcast search may represent a more efficient method of transferring "sticky" information from academia into applied contexts. However, research is necessary to understand under which conditions researchers will participate in innovation contests. The concept of knowledge distance is likely to play an important role in this regard. In order to exploit broadcast search as efficient transfer mechanism, the role of knowledge distance in motivating participation in crowdsourced contests as well as the role with regard to the quality of submitted solutions needs to be determined. Chapter 7 describes the topic in more detail and discusses the results of a project that aimed to better explain the participation behavior of scientists in innovation contests.

Regional Perspective on Collaborative Research: Collaborations between industry and universities in research projects represent a transfer channel that can be measured with bibliographic data. The knowledge production function is often used as a tool for measuring the impact of collaboration on the production of new knowledge by industry. Hence this framework allows to measure the impact of technology transfer in addition to the effect of various distance types on the transfer channel's effectiveness against the theoretical backdrop of Marshall-Arrow-Romer and Jacobian externalities. Chapters

8 and 9 will investigate some of the aspects that lead to a positive effect of collaborative research on knowledge production.

Chapter 3

Data

This chapter introduces the main data sources for the projects outlined in the following sections and briefly explains some of the preprocessing steps that were required to prepare the data for analysis.

The main data sources for this work were patent and publication databases. These were combined with additional primary data obtained from surveys and secondary data from official sources such as the federal statistical office of Germany. The main data sources were used in all of the projects described in subsequent chapters. Hence they will be introduced here in greater detail. The other data sources, which are specific to the projects, are described in this chapter as there are some similarities in the way they were collected. More information on these sources is included in the chapters describing their analysis.

The primary databases used for this thesis are the Web of Science (by Thompson Reuters) and the European Patent Offices's Statistical Database (Patstat). Both types of databases contain information particularly relevant for bibliometric analysis, i.e. meta-information typically saved during the publication process such as author names, author locations or publication titles. This type of information is useful for descriptive analysis and in network analysis (e.g. in the context of networks of co-authors of publications). The databases also contain abstracts, a potentially rich source of information for automatic information extraction. Advances in computer science, specifically in machine learning and computational linguistics allow extraction of information from large numbers of abstracts within reasonable time-frames. However, in order to extract any information from the databases, some challenges have to be overcome: data is often incomplete or of mixed quality, which necessitates complex data-cleaning procedures to prevent such errors from distorting the subsequent analysis. The size of database tables can make it challenging to apply queries unless the limitations of available hardware are understood. The following sections describe the two main databases in more detail.

3.1 Patent data

Patent data has been a mainstay for econometric analyses for several decades (Zuniga et al., 2009). The availability of patent data, a useful byproduct of the procedures required to maintain the patent system, allowed to derive valuable indicators for scientists in economics. Aggregated patent counts indicate country's, region's or firm's research and development capabilities. As each patent is required to be novel and non-obvious and as the patenting process can be costly, each patent is a signal that knowledge has been created. A downside of using patent data is that patents represent only a fraction of the inventive output of an economy. Since patenting is time-intensive and expensive, inventions of low value are not likely to receive protection. However, large companies may want to protect important technologies by filing large numbers of patents that are similar to the one patent that covers their key technology. This process of using "patent thickets" protects the company from other companies that may try to circumvent the original patent by filing their own patents with similar technologies that achieve the same purpose (Zuniga et al., 2009). The result is an inflation of relatively low-value patents that further complicates the assessment of patent counts. However, since valuable inventions are likely to be patented, patent counts are a useful measure when determining the knowledge output of larger aggregates such as regions or states.

The main database for patent data used in this work is the PATSTAT (patent statistics) database (version of 2012) by the European patent office. To complement PATSTAT, the database of the German patent office as well as databases by the OECD (Organisation for Economic Co-operation and Development) were used. This section introduces PATSTAT, its structure, the information contained within, the challenges in extracting information and how supplementary datasources can be used for that purpose.

3.1.1 An introduction to PATSTAT

PATSTAT is a snapshot of the European Patent Offices' (EPO) master documentation database (DOCDB). It is designed for statistical analysis of patents by intergovernmental organizations and academic institutions and contains data from more than 100 national and international patent offices (although the degree of coverage differs) (Tarasconi and Kang, 2015). The central table of the database contains information on patent applications as shown in Figure 3.1. The table records applications for patents with information on the application date and what kind of protection was applied for (patent, utility model, design patent or other). For the 2012 version the table contains approximately 73 million entries. The ID of individual applications is used to link the application table to other tables with additional information. Of those the tables containing information on persons, application classification, abstracts and titles are most relevant to this work. The person table lists individuals associated with one application. Persons are divided into applicants and inventors. Typically applicants are assumed to be owners of the patent whereas inventors are the individuals who carried out the inventive work. Legal persons such as companies or governmental institutions can occur as either applicant or inventor, but are not specifically marked. The person table contains columns with the person's name, address and country code. Since the data are aggregated from many national patent offices, the quality and coverage are not homogeneous. The address field is often empty, hence additional data sources were required to localize some patents. Since the person names are derived from applications, one individual or company may occur many times in the table. Title and abstract are saved in two tables as raw text. These entries describe the patent as mandated by the respective patent office's rules.

Since one patent may be linked to multiple applications, for example when the applicant files for a patent at another patent office to extend the protection to cover several countries, it is possible to double count patents. PATSTAT contains tables on patent families

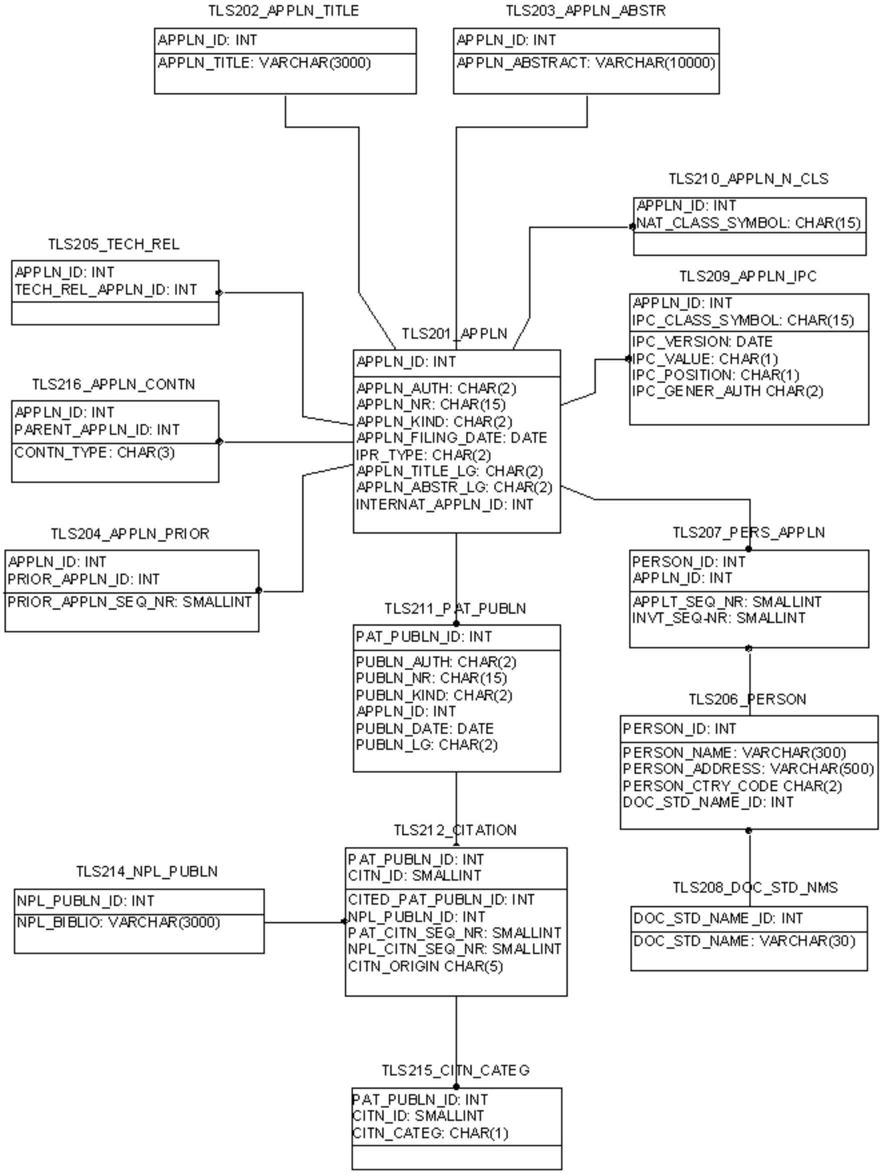


FIGURE 3.1: PATSTAT Physical Model, *CEMI's PATSTAT knowledge base*

which combine applications for the same patent. Several versions of the concept of patent families exist, the most common links patents by their priority date, i.e. the earliest application date of the applications in a family.

Furthermore, the database contains tables with information on the patent's classification. The international patent classification

(IPC) is a hierarchical classification scheme that assigns patents to specific sub-fields of one of the following industry sectors:

A: Human Necessities
B: Performing Operations, Transporting
C: Chemistry, Metallurgy
D: Textiles, Paper
E: Fixed Constructions
F: Mechanical Engineering, Lighting, Heating, Weapons
G: Physics
H: Electricity

A patent is assigned to an IPC class by its function or by its field of application. One patent can be assigned to multiple IPC codes. In a separate table the EPO's European Classification (ECLA) scheme is available, which is a more refined scheme with 140000 categories compared to the IPCs 70000 (*OECD Patent Statistics Manual*; 2009). The classification schemes are primarily used for identifying prior art, i.e. enabling patent offices to determine whether a new patent application fulfill the novelty requirement. However, the classification also proves useful in the analysis of patent data, for example when determining technology life cycles or cross-technology fertilization (*OECD Patent Statistics Manual*; 2009). However, when applied to newly emerging fields of technology, existing classification schemes may be insufficient as will be shown in the next section.

PATSTAT is a popular choice among researchers for a variety of analyses such as patent citation analysis, patent count analysis, inventor analysis or technology class analysis. But it also comes with some downsides: data may be missing from some tables, most importantly person's address data and the coverage of patent offices differs with a bias in favor of western European nations (Tarasconi and Kang, 2015). The next section highlights some of the steps that were undertaken to retrieve, clean and extend data from PATSTAT.

3.1.2 Preprocessing patent data

Initially the prospective user has to decide whether to use the online or offline version of PATSTAT. The offline version was preferred as it offered greater flexibility with regard to some of the steps described in this section. The offline version of PATSTAT consists primarily of three DVDs with raw data (additional products on patent legal status are also availalbe). These contain approximately 15GB of compressed data that can be uploaded into an SQL (structured querly language) database of about 150GB. As described in chapter 2, we analyze patents from the field of nanotechnology filed in Germany. While limiting the dataset to applications filed in Germany is simple, identifying nanotechnology patents is not as easy. At the time of writing, both the IPC and ECLA classificaiton schemes had introduced classes for nanotechnology, but these are not yet a good tool for identifying respective patents, as shortly after their introduction not many past patents were categorize according to the new classificaiton system. Hence a keyword-based search based on (Porter et al., 2008) was used to identify nanotechnology patents. Compared to a simple keyword search such as "nano*" (where "*" is a wildcard that refers to any combination of additional characters following "nano"), the strategy offers some benefits. The simple search also returns false positives such as patents mentioning the chemical formula of sodium nitrate ($NaNO_2$). In addition to searching for combinations of relevant keywords, the search strategy has exclusion terms that remove false positives.

Research on academic patenting has been complicated by the fact that academic patents are not always easy to identify. Patent databases usually list the owning individual and company along with inventors. Information on the inventor's occupation is often missing. Hence, when a scientist files a patent with university, it is possible that the scientist's name along with the name of the research institution are recorded on the patent application. However, when the patent is filed as a result of industry-university cooperation, the title is owned by a company. Since inventors do not usually give their occupation and the university name is missing, these patents are difficult to identify, although some researchers have successfully used publicly available information on university staff to circumvent this issue. For the purpose of this thesis, the problem could be partially solved using a complex query of patent databases in combination with survey data, as will be described in greater detail in subsequent chapters.

3.2 Publication data

Scientific articles, conference proceedings, book chapters and similar publications are increasingly available through online databases such as Web of Science, Scopus or Google Scholar. These databases are useful tools for scientists attempting to search specific literature and thus prepare future research projects, as documents are typically indexed and the databases come with convenient user interfaces for keyword-based searches. These databases usually also provide information not directly contained in a publication but relevant to researchers: citation counts serve as basis for measures regarding an article's impact. Links to other articles (inbound and outbound citations) are useful for scientific meta-analyses that can identify structures within scientific fields. Bibliographic information (e.g. author names and affiliations) can be used to construct networks that are useful in a variety of research projects. The importance of such databases was anticipated in a seminal article by Bush (1945) some 50 years before the advent of the Internet.

3.2.1 Introduction to the Web Of Science

The Web of Science is a scientific citation indexing service and database based on the science citation index introduced by Garfield (1955), who recognized the value of citations as indicator for research with high impact and as tool for finding related research articles. The Web of Science covers some 12.000 journals and documents published as early as 1900 in multiple disciplines. The database contains over 90 million records and has indexed over a billion citations (according to the Web of Science website). Aside from indexed citations and bibliographic information on articles, the database also provides publication abstracts (coverage varies with publication date) and a user interface that facilitates the download of large amounts of data. This makes the database a good choice as data source for the research projects described in the following chapters.

3.2.2 Preprocessing publication data

Retrieval of publications using the Web of Science is straightforward. A keyword-based query with some options for wildcards (i.e. special characters that allow for more complex search strategies by matching, for example, arbitrary characters in addition to predefined keywords) enables a selection of records. The selection can either be analyzed using the web interface, for example to obtain frequency counts for publications by journal in a field, or downloaded. Various file formats are available. We chose the ISI flat file format due to its compatibility with tools used in some of the projects described later (e.g. Chapter 5). For other projects the data had to be converted into other formats. For example, geocoding articles involved extracting information on affiliations recorded in the article (relevant in Chapter 8). Information related to individual researchers also required writing custom scripts for name disambiguation, i.e. merging names when one author is referenced in multiple documents. Initially data was cleaned to remove non-alphabetical characters such as accents or diacritical marks. The similarity between two names was computed using string-based similarity metrices such as Jaro-Winkler or Levenshtein distance (Cohen, Ravikumar, and Fienberg, 2003). As the information available differed from case to case (older publications usually recorded author last names and first name initials, whereas more recent publications contained full name information), a custom script was written that takes into account available information and uses name-similarity as well as other available information (such as affiliation and co-authors) to disambiguate the records. The resulting set of matches was manually checked.

3.3 Other data sources

The research projects presented in the next chapters use additional data with the source differing from project to project such as governmental databases and data based on publicly available requests for proposals published by innovation intermediaries. This section provides a very brief overview of the main types of data as shown in Table 3.1. For details see the respective chapters.

Chapter/Datasource	Patent data	Publication Data	Innovation intermediary data	Statistical data on German regions
Chapter 5		X		
Chapter 6	X	X		
Chapter 7		X	X	
Chapter 8	X	X		X
Chapter 9	X	X		X

TABLE 3.1: Data overview

Chapter 4

Methods

This chapter introduces the reader to Data Science, an emerging discipline with the potential for significant impact on methodologies for empirical research. The second sub-chapter describes discrete choice experiments. The methods discussed in this chapter are applied in several of the subsequent projects and thus introduced here in some detail to avoid redundancies in later chapters.

4.1 Data science & empirical research

Data Science, an emerging field of occupations in data centric companies, is concerned with the collection, processing and analysis of data in order to derive insights that translate into value for a business (Barga, Fontama, and Tok, 2014). Data scientists need to be proficient with skills from several fields such as statistics, programming and solid understanding of theoretical aspects relevant to the subject. The theoretical foundation is required to hypothesize on possible relations between objects of interest. Theoretical groundwork enables scientists to define which data are necessary to test hypotheses. The required data are often not directly available in a form that allows analysis. Instead the information is often contained in unstructured data or dispersed over data bases. Programming skills enable data scientists to collect and merge data, and to transform it into a suitable format for analysis. Statistical methods are used to model data and test hypotheses. Because of the importance of statistical methods in the analysis of data, data science is sometimes defined as an extension of statistics (Cleveland, 2001). However, data science also encompasses methods from computer science, specifically from machine learning. Machine learning, a sub-field of artificial intelligence, is concerned with algorithms that enable computers to learn.

Learning, in this context, is the automatic identification of patterns in data and the generalization of these patterns for the purpose of application to new data (Segaran, 2007). Some typical applications for machine learning algorithms are classification, clustering or building of recommendation systems (collaborative filtering). Machine learning is related to statistics in that is uses some of the same methods with different terminology (e.g. regression analysis can be considered as a form of machine learning), but it also encompasses methods that are not commonly used by statisticians (such as clustering algorithms). Due to the availability and importance of data to companies, data science has significant impact in applications in various sectors (Piatetsky-Shapiro, 2012). In the scientific world the methods used in data science are subject of research in the field of computer science. In addition, the application of these methods to real world problems is subject of research in some specialized data science journals such as the Journal of Data Science.

With regard to this thesis data science is of interest as its methods allow to derive insights from data sources that are rarely used in scientific analyses in the field of economics. Unstructured data in the form of texts (e.g. scientific publications) are a potentially rich source of information. Accessing and manipulating data contained in various databases requires at least some understanding of relational databases, including relevant programming languages such as the Structured Query Language (SQL). The larger the database and the worse the quality of the contained data, the larger is the required skill-level: given a certain size of databases, simple operations can take up significant amounts of time unless the operator is aware of advanced concepts which allow processing datasets in acceptable time frames. Additionally, if the data contained within the database needs to be cleaned or harmonized, the database software often has to be extended with special purpose functions. Given unstructured text, it is necessary to quantify the information contained within. In computer science this task is termed "feature extraction".

In traditional science, particularly in economics, there is a strong focus on theoretically grounded research. Empirical scientists tend to incorporate complex statistical models into theoretically driven research. However, the collection and preparation of data is not typically elaborated on. This is regrettable for several reasons. If high quality journals do not provide incentives for high quality data collection and preparation, the overall quality of research suffers. Expanding the focus of academic research in economics to include the aspects of programming, data cleaning and machine learning offers two benefits: 1) transparency on methods used makes it easier to reproduce and to connect to existing research, 2) machine learning enables quantification of data that would otherwise require prohibitive amounts of manual preparation. The following chapters briefly describe the machine learning algorithms and statistical methods that are used in the subsequent chapters to extract information from various data sources. Some of these algorithms were used in multiple projects and thus deserve discussion in some detail.

4.2 From raw text to topic models

The availability of large sets of unstructured data and of cheap processing power enables companies and research institutions to let computers extract information that is valuable in later stages of analytical projects. Initial attempts to characterize texts focused on reducing the amount of data by distilling texts into a few representative data points (Blei et al., 2003). One such method involves calculating simple frequency counts. These are useful for gaining a quick overview of the focus of a text. Simple frequency counts also show the necessity for pre-processing as the most common words in a text, such as articles, contribute little to the text's message. Hence a common pre-processing step is to compare a text to a list of "stop-words" and to filter out words that do not contribute meaning. The remaining frequency distributions are only a rough approximation of the text's content and are not very useful when comparing sets of documents. A more advanced method normalizes the frequency counts of words in one document with the frequency count of the same word over all documents in a corpus. Thus the "term frequency – inverse document frequency" (tf-idf) method weights words according to how characteristic they are of individual texts (Salton and McGill, 1986). Given a collection of texts in a corpus ($d\epsilon D$), the tf-idf weight of a word for one text can be calculated as the product of the frequency of that word in the current text ($f_{w,D}$) and the inverse document frequency. The inverse document frequency is the logarithm of the number of texts in a corpus ($|D|$) divided by the number of texts containing the word to be weighted ($f_{w,D}$):

$$tfidf = f_{w,D} \log \frac{|D|}{f_{w,D}} \tag{4.1}$$

Tf-idf is a robust method (which will be employed for some tasks in subsequent chapters) but still lacks complexity to accurately describe larger text collections and the interdependencies between documents. An advancement over tf-idf was achieved with latent semantic indexing (LSI), also sometimes referred to as latent semantic analysis (Deerwester et al., 1990). LSI is used to find a word with related meaning by analyzing co-locations within texts. LSI starts with a term-document matrix that is filled with weighted frequencies such as those derived from tf-idf. The matrix then undergoes a singular value decomposition (SVD), i.e. decomposing a matrix as product of three matrices, one of which contains the original matrice's singular values. In the context of LSI, SVD is used as a method of dimensionality reduction: the term document matrix with its high dimensionality arising from the large number of words in a corpus is reduced to a small number of vectors that preserve semantic information on related words. The linear subspace identified by LSI is able to capture some linguistic features of text such as polysemy and synonymy (i.e. one word with different meanings or different words with similar meaning). In order to verify the ability of LSI to recover semantic aspects of texts, generative models were created. Probabilistic LSI could be used to recover features of the generative model from text collections (Hofmann, 1999). Probabilistic LSI samples words in a text from a mixture of multinomial distributions over all words in a text collection. These distributions are labelled "topics" as they tend to contain semantically related words. However, probabilistic LSI does not provide a probabilistic model at the level of documents, which precludes the model from being applied to "unseen" documents, i.e. documents that were not included in the data used to estimate the model. This shortcoming was addressed with the Latent Dirichlet Allocation (LDA) model (Blei et al., 2003). LDA is a generative model that represents documents as mixture of latent probability distributions over words ("topics"). For the generation of one document, topics are sampled repeatedly from the dirichlet distribution and subsequently words are sampled from each of these multinomial distributions. This allows one document to be associated with several topics, which is a defining feature of LDA. The latent variables of the various distributions are uncovered from texts using Bayesian inference. The generative process for a document with N words is:

1: Choose $\Theta \sim Dir(\alpha)$

where Θ is a distribution of topics in a document obtained from a dirichlet distribution with parameter α.

2: For each of the N *words:*

a. choose a topic $z_n \sim Multinomial(\Theta)$
b. choose a word w_n *from* $p(w_n|z_n, \beta)$

where β is a $K{*}V$ (number of topics times vocabulary size) matrix where each row contains the probability distribution of words for one topic. The joint probability for a set of N topics z and N words w and a distribution over topics Θ is given by:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \tag{4.2}$$

LDA estimates the posterior distribution of hidden variables from the observed words of a document:

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)} \tag{4.3}$$

This distribution can be estimated with Markov chain Monte Carlo methods such as Gibbs sampling.

LDA allows the representation of documents in a reduced vector space similar to tf-idf, captures semantic structures as LSI and can be applied to documents not in the training set. The vector representation allows the comparison of documents with similarity metrics for such purposes as categorization or novelty detection.

4.3 Discrete choice experiments

Discrete choice models are used to predict people's decisions for one of several discrete alternatives. Discrete choice experiments encompass the collection of data for discrete choice models and their analysis. In order to gain a better understanding of variables that may affect a decision making process in a given context a discrete choice experiment confronts subjects to scenarios with differences in the levels or amount of some variables believed to be relevant. The subjects are then asked to choose the preferable scenario. Through repeated observation and systematic choice of variable levels we gain an understanding which variables contribute more to the decision making. Since discrete choice experiments are used in two of the subsequent chapters of this thesis, they deserve some elaboration. The following sections describe the experimental setup for discrete choice experiments as well as the statistical methods used to analyze collected data.

4.3.1 Experimental design

Discrete Choice (DC) experiments belong to the group of Stated Choice (SC) experiments, a popular framework for the design of experiments. SC experiments differ from traditional experiments in that they are intended to elicit a choice between alternatives rather than being based on revealed preferences, i.e. preferences that can be observed in real data (Louviere, Flynn, and Carson, 2010). The purpose of discrete choice experiments is to determine the influence of attributes on choice alternatives by surveying participants on their preferences for alternatives. A drawback of stated choice experiments is that when multiple attributes exist, a large sample size is required to estimate parameters. This makes it necessary to pool the obtained responses (Rose and Bliemer, 2009). Hence, in a typical DC experiment each respondent will be confronted with several choice situations, i.e. combinations of attribute levels, and asked to choose from one of the available alternatives. The resulting dataset contains several observations for each participant for a given combination of attribute levels. The objective of experimental design is to determine which attribute levels should be used for the various choice situations in order given that the number of participants to the survey is likely to be relatively small. For this purpose DC experiments are often described with a design matrix, which lists choice situations as rows and choices as well as choice-specific attributes in columns as shown in Table 4.1. To determine which attribute levels should be assigned to cells in the design matrix, researchers have traditionally used the orthogonal design approach. However, orthogonal designs have been subjected to some criticism (Kessels, Goos, and Vandebroek, 2006). Orthogonality refers to the correlation between attributes with an orthogonal design exhibiting zero in-between attribute correlation. This property of designs is important for the determination of independent effects in linear models (Rose and Bliemer, 2009). However, discrete choice models typically are not linear but instead assume a logit or probit model. With non-linear models the correlation of differences between attributes is more relevant to the design of choice experiments (Train, 2009). Designs that attempt to minimize the asymptotic standard errors of the parameter estimates (i.e. the square roots of the diagonal elements of the asymptotic variance-covariance matrix) have been found to yield more stable parameter estimates or reduce the required number of observations to reach stable estimates (Huber and Zwerina, 1996). Designs that are optimized to reduce the asymptotic standard errors of parameter estimates are labeled efficient designs (Rose and Bliemer, 2009).

	Choice 1 - Attribute 1	Choice 1 - Attribute 2	Choice 2 - Attribute 1	Choice 2 - Attribute 2
Choice Situation 1	0	2	1	2
Choice Situation 2	3	1	3	2

TABLE 4.1: Sample Design Matrix

The creation of an experimental design involves the characterization of an econometric model, the choice of a suitable design method, the generation of the design matrix and finally the incorporation of the design matrix into a survey. The econometric model is typically a multinomial logit model as described in section 4.3.2, which measures the preference for a choice alternative as function of choice attributes multiplied by parameters. Ideally the econometric model is fully specified before development of the experimental design. The specification includes the number of choice alternatives, the attributes and their levels as well as a distinction between choice-specific attributes and those that are constant across alternatives.

The next step concerns the generation of the experimental design and the associated design matrix. The required number of choice situations in a design is a function of the attributes and their levels. The smallest number of choice situations is the smallest number that is divisilbe by all attribute levels of the various attributes (Rose and Bliemer, 2009). The choice of design type is typically determined by practicality. A full factorial design, which requires all possible choice situations, is usually not feasible (as the number of required observations increases exponentially with the number of attributes). Instead, fractional factorial designs are used to limit the number of required observations. Fractional designs can limit the number of scenarios per survey participant by randomly selecting choice situations. However, the resulting parameter estimates may be biased. One popular fractional factorial design method which avoids this issue is the orthogonal design. The orthogonal design enables independent parameter estimation by selecting a subset of the full factorial design, while minimizing the correlation between attribute levels in choice situations. If the resulting design is still too large for an experiment (i.e. too many choice situations need to be considered by one participant, which would adversely affect the response rate), the design can be blocked. That is, different respondents are presented with different subsets of the orthogonal design. The blocks do not need to be orthogonal, but they need to satisfy attribute level balance (i.e. one respondent is not confronted with only high or only low levels for one attribute). However, an orthogonal design does not necessarily lead to an orthogonal data set: non-responses or varying response rates in blocked designs can cause the loss of orthogonality in the resulting data. Orthogonality may also be lost if additional variables, such as demographic data, is used in the estimation. These variables are often constant for individual respondents and thus introduce correlation between attribute levels into the data set (Rose and Bliemer, 2009).

Efficient designs provide an advantage over orthogonal designs in that they maximize the information obtained from each choice situation. For this purpose prior information on parameters are used to create a design that minimizes the expected standard errors of parameter estimates. The standard errors are the off-diagonal values in the asymptotic variance-covariance matrix obtained as the second derivative of the log-likelihood function. The matrix can be determined by Monte Carlo simulation. The orthogonality attribute is implicitly considered in efficient designs if the attribute has a negative impact on parameter standard errors (Rose and Bliemer, 2009).

The final step of a discrete choice experiment is to use the design matrix to create a questionnaire. The cell values of the design matrix are converted to descriptions of the attribute levels and inserted into a survey. The resulting survey data on choice preferences can subsequently be used in regression analysis.

4.3.2 Logistic regression and related models

In the simplest case a discrete choice can be modeled using logistic regression. In this case the choice between two alternatives is modeled using a linear combination of independent variables. Since the dependent variable is discrete, the conditional distribution of the dependent variable given the predictors is a Bernoulli trial with logistically distributed errors (in contrast to probit regression with its normally distributed errors the logistic distribution has higher kurtosis). Logistic regression models the probability of observing an event. The combination of predictors needs to be converted from continuous into discrete space using the logistic function:

$$F(x) = \frac{1}{1 + e^{-g(x)}} \tag{4.4}$$

with $F(X)$ as the probability of observing one of the two possible outcomes and $g(x)$ as linear combination of predictors and their coefficients:

$$g(x) = \beta_0 + \beta_1 x \tag{4.5}$$

This corresponds to taking the logit (the inverse of the logistic) function of the odds of the dependent variable. Taking the logarithm of the odds transforms the discrete dependent variable into a continuous space and establishes a link between the independent variables and the probability of observing an outcome.

$$\ln\left(\frac{F(x)}{1-F(x)}\right) = \beta_0 + \beta_1 x \tag{4.6}$$

When exponentiated, this allows to easily interpret the effects of coefficients by taking the odds ratio, i.e. the ratio of two odds: the relation of the odds of $f(x)$ and $f(x+1)$ results in $e^{\beta}x$. In other words, increasing regressor x by one results in a multiplicative change of the odds ratio of e^{β_1}.

Binomial logistic regression can be extended to multinomial logistic regression (also known as softmax regression or maximum entropy classifier) in cases where the dependent variable can take on more than two outcomes. In discrete choice analysis, multinomial logistic regression is popular as questionnaires often include a "none" option when asking for a choice between two options. Given the assumption that adding additional alternatives does not change the preference in a choice between two given alternatives, it is possible to model a multinomial dependent variable with k classes as function of $k-1$ binary choices with one alternative as reference to which the $k-1$ alternatives are compared:

$$\ln\left(\frac{F(Y_i=1)}{F(Y_i=K)}\right) = \beta_{1,0} + \beta_{1,1}x_i$$

$$\ldots.. \tag{4.7}$$

$$\ln\left(\frac{F(Y_i=K-1)}{F(Y_i=K)}\right) = \beta_{k-1,0} + \beta_{k-1,1}x_i$$

which, given that the probabilities of each outcome sum to one, can be transformed into the expression:

$$F(Y_i=K) = \frac{1}{1+\sum_{k=1}^{K-1} e^{\beta_k x_i}} \tag{4.8}$$

which can be used to estimate the probability of each outcome by dividing the exponentiated linear combination of predictors and coefficients for the outcome of interest by the denominator given in equation 4.8. As indicated in equation 4.7, each outcome category is associated with a set of coefficients. These are estimated using a modification of the maximum likelihood estimator that uses L2-regularization to avoid large coefficient estimates. In the context of discrete choice experiments it is often the case that in addition to variables that differ between subjects there are variables that characterize the choices. To account for choice-specific independent variables the multinomial logistic regression has been extended by the conditional logit model (McFadden, 1973). Another common issue in discrete choice situations is unobserved heterogeneity between subjects. To account for unobserved variables that may affect a subject's choice preference it is possible to add random effects to logit models. However, while this allows to correct for bias introduced due to unobserved heterogeneity, it is not very useful for inference. One way to explicitly model unobserved heterogeneity is to assign subjects to one of several groups of subjects. This latent class (LC) model allows to take into account individual and choice specific variables while providing information on whether there are clusters (latent classes) of subjects with different preferences (Lazarsfeld, Henry, and Anderson, 1968). In contrast to other latent variable models (such as random effects regression), the LC model's latent variable is categorical. The resulting assignment of subjects into categories with differing preferences makes LC regression useful for the interpretation of results arising from discrete choice experiments as the probability of a subject belonging to a given class can be linked to variables of interest (Hess et al., 2011). For the latent class logit model the probability of subject n who belongs to class s with a probability of π_{ns} choosing choice alternative i given a set of class-specific parameters β_s is given by:

$$P_n(i|\beta_1, ..., \beta_S) = \sum_{s=1}^{S} \pi_{ns} P_n(i|\beta_s), \tag{4.9}$$

which represents a finite mixture model that consists of weighted sums of probabilities for the distributions contributing to the mixture. The class probabilities can be estimated with a multinomial logit model:

$$\pi_{ns} = \frac{e^{g(\lambda_s, z_n)}}{\sum_{s=1}^{S} e^{g(\lambda_s, z_n)}}, \tag{4.10}$$

where z_n are variables that may affect class probability with the estimated parameters λ_s. To determine the optimal number of classes the model is estimated with several estimates for S and a likelihood criterion (such as AIC or BIC) is used to choose the best solution.

Chapter 5

Literature review using machine-learning methods

This chapter presents the findings of a project[1] intended to complement traditional methods with machine learning in order to identify research streams in Entrepreneurship and to determine what drives the impact of articles in the field. We studied the conflict between two mechanisms influencing impact of scientific articles using an extensive dataset derived by methodological innovations of proven methods. The project covers contributions published in journals, indexed by the web of science database resulting in 15.598 individual articles.

> *Professionally our methods of transmitting and reviewing the results of research are generations old and by now are totally inadequate for their purpose. (Bush, 1945)*

The quote from a famous article by Bush refers to a problem in scientific work that is still an issue: the time required to process existing research, in order to identify gaps, increases with the constantly expanding volume of information. Bush imagined a record-keeping machine that would solve these issues. Publication databases are now well established and valuable tools that mitigate the problem described by Bush to some extend. However, the ever increasing specialization increasingly requires scientists to bridge disciplines to arrive at new insights. Hence the quote by Bush is still relevant.

This chapter shows how the methods described in Chapter 4 can be applied to complement traditional literature reviews. Information obtained in the process in then used for a science of sciences-type

[1] The study is the result of joint work with Hannes Lampe (Hamburg University of Technology).

analysis of the field of entrepreneurship. This field contains literature on the topic of university-industry knowledge transfers, hence the results of literature reviews are helpful as preparation for subsequent studies. Also, entrepreneurship science, a relatively new field with influences from several other fields, is an interesting choice for a meta-scientific analysis.

It has been noted in various disciplines that the number of published scientific articles increased exponentially in the last 20 - 30 years. This trend presents both opportunities and obstacles for the budding scientist: scientific work traditionally encompasses a phase of catching up, i.e. understanding the status quo in a given discipline in order to identify gaps in existing research. If the amount of papers that need to be considered increases exponentially, the only viable strategies involve either decreasing the amount of time spent on one article or to specialize on a narrow subset of existing publications. Whichever filtering approach is used, there is a risk that the selection of papers has a strong influence on the scientist's ability to identify gaps and formulate research accordingly. For example, a focus on recent publications may eclipse opportunities that might arise from the combination of a recent and an older idea. A focus on a narrow field within one discipline may lead to blindness towards important information from adjacent fields. Hence, in addition to the conventional approach to literature reviews, this thesis investigates how advances in automatic text processing can be used to improve the efficiency of literature reviews. The intention is not to replace the conventional literature review but to complement it by visualizing the progress in a discipline over the last few decades.

The following sections give a brief introduction to prior literature on the field of Entrepreneurship and present methods used to derive our results. This chapter is concluded by a section discussing the findings.

5.1 Introduction

Entrepreneurship science as a field of research has so far been characterized predominantly by literature reviews focusing on smaller subsets (Schildt, Zahra, and Sillanpää, 2006) or by qualitative articles that focus on the research framework in comparison to other disciplines (Shane and Venkataraman, 2000).

There is a research gap as regards quantitative analysis of the structure of the entirety of entrepreneurship science and the confluence of these two streams. There are a number of relevant applications to an exhaustive structure of entrepreneurship science: budding scientists require an overview of existing research directions and tools to identify relevant articles. Available tools, such as search engines for scientific databases, are restricted in their utility by the coarseness of parameters. It is easy to omit relevant research with a given search strategy or, at the other extreme, to receive a flood of irrelevant publications. Further, exhaustive information on the status quo enables us to estimate the impact of articles as a function of characteristics of entrepreneurship science. This is relevant for two reasons: scientists may want to know which type of article pays the highest dividend in form of citations. Perhaps more importantly, being aware of the factors that shape scientific impact is relevant to the design of the framework of a scientific field: identifying possible mis-matches between attractive areas and those with high potential may enable policy that helps the field to live up to its potential rather than taking a more or less random direction influenced by career aspirations of individual scientists.

Entrepreneurship research has been described as a relatively fragmented field of science (Gartner, 1990; Shane and Venkataraman, 2000; Schildt, Zahra, and Sillanpää, 2006). With regard to formulating an optimal research framework, the question needs to be asked to what degree fragmentation (or, as we label it in this study, diversity) presents a benefit or detriment to scientific progress. We show how trends in entrepreneurship research influence downstream research and in particular how diversity of upstream work influences the impact of appendant publications using an extensive dataset obtain with methodological innovations to proven methods and derive implications for individual researchers as well as the larger discussion of designing a research framework for entrepreneurship science.

5.2 Literature reviews on entrepreneurship science

Scientific impact is often equated to or derived from counts of citations (Martin and Irvine, 1983; Wang, 2014) although the limitations of citations as measure for the distinct concepts of quality (Martin and Irvine, 1983) or novelty (Lee, Walsh, and Wang, 2015) have been acknowledged. Citation measures can be aggregated at the journal level (Moed, 2010), which indicates that aggregating citations for a subfield of entrepreneurship research (cluster) may be viable. As impact measures are important to scientist's careers, it is likely that scientists will tend to publish in areas that promise a large return on their investment. A simple heuristic explaining participation would be to look at research areas that have delivered above average impact and to attempt to contribute to such an area. If researchers follow this heuristic, the expected payout of participating in a popular research field increases due to the large number of scientists attracted by the popularity of the field. Therefore, building up on relatively often cited research and thus appending to an established highly cited research stream, might increase own paper citations. Hence we formulate:

> *Hypothesis 1: The higher a cluster's citations per year (popularity), the higher downstream paper's citations.*

The concept of diversity has been central to many studies on scientific impact (Lee, Walsh, and Wang, 2015; Van Noorden, 2015). Since the purpose of scientific literature is the advancement of a field of science and since advancement is closely related to innovation, it seems logical to investigate a correlation between factors that enable innovation and scientific impact. Innovation has been defined as recombination of existing knowledge (Schumpeter, 1934). Knowledge derived from this process can be identified by some form of diversity. Previous literature has, for example, investigated diversity in terms of individual characteristics, such as gender or nationality, or in terms of content of texts describing the knowledge (Harrison and Klein, 2007).

The diversity of information has also been previously proxied by educational background or speciality (Williams and O'Reilly III, 1998). We conceptualize diversity in terms of variety in subject labels of scientific clusters/ sub-research fields derived from bibliometric data of scientific publications in Entrepreneurship science.

Typically, as innovation is regarded as desirable, recombination and its indicator diversity are understood as indicators of high research impact. However, recent studies have demonstrated that individuals may interfere with this relation due to cognitive limitations: if research impact depends on citations, which in turn depend on subjective evaluation of science, then research impact is influenced by human biases towards diversity. Boudraeu et al. (2016) find that the evaluation of research proposals is negatively biased if the proposal's content is novel. Piezunka and Dahlander (2015) find that companies seeking innovative solutions to R&D problems tend to blend out submissions that are more distant to their own domain in some situations. Noteboom et al. (2007) find an inverted u-shape relation between innovation performance of companies and the cognitive distance to partnering firms. A closely related finding by Uzzi et al. (2013) shows that high impact science derives for the most part from conventional recombination of knowledge. Van Noorden (2015) shows that interdisciplinary research has lower impact in the short term and higher impact in the long term, and that the field of economics is less interdisciplinary relative to other sciences.

Based on prior research we predict that the impact of an article depends to some degree on the diversity of the sub-research field it is supposed to extend. Higher diversity will be more difficult to assess for the audience, leading to lower acceptance. For example, if literature reviews are characterized by a high content and methodological diversity, we expect the average respondent to be negatively biased towards the more distant information and therefore the number of citations to decrease. Another mechanism that may lead to the same conclusion would be that a research field with high diversity offers little opportunity for novel recombination of knowledge.

Hypothesis 2: A cluster's subject variety decreases downstream article's citations.

Furthermore, we are interested in the interaction of these two effects. An increase in cluster diversity most likely facilitates contributing to respective cluster due to the exponential relation of possible recombination's and number of subject categories. Given the mechanism defined for Hypothesis 1 leads us to predict that contributions to highly diverse clusters that are also highly popular leads to an increase in downstream article impact.

> *Hypothesis 3: The effect of a cluster's citations per year (popularity) on downstream article's impact is positively moderated by the cluster's subject variety.*

To answer our research questions we need detailed and comprehensive data. The next chapter explains how we compose this data basis.

5.3 Data and methods

This section discusses data and methods used for this project, including the methods used for collecting and processing the data, a description of co-citation analysis, keyword extraction, classification and regression methods.

5.3.1 Data

To analyze Entrepreneurship research we use Thomson Reuters Web of Science (WOS) to retrieve bibliometric data on corresponding publications. WOS is a prominent citation database, covering over 10.000 high impact journals and 120.000 international conference proceedings. As customary in prior research (Schildt, Zahra, and Sillanpää, 2006), we capture a broad selection of potentially relevant articles by using the search term "entrepre*". The query was applied to paper titles, abstracts as well as keywords (both original keywords and keywords generated by WOS). The search was conducted in August 2014 including a timespan from 1945 to August 2014 resulting in 21.973 unique WOS Records. Excluding all non-articles (e.g. book chapters or proceeding papers) resulted in 16.683 documents; leaving out all articles written in another language than English left us with 15.598 documents.

Figure 5.1 shows the total publications of Entrepreneurship from 1945 to 2013. A strong increase of publications since the 1990s is quite apparent. Presumably, the increasing number of publications gives rise to more diversity in the research field that may therefore benefit from this systematic analysis. This lies in accordance with Ireland, Reutzel and Webb (2005), who point out the continuing evolution of entrepreneurship research as a viable research paradigm.

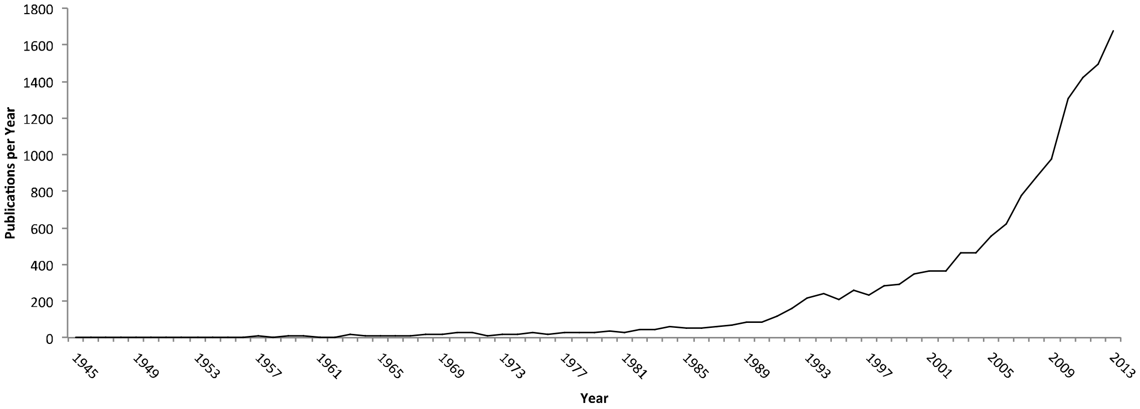


FIGURE 5.1: Entrepreneurship Publications Per Year

In contrast to earlier studies, we do not remove articles based on a selection of journals according to their relevance to the field of entrepreneurship. While such a selection reduces the number of false positives that inevitably result from a simple search strategy such as "entrepre*", it potentially also removes articles which may be of interest as they may be located at an intersection of entrepreneurship and other research fields. The problem of false positives is mitigated to some extent by the parameter selection for the co-citation analysis, described in the next section. Table A1 shows the top 43 journals with the most publications of articles included in this study. Schildt et al. (2006) analyze the research field of entrepreneurship on the basis of publications in 27 Journals (denoted by *). Several journals, which appear to be highly relevant to entrepreneurship by now, are not included. Excluding these journals would remove 74,62% of the articles included in our dataset. It appears that the smaller temporal scope of previous research in combination with changing publication trends has made previous lists of influential journals a poor choice to delineate the research field today.

5.3.2 Methods

Dividing entrepreneurship research into clusters/sub-research fields and studying how these clusters' characteristics influence the impact of downstream articles requires a combination of diverse methods. Hence, the following sections give a brief introduction to co-citation analysis and our regression analysis. Linking the two methods is facilitated using automatically extracted keywords.

Document Co-Citation Analysis. Citation analysis is a growing research area, applied in several research domains. It enables a quantitative analysis of citations and is therefore more and more adopted as a state of the art tool to overcome subjectivity (Lampe and Hilgers, 2015; Schildt, Zahra, and Sillanpää, 2006).

The most popular approach is bibliographic coupling, occurring when two works reference a common third work in their bibliographies. Document co-citation, in contrast, is defined as the frequency with which two documents are cited together by other documents. The distinction of these two semantic similarity measures is illustrated in Figure 5.2.

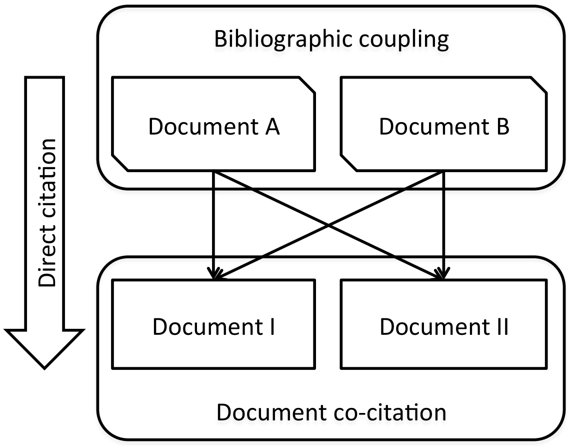


FIGURE 5.2: Distinction of Document Co-citation and Bibliographic Coupling

As document co-citation analysis (DCA) connects articles cited in the same paper, it is a measure for their relatedness due to belonging to the same topic or because their topic areas are closely connected

(Cawkell, 1976; Garfield, Malin, and Small, 1983; Small, 1973). As some co-citations may be unrelated, we included a comprehensive database in this analysis to overcome this noise.

Data was preprocessed using the software Sci2 (Team, 2009) in accordance with prior research (Lampe and Hilgers, 2015). This process can be distinguished into four steps. First, converting all letters to lower case and thus enabling case-insensitive algorithms to detect similarity in authors. Second, we merged identical authors using the Jaro-Winkler metric (Jaro, 1989; Winkler, 1999). The Jaro-Winkler algorithm estimates the similarity of two pieces of text by counting and weighting disparate letters, so that deviations closer to the beginning of a text receive a larger penalty to the similarity score compared to deviations later in the text. Potential matches with high similarity were matched automatically. In cases with lower similarity score, the pairings of author names were controlled manually. Third, based on an "Authoritative Journal Merging List" provided by the Sci2 Team (2009), we merged identical journals to account for misspellings in the references of articles. Fourth, citations were matched to documents.

We then excluded all papers with less than four references to only include research articles, resulting in 14.657 papers (after visual analysis of the distribution of citations four appeared like a reasonable cut-off to delineate research articles from other data). In the next step, we only kept articles with 15 or more citations in order to focus on articles by specialists in this research domain (resulting in 3.358 articles). This enables our findings to be built upon a wide range of expert opinions. After building the DCA-network, we deleted isolates (i.e. papers not linked in the DCA-network to other articles); resulting in the final dataset, including 2.117 articles and 62.511 co-citation links. These steps enable a robust citations analysis minimizing the possible effect of noise (Lampe and Hilgers, 2015). Following earlier research, we adopt the Jaccard index (Jaccard, 1901) as a normalized measure for the connectivity of co-cited articles (Small and Greenlee, 1980). This index describes the ratio of the number of co-citations to the total citations minus their common citation (co-citations) (Gmür, 2003). The value of the Jaccard index (S) ranges from 0 (no co-citations) to 1 (representing perfect co-citation) and is stated as follows:

$$S = \frac{|A \cup B|}{|A \cap B|} \tag{5.1}$$

with A and B as the set of citations received by two articles.

To define co-citation clusters and distinguish them from each other, we chose a rather crude method. We exclude articles' links representing a weak connectivity of articles (Jaccard value lower than 0.2). The cut off value of 0.2 results from a comparison of various cut-off values and the resulting number of disconnected components in the network. We tried to find a value where the number of clusters would not change with a slight change of this value (Lampe and Hilgers, 2015).

Compared to previous research, this cut off value is quite small (Schildt, Zahra, and Sillanpää, 2006), a necessity following the larger number of articles considered in the dataset. The issue of false positives showing up in the dataset due to the basic search query is mitigated by this step: papers that do not belong to the field of Entrepreneurship are unlikely to have been highly co-cited by those papers that do belong to the field.

Automatic Keyword Extraction. Inconsistency of available data also affects the keywords provided in the dataset. Both, "new ISI keywords" (keywords generated by Thompson Reuters WOS) and "original keywords" (keywords supplied by the papers' authors) are only present for about 50% of all texts. The value of these keywords is questionable as well: the original keywords are not standardized, whereas the ISI keywords fail to capture the information relevant for identifying research clusters. They seem to be more suitable for a more abstract classification. Hence, we created new sets of keywords regarding the clusters of the DCA. A popular method in data analysis for weighting the importance of words in text collections is tf-idf (see Chapter 4) (Robertson, 2004).

To enable the definition and allocation of keywords to each cluster determined by DCA, available abstracts and titles for one cluster are added to a new document. After appropriate pre-processing, the term frequencies are calculated per cluster and over the collection of clusters. Subsequently, tf-idf scores are obtained using the Python package Gensim (Rehurek and Sojka, 2010). Pre-processing includes stop-word removal, i.e. the filtering of words that add little meaning to a text such as articles or pronouns. Remaining words are stemmed (i.e. reducing words to their stem by removing word endings that result from grammatical cases such as the plural or genitive *-s*) using the Python library Natural Langage Toolkit (NLTK) (Bird, Klein, and Loper, 2009).

This allows for aggregating different forms of words (e.g. plural and singular of a word). Furthermore, numbers, very short words (less than three characters) and non-alphabetic characters are removed, and capital letters are replaced by their underscore equivalents. This results in a list of tf-idf-weighted words for each cluster, where the highest ranked are kept. To facilitate comparison to existing (original and ISI) keywords, the three most frequent keywords for each cluster are extracted.

From keyword to classification. Co-citation analysis, while established in the literature as a useful tool, is limited in its utility by the fact that it requires articles with a certain number of citations and results in clusters of high impact publications. New publications and those with few citations are omitted. To avoid this problem, one could attempt to find clusters using alternative methods, for example using machine learning on semantic data to estimate the similarity between papers. However, machine learning algorithms present their own drawbacks: supervised algorithms require correctly classified data. Unsupervised algorithms return results that are highly susceptible to the choice of algorithm parameters. In order to include a large sample of publications in our study, we opted for a compromise – extending the co-citation analysis with machine learning algorithms based on semantic data. Using cluster-specific keywords, we initially classified publications using a simple heuristic: if the title and abstract of a publication contain more than three of the four most significant keywords (defined in the previous section), we attribute that publication to the respective cluster. This increased the sample of labelled data (i.e. publications that could be assigned to one cluster) by 1667 publications. Newly labelled data were checked manually by comparing their titles and abstracts to the cluster definitions derived from co-citation analysis. Approximately 70% of the publications classified with tf-idf-based cluster keywords appeared to be labelled correctly. New labels and original labels from co-citation analysis could then be used to train an unsupervised machine learning algorithm.

This algorithm, available in Scikit-learn (2011), a machine learning package for the Python programming language, comprised several data pre-processing steps and a support vector machines (SVM) classifier trained with a stochastic gradient descent function. Parameters for the various steps were obtained by searching a parameter space for optimal values whereby the optimum was defined as highest cross-validated accuracy. The resulting classifier has an accuracy of 79%. Finally, the classifier was used to assign labels to all publications in our dataset for which title and abstract were available (14,053 publications).

Technically it is of course possible that these articles are written before the cluster has been published. This might be due to articles analyzing similar topics but not being recognized as fundamental articles for this research stream. To overcome this problem we further neglect all articles belonging to a sub-research field and being published before the last article of this underlying cluster. Furthermore, neglecting articles with missing data, our resulting data set for classification consists of 9.846 articles. All of these do not belong to the cluster detected by DCA. These article are further referenced as downstream or appendant literature.

Regression analysis. The dependent variable "impact" is operationalized by citation counts in Web of Science. This measure is commonly used by scholars when analyzing patents or publications (Martin and Irvine, 1983; Moed, 2010; Wang, 2014; Lee, Walsh, and Wang, 2015).

Two independent variables are incorporated into our model. First, citations per year in the underlying cluster used to measure the impact of the underlying sub-research field an article belongs to. Second, the variety of a cluster's articles are taken into consideration. We measure a cluster's variety using articles' subject categories listed by Web of Science. We create a Blau index as a measure for a cluster's variety (Harrison and Klein, 2007; Lee, Walsh, and Wang, 2015):

$$blauindex = 1 - \sum p_k^2 \tag{5.2}$$

where p_k is the proportion of members in the k-th field category. In our analysis the 35 clusters consist of 335 articles and 17 different subject categories (the distribution of the association to Web of Science subject categories is shown in Table A2). As our first independent variable is correlated to cluster size (number of articles), we did not explicitly control for cluster size. Furthermore, the correlation between a cluster's size and its blue index is almost negligible (Pearson correlation: 0.1274) due to the definition of the blue index.

Control variables are the age of the paper (in years compared to 2015), the number of authors, the amount of included references in a paper and the number of pages. The descriptive statistics are shown in Table 5.1.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1) Citations	1						
(2) Variety of underlying Cluster	-0.04	1					
(3) Citations of underlying cluster (per year)	-0.11	0.02	1				
(4) Age of paper	0.41	-0.03	-0.26	1			
(5) Number of Authors	0.02	-0.05	0.03	-0.1	1		
(6) Number of Pages	0.08	-0.03	-0.01	-0.00	-0.04	1	
(7) References	0.07	-0.04	0.06	-0.21	0.0	0.38	1
Mean	7.83	0.32	152.59	4.98	2.21	18.13	56.49
SD	21.7	0.18	136.45	3.59	1.41	8.98	33.86
Min	0	0	4.28	1	0	0	0
Max	618	0.61	510.97	25	68	216	476

Notes: N = 9846.

TABLE 5.1: Descriptive Statistics

The output is measured as the amount of citations per paper. As the dependent variable cannot assume values smaller than zero and is initially treated as integer values, we are dealing with count data. Furthermore, our dependent variable is skewed suggesting that a Poisson regression may be appropriate (Atkins et al., 2013). In addition, our dependent variable is over dispersed, consequently the negative binomial estimation framework was chosen for our analysis. In this case, a generalized linear model (GLM) is adopted, using the logarithm as a link function.

5.4 Results and discussion

This section is divided into two parts, covering the results of the methods described above. First, we present 35 research clusters identified by using document co-citation. Second, the results of our downstream article impact regression are discussed.

5.4.1 Research Clusters

This study aims to determine a quantitative categorization of entrepreneurship sub-research field. Therefore, we conduct a document co-citation analysis akin to that of Schildt et al. (2006) to reveal the different clusters of these research areas. In contrast to Schildt et al. (2006), we include a much broader dataset. Thus, we give a more extensive overview of entrepreneurship research. Given that we are interested exclusively in the most cited and coherent groups of articles, obviously some of the highly cited articles will be excluded from this analysis due to their lacking affiliation to a cluster (the most cited articles are stated in Table A3).[1] The top ten cited clusters are displayed in Figure 5.3. Each document is represented by a node and its size simulates the number of citations a document has. Edges represent the co-occurrence of articles in the reference of an article and its strength corresponds to the value of the Jaccard Index. Clusters smaller than or equal to 3, as well as clusters represented by a star, are neglected. In total we found 35 clusters stated in Table A4. Next to a cluster's total citation count, the amount of associated articles, and the average cites per article with regard to each cluster are stated. Additionally, we state the publication date of the earliest and youngest article of each cluster. A concise overview of each cluster is given referring to its most cited articles. Headings for each cluster are created by screening articles manually.

[1] Table A3 lists the top 40 cited research articles of entrepreneurship science, ranked by total citations and citations per year.

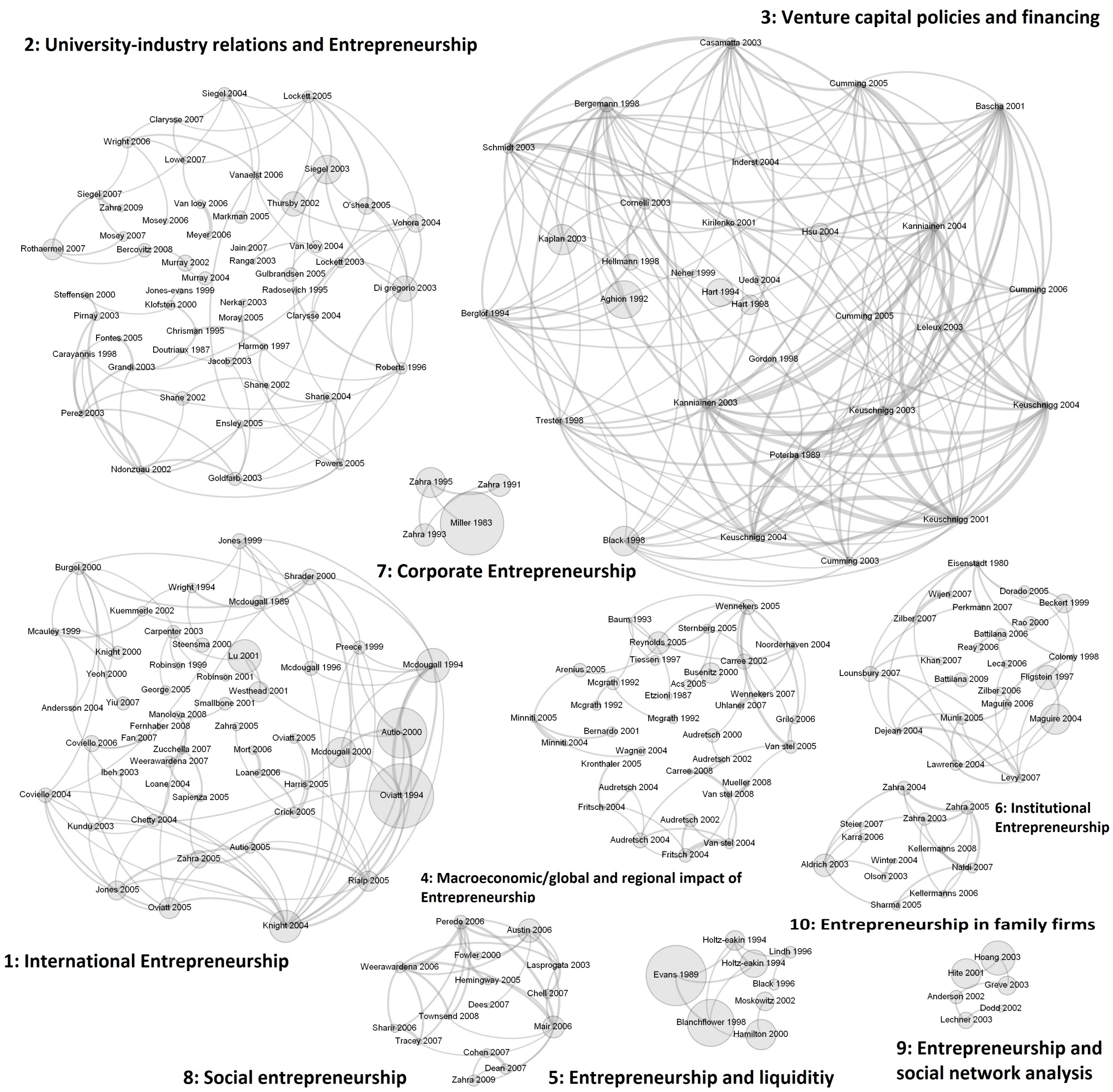


FIGURE 5.3: Top Ten Clusters by Sum of Citations

Interestingly, some of the top clusters (five and seven) contain only very few papers with high citation counts, whereas other clusters consist of many publications with smaller and more evenly distributed citation counts per article (clusters one to three). Cluster three (venture capital policies and financing) seems to differ from other clusters: even though it contains fewer papers than cluster one or two, it contains many more edges, resulting in a higher density compared to other clusters. Visually, this is represented by the cluster taking up more space even though it contains fewer papers.

It appears that the top clusters are represented mostly by papers published in the last two decades. Very recent papers may not be available in Web of Science or may not have been cited often enough for co-citation patterns to emerge. But it is surprising that older papers do not seem to be part of these clusters. Intuitively, papers that had more time to be cited and that are upstream in a field of research should receive many co-citations and therefore show up in clusters. A possible explanation is that this intuitive reasoning applies; but is moderated by the small yearly publication numbers before 1990, which may in turn be influenced by data coverage of the Web of Science database. Concerning the issue of false positives mentioned above, it appears that most clusters belong to the field of economics with only a few exceptions from other research areas: cluster 18 seems to be closer to the field of arts and humanities.

5.4.2 The Effect of Cluster-Variety on Research Impact

Given the importance of scientific impact, especially concerning career opportunities, Table 5.2 estimates the effect of several variables on articles' impact (measured via citations). We analyze downstream articles, defined as contextually similar and published later. This enables us to analyze the importance of a scientists' choice of a research domain (in Entrepreneurship science) on their research impact.

Our first hypothesis argues that a cluster′s citations per year affect a downstream article's citations positively. Model 2 shows a significant positive effect of the cluster′s citations per year, which supports Hypothesis 1. Thus, we find a positive relation between the popularity of a cluster and the impact of its downstream articles. Apparently popular clusters attract a larger number of scientists. This increases the pool of potential citation sources influencing the expected impact of an article for downstream papers.

This mechanism is likely to incentivize scientists to join already large research streams. The downturn of this mechanism thus lies in a choice of a research field not due to necessity from the perspective of an optimal research framework but rather driven by trends. This might even inflate certain research streams leading to research bubbles.

In Hypothesis 2 we predict that downstream research appended to clusters' characterized by high variety of subject labels reduces downstream paper impact. We find highly significant support for this hypothesis in the negative parameter of the variety of underlying cluster in model 3. Thus a relatively high variety is rather counterproductive for a scientist's impact. It seems that clusters where the potential for recombination has already been exploited to a greater extend offer fewer opportunities for high impact downstream research.

	Model 1	Model 2	Model 3	Model 4
Constant	-1.5373***	-1.637***	-1.5554***	-1.5279***
	(0.05)	(0.0549)	(0.0615)	(0.0636)
Main Effects				
A: Variety of underlying Cluster			-0.2136***	-0.368***
			(0.0784)	(0.1163)
B: Citations per underlying cluster (per year)		0.0005***	0.0005***	0.0003*
		(0.0001)	(0.0001)	(0.0002)
A * B				0.0008*
				(0.0005)
Controls				
Age of paper	0.3699***	0.3755***	0.3746***	0.3757***
	(0.004)	(0.0042)	(0.0042)	(0.0042)
Number of Authors	0.1416***	0.1406***	0.138***	0.1379***
	(0.0099)	(0.0099)	(0.0099)	(0.0099)
Number of Pages	0.0049***	0.0052***	0.005***	0.0051***
	(0.0017)	(0.0017)	(0.0017)	(0.0017)
References	0.0126***	0.0125***	0.0125***	0.0125***
	(0.0005)	(0.0005)	(0.0005)	(0.0005)
Log Likelihood	-24983.553	-24973.7735	-24969.944	-24968.3065
theta	0.5954	0.5966	0.5974	0.5976
	(0.0106)	(0.0106)	(0.0107)	0.0107
Akaike Inf. Crit.	49979	49962	49956	49955

Note: N = 9846. Standard errors are in parentheses; *p<0.1; **p<0.05; ***p<0.01

TABLE 5.2: Regression Results: Article Impact

In comparison to the inflationary effect of cluster popularity (Hypothesis 1), the negative effect of cluster diversity seems to present a self correction mechanism from the perspective of an optimal research framework. The attractiveness of popular clusters diminishes if there are fewer opportunities for knowledge recombination.

This raises the question if the self correcting effect of cluster diversity outweighs the effect of cluster popularity. Model 4 includes both effects as well the interaction effect. The interaction effect is positive and significant. It is depicted in Figure 5.4 and shows that the inflationary effect of popularity is only partially corrected by diversity. The correction is more pronounced for clusters with lower citations per year (lower popularity) but rather small for the popular clusters.

The latter are more likely to attract scientists using the popularity heuristic to maximize utility/impact of their publications.

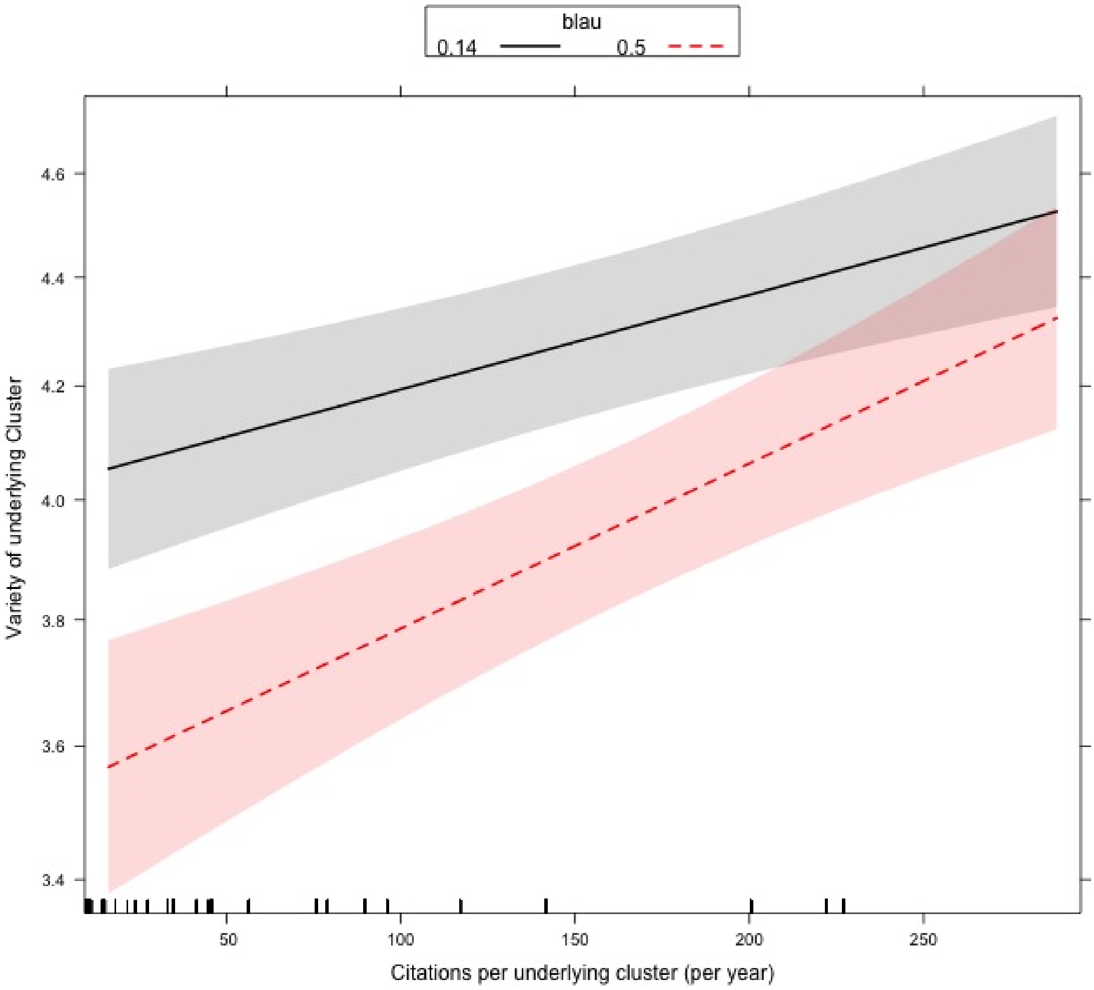


FIGURE 5.4: Interaction Effect of Variety and Citations per Year

Almost all of our four control variables are strongly significant as well. Articles' age unsurprisingly has a positive significant effect, the number of authors also has a significant positive effect on a paper's citations. This might be due to networking effects or a higher quality due to more knowledge (authors) engaged in the process. The number of pages and article's references positively affect its citations as well.

5.5 Conclusion

Our study contributes to Entrepreneurship research in several ways. We extend a proven method for structuring publication data / research fields with innovations from the machine learning domain. This resulted in a comprehensive dataset enabling us to match downstream literature to sub-research fields of entrepreneurship science.

This type of analysis may, if expanded upon by future research, improve the production of new knowledge by enabling researchers to aggregate information on topics in the field of Entrepreneurship research effortlessly. Structuring the data may also be used as starting point for a more detailed analysis on how various research streams from other disciplines affect Entrepreneurship research and uncover gaps where potentially useful combinations with other disciplines have yet to be explored. Gaining an understanding of the dynamics of research impact further helps researchers with directing effort to areas that may remain under-explored due to trends and the general development of the field.

Clustering coherent fields of research, we have presented an extensive quantitative differentiation that indicates which subjects play a major role in entrepreneurship science. By conducting a document co-citation analysis, we have identified 35 sub-research areas. Based on their citation counts, we worked out the most influential clusters. The top four clusters are (1) International entrepreneurship, (2) University-industry relations and entrepreneuship, (3) Venture capital policies and financing, and (4) Macroeconomic/global and regional impact of entrepreneurship. These clusters enable researchers new to the field of entrepreneurship science to quickly gain an overview of existing topics and trends. In contrast to prior bibliometric literature analyses for this field we use a more extensive dataset and apply state of the art methods.

Implementing machine learning approaches, we matched downstream literature to their associated sub-research fields / clusters. This enables us to analyze the effect of the underlying clusters' characteristics on downstream articles' impact. First, we show that entering more popular research domains leads to a higher impact of own research. This "inflation effect" may attract scientists to clusters irrespective of the cluster's real unobserved importance to entrepreneurship science. Second, we show that a higher diversity of an underlying cluster counteracts the effect of its popularity. Third, we find that this "self correction effect" does not completely mitigate the "inflation effect". Switching from a high to a low diversity cluster only results in higher impact for low popularity research streams.

Our results highlight the importance of clearly defined research frameworks as policy instruments. To correct for the "inflation effect", additional incentives are required to attract researchers to less popular research streams.

There are several limitations to this study. First, the dataset is taken from Web of Science. Although it is the largest academic database available, of course some articles are published in journals not listed on Web of Science. Interpretations of the results should incorporate a warning on the limitations of the data source.

Additionally, potential effects might include citation networks among researcher groups, or citations made to please the potential reviewer or editor of the target journal. Citations are equally weighted even though their importance may vary. Even though these limitations could falsify our results, the extent of the deployed dataset should mitigate these effects.

As future work, we propose to add an exploitation dimension to the discussion of the research framework. Furthermore, similar to the above analysis, other research domains, e.g. Marketing or Finance, might be analyzed by our approach to enable a comparison of our findings to more mature research domains.

Chapter 6

Motivational aspects of academic patenting

This chapter presents a study[1] on the motivational aspects of academic patenting. Using data collected from patent databases as well as survey data used in prior publications we aim to elaborate on research findings regarding the role of incentives in the filing of academic patents and how they interact with scientist's beliefs. Based on a theoretical framework that explains motivation we take into account personal characteristics and add insights from prior research regarding the effects of the institutional context, i.e. the role of colleagues and the faculty. Prior research has been concerned with crowding-effects resulting from incentives, hence we try to identify such effects as well.

After an introduction to the topic, relevant literature is discussed and a research framework derived. The discussion of experimental methods, data and the respective analysis follow. Finally, this chapter concludes with a discussion of results.

6.1 Introduction

The topic of academic entrepreneurship has received increasing attention in recent years. Scientists, formerly confined by normative restrictions associated with their work, have been identified as potentially valuable source of knowledge for recombination and application in applied contexts (Etzkowitz, 2004). Science policy makers, university administrators and scholars alike have struggled with

[1] The study is the result of joint work with Christoph Ihl (Hamburg University of Technology.)

the question whether to induce scientists to seek intellectual property protection (especially patents) for their research results prior to a publication effort or not. The controversy surrounding this issue stems from the fact that university science no longer follows the disinterestedness postulate (Merton, 1973), which originally imposed a certain detachment of the scientific discovery process from commercial matters in general and of personal economic gain from inventions made at universities in particular.

This departure from Mertonian norms is largely driven by a new 'ambidextrous' generation of university scientists who pursue academic excellence and explore potential applications of their research results in parallel (Sauermann and Roach, 2012). Empirical research on the phenomenon confirmed that only a fraction of the university scientist population chooses to participate in such technology transfer activities (see, for example, Baldini, 2009). However, it could also be demonstrated that the decision-making process can be influenced by specific sets of incentives affecting the underlying motivational drivers of the patenting decision at the level of the individual scientist (Walter et al., 2013). It was during this investigation that a base level motivation to participate in university patenting was detected, whose presence appeared to be robust and independent of any of the hypothetical incentive bundles offered. This finding ties in well with earlier results of Sauermann and Roach (2010; 2012), who reinforced the notions of a “taste for science” as well as a “taste for commercialization” as strong intrinsic, personal-level motives to explain self-selection behavior with regard to university science career options and the participation in technology transfer and entrepreneurship activities.

The detection of this strong intrinsic motivation sparked the interesting question if motivation crowding, i.e. the systematic displacement of an existent intrinsic motivation by extrinsic incentives (Frey and Jegen, 2001), could occur in the research setting at hand. The presence of this effect would imply that university administrators need to reconsider the use of incentives even if their general effectiveness has been substantiated in advance. Another relevant stream of research has emerged in the form of self-determination theory (SDT) as championed by Deci and Ryan (2000). Viewed from an SDT perspective, university patenting could be interpreted as an act of personal choice in accordance with personal interests, beliefs and values (rather than an obligation imposed by an employer). Notions of self-actualization or self-realization found by Wilhelm (2010) in qualitative field studies point in the same direction.

Due to the importance of entrepreneurial activity to economic growth (King and Levine, 1993) and the potential for commercialization of academic knowledge, prior research has investigated whether there is a trade-off between traditional academic knowledge production and the less traditional efforts to commercializing that knowledge (Crespi et al., 2011; Glenna et al., 2011).

More recent research investigates how commercialization efforts can be regulated and in particular which tools are most suitable for encouraging commercialization of knowledge produced by university scientists (Conti and Gaule, 2011; Baldini, Grimaldi, and Sobrero, 2007; Baldini, 2010). This particular stream of research deals primarily with incentives, a set of policy tools aimed at encouraging or easing commercialization of academic knowledge. A reoccurring basic assumption in economic research is that an increase in expected profit results in increased productivity. Hence the obvious policy tool is to incentivize commercialization with monetary rewards such as fixed payouts or royalty shares. However, using monetary rewards as incentive can prove counter-productive in the long term: a crowding-out effect has been observed where such rewards lead to lower levels of productivity, especially in contexts where a monetary reward was introduced and later rescinded (Deci, 1971). These counter-intuitive results suggest that regulating a commercialization process with incentives requires understanding scientists' motivation and the interaction of incentives and motivation. Hence we try to define an attitude in scientists towards invention disclosure and study its composition, relation to disclosure intentions as well as interactions with incentives drawing upon SDT as theoretical framework.

6.2 Theory and hypotheses

The following sections describe prior literature with regard to what motivates scientists to patent, the role of incentives in academic patenting and the role of the individual and institutional contexts of scientists.

6.2.1 The role of self determined motivation for academic patenting

What motivates scientists? Scientific work has a specific appeal on scientists, a “taste for science”, characterized by the three aspects autonomy (the academic freedom to choose to work on interesting problems), reputation (recognition of a scientist’s discoveries by peers through citations) and, to a smaller degree, money (Merton, 1973; Stephan and Levin, 1992).

The latter has been found to be less important than the first two motivators; in fact, there is some evidence that scientists incur monetary opportunity costs when they decide to pursue academic work (Stern, 2004). Merton (1973) suggests that scientific work may entail a disregard for personal monetary rewards as conflicting with scientific norms. The taste for science is subject to ongoing analysis (Agarwal and Ohyama, 2013; Lacetera and Zirulia, 2008; Roach and Sauermann, 2010; Sauermann and Stephan, 2010).

Recent work extends this research by investigating scientists’ motivation to commercialize their work (D’este and Perkmann, 2011) contrasting Merton’s taste for science with a taste for commercialization (Sauermann and Roach, 2012).

In that context Lam (2011) applied findings from SDT (Deci and Ryan, 2000), which proves to be a framework capable of explaining motivational aspects of scientists with regard to their norms and beliefs. We extend prior research by investigating whether there exists in scientists an attitude specific to invention disclosure and apply SDT to explain the interactions of this attitude, actual disclosure rates and various incentives.

SDT posits that humans are growth-oriented beings with a natural inclination to a sense of self and to integration into social groups (Deci and Ryan, 2000). This behavior entails engaging in activities that are interesting or important to an individual so long as a set of preconditions, termed "needs", is met. The three needs defined by SDT are autonomy, competence and relatedness. Autonomy refers to a desire for volition, a need to experience activities congruent with a sense of self. Competence is a propensity to have an effect on one's environment and obtain valued outcomes from one's actions. Relatedness is the desire to feel connected to others. Satisfying these needs leads to many positive effects such as better performance at a given task or enhanced well-being. Exposing individuals to contexts opposed to these needs such as a controlling, over-challenging or rejecting environments leads to sub-optimal results and in some cases to persistent or self-reinforcing negative behavior. Building on the basic concept of needs, SDT distinguishes between types of motivations and associated regulatory processes: amotivation, intrinsic and extrinsic motivation (the latter of which is divided into several sub-categories).

Amotivation is defined as a lack of motivation and associated with a lack of regulation as well as a lack of the satisfaction of the three needs mentioned above. Intrinsic motivation describes motivation linked to activities that are interesting to an individual and provide an optimal challenge so that the individual will freely engage in the activity (i.e. the type of regulation is also intrinsic). Autonomy and competence are preconditions to intrinsic motivations, while the need for relatedness is not necessary but provides a positive effect. Extrinsic motivation causes individuals to pursue activities even though the activity does not intrinsically appeal to them. The three needs also affect extrinsic motivation: whereas competence is a requirement, the need for autonomy can be partially or completely absent.

The degree of autonomy associated with external motivation allows the identification of several sub-types of external motivation. SDT assumes that individuals have a tendency to adopt external regulations through a process called internalization, which is contingent on autonomy (Deci and Ryan, 2000). This process transforms social norms or requests into values and self-regulations, thereby allowing individuals to feel self-determined when enacting external regulations. SDT distinguishes four types of external motivation that differ in their degree of internalization in accordance with the degree of control or autonomy associated with the type of regulation: external regulation, introjection, identification and integration (Deci and Ryan, 2000).

External regulation is associated with a high degree of control, behavior is encouraged by tangible rewards or the avoidance of sanctions. It generally leads to poor results and a low chance that the controlled behavior will continue once the regulation has been removed.

Introjected regulations are only partially internalized. In contrast to external regulation, they are a form of self-control, but they are not congruent with the individual's values.

Identification is defined as process with higher degree of internalization; the value of an idea is accepted but the regulation remains instrumental. As example, exercising for health benefits in contrast to engaging in sports out of interest or enjoyment indicates identification (Deci and Ryan, 2000).

The highest degree of internalization is provided by integration, whereby individuals fully identify with the value of an idea. This type leads to results (i.e. performance, personal well-being) comparable to those achieved when engaging in intrinsically motivated behavior.

Furthermore, SDT finds a correlation between goals, motivational processes and needs: some life goals are found to be closer to the three needs (e.g. self-fulfillment), whereas others are more distant (pursuit of wealth). Goals closer to the three needs are associated with better performance and well-being. The motivational process used to attain the goal appears to moderate the effect of goal pursuit on the individual; a goal leading to tangible rewards pursued due to well internalized motivation leads to better performance than when the same goal is pursued due to external regulation.

If we apply the findings of SDT to the analysis of scientist's attitudes towards commercialization, we expect to find that the individuals' attitudes can be divided along the lines of motivational types described in SDT. Researchers may be intrinsically motivated to commercialize their research as the challenge of reducing a theoretical problem to practice may appeal to them: they may be convinced that the commercialization is a valuable process in itself or that commercialization has a positive effect on the application of research deemed to be important. In this case the scientist's motivation to commercialize would be completely self-determined. As the perceived value of research can transcend its commercial potential (Lam, 2011), we expect to find differences in the degree of intrinsic motivation between various branches of research. For instance, in the life sciences the commercialization of novel drugs, often associated with a higher potential for commercialization due to their monetary worth (Glenna et al., 2011), may also be positively affected by the researcher's awareness of the fact that commercialization of a drug will provide health benefits to others (Lam, 2011). Alternatively, researchers may be extrinsically motivated with a high degree of internalization, in which case their motivation would be mostly self-determined. This is likely to occur when researchers do not identify with the process of commercialization but perceive that commercialization of their work may have positive effects on their institution and thereby on their more intrinsically motivated activities. For example, successful commercialization of a technology may provide funds to the institution or prestige to the scientists enabling further research. Less well integrated external motivation, such as researchers acknowledging that commercialization is wanted but not agreeing with the concept or finding the process tedious, would likely lead to less commercialization activities or at best symbolic but largely ineffective efforts.

Since the scale of self-determined motivation types defined by SDT provide a basis for understanding the antecedents of academic invention disclosure, we formulate that

> *Hypothesis 1: Researchers with high levels of self-determined motivation towards academic patenting (a taste for patents) are more likely to disclose their inventions to university.*

6.2.2 The role of incentives for academic patenting

Which tools are appropriate to incentivize commercialization of academic knowledge is subject to ongoing debate (Conti and Gaule, 2011; Baldini, 2009; Göktepe-Hulten and Mahagaonkar, 2010; Macho-Stadler and Pérez-Castrillo, 2010; Rothaermel, Agung, and Jiang, 2007). Unlike motives, which are stable and trait-like, incentives are externally provided and vary in their effect depending on circumstance (Deci and Ryan, 2000). Incentives can be devised as “positive incentives” in that they lead to some form of reward or as a reduction of “negative incentives”, i.e. mitigation of barriers (Bamard, 1938). A comprehensive set of incentives relevant to academic patenting has been investigated by Walter et al. (2013): monetary incentives provide tangible rewards to individual scientists, groups of scientists or faculty. At the individual level royalty shares are often mandated by law (Harhoff and Hoisl, 2007; Baldini, 2010; Will and Kirstein, 2002), whereas one-off payments are implemented by some universities as additional reward (DiMasi, 2002; Walter et al., 2013).

At the team or faculty level possible incentives are a payout of some revenue to the workgroup (Lach and Schankerman, 2008) or royalty shares paid to the faculty (Baldini, 2010; Walter et al., 2013), which encourage scientists by the prospect of additional funding for their future work. Non-monetary rewards are often aimed at reducing barriers such as paperwork or legal limitations inherent to the patenting process. Examples for such incentives are the introduction of a grace period, a legal mechanism that loosens the novelty requirement for patent filings by allowing researchers more time to file a patent after publishing their work in journals or at conferences (Azoulay, Ding, and Stuart, 2009; Franzoni and Scellato, 2010; Straus, 2000) or technology transfer offices (Nilsson, Rickne, and Bengtsson, 2010; Saragossi and Potterie, 2003; Owen-Smith and Powell, 2001). TTOs have been found to affect the quality (Jensen, Thursby, and Thursby, 2003; Siegel, Waldman, and Link, 2003) and sustainability (Kenney and Patton, 2011; Markman et al., 2005) of commercialization activities, although they represent a significant investment (Baldini, 2009; Bercovitz and Feldman, 2008).

Finally, incentives may appeal to the same aspect as the “puzzle” and “ribbon” categories identified by Stephan and Levin (1992): patent counts can be included in academic performance assessments (Aldridge and Audretsch, 2011; Dietz and Bozeman, 2005). Also, while largely ceremonial, awards for granted patents confer reputational advantages to scientists (Baldini, Grimaldi, and Sobrero, 2007) and provide a feeling of accomplishment (Giuri et al., 2007). Some research confirms the positive effect of such awards (Haeussler and Colyvas, 2011; Frey, 2010; Neckermann, Cueni, and Frey, 2009).

In accordance with prior literature these incentives are likely to have a positive effect on the intention to disclose knowledge through patent filings (Walter et al., 2013). There are, however, counter-intuitive effects associated with some types of incentives, in particular of monetary rewards. In economics, a rationally acting agent would generally be expected to perform better when offered additional monetary rewards. In certain conditions the opposite has been observed: offering monetary rewards reduced performance, in particular when compared to the performance level observed before the rewards were implemented (Deci and Ryan, 2000). SDT explains this effect with a shift in the perceived locus of causality from internal to external (Deci and Ryan, 2000); the monetary incentive is perceived as controlling and reduces the perceived autonomy of the subject. Denial of the need of autonomy rules out the more effective self-determined types of motivation and leads to sub-optimal results. This effect is alternatively known as crowding-out of intrinsic motivation in motivation crowding theory (Frey and Jegen, 2001) or as overjustification effect (Carlson, Heth, and Miller, 2007).

Frey and Jegen (2001) suggest that a lowering of self-esteem, aside from the loss in autonomy, is responsible for the shift in motivation: in SDT terms the need for competence is undermined by the incentives. Additionally, the denial of autonomy may also inhibit the need for relatedness; if a scientist's intrinsic motivation is connected to the exhibition of this intrinsic behavior to others, incentives may undermine this connection (Frey, 2012). The negative effect is particularly relevant to tangible incentives such as royalty shares or payments to individuals (Deci and Ryan, 2000). Whether this effect can be observed hinges on two aspects: the individual attitude towards patenting, which has an effect on internalization, as described above, and the combination effect of others incentives. The total effect of a set of incentives may differ from the sum of individual incentives (Holmstrom and Milgrom, 1994).

Other forms of incentives may be more in line with scientists' values and beliefs and therefore unlikely to be affected by the negative effects mentioned above: non-tangible incentives such as reputational awards or a reduction in barriers that make patenting uninteresting do not conflict with Mertonian norms a scientist may be used to. This includes the barrier-mitigating grace period and

TTOs; by reducing the hassle associated with patent filing, the activity may appear more interesting and therefore correspond better to self-determined motivation. Awards for granted patents are less likely to be considered controlling. Instead, they confer reputational advantages to scientists (Baldini, Grimaldi, and Sobrero, 2007) and provide a feeling of accomplishment (Giuri et al., 2007) that is likely to satisfy SDT's need for competence.

An intermediate type of incentive is not explicitly of monetary nature but still leads to tangible results. Accordingly they may intrude into self-determined motivation more than intangible rewards but less than purely pecuniary measures. Examples of this type of incentive are the inclusion of patent counts in academic performance assessments or royalty shares paid to the faculty or work group. These incentives conform to the instrumental nature of the more autonomous type of extrinsic motivation "identification" defined by SDT.

To correctly assess the impact of incentives, we need to know the scientist's attitude towards disclosure in order to determine the degree of internalization of the motivation (Deci and Ryan, 2000). If scientists regard invention disclosure as important to society, the incentive is likely to coincide with a high degree of autonomy and therefore a higher degree of internalization (leading to identification with commercialization or even integration of the concept). This may mitigate the negative effect of incentives that would ordinarily be perceived as controlling. If the taste for patents is less pronounced, the incentive may be perceived as more controlling and consequently be less well internalized (leading to the categories external or introjected motivation), likely to cause relatively poor results.

So, in accordance with the findings of SDT, it is possible that scientists with highly self-determined motivation to patent may consider incentives, especially the offering of tangible rewards, as intruding. The incentive causes a shift in the perceived locus of causality, thereby leading to less positive or perhaps even negative effects. However, the degree to which this effect changes the scientist's motivation depends on the scientist's attitude, so that scientist with well internalized motivation, i.e. a high taste for patents, is likely to approve even of tangible incentives. This may take the form of a substitution effect, whereby a more efficient self-determined motivation is replaced by a motivation characterized by low internalization. Thus we formulate Hypotheses 2a and 2b as competing hypotheses:

> *Hypothesis 2: The relationship between researchers' self-determined motivation towards academic patenting and their probability of invention disclosure is: a) stronger under high incentive conditions, b) weaker under high incentive conditions.*

6.2.3 The impact of individual and institutional context on academic patenting

In the previous chapter we stated that the degree of internalization of external motives changes the effect of incentives and that internalization is closely connected to the attitude of scientists towards commercialization, i.e. their taste for patents. To understand the effect of incentives, it is therefore necessary to look into the antecedents of this taste. By identifying factors that positively or negatively affect taste, it should be possible to predict conditions under which incentives are more likely to be effective, i.e. conditions where incentives have less of an undermining effect on self-determined motivation. Factors affecting a taste for patents can be divided into variables related either to an individual's background or to the institutional context of the respective scientist. The following paragraphs illustrate findings of prior research regarding these variables.

Individual factors with an influence on the taste for patents are the number of scientists' publications, the number of patent applications, the degree of industrial involvement, gender, nationality, whether a scientist is tenured and the time since tenure. The behavior of individual scientists is indicative of the focus of their work. The number of publications shows to what degree a scientist engages in basic research. Hence it could be argued that a scientist with many publications has less interest or time for applied research, which patents are usually regarded as indicator of. However, both publications and patent filings may be the result of a scientist discovering a promising research venue (Azoulay, Ding, and Stuart, 2007), in which case patenting and publishing would be complementary.

The number and type of publications also indicate a scientists' productivity, at least as far as their endeavors in basic research are concerned. If an academic patent fulfills the same function with respect to a scientist's productivity as regards applied research, it is to be expected that highly productive scientists will have high counts of both publications and patents. Zucker, Darby and Brewer (1998), for example, find that "star scientists" are important to the emergence of commercial projects. While publications may antecede a patent application, they may also have an effect on the scientist's publishing behavior following the patent filing (Breschi, Lissoni, and Montobbio, 2005).

To sum up, publications may be relevant to academic patenting in three ways: a stock of publications may influence the probability of a patent being filed. Secondly, the filing of a patent may influence the number of following publications. Thirdly, both publications and patent applications may be the result of an unobserved event such as the discovery of a new research opportunity. A portfolio of prior patent applications indicates that the scientist is familiar with the application process. This may be beneficial for the intention to disclose more inventions at a university as learning effects are likely to reduce the costs of the process. However, this learning effect may also make cooperation with a university in disclosing unnecessary, thereby increasing the likelihood that the scientist will attempt to file for a patent either alone or in cooperation with industry. The latter point has been described as less likely because licensees are wary of potential conflicts between employer and employee (Bercovitz and Feldman, 2008).

The degree to which a scientist is involved with industry may provide opportunities for applied research and access to industrial resources (Calderini, Franzoni, and Vezzulli, 2007), although a high degree of industrial involvement may also be associated with legal restrictions that make it more difficult for scientists' to patent their work. Prior research by Ding et al. (2006) indicates that the gender gap found in publication numbers of scientists also applies to academic patenting, which may be explained by lack of industry contacts. This could increase the costs associated with invention disclosures as well as concern about the impact of the costs of patenting on a scientific career. Gender may also affect the degree to which institutional variables or incentives (described below) determine the taste for patents (Ding, Murray, and Stuart, 2006).

The scientist's nationality may also indicate differences with an effect on patenting behavior: migrating scientists are likely to be very productive, which in turn influences their taste for patents (Zucker and Darby, 2006). Tenure and time since tenure are representative of the scientist's experience. It has been found (Azoulay, Ding, and Stuart, 2007) that younger and older scientists patent less than compared to their mid-career peers, most likely due to changing attitudes towards academic patenting in recent years. Bercovitz and Feldman (2008) find that younger scientists are less affected by imprinting effects from an environment where academic patenting was not common and therefore patent more than their elder peers.

The institutional context, i.e. the behavior of other scientists working in the same team, institute or faculty, has been found to affect the behavior of the individual scientist. Higher faculty quality (e.g. higher scientific productivity) has been linked to increased patenting activity (Perkmann, King, and Pavelin, 2011; Van Looy et al., 2011).

Generally scientists are assumed to adapt their own behavior to that of the team, although the reason for this compliance can differ. Scientists may recognize the norms and regulations of their workplace but only comply with them symbolically, i.e. as much as necessary but as little as possible (Bercovitz and Feldman, 2008).

This may be due to a lack of identification with the required behavior and would correspond to an extrinsic motivation in SDT that has not been well internalized. Alternatively, a scientist may identify with the idea behind regulations encountered at work and comply with the regulation, being motivated by a well internalized external motivation. Stuart and Ding (2006) find that the behavior of peers is instrumental in this regard, it may not only change the individual's beliefs but also provide useful information that facilitates patenting. Finally, a scientist with an intrinsic interest in the activity may self-select into an environment supportive of this activity (Lam, 2011).

The impact of an institutional context presents itself as a possible solution to the problem of tangible incentives undermining intrinsic motivation: if peers with high intrinsic motivation towards commercialization change the attitudes of their colleagues, a viable policy tool may be a hiring strategy that focuses on scientists with a propensity to patent. Adoption of institutional norms can be distinguished from mere symbolic compliance by comparing the effect of the context on the individual's attitude and his patenting activity: symbolic compliance is unlikely to change the individual's attitude and lead only to a small increase in patenting, whereas internalization of the faculty's norms would affect both attitude and knowledge disclosure. As measure for scientific output we use the faculty's publications over a five year timespan. A higher scientific output is generally associated with a high focus on basic research (Azoulay, Ding, and Stuart, 2007), although higher output may also indicate a larger stock of knowledge and therefore additional opportunities for commercialization (Azoulay, Ding, and Stuart, 2007).

In contrast, a faculty's patent stock over a five year timespan indicates a focus on applied research as well as available experience in commercialization activities (Bercovitz and Feldman, 2008). To complement these variables, we use governmental and industry funding as indicators of faculty basicness vs. appliedness (Bozeman and Gaughan, 2007).

We can expect a negative effect of an institution's basic research orientation on a scientist's attitude towards commercialization due to a shift in the scientist's beliefs away from approval of commercialization and towards Mertonian norms. However, a stronger basic research performance is likely to affect the individual's research performance, which may then result in additional patenting opportunities. The opportunity effect and the attitude-changing effect of basic research orientation with regard to knowledge disclosure rates may cancel each other out. A focus on applied research by a faculty can be expected to have a positive effect on a researcher's taste for patents as this context affects workplace norms and reduces barriers through experience sharing. Likewise, individual behavior that is indicative of interest in or experience with applied science is likely to positively affect attitudes towards disclosure.

> *Hypothesis 3: A more applied (as opposed to a more basic) research orientation of academic researchers (a) institutional context and (b) individual background leads to a higher self-determined*

motivation towards academic patenting.

The combination of our hypotheses in one framework is shown in Figure 6.1: taste, formed by individual and peer effects, affects disclosure intention after being moderated by incentives.

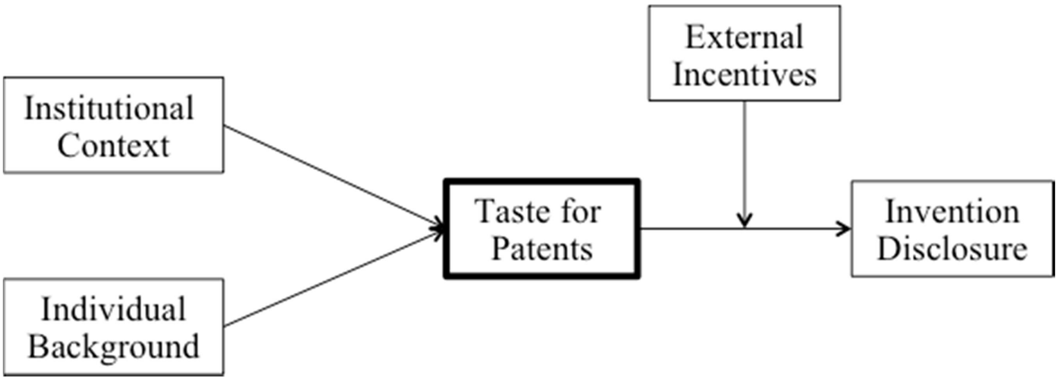


FIGURE 6.1: Conceptual Framework: Taste for Patents

6.3 Empirical setting

The following sections describe the collection of data in surveys as well as secondary data sources. This sub-chapter also illustrates the estimation methods used to test our hypotheses.

6.3.1 Survey measures

Data was collected using an online survey conducted between December 2010 and March 2011 that has previously proven useful for research into incentives (Walter et al., 2013). We invited researchers from several disciplines (natural sciences, engineering, mathematic/ computer science) at faculties from the nine major technical universities of Germany. 1.408 researchers out of 17.178 invited participants completed the survey. Of these, 10,4% were full professors, 17,3% were post-docs or junior professors and the remaining 72,2% were research associates. Most participants belonged to an engineering faculty (63%). The focus on MINT subjects may explain the relatively high proportion of male participants (77,5%). As these proportions do not differ significantly to the proportions of the invited population, there is no indication of a possible selection bias.

Our focal construct "taste for patents" is conceptualized as self-determined motivation to engage in academic patenting. Operationalization of motivation and attitudes (Ajzen, 1988) served as a starting point. However, we frame motivation as beliefs about expected consequences rather than value or importance (cf. Sauermann and Roach, 2012) because of better predictive capabilities of the former (e.g. Ajzen, 1988; Bagozzi, 1984; Valiquette et al., 1988; Pieters, 1988).

In order to derive items for a multi-items scale that tap into expected consequences of academic patenting, we draw on 20 interviews and eight in-depth case studies with patent-experienced university officials and researchers at universities conducted between January and August 2008. Table 1 gives an overview of the items retained in the final scale. For illustrative purposes, we have grouped the items according to the relevant motivational dimensions that we identified in the theoretical derivation of the measure. Obviously, academic patenting is rather seen as a constraint to autonomy. Therefore, these items have to be reversed before entering the scale. Despite different motivation dimension, the scale shows sufficient internal reliability (Cronbach $\alpha = 0.86$) so that we average the items in further analyses to arrive at our measure of a "taste for patents". The scale is shown in Table 6.1.

The second central element from the survey is a scenario-based conjoint experiment to elicit scientists' preferences for incentives to stimulate the submission of invention disclosure filings and to quantify their relative impact (see Walter et al., 2013).

Despite most applications in marketing (Netzer et al., 2008; Green, Krieger, and Wind, 2001), conjoint analysis has recently been used in management research as well (Franke et al., 2008; Fischer and Henkel, 2013; Leptien, 1995; Monsen, Patzelt, and Saxton, 2010). By experimental manipulation and random assignment, this method allows to create 'bundles' of incentives that are not yet implemented in researchers' actual university setting and to disentangle their relative effects as well as to overcome a potential selection bias of researchers into favorable university contexts.

Preceding the conjoint scenarios expert interviews were used to

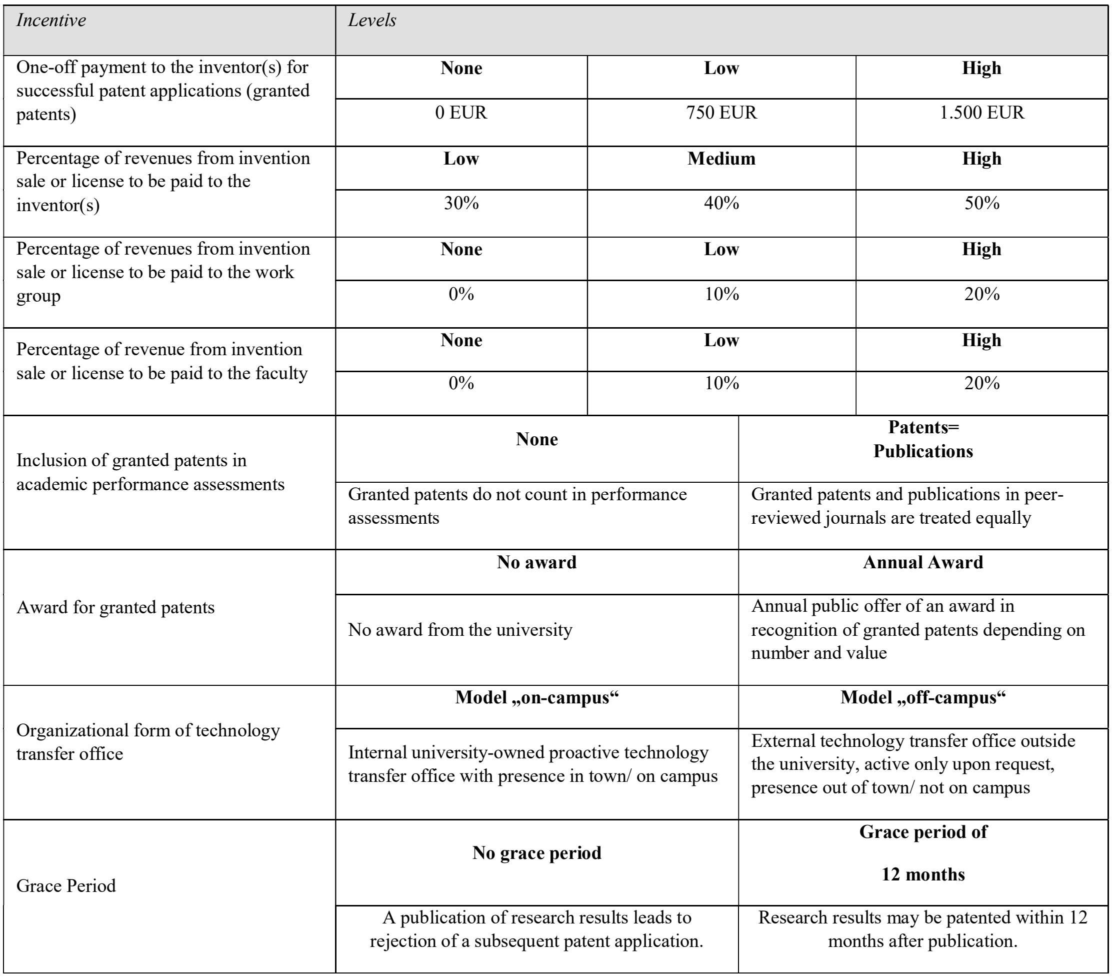
Category	**Item**	**"I expect that patenting my academic work will …"**
Autonomy	Constraints in topics	… require a shift in my research agenda. *(reverse)*
	Constraints in disclosure	… require a delay for the publication of my research results. *(reverse)*
	Constraints in time	… require additional effort that keeps me from doing research. *(reverse)*
Recognition	Applicable results	… ensure that my results applied in actual products and services.
	Contribution to public	… ensure that my results make a contribution to public.
	Reputation	… lead to higher reputation among my colleagues.
	Career perspectives	… improve my carrear opportunities.
	Access to experts	… lead to contacts with experts in commercialzing the invention.
Money	Access to funding	… lead to additonal funds for my unit.
	Personal income	… lead to a financial benefit for me.

TABLE 6.1: Scale to Measure the Taste for Patents

validate the selection of incentives. The scenarios were tested and adjusted before invitations to the main study were sent out. A description of the scenario attributes is shown in Table 6.2.

A blocking factor with three levels was used to reduce the number of scenarios per participant to twelve. The order of scenarios within each block was randomized for every participant to avoid order effects. As dependent variable we chose the participants' agreement to the statement: "This combination of incentives motivates me to have my research results checked for commercial applications by means of invention disclosure filings." The variable was measured using a Likert scale ranging from "strongly disagree" (0) and "strongly agree" (6). An example conjoint scenario is shown in Table 6.3.

Five independent variables were derived from the survey in order to characterize researchers' individual background. Industrial involvement is measured using the weighted scale proposed by Bozeman and Gaughan (2007), which we normalize to be between 0 and 1. Four variables from the survey serve as controls: (1) tenure measured in terms of years that respondents worked in research; (2) a dummy variable that indicates whether respondents' working contract has tenure; (3) a dummy variable equaling one if the respondent has a foreign (i.e. non-German) nationality; (4) a dummy variable indicating the participant's gender (one indicates



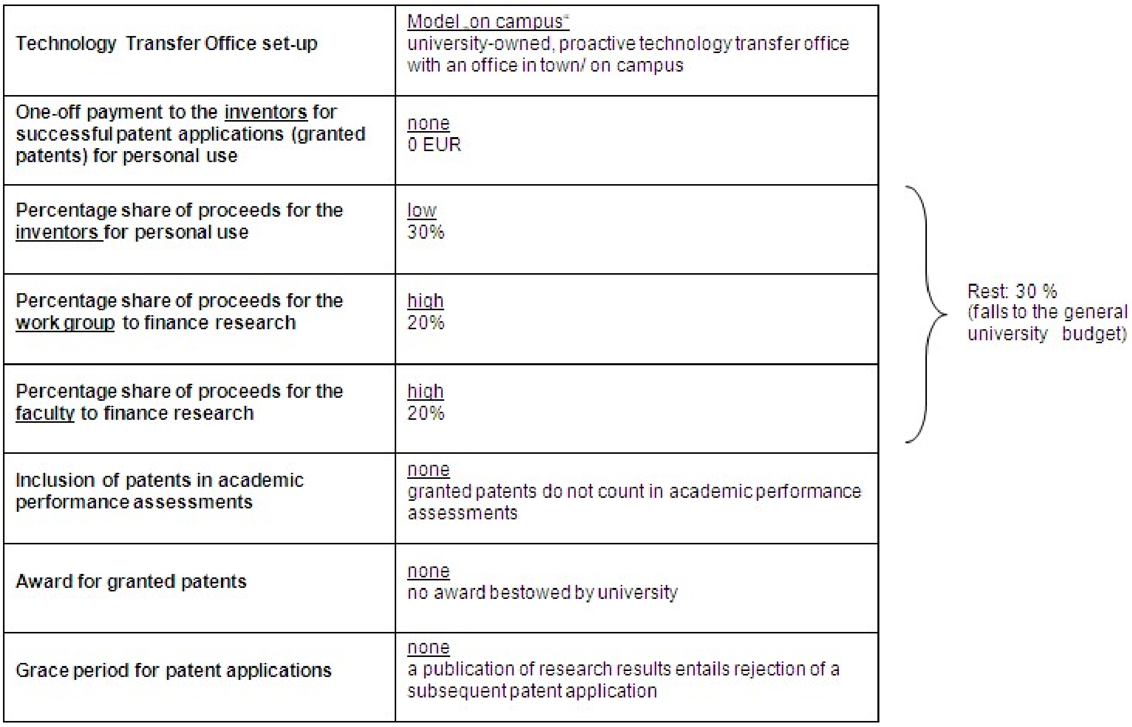
Incentive	*Levels*		
One-off payment to the inventor(s) for successful patent applications (granted patents)	**None**	**Low**	**High**
	0 EUR	750 EUR	1.500 EUR
Percentage of revenues from invention sale or license to be paid to the inventor(s)	**Low**	**Medium**	**High**
	30%	40%	50%
Percentage of revenues from invention sale or license to be paid to the work group	**None**	**Low**	**High**
	0%	10%	20%
Percentage of revenue from invention sale or license to be paid to the faculty	**None**	**Low**	**High**
	0%	10%	20%
Inclusion of granted patents in academic performance assessments	**None**	**Patents= Publications**	
	Granted patents do not count in performance assessments	Granted patents and publications in peer-reviewed journals are treated equally	
Award for granted patents	**No award**	**Annual Award**	
	No award from the university	Annual public offer of an award in recognition of granted patents depending on number and value	
Organizational form of technology transfer office	**Model „on-campus“**	**Model „off-campus“**	
	Internal university-owned proactive technology transfer office with presence in town/ on campus	External technology transfer office outside the university, active only upon request, presence out of town/ not on campus	
Grace Period	**No grace period**	**Grace period of 12 months**	
	A publication of research results leads to rejection of a subsequent patent application.	Research results may be patented within 12 months after publication.	

TABLE 6.2: Incentives and respective levels

female gender). In line with previous research (Azoulay, Ding, and Stuart, 2007; Bercovitz and Feldman, 2008), we expect researchers to be less interested in patenting when they are more experienced, tenured, foreign and female.

6.3.2 Secondary data

Each responding researcher was manually assigned to one of the following four academic disciplines, based on their affiliation, to better account for different patentability across disciplines (Jaffe, 1989; Zucker and Darby, 2006): (1) information and communication technology (ICT), (2) life science, (3) physics and electrical engineering, (4) other engineering. Given the nine different technical universities



Technology Transfer Office set-up	Model „on campus“ university-owned, proactive technology transfer office with an office in town/ on campus
One-off payment to the inventors for successful patent applications (granted patents) for personal use	none 0 EUR
Percentage share of proceeds for the inventors for personal use	low 30%
Percentage share of proceeds for the work group to finance research	high 20%
Percentage share of proceeds for the faculty to finance research	high 20%
Inclusion of patents in academic performance assessments	none granted patents do not count in academic performance assessments
Award for granted patents	none no award bestowed by university
Grace period for patent applications	none a publication of research results entails rejection of a subsequent patent application

Rest: 30 %
(falls to the general
university budget)

TABLE 6.3: Example Scenario

and four different academic disciplines, we consider 36 different faculties. Two faculties did not yield any answers, so that we are left with 34 faculties with an average of 41,4 respondents per faculty.

On this premise, we complement the survey data with publication from the Web of Science database and with patent data from the PATSTAT database, both on individual researcher and faculty level for the period of 2005-2010 just prior to the survey. We searched for the individual researchers' names in these databases along with their affiliations. The resulting number of publications and patents for individual researchers is highly skewed, so that we logged the variables for the analyses. To determine the number of patents and publications on faculty level, we first searched for all documents with relevant versions of the university names in the affiliations. To arrive at the faculty level, the resulting publications and patents were then assigned to the four academic disciplines by the means of concordance tables (Jaffe, 1989; Zucker and Darby, 2006). Additional information on faculty level was obtained from the TU9 association as well as the Research Ranking 2009 of the Center for University Development (CHE) on the following items: (1) number of tenured academic staff as a proxy for size, (2) funding for basic research from the German National Science Foundation, (3) funding for applied research from industry. Table 6.4 shows the descriptive statistics of the independent variables used in this study.

6.3.3 Estimation

We use an ordered logit random-effects model for testing our first two hypotheses (Greene, 2003; Agresti, 2010). The dependent variable estimates the unobserved intention to submit an invention disclosure filing Y^*_{fij} of individual i from faculty f in situation j:

$$Y^*_{ij} = \alpha_0 + \alpha_1 TASTE_i + \beta' INC_j + \gamma'(TASTE_i INC_j) + \delta' CON_i + \mu_i + \omega_f + \epsilon_{fid} \tag{6.1}$$

where α_0 is a constant, $TASTE_i$ is the individual-specific taste for patents with the corresponding coefficient α_1, INC_j is a vector of the incentive levels presented to individual i in scenario j with the corresponding coefficient vector β, $TASTE_i INC_j$ is the interaction of incentives and taste with the corresponding coefficient vector γ, and CON_i is a vector of control variables with the coefficient vector δ. Due to the structure of our data, with repeated observations per individual that are nested in faculties, we use a faculty-level fixed effects ω_f and an individual-specific random effect μ_i to control for unobserved heterogeneity and correlation across observations from the same faculty and respondents.

To test our third hypothesis, we estimate a random-coefficient linear regression model of the following form:

$$TASTE_i = \alpha_0 + \beta' IND_i + \gamma'(FAC_f) + \omega_f + \epsilon_{fi} \tag{6.2}$$

where α_0 is a constant, IND_i is a vector of individual-level determinants of taste with the corresponding coefficient vector β, FAC_f is a vector of faculty-level determinants of taste with the corresponding coefficient vector γ, and ω_f is a random effect that captures unobserved heterogeneity (and correlation) across the repeated observations from the same faculty. We estimate both models by a simulated maximum likelihood procedure based on 100 Halton draws for the random effects in each model (Train, 2009).

Variable	Mean	Std.Dev.	Min	Max	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)	(25)	(26)
(1) Taste for Patents	5	1,17	1	7	0,09	0,03	-0,16	-0,17	-0,17	0,06	0,02	0,00	0,09	0,13	0,15	0,15	-0,14	-0,06	-0,02	0,18	0,00	0,02	-0,03	-0,01	-0,02	-0,04	0,08	0,00	0,01
(2) Industrial Involvement	0	0,22	0	1		0,18	0,11	0,40	0,35	-0,10	-0,16	-0,08	0,00	0,19	0,19	0,18	-0,03	-0,13	-0,07	0,18	-0,01	-0,03	0,02	-0,02	0,00	-0,04	0,03	0,02	0,01
(3) ln(Ind. Patents)	0	0,69	0	5			0,37	0,13	0,12	-0,04	-0,13	-0,02	0,00	0,05	0,08	0,09	-0,05	-0,07	0,04	0,06	-0,01	-0,04	0,01	0,01	0,00	0,00	0,01	0,01	0,00
(4) ln(Ind. Publications)	1	0,92	0	6				0,36	0,30	0,06	-0,06	0,23	0,14	-0,11	-0,13	-0,19	-0,06	0,27	0,05	-0,20	0,00	-0,04	0,04	0,07	0,04	-0,02	-0,05	0,03	-0,06
(5) Tenure	7,7	7,55	0	48					0,68	0,06	-0,12	0,04	0,05	0,00	-0,03	-0,02	0,00	0,07	0,01	-0,05	0,00	-0,02	0,02	-0,01	-0,01	0,00	0,05	-0,04	0,01
(6) Tenured	0,2	0,37	0	1						-0,06	-0,14	-0,01	0,00	0,02	0,02	-0,02	0,00	0,04	-0,02	-0,01	-0,02	0,00	0,01	-0,04	0,05	-0,01	-0,01	-0,04	0,05
(7) Foreign Nationality	0,1	0,29	0	1							0,05	0,01	-0,02	-0,02	0,02	0,01	0,02	-0,02	0,00	0,00	0,00	-0,04	0,04	0,02	-0,02	0,03	-0,04	0,03	-0,04
(8) Gender(=Female)	0,2	0,42	0	1								0,12	0,12	-0,02	-0,09	-0,10	-0,07	0,21	-0,03	-0,08	-0,01	0,00	0,01	0,01	0,00	0,00	0,04	-0,03	0,00
(9) Faculty Patents/100	2,40	1,86	1	9									0,64	-0,06	-0,07	-0,26	-0,38	0,63	0,27	-0,35	0,03	-0,14	0,08	-0,03	-0,20	-0,15	-0,08	0,41	-0,07
(10) Faculty Publications/1000	0,70	0,51	0	2										0,23	0,05	0,19	-0,53	0,46	0,01	0,08	0,07	-0,17	0,04	-0,04	-0,18	-0,35	0,53	0,11	-0,11
(11) Basic Research Funding (m)	5,86	4,50	1	16											0,77	0,64	-0,41	-0,18	-0,37	0,74	-0,04	0,02	0,36	-0,07	-0,19	-0,12	-0,02	-0,18	0,11
(12) Applied Research Funding (m)	6,90	6,23	0	20												0,68	-0,47	-0,31	-0,18	0,75	-0,16	-0,11	0,15	-0,07	-0,12	-0,01	-0,15	0,12	0,23
(13) No. of Tenured Faculty	57,54	19,39	27	92													-0,28	-0,51	-0,24	0,80	-0,05	-0,30	0,06	0,05	-0,20	-0,08	0,34	0,08	0,04
(14) ICT	0,22	0,41	0	1														-0,24	-0,22	-0,49	-0,01	-0,07	0,08	-0,08	0,03	-0,03	-0,07	0,09	-0,05
(15) Life Science	0,17	0,38	0	1															-0,19	-0,42	-0,09	0,03	0,04	-0,07	0,07	-0,08	-0,09	0,15	-0,05
(16) Physics & Electr. Engin.	0,15	0,36	0	1																-0,39	0,10	0,02	-0,07	0,16	-0,05	0,04	0,03	-0,12	0,04
(17) Other Engineering	0,46	0,50	0	1																	0,01	0,02	-0,05	0,01	-0,04	0,05	0,10	-0,10	0,05
(18) Univ. 1	0,08	0,27	0	1																		-0,08	-0,17	-0,05	-0,09	-0,11	-0,11	-0,13	-0,09
(19) Univ. 2	0,08	0,26	0	1																			-0,16	-0,04	-0,09	-0,11	-0,10	-0,13	-0,09
(20) Univ. 3	0,24	0,43	0	1																				-0,09	-0,18	-0,21	-0,20	-0,25	-0,17
(21) Univ. 4	0,02	0,15	0	1																					-0,05	-0,06	-0,06	-0,07	-0,05
(22) Univ. 5	0,09	0,29	0	1																						-0,12	-0,12	-0,14	-0,10
(23) Univ. 6	0,12	0,33	0	1																							-0,13	-0,17	-0,11
(24) Univ. 7	0,11	0,32	0	1																								-0,16	-0,11
(25) Univ. 8	0,17	0,37	0	1																									-0,13
(26) Univ. 9	0,08	0,27	0	1																									

TABLE 6.4: Descriptive Statistics

6.4 Results

Table 6.5 shows the estimation results for the dependent variable "invention disclosure", which allow for testing Hypotheses 1 and 2. Model 1 shows that taste for patents has a strongly significant effect in support of hypothesis 1. However, Model 2 shows that incentives alone seem to better explain intended invention disclosure. Model 3 includes both taste and incentives separately, while our proposed Model 4 also includes the interaction effects between these two. It shows that there are 5 positive interaction effects compared to one negative. From these first results one would conclude that there is a complimentary rather than a substitutive effect between taste for patents and external incentives.

Model	**I**		**II**		**III**		**IV**	
Independent variables	Parameter	(S.E.)	Parameter	(S.E.)	Parameter	(S.E.)	Parameter	(S.E.)
Focal Variable								
Taste for Patents	0,644 ***	(0,013)			0,745 ***	(0,014)	0,109 *	(0,063)
Incentives								
Share Inventor			0,051 ***	(0,001)	0,051 ***	(0,001)	-0,003	(0,007)
Share Group			0,035 ***	(0,002)	0,036 ***	(0,002)	0,021 ***	(0,008)
Share Faculty			0,015 ***	(0,002)	0,015 ***	(0,002)	0,008	(0,009)
One-off Payment			0,093 ***	(0,018)	0,093 ***	(0,002)	0,028 ***	(0,008)
Performance Eval.			0,574 ***	(0,025)	0,575 ***	(0,025)	0,084	(0,110)
Award			0,339 ***	(0,031)	0,339 ***	(0,031)	0,029	(0,138)
TTO on Campus			0,206 ***	(0,027)	0,207 ***	(0,027)	0,177	(0,131)
Grace Period			0,648 ***	(0,023)	0,649 ***	(0,023)	1,131 ***	(0,098)
Interactions								
Taste * Share Inventor							0,011 ***	(0,001)
Taste * Share Group							0,003 **	(0,002)
Taste * Share Faculty							0,002	(0,002)
Taste * One-off Payment							0,014 ***	(0,002)
Taste * Performance Eval.							0,102 ***	(0,023)
Taste * Award							0,065 **	(0,028)
Taste * TTO on Campus							0,006	(0,026)
Taste * Grace Period							-0,101 ***	(0,021)
Individual-level Controls								
Industrial Involvement	0,907 ***	(0,193)	2,814 ***	(0,195)	1,081 ***	(0,198)	0,811 ***	(0,197)
Industrial Involvement^2	-1,206 ***	(0,243)	-2,951 ***	(0,249)	-1,421 ***	(0,251)	-1,103 ***	(0,250)
ln(Ind. Patents)	0,022	(0,023)	0,135 ***	(0,023)	0,023	(0,023)	0,053 **	(0,023)
ln(Ind. Publications)	-0,054 ***	(0,019)	-0,172 ***	(0,019)	-0,055 ***	(0,019)	-0,064 ***	(0,019)
Tenure	-0,019 ***	(0,003)	-0,039 ***	(0,003)	-0,232 ***	(0,053)	-0,023 ***	(0,003)
Tenured	-0,213 ***	(0,052)	-0,416 ***	(0,053)	-0,022 ***	(0,003)	-0,207 ***	(0,053)
Foreign Nationality	0,119 **	(0,049)	0,496 ***	(0,050)	0,137 ***	(0,050)	0,122 **	(0,050)
Gender(=Female)	-0,275 ***	(0,035)	-0,358 ***	(0,035)	-0,302 ***	(0,035)	-0,347 ***	(0,035)
Faculty-level Controls								
Discipline Dummies	*(Incl.)*		*(Incl.)*		*(Incl.)*		*(Incl.)*	
University Dummies	*(Incl.)*		*(Incl.)*		*(Incl.)*		*(Incl.)*	
Constant	0,067	(0,093)	-0,747 ***	(0,098)	-4,115 ***	(0,120)	-1,033 ***	(0,322)
Random Effect (σ)	1,488 ***	(0,015)	1,942 ***	(0,017)	1,734 ***	(0,017)	1,786 ***	(0,017)
5 Ordered Thresholds	*(Incl.)*		*(Incl.)*		*(Incl.)*		*(Incl.)*	
No of obs. (1.408*12)	16.896		16.896		16.896		16.896	
Parameters (k)	27		34		35		43	
Log likelihood (6)	-31.678		-31.678		-31.678		-31.678	
Log likelihood (k)	-28.651		-26.965		-26.830		-26.776	
Chi-square	6.054 ***		9.427 ***		9.698 ***		9.805 ***	
McFadden R^2 (adj.)	0,095		0,148		0,152		0,153	

Two-tailed t -tests; * < 0.1, ** p < 0.05, *** p < 0.01

TABLE 6.5: Estimation Results Invention Disclosure

To further visualize the net effect of the interactions obtained from Model 4, we plot the predicated probability of a very high intention to file an invention disclosure under low, average and high incentive conditions. This analysis is depicted in Figure 6.2. It shows that the effect of taste for patents under low incentives is almost negligible. A self-determined taste for patents can only unfold its effects under moderated levels of incentives. The average partial effect under average incentive conditions is two percentage points (s.e. = 0.00069). Under high incentive conditions the extremely complimentary nature between self-determined taste and external incentives becomes apparent. Instead of a maximum probability of 10%, very high levels of taste give rise to 70% probability under high incentive conditions. The average partial effect under high incentive conditions is 14 percentage points (s.e. = 0.00550). All in all, we take this as support for hypothesis 1, the direct effect of self-determined taste for patenting, as well as support for hypothesis 2a, the crowding-in effect of taste for patents under high incentive conditions.

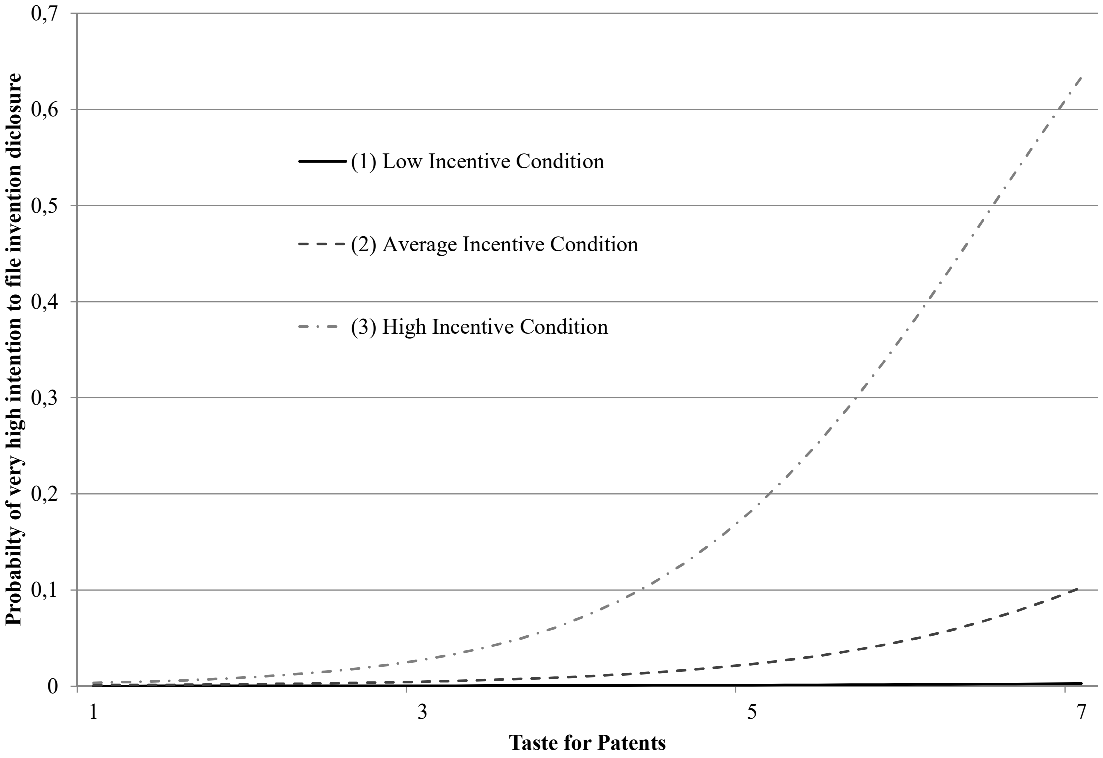


FIGURE 6.2: Effect Of Taste For Patents under Different Incentive Conditions

Table 6.6 shows the estimation results in order to test the hypotheses 3 a and b regarding individual and contextual determinants of a self-determined taste for patents. Model 5 includes the faculty level determinants first, whereas model 6 enters individual variables only. Model 7 shows the combination of both. Model 5 shows that the variables that reflect an applied research orientation of the institutional context indeed lead to a higher taste for patents among researchers. Model 6 shows that the variables reflecting an applied research orientation of the individual have a positive effect on taste, while the variables reflecting basic research orientation have a negative effect on taste. In the combined model the individual level results hold, while the faculty level results weaken. All in all we take this as partial support for Hypothesis 3a and full support for Hypothesis 3b.

Model	**V**		**VI**		**VII**	
Independent variables	Parameter	(S.E.)	Parameter	(S.E.)	Parameter	(S.E.)
Individual-level Variables						
Industrial Involvement			2,186 ***	(0,484)	2,174 ***	(0,467)
Industrial Involvement^2			-1,705 **	(0,701)	-1,717 **	(0,674)
ln(Ind. Patents)			0,143 **	(0,058)	0,144 **	(0,067)
ln(Ind. Publications)			-0,166 ***	(0,051)	-0,175 ***	(0,056)
Tenure			-0,020 ***	(0,005)	-0,020 ***	(0,005)
Tenured			-0,300 **	(0,134)	-0,296 **	(0,138)
Foreign Nationality			0,354 ***	(0,099)	0,362 ***	(0,104)
Gender(=Female)			0,038	(0,074)	0,036	(0,074)
Faculty-level Variables						
Faculty Patents/100	0,250 **	(0,101)			0,311 ***	(0,121)
Faculty Publications/1000	-0,031	(0,029)			-0,022	(0,033)
Basic Research Funding (m€)	-0,001	(0,012)			-0,003	(0,016)
Applied Research Funding (m€)	0,017 *	(0,009)			0,017	(0,012)
No. of Tenured Faculty	0,003	(0,003)			-0,001	(0,003)
Constant	4,336 ***	(0,139)	4,542 ***	(0,067)	4,353 ***	(0,151)
STD. of Random Effect	0,182 ***	(0,035)	0,216 ***	(0,034)	0,092 ***	(0,031)
STD. of Error Term	1,138 ***	(0,014)	1,091 ***	(0,019)	1,095 ***	(0,020)
No of obs.	1.408		1.408		1.408	
Parameters (k)	8		11		16	
Log likelihood (3)	-2.196		-2.196		-2.196	
Log likelihood (k)	-2.191		-2.135		-2.130	
Chi-square	10 ***		123 ***		133 ***	
McFadden R^2	0,002		0,028		0,030	

Two-tailed t -tests; * < 0.1, ** p < 0.05, *** p < 0.01

TABLE 6.6: Estimation Results Taste For Patents

6.5 Discussion and conclusion

Empirical results confirm Hypothesis 1: an attitude in favor of commercialization translates into higher intention to disclose inventions. There exists a self-determined motivation in researchers to commercialize their knowledge with consequences to their subsequent behavior. This finding supports and extends existing research on motivational aspects of scientists and is a requirement for the subsequent hypotheses dealing with the interaction of incentives and motives with regard to disclosure intentions.

The analysis of the role of incentives on the link between attitude towards patenting and intention to patent reveals that the bundle of incentives has a positive effect on disclosure intention under high incentive conditions. The effect under low incentive conditions is weaker, but still significant. As high incentive conditions imply a bundle of incentives that includes measures sometimes considered to be more controlling, the lack of a crowding-out effect suggests that researchers either consider commercialization a valid activity or at least as instrumental to their own work, i.e. their motivation to disclose patents is well internalized. With regard to policy the finding reveals that incentives scale well and that future research may benefit from a stronger focus on the relation of the incentive to the commercial value of the disclosure (i.e. fairness) than possible crowding-out effects of monetary rewards.

Hypotheses with regard to antecedents to the taste for patents could also be confirmed. As suggested by literature, a stronger basic research focus tends to negatively affect the attitude towards commercialization. This is more pronounced for factors relating to the individual, such as a scientists' publications, than for contextual measures, such as publications by faculty, which suggests that self-determined activities act stronger on attitudes than imprinting or social learning effects. Hypothesis 3b, which suggests that individual and contextual factors indicative of approval of applied research have a positive effect on the taste for patents was also confirmed. In this case both individual and contextual factors are significant. The relation of weaker significance for contextual compared to individual measures found for measures indicating basicness also applies to measures indicating an applied focus, confirming the idea that self-determined actions are more important than social learning. However, the positive and significant effect of an applied context implies that the context should not be neglected when designing an incentive system, especially as we cannot exclude the possibility of a crowding-out effect that applies as consequence of a later change in the incentive system. In that case an institutional context that highlights the importance of commercialization activities and lowers barriers through experience sharing may mitigate negative effects on motivation. As policy implication, taking into account the effects of the institutional context, for example when making hiring decisions, translates into higher efficiency of incentives.

To conclude, the study presented in this chapter extends existing research on academic patenting by defining a taste for patents in the fashion of previously defined tastes for science or commercialization. We explain the effect of incentives on the relation between this taste and the intention to disclose inventions using SDT and find indications that incentives work well, if the motivation to commercialize is well internalized as a result of individual and institutional antecedents to the taste for patents. Our results suggest that the mix of incentives commonly applied works as intended. The lack of crowding-out effects suggests that strengthening these incentives may be beneficial. Additionally, we find that the efficiency of incentives can be improved by considering attitudes of aspiring researchers in the hiring process and by cultivating a culture of commercialization that prevents institutional barriers from interfering with self-determined motivation to translate research results into commercial ventures.

Our study is limited by the possibility that an undermining of motivation through external incentives may take place only when an imposed set of incentives is revoked at a later point. Furthermore, the quality of secondary data on publications and patents leaves something to be desired – incomplete personal data is likely to introduce some error into the attribution of patents and publications to individuals. Future research may investigate the relation between the value of innovations and the rewards typically allocated to university scientists. A mismatch in this respect may have stronger motivational implications than the nature of monetary rewards.

Chapter 7

Broadcast search and knowledge distance

Chapter 7 presents the results of a study [1] on crowdsourcing of R&D problems using innovation intermediaries. Since Chapters 7 and 8 are set in the context of nanotechnology a brief introduction to this technology is provided here.

> *"...there is a device on the market, they tell me, by which you can write the Lord's Prayer on the head of a pin. But that's nothing; that's the most primitive, halting step in the direction I intend to discuss" (Feynman, 1960)*

For this study the field of nanotechnology is used to reduce the amount of data which needs to be preprocessed. Nanotechnology is, as cross-sectional and general purpose technology (GPT), a good choice as a means to reduce the size of datasets: it covers both applied and basic research. It also intersects with several other scientific disciplines, such as material sciences and physics. GPTs are defined as technologies that, due to offering a wide field of applications, have a stronger effect on the economy than more specialized technologies (Bresnahan and Trajtenberg, 1995). According to Bresnahan and Trajtenberg (1995), GPTs are characterized by a potential for spread use in broad range of sectors as well as high technological dynamism (i.e. a potential for further improvement). A few existing GPTs are drivers of economic development and lead to large numbers of innovations in terms of applied technologies. Furthermore, GPTs may lead to advances that in turn have an effect on downstream research and development efforts, an effect termed "innovational complementarities" by Bresnahan and Trajtenberg (1995). Hence GPTs provide

[1] The study is the result of joint work with Christoph Ihl (Hamburg University of Technology) and Robin Kleer (TU Berlin).

substantial incentives for investment from both private and public sectors. A typical example for a GPT is the steam engine with its wide-spread economic and societal impact in the industrial revolution (Crafts, 2004). Nanotechnology is defined as any technology that involves manipulation of matter at the scale of one to 100 nanometers (a range that includes objects the size of molecules up to viruses) (Hornyak et al., 2008). The scale restriction implies that quantum-mechanical effects are relevant for nanotechnologies: altering matter, specifically its shape or size, at the nanoscale can change the material's properties such as its electric conductivity, fluorescence or melting point (Hornyak et al., 2008). This enables the creation of essentially new materials with applications in fields such as medicine or material sciences. A popular example for nanomaterials is graphene, a one-atom thick honey-comb lattice of carbon with interesting properties: it is 100 times stronger than steel, conducts electricity and heat efficiently and is nearly transparent. Graphene is also the basic building block for other nanomaterials such as carbon nanotubes (Gogotsi and Presser, 2013). Some of the significance of nanotechnology lies in the potential risks associated with its derivative products. There is uncertainty with regard to the toxicity of nanomaterials (Donaldson et al., 2004). For example, carbon nanotubes may exhibit some of the same similar detrimental effects on health as those associated with asbestos (Poland et al., 2008).

Open innovation is exemplified in crowdsourcing platforms that allow firms to broadcast R&D problems to a wide range of potential solvers. A few studies so far indicate that especially solvers from distant fields have higher chances to make the winning contributions in crowdsourcing contests. It is not fully understood, however, what generally attracts potential solvers to crowdsourcing in the first place and how solvers' knowledge distance towards the broadcasted innovation problem in particular affect their initial interest and adoption. To investigate this question, we situate our study in the field of nanoscience and nanotechnology. By the means of topic modeling with over 900.000 scientific papers and 35 real requests for proposals (RfPs), we are able to locate solvers and problems within a knowledge space and measure the distance between them. In a field experiment, we invite scientists to inspect randomly assigned RfPs of high and low distance. In a subsequent discrete choice analysis, we measure their willingness to engage in solving the assigned R&D problem conditional on contractual arrangements. Our findings lend support to the conjecture that knowledge distance reduces scientists' attention paid towards broadcasted innovation problems and their willingness to solve them. Contractual arrangements can only partially mitigate this effect. Solvers that are more closely linked to the problem are also more responsive to contract attributes. More distant solvers can best be incentivized by higher award money and by the right to license the invention also to third parties. Overall, we shed light on managing an important trade-off in innovation crowdsourcing: while more distant solvers could make valuable contributions, they are more difficult to contract.

The chapter is organized as follows: in the next sub-chapter we introduce the topic and continue to describe the theoretical background on broadcast search and knowledge distance. We then describe our data gathering process and the employed empirical methodology. Results are shown and discussed in the following sections. The chapter concludes with a summary of the findings and a discussion of their implications for future research.

7.1 Introduction

Many companies open up their innovation process to gain access to external knowledge from different domains (Laursen and Salter, 2006). Open innovation characterizes an innovation process that operates as an open search and solution process beyond technical and organizational boundaries (Chesbrough, 2003; Dahlander and Gann, 2010). The rationale behind open innovation is to overcome problems of local search and industry blindness (Stuart and Podolny, 2007; Rosenkopf and Nerkar, 2001). While locally bounded search may be advantageous when current problems are similar to old ones, a limited search space only leads to obvious solutions and rarely to radical advancements (Rosenkopf and Almeida, 2003).

As a management approach, open innovation offers different methods and practices which support innovating companies to identify and integrate relevant external knowledge. Next to conventional arrangements, such as innovation alliances or contract research, Internet technology has enabled new forms of distributed problem solving, subsumed under the terms "tournament-based crowdsourcing" (Afuah and Tucci, 2012) or "broadcast search" (Jeppesen and Lakhani, 2010). These new forms are considered to be especially well suited to overcome local search biases and to tap into unobvious knowledge domains.

Prior research on broadcast search platforms has mainly focused on the characteristics of participants and its effect on the contest's outcome (Jeppesen and Lakhani, 2010) as well as on perceptions of the contests after participation (Franke, Keinz, and Klausberger, 2013).

Within this chapter, we examine the barriers researchers face in contributing to such an open call. We focus on the importance of knowledge distance in determining participation in innovation contests. Knowledge distance is closely related to the classical trade-off that organizations face during the knowledge transfer process: if knowledge is too far away, it is difficult to transfer; if it is too close, there is little new information (Gilsing et al., 2008).

We use an empirical study in the field of nanotechnology to analyze the relation of knowledge distance and participation likelihood. Using a topic modeling approach, we match researchers in nanotechnology with real-world contests broadcasted by innovation intermediaries. Participants first report their perception of the contest in a survey; subsequently they are confronted with variants of contractual arrangements in a conjoint study to analyze the drivers of their willingness to participate.

We find that knowledge distance negatively affects participation likelihood, a relation that can only be partially mitigated by contractual design parameters. This finding is interesting from the perspective of innovation intermediaries (or more generally the innovation seekers), as former studies argue that more distant solvers are likely to submit more innovative solutions (Jeppesen and Lakhani, 2010). Investigating further into the effects of the contractual design parameters, we find that these more distant solvers response most positively to an increase in monetary incentives and more retained patenting rights. Hence, this study contributes to our understanding of the optimal distance of solvers in broadcast search and drivers of solvers' participation.

7.2 Theoretical background

In this section we describe the theoretical context of our research. First, we summarize existing research on broadcast search and identify knowledge distance as the central construct affecting the willingness of scientists to participate in innovation contests. Subsequently we derive hypotheses regarding the effect of this construct on participation likelihood and its interaction with contractual settings.

7.2.1 Innovation crowdsourcing

Innovation crowdsourcing describes a search mechanism where a seeker's (typically a company) technical problem is announced broadly via web-based platforms to a large and diverse group of potential external solvers in form of an open request for proposals (RfP) (Jeppesen and Lakhani, 2010). The idea is to spread the problem as widely as possible to attract solvers even from unobvious knowledge domains and fields of expertise. Potential solvers screen the problem description and self-select whether to invest in solving the problem and to submit a solution proposal. The seeker then selects among all submissions the solutions that meet pre-defined performance criteria best and either awards a pre-defined prize money or negotiates terms of collaboration with the identified solution providers (Spradlin, 2012).

Very often, this process is facilitated by specialized intermediaries who provide broadcast search as a service to connect solution seeking clients with external solvers (Lakhani et al., 2007). Established intermediaries in this domain include NineSigma, InnoCentive, YourEncore, Atizio, or Yet2.com. Their success is greatly dependent on the ability to match seekers with solvers. Hence, most intermediaries maintain a web-based community of pre-registered solvers. In addition, intermediaries support clients in terms of drafting good problem statements, maintaining client anonymity, preselecting appropriate solutions, and monitoring fair play to prevent exploitation of solution proposals without the acquiring the underlying intellectual property (Diener and Piller, 2010).

Research on innovation crowdsourcing to assist technical problem solving is still scarce. It has mostly focused on the efficient design of contests and the effectiveness of the mechanism. Terwiesch and Xu (2008) propose a theoretical model that shows that an increase in the solver base results in a trade-off between the overall solution diversity and quality as well as solvers' problem-solving effort. This trade-off can be shifted in favor of the former through performance-contingent rather than fixed-price rewards. Boudreau et al. (2011) empirically study the effects of an increased solver base for innovation contests on the TopCoder platform. They find that benefits of a larger solver base, i.e. higher solution quality, outweight the costs of fiercer competition and solvers' reduced effort, especially for complex problems. This is in line with Franke et al. (2014) who argue that a promising measure to increase chances of discovering a successful solution is to invite more candidates. However, with the number of invited contestants the required effort for solution screening is likely to increase (Piezunka and Dahlander, 2015) as there will be a wide range of unfitting proposals. In addition, the number of scientists and their available time is finite, hence there exists a natural limit for this way of improving a contest's outcome.

In order to overcome this problem, pre-selecting candidates based on the RfP may be a suitable solution. In a conceptual paper, Afuah and Tucci (2012) derive a number of testable propositions regarding the effect of the characteristics of problems, type of knowledge transfer and crowd characteristics on the effectiveness of the crowdsourcing mechanism. In particular, they argue that a seeker is more likely to crowdsource a problem if the distance to the knowledge needed is large. Jeppesen and Lakhani (2010) study the effect of distance in their study of crowdsourcing at InnoCentive. They define two types of marginality as predictors of innovation success: a technical marginality that indicates a difference in professional background between the solution seeker and the solver as well as a more social marginality which encompasses gender-related biases. For our study, it is particularly interesting that the effect of technical marginality is positive. The likelihood of submitting a winning solution increases with solvers' perceived technological distance between the problem domain and the solvers' field of expertise, which is attributed to a changed perspective of researchers. These findings raise the question how solvers with a high distance to the problem react to an RfP.

We can conclude that the outcome of innovation crowdsourcing is driven by the number of participants and their proposal quality. Both these factors are influenced by the distance of the potential solver to the problem. Hence, this necessitates a closer analysis of the distance concept and its quantification.

7.2.2 Knowledge distance

Distance can be understood as knowledge heterogeneity that arrives from diverse knowledge resources that each individual exhibits. It has been defined as the distance between different persons in terms of their mental perception function and ability (Nooteboom et al., 2007; Wuyts et al., 2005). The way this distance is structured arises from past behavior and experiences of persons and is therewith unequal for each individual. Due to different backgrounds, people interpret, comprehend and judge the world in various ways (Nooteboom et al., 2007).

Resulting from cognitive inequality, the capability of solving one specific problem is diverging from one solver to another and explains why different scientists have diverse distances to diverse topics (Nooteboom et al., 2007; Wuyts et al., 2005).

Alternatively distance can be defined in terms of technological knowledge among potential partners of a knowledge exchange (Nooteboom et al., 2007). Then it is a construct of how large the knowledge bases and fields of expertise deflect among organizations and individual (Hartig, 2011) and aims at the interspace between organizations in terms of technological assets (Benner and Waldfogel, 2007). Technological distance is a determining factor when transferring knowledge. If the distance in terms of technological knowledge between the communicating instances becomes too large, a mutual understanding between those is precluded and the transfer is likely to fail (Nooteboom et al., 2007) If the technological distance is too close, the technological familiarity takes "out the innovative steam" and dramatically decreases the likelihood of novelty creation, which is the actual purpose of knowledge transfer (Gilsing et al., 2008). This view stresses the mutual learning aspect of collaborations. In contrast, Mowery et al. (1998) argue that firms are often not learning from one another in collaborations but are rather accessing or acquiring specific information. This perspective, which has also been confirmed in empirical studies by Grant and Baden-Fuller (2004) and Nielsen and Nielsen (2009), argues that larger distances are useful as the new knowledge only has to be integrated, which avoids the cost of learning (Balconi et al., 2013). In the context of broadcast search, this knowledge assessing view reflects the perspective of the seeker. Thus, it can be used to explain why seekers aim at getting proposals from distant solvers. However, the solvers' incentives and benefits are neglected in analysis by Balconi et al. (2013).

While prior research on innovation platforms considered the positive effects of distance in performance, the effect of distance on participation likelihood has so far been neglected. Intuitively, a perceived distance to a subject decreases the probability of participation as scientists are likely to spend most of their attention on their specialty. Prior research has confirmed this intuition: in research on decision theory under uncertainty it has been shown that unfamiliarity reduces the likelihood of acting on an opportunity, there exists a bias towards the status quo (Samuelson and Zeckhauser, 1988). Constant et al. (1996) find that higher expertise leads to more contributions in online discussion. This is confirmed by Wasko and Faraj (2000) who show that individuals tend to respond more frequently in crowdsourcing situations when they feel to have sufficient expertise in the respective field. Haas et al. (2015) find additional evidence that individuals allocate more attention to a problem that is closer to their field. The costs and benefits of participation can serve as an explanation for such behavior. In addition to monetary rewards, solving RFPs is also expected to enhance reputation or to encourage future reciprocity (Chiu et al., 2011; Constant, Sproull, and Kiesler, 1996; McLure Wasko and Faraj, 2000).

These benefits are more likely and higher in expectation if the contest is closer to the participants' interests, thus expected benefits are higher. Moreover, similar fields increase the chances that participants have the necessary absorptive capacity to understand the RFP (Cohen and Levinthal, 1990). As Kotha et al. (2012) argue, overlapping expertise fosters mutual knowledge and reduces communication costs. Therefore, it is easier for potential solvers to understand and make sense of a problem, capture its special characteristics and dependencies, and finally identify and submit a solution (Thomas, Sussman, and Henderson, 2001; Tsai, 2001). Boudreau et al. (2011) add that greater knowledge distance can be interpreted as being less well informed, which is regarded in risk and decision theory as an indicator for greater uncertainty. It is thus likely that potential participants discount their chances of succeeding in the crowdsourcing contest on the basis of "ambiguity aversion" (Fox and Tversky, 1995). Thus, costs for participation are higher in expectation if knowledge distance is higher. Piezunka and Dahlander (2015) find that inviting many solvers to contribute overburdens the inviting company which then resorts to focus on input that is not distant. So even if distant solvers participate there is a chance that the seeker may be biased against distant knowledge. Additionally, the probable lack of prior experience with innovation platforms may introduce additional bias. This implies that scientists unfamiliar with innovation platforms and/or a given technical problem from an innovation platform are less likely to participate in a contest than scientists experienced with platforms and/or the sort of technical problem. The impact of knowledge (dis-)similarity has also been researched in the field of alliance formation: companies active in similar technological contexts are more likely to cooperate (Mowery, Oxley, and Silverman, 1998; Rothaermel and Boeker, 2008) as common knowledge stocks increase absorptive capacity, enabling firms to assimilate knowledge at lower cost (Lane and Lubatkin, 1998).

The concept of absorptive capacity as explanatory mechanism for a negative effect of distance on participation has been studied by Haas et al. (2015) in the context of company internal forums for problem solving. Consequently, we formulate our first hypothesis:

> *Hypothesis 1: The larger the knowledge distance of a scientist towards an RFP, the less likely is s/he willing to participate.*

Next to the knowledge distance, it is likely that contractual details influence participation behavior (Franke, Keinz, and Klausberger, 2013).

While the expected direct effect of contractual details is obvious in many cases (e.g., increased participation for higher monetary payment), the interaction of knowledge distance and contractual details is interesting (as these details can be influenced by the seeker / intermediary). In particular, it is interesting to know what measures an intermediary can employ to attract more distant solvers. The various business models of innovation intermediaries hint to the fact that different sets of incentives or platform characteristics are used to attract different kinds of solvers. We can group the effect of contractual details in two categories: Safeguards and effects on costs and benefits of the solver. We expect more distant solvers to require additional contractual safeguards as their distance is likely to result in higher uncertainty. Hence, given ambiguity aversion, distant solvers would be less likely to contribute given the same set of safeguards. Since distant solvers are less likely to benefit from potential spillovers, such as reputational effects or industry connections, when they participate in broadcast search, additional compensation may be required to entice their participation.

7.3 Data and methods

7.3.1 Identification of researchers and deriving a distance measure

The distance between a scientists' experience and the technical problem is an important aspect in the design of our study. For the purpose of our study it would prove useful to be able to measure distance as a first step to pre-select respondents. In contrast to interpersonal distance, a more technological distance may be quantified by careful analysis of the texts describing the technical problem, on the one hand, and the texts written by the potential solver, on the other. A manual approach to this problem, ideally by experts in the respective fields, may yield high quality results but seems impractical given the large amount of data. Instead we use machine learning algorithms to compare texts.

Our research focusses on RFPs and researchers in the field of nanotechnology. While this limits the scope of our study to a certain extent, nanotechnology is a general purpose technology and therefore possible distance values are not too restricted. As a first step we gathered information on RFPs published by innovation intermediaries related to nanotechnology by searching the web. We found some 4.700 RFPs from two of the leading broadcast search platforms: NineSigma and Innocentive. RFPs related to nanotechnology were identified by a simple search for the keyword "nano*" (i.e. words containing "nano"). The results were checked using a more complex search based on Arora et al. (2013) to rule out false positives (such as "nanoliter", which contains the letters "nano" but does not necessarily imply that the article is about nanotechnology), resulting in 110 nanotech RFPs. These were manually checked for suitability. RFPs that, despite the keyword-based searches, were not clearly related to nanotechnology were removed (in some instances nanotechnology was only mentioned very briefly among possible approaches to a viable solution). We also removed RFPs that differed significantly in text length (i.e. data available to automatic processing), time required for solution as well as required team size. The resulting set of 38 RFPs was processed to remove irrelevant information (e.g. specifics on the submission process, formatting) as well as information related to the treatments of the planned conjoint analysis (e.g. firm identity). We noticed that inactive RFPs (those that have been withdrawn or where a winner has been awarded) contained less information compared to active RFPs, with inactive RFPs representing the majority of RFPs in the downloaded dataset. However, the information contained in inactive RFPs corresponds well to the information contained in a publication abstract. While we lose information in comparison to an analysis based on full RFP descriptions and full papers the smaller amount of data significantly eases preprocessing. Once pre-processed, the RFPs were integrated into an online survey based on the estimated distance between RFP topic and scientists' field of expertise. The data source for estimating this distance was publications relating to nanotechnology from 2000 to 2011 downloaded from the Web of Science. The bibliographic information from the Web of Science articles was searched for author e-mail addresses. To increase the expected response rate we only kept e-mails from researchers who published in the years 2010 or 2011. Starting with this set of e-mails, author names were disambiguated using a custom Python script: word similarity metrics based on name components (last name, first name and initials), location information and co-author information were taken into account to find all nanotechnology publications in the dataset for each of the authors identified by searching for e-mail addresses. As a result, we obtained data on approximately 24.000 scientists.

7.3.2 Deriving distance measures

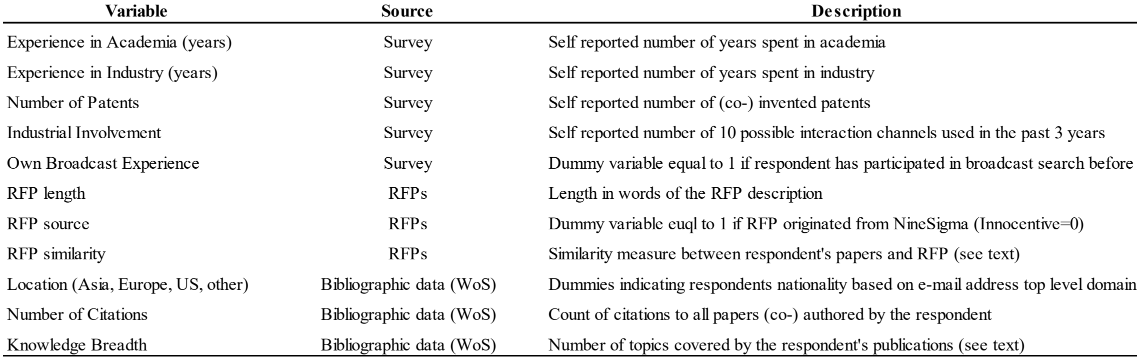
To estimate the proximity of a scientist's prior work to an RFP, we used Latent Dirichlet Allocation (LDA), a generative model for text data also known as topic model (Blei et al., 2003). Topic models are generative statistical models that return probability distributions of groups of words that tend to occur together in texts (topics). As a topic model translates documents into vectors of topics, the model can be used to compare two texts using similarity measures that calculate the distance between two vectors. For details on LDA see Chapter 4.

To make sure that our model accounts for the variance in scientific and technical texts we used a large number of scientific paper abstracts as well as patent abstracts from the field of nanotechnology published in the years 2000-2012 (approximately 850.000). After pre-processing the abstracts using the Python library NLTK (Bird, Klein, and Loper, 2009) (removing irrelevant information such as copyright data, stemming the remaining text and removing stopwords), a number of models were calculated by varying the topic parameter. We compared two implementations of LDA: Gensim and Mallet (McCallum, 2002; Rehurek and Sojka, 2010). Model perplexity and subjective tests of similarity scores obtained with various models were used to select a model estimated with 250 topics in Mallet. We proceeded to calculate the similarity of each paper for each author to the RFPs in our sample. The similarity measure employed is cosine similarity, i.e. the cosine of the angle between two vectors. The average of similarity scores for one author's papers to all RFPs as well as the highest similarity was used to determine the proximity of each author's knowledge stock to the various technical challenges described in the RFPs. For each author three RFPs were selected, two that are conceptually close (highest and second-highest similarity score between author's papers and one RFP) and one that is more distant to the author's work (RFP where similarity score is close to the mean similarity of all author-RFP pairings). For the online survey either the closest or the medium distance RFP was randomly allocated to candidates. The second-closest RFP was kept as a reserve; in the event that a respondent decides not to complete the survey after reading the initial RFP, the scientist is given the option to complete the study with the second-best RFP instead.

Since prior art conceptualizes knowledge disparity as distance, we convert the topic model cosine similarity measure to a knowledge distance measure using the approach described in Goldberg et al. (2016) of using an exponential link between negative distance and similarity.

7.3.3 Survey

We invited the scientists identified with Web of Science articles to a survey in which they were confronted with the RFP, which were selected as described above. In the questionnaire to the RFP we retrieve some general information on researchers. Table 7.1 shows variables obtained either from the survey or by analysis of bibliographic data.

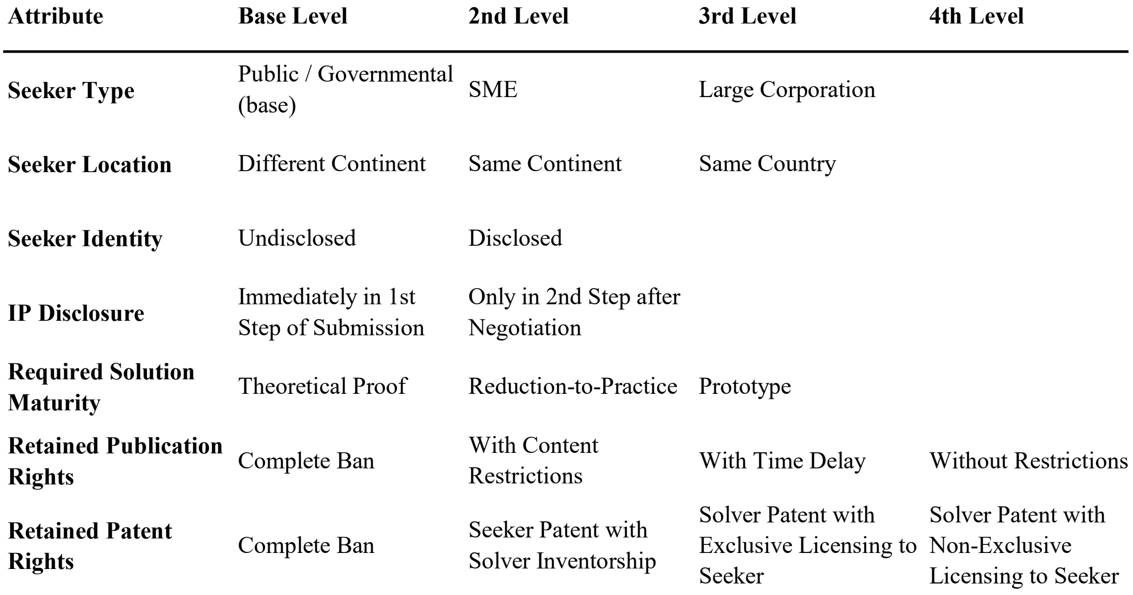


Variable	Source	Description
Experience in Academia (years)	Survey	Self reported number of years spent in academia
Experience in Industry (years)	Survey	Self reported number of years spent in industry
Number of Patents	Survey	Self reported number of (co-) invented patents
Industrial Involvement	Survey	Self reported number of 10 possible interaction channels used in the past 3 years
Own Broadcast Experience	Survey	Dummy variable equal to 1 if respondent has participated in broadcast search before
RFP length	RFPs	Length in words of the RFP description
RFP source	RFPs	Dummy variable euql to 1 if RFP originated from NineSigma (Innocentive=0)
RFP similarity	RFPs	Similarity measure between respondent's papers and RFP (see text)
Location (Asia, Europe, US, other)	Bibliographic data (WoS)	Dummies indicating respondents nationality based on e-mail address top level domain
Number of Citations	Bibliographic data (WoS)	Count of citations to all papers (co-) authored by the respondent
Knowledge Breadth	Bibliographic data (WoS)	Number of topics covered by the respondent's publications (see text)

TABLE 7.1: Variable Description

Potential solvers were asked to report their experience (in years) in academia and industry respectively. Solvers were also asked to report how many patents they (co-) invented. To measure the degree of involvement with industry, we used the industrial involvement index developed by Bozeman and Gaughan (2007) and normalized it to the range 0-1. Own Broadcast Experience is a dummy variable that takes the value one if potential solvers already had some experience with broadcast search platforms in the past.

Subsequently researchers participated in a conjoint experiment where the following variables were modified across a set of five scenarios (i.e. each respondent was presented with five variations in levels of the following variables): the seeker type was set to either small company, large company or governmental organization. The identity, i.e. the name of the company or organization, is either revealed or withheld. Seeker location could be either Asia, Europe or the US. Incentives and barriers consisted of one variable for the required technical maturity, a variable for the timing of IP disclosure by the solver as well as one variable for retained publication and patenting rights respectively. Finally, four levels of financial rewards for submitting a winning solution were defined, ranging from US$ 10.000 to US$ 75.000. Details on these variables are shown in Table 7.2. Following each scenario, respondents were asked to choose for one of two contractual designs differing in the variables described. A none-option was included.



Attribute	Base Level	2nd Level	3rd Level	4th Level
Seeker Type	Public / Governmental (base)	SME	Large Corporation	
Seeker Location	Different Continent	Same Continent	Same Country	
Seeker Identity	Undisclosed	Disclosed		
IP Disclosure	Immediately in 1st Step of Submission	Only in 2nd Step after Negotiation		
Required Solution Maturity	Theoretical Proof	Reduction-to-Practice	Prototype	
Retained Publication Rights	Complete Ban	With Content Restrictions	With Time Delay	Without Restrictions
Retained Patent Rights	Complete Ban	Seeker Patent with Solver Inventorship	Solver Patent with Exclusive Licensing to Seeker	Solver Patent with Non-Exclusive Licensing to Seeker

TABLE 7.2: Conjoint Experiment Variables

In order to account for the possibility of different RFPs containing varying amounts of information despite pre-processing, we control for the length of the RFP description in words. We also control for the RFP source (Ninesigma or Innocentive). The RFP similarity was obtained using the topic modeling approach described above.

The location of the solver was estimated from the top-level domain of author e-mail addresses obtained from bibliographic data.

Bibliographic data was also used to determine the citation count for each author. We further defined a variable indicative of the breadth of a scientist's knowledge by comparing a scientist's publication abstracts to all RFPs in our dataset. The average similarity of one scientist's abstracts to the topics obtained by topic modelling, relative to the average similarity of all scientists' publications to these topics, was used to decide whether a scientist is familiar with a given topic. We then summed the familiar topics for each author to obtain a breadth measure ranging from 0 to 250. We finally asked respondents to choose between two contracts with the additional option of not participating in either. Based on this question we form our dependent variable participation. In total, we received 249 responses to the survey and the conjoint analysis with five contractual experiments per respondent.

7.3.4 Latent class estimation

In order to analyze participants' responses in the conjoint study, we use latent class regression, an extension of the logit model. Logit models are a type of generalized linear model that estimate the utility of a choice as linear function of parameters that is linked to the categorical dependent variable through the logit function. The model can be extended for the case of more than two outcome categories (in our case respondents can opt for one of two contractual arrangement or a none-option) with a multinomial logit model. If in addition to choice-variant attributes choice-invariant variables are to be included, a conditional logit model is employed. The latent class model is a conditional logit model that allows correcting for unobserved preference heterogeneity with latent classes: observations are grouped along similar utility parameter estimates. For a logit model with k parameters each latent class adds another k parameters, hence there is a risk of overfitting. We use the Bayesian Information Criteria (BIC) to calculate a trade-off between the number of additional classes and the gain in log-likelihood. The BIC penalizes the use of additional parameters more strongly than the Akaike Information Criterion (AIC), thereby allowing us to optimize the trade-off between more complex model specification and model (over)-fit.

7.3.5 Selection bias

Bias may be introduced at two points in our study: after receiving an invitation e-mail a potential responder may choose to open the survey. Subsequently the potential responder chooses whether to complete the survey. It is possible that these two decisions are functions of the characteristics of the potential respondent. Hence the remainder of completed surveys used for further analysis may differ from the invited (random) population systematically. We correct for this issue using a modified two-step Heckman correction. According to Heckman (1979), selection effects can be regarded as instances of truncated data. Including an inverted Mills ratio calculated from an initial probit regression in a subsequent linear regression corrects the introduced bias. The Heckman correction has since been extended for cases of double selection (Mohanty, 2001), i.e. two subsequent selection effects. In this case the correction involves calculating two Mills ratios from a bivariate probit regression to correct the bias in both selection stages. The equation for a latent class regression is given as:

$$p(y_{i,t} = j|c) = \frac{e^{\beta_c x_{jit}}}{\sum_{j=1}^{J} e^{\beta_c x_{jit}}},\ if\ COM_i = 1, \tag{7.1}$$

$$= 0\ otherwise$$

i.e. the probability of outcome y given class c is a function of class specific parameter vectors β and attribute x for individual i, choice alternative j and choice situation t. The choice yit is observable only when an individual has completed the survey and choice experiment, so we let COM_i denote whether a participant completed the survey and PAR_i whether the individual participated:

$$PAR_i = 1,\ if\ y_{1i} > 0 \tag{7.2}$$

$$= 0\ otherwise$$

$$(COM_i|PAR_i) = 1,\ if\ y_{2i} > 0 \tag{7.3}$$

$$= 0 \; otherwise$$

Where y_{1i} and y_{2i} denote the outcome of two probit models used to estimate whether individuals participate in the survey and whether they subsequently complete the survey:

$$y_{1i} = x_{1i}\beta_i + \epsilon_{1i} \tag{7.4}$$

$$y_{2i} = x_{2i}\beta_i + \epsilon_{2i} \tag{7.5}$$

from which we can calculate selectivity variables:

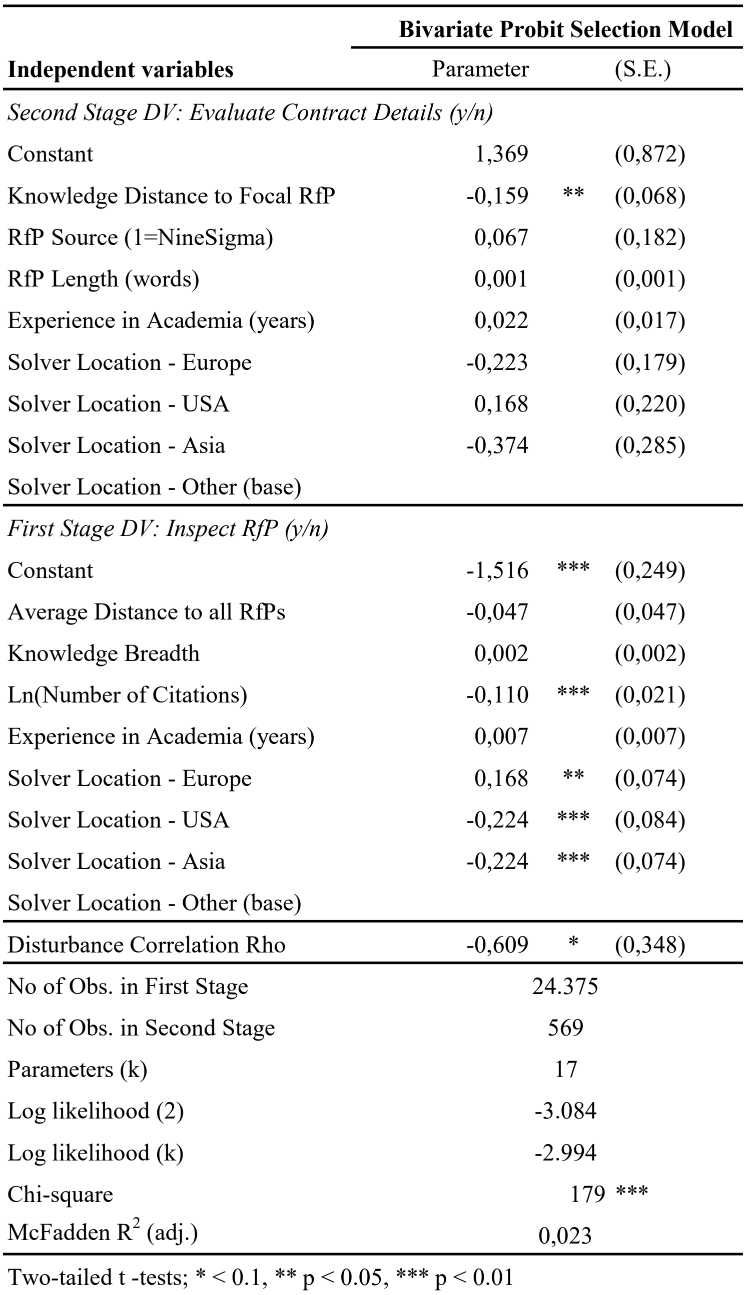
$$\lambda_{1i} = \frac{\phi(a_i)\Phi(A_i)}{F(a_i, b_i, \rho)} \tag{7.6}$$

$$\lambda_{2i} = \frac{\phi(b_i)\Phi(B_i)}{F(a_i, b_i, \rho)} \tag{7.7}$$

with $a_i = x_{1i}\beta_1$, $b_i = x_{2i}\beta_2$, $A_i = \frac{(b_i - \rho a_i)}{\sqrt{1-\rho^2}}$, $B_i = \frac{(a_i - \rho b_i)}{\sqrt{1-\rho^2}}$, ϕ as univariate standard normal density, Φ as cumulative standard normal density and F as bivariate standard normal distribution function. Including the Mills ratios from 7.6 and 7.7 as control variables in the latent class regression 7.1 corrects the selection bias.

7.4 Results

Table 7.3 shows the results from the selection model. We account for the selection process of scientists into our sample and potential selection biases in three stages. During the first stage 24.374 invited scientists decide whether or not to inspect the assigned RfP. During the second stage 569 scientists who inspected the assigned RfP decide whether or not to evaluate the contractual details of the assigned RfP. During the third stage 229 scientists decide whether or not to accept certain contracts with respect to the assigned RfP. In the statistical analysis of this double selection process (Mohanty, 2001), a scientist's average distance to all 35 RfPs in our study serves as an exclusion restriction from stage one to stage two, whereas RfP length and source serve as exclusion restrictions from stage two to stage three. The bivariate probit selection model covering the first and second stage shows that the first stage decision to inspect the RfP is less likely for scientists with higher number of citations after controlling for other background variables. Thus, academic achievement in terms of citations has a negative effect on the interest in innovation crowdsourcing. In the second stage, conditional upon inspecting the RfP, a further interest in judging the contractual details of the assigned RfP is less likely for scientists with a greater knowledge distance towards the RfP. This lends initial support to our overall conjecture that knowledge distance reduces scientists' attention paid towards crowdsourced innovation problems. The significant correlation of error terms ρ reveals a significant selection effect such that unobserved factors positively (negatively) affecting the decision to inspect the RfP in the first stage negatively (positively) affect the decision to further evaluate the contractual details in the second stage.



	Bivariate Probit Selection Model		
Independent variables	Parameter		(S.E.)
Second Stage DV: Evaluate Contract Details (y/n)			
Constant	1,369		(0,872)
Knowledge Distance to Focal RfP	-0,159	**	(0,068)
RfP Source (1=NineSigma)	0,067		(0,182)
RfP Length (words)	0,001		(0,001)
Experience in Academia (years)	0,022		(0,017)
Solver Location - Europe	-0,223		(0,179)
Solver Location - USA	0,168		(0,220)
Solver Location - Asia	-0,374		(0,285)
Solver Location - Other (base)			
First Stage DV: Inspect RfP (y/n)			
Constant	-1,516	***	(0,249)
Average Distance to all RfPs	-0,047		(0,047)
Knowledge Breadth	0,002		(0,002)
Ln(Number of Citations)	-0,110	***	(0,021)
Experience in Academia (years)	0,007		(0,007)
Solver Location - Europe	0,168	**	(0,074)
Solver Location - USA	-0,224	***	(0,084)
Solver Location - Asia	-0,224	***	(0,074)
Solver Location - Other (base)			
Disturbance Correlation Rho	-0,609	*	(0,348)
No of Obs. in First Stage	24.375		
No of Obs. in Second Stage	569		
Parameters (k)	17		
Log likelihood (2)	-3.084		
Log likelihood (k)	-2.994		
Chi-square	179		***
McFadden R^2 (adj.)	0,023		

Two-tailed t -tests; * < 0.1, ** p < 0.05, *** p < 0.01

TABLE 7.3: Selection Model

Table 7.4 shows descriptive statistics for variables used in model estimation, across the three stages (invited population, partial completion of survey, completed survey). Cross correlation does not appear to be a concern. Noticeable are some outliers as regards the number of patents, citation count and years of experience in industry and academia.

Variable	Source	N	Mean	Std.	Min	Max	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
1st Stage - Invited Scientists: N=24,375																					
(1) Solver Location - Other (base)	WoS	24.375	0,1	0,2	0,0	1,0	-0,22	-0,13	-0,18	-0,09	-0,06	-0,06	0,072								
(2) Solver Location - Asia	WoS	24.375	0,4	0,5	0,0	1,0		-0,42	-0,59	-0,11	-0,06	-0,18	0,155								
(3) Solver Location - USA	WoS	24.375	0,2	0,4	0,0	1,0			-0,33	0,039	0,057	0,201	-0,22								
(4) Solver Location - Europe	WoS	24.375	0,3	0,5	0,0	1,0				0,128	0,046	0,048	-0,02								
(5) Experience in Academia (years)	WoS	24.375	8,2	2,9	0,0	12,0					0,234	0,343	-0,21								
(6) Knowledge Breadth	WoS	24.375	62,5	10,6	18,0	129,0						0,15	-0,44								
(7) Ln(Number of Citations)	WoS	24.375	4,5	1,0	0,0	9,3							-0,47								
(8) Average Distance to all RfPs	RfPs/WoS	24.375	2,6	0,5	1,3	4,8															
2nd Stage - Inspected RfP: N=569																					
(1) Solver Location - Other (base)	WoS	569	0,1	0,3	0,0	1,0	-0,18	-0,10	-0,29	-0,16	-0,06	-0,07	0,05	0,00	0,06	-0,02	0,00	0,00			
(2) Solver Location - Asia	WoS	569	0,3	0,5	0,0	1,0		-0,23	-0,67	-0,17	-0,16	-0,08	0,17	-0,07	0,04	0,03	0,23	0,00			
(3) Solver Location - USA	WoS	569	0,1	0,3	0,0	1,0			-0,38	0,01	0,04	0,10	-0,22	-0,03	-0,03	-0,17	0,14	0,00			
(4) Solver Location - Europe	WoS	569	0,5	0,5	0,0	1,0				0,23	0,15	0,04	-0,04	0,08	-0,05	0,09	-0,30	0,00			
(5) Experience in Academia (years)	WoS	569	8,3	2,9	2,0	12,0					0,30	0,36	-0,22	0,03	0,02	-0,04	-0,05	0,00			
(6) Knowledge Breadth	WoS	569	63,1	10,4	32,0	95,0						0,17	-0,46	0,03	-0,05	-0,07	-0,05	0,02			
(7) Ln(Number of Citations)	WoS	569	4,3	1,0	1,4	7,1							-0,36	-0,06	-0,01	-0,16	0,14	0,01			
(8) Average Distance to all RfPs	RfPs/WoS	569	2,7	0,5	1,5	4,3								-0,05	0,09	0,31	-0,01	-0,02			
(9) RfP Length (words)	RfPs/WoS	569	141,2	66,5	74,0	439,0									-0,31	0,02	-0,04	0,00			
(10) RfP Source (1=NineSigma)	RfPs/WoS	569	0,9	0,3	0,0	1,0										0,03	0,02	0,00			
(11) Knowledge Distance to Focal RfP	RfPs/WoS	569	1,8	0,9	0,2	3,7											-0,04	0,00			
(12) Inverse Mills Ratio (Lambda) 1 Stage	Estimate	569	2,3	0,5	1,4	3,6												0,94			
(13) Inverse Mills Ratio (Lambda) 2 Stage	Estimate	569	0,0	0,9	-1,3	1,8															
3rd Stage - Completed Contract Decisions: N=229																					
(1) Solver Location - Other (base)	WoS	229	0,1	0,3	0,0	1,0	-0,15	-0,13	-0,35	-0,18	-0,05	-0,09	0,12	-0,02	0,10	0,05	-0,10	-0,21	-0,02	-0,06	-0,08
(2) Solver Location - Asia	WoS	229	0,2	0,4	0,0	1,0		-0,21	-0,55	-0,11	-0,15	-0,04	0,00	-0,07	0,15	0,03	0,79	0,74	-0,04	-0,03	0,28
(3) Solver Location - USA	WoS	229	0,2	0,4	0,0	1,0			-0,49	0,00	0,07	0,14	-0,17	-0,03	0,05	-0,15	0,16	-0,41	0,15	0,01	0,02
(4) Solver Location - Europe	WoS	229	0,6	0,5	0,0	1,0				0,19	0,09	-0,02	0,06	0,09	-0,21	0,06	-0,69	-0,17	-0,07	0,05	-0,20
(5) Experience in Academia (years)	WoS	229	8,6	2,8	2,0	12,0					0,30	0,40	-0,20	0,05	-0,01	-0,11	-0,22	-0,32	0,04	0,06	0,00
(6) Knowledge Breadth	WoS	229	64,1	10,6	37,0	90,0						0,20	-0,47	0,04	-0,07	-0,03	-0,24	-0,24	0,16	0,22	0,09
(7) Ln(Number of Citations)	WoS	229	4,3	1,0	1,6	6,7							-0,40	-0,05	-0,05	-0,22	0,29	-0,05	0,14	0,18	0,12
(8) Average Distance to all RfPs	RfPs/WoS	229	2,6	0,5	1,6	3,8								-0,09	0,13	0,34	0,03	0,22	-0,23	-0,27	-0,23
(9) RfP Length (words)	RfPs/WoS	229	144,0	70,4	74,0	439,0									-0,44	0,10	-0,14	-0,11	-0,05	0,10	-0,07
(10) RfP Source (1=NineSigma)	RfPs/WoS	229	0,9	0,3	0,0	1,0										0,00	0,14	0,04	0,10	0,02	0,09
(11) Knowledge Distance to Focal RfP	RfPs/WoS	229	1,7	0,8	0,2	3,5											0,19	0,54	-0,09	-0,07	-0,18
(12) Inverse Mills Ratio (Lambda) 1 Stage	Estimate	229	2,9	0,3	2,4	3,6												0,73	0,01	-0,03	0,23
(13) Inverse Mills Ratio (Lambda) 2 Stage	Estimate	229	1,1	0,2	0,7	1,8													-0,11	-0,06	0,11
(14) Own Crowdsourcing Experience	Survey	229	0,2	0,4	0,0	1,0														0,25	0,23
(15) Industrial Involvement	Survey	229	3,7	2,4	1,0	10,0															0,37
(16) Ln(Number of Patents)	Survey	229	1,4	1,2	0,0	5,4															

TABLE 7.4: Descriptive Statistics

Table 7.5 shows a minimal BIC for a specification with two latent classes, compared to those with one or three classes. Based on BIC, two latent classes clearly yield the best model fit. To make the model even more parsimonious, we test for the restriction whether the control variables that affect the baseline adoption utility do not vary across the two classes (model 4). We accept this restriction because it improves model fit in terms of BIC. In model 5, we include our focal variable knowledge distance as an active covariate to predict class membership of scientists. Since the constant utility also varies with latent classes, this specification allows knowledge distance to exert both a direct effect on baseline adoption utility and a moderating effect on contract preferences. This specification yields a lower BIC, albeit with only little improvement. Based on this specification, we test a further restriction of a linear effect of prize money in model 6, which needs to be rejected. Instead, the specification of a monotonic increasing effect of award money in model 7 gives a more parsimonious model with better approximation to the data.

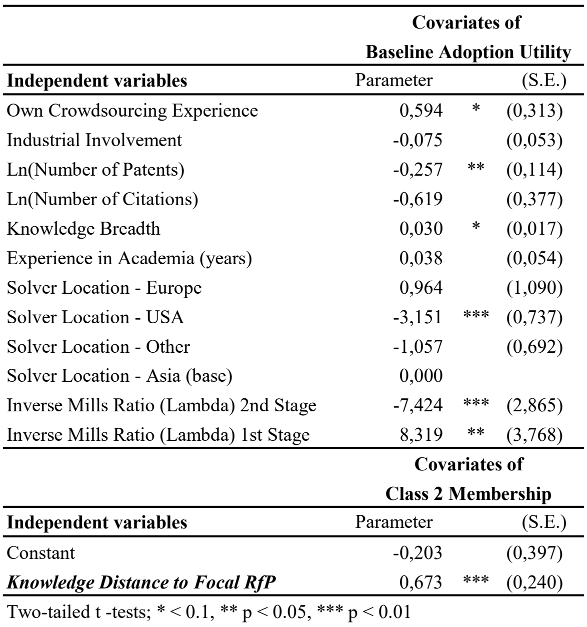
Model	Specification	Npar	LL	BIC(LL)	McFadden R^2
(1)	1 latent class	29	-1.107,96	2.373,51	0,111
(2)	2 latent classes	59	-984,43	2.289,44	0,210
(3)	3 latent classes	89	-910,43	2.304,45	0,269
(4)	2 latent classes & restricted covariates of baseline adoption	48	-995,15	2.251,12	0,201
(5)	... & including knowledge distance as a covariate of class membership	49	-992,09	2.250,44	0,204
(6)	... & restricted award money attribute to me metric	45	-1.005,35	2.255,21	0,193
(7)	... & restricted award money attribute to be monotonic increasing	47	-993,86	2.243,10	0,203

TABLE 7.5: Latent Class Model Selection

Table 7.6 shows covariates of the baseline adoption utility. Inverse Mills Ratios from the bi-variate probit were included to correct for the double selection effect. As both lambda parameters are significant, we can reject the null hypothesis that no selection effect occurs. In particular, this means that the 229 sampled scientists positively select themselves into the adoption of crowdsourcing contracts in the third stage compared to a random sample from the 24.374 invited scientists due to unobserved factors affecting the first stage decision to inspect the RfP itself. And the 229 sampled scientists negatively select themselves into the adoption of a crowdsourcing contract in the third stage compared to a random sample from the 569 scientists who inspected the RfP due to unobserved factors affecting the second stage decision to evaluate the contractual details of the RfP. The latter effect could be interpreted that scientists who would benefit from entering into a crowdsourcing contract are deterred from thinking about contractual details in the first place.

Conditional upon selection into the contract decision stage and controlling for (unobserved) factors affecting the double selection into this third stage, the sampled scientists' number of citations do not exert any further (negative) effect on baseline adoption of innovation crowdsourcing. Instead, previous experience with innovation crowdsourcing as well as knowledge breadth in the problem domain have a positive impact on the baseline utility of accepting a crowdsourcing contract, whereas the number of patents in the problem domain as a proxy for scientists' own commercialization potential has a negative effect.

Our focal variable knowledge distance is highly significant in predicting membership of the second class, which comprises over 70% of our sample. Scientists in this distant class have a significantly lower baseline adoption utility, thus, again lending support to our overall conjecture that knowledge distance reduces scientists' attention paid towards crowdsourced innovation problems. Furthermore, they have on average a lower responsiveness to all contracting attributes except for award money. The second important difference is that they benefit much more from being granted the right to apply for an own patent with a non-exclusive licensing.



	Covariates of Baseline Adoption Utility		
Independent variables	Parameter		(S.E.)
Own Crowdsourcing Experience	0,594	*	(0,313)
Industrial Involvement	-0,075		(0,053)
Ln(Number of Patents)	-0,257	**	(0,114)
Ln(Number of Citations)	-0,619		(0,377)
Knowledge Breadth	0,030	*	(0,017)
Experience in Academia (years)	0,038		(0,054)
Solver Location - Europe	0,964		(1,090)
Solver Location - USA	-3,151	***	(0,737)
Solver Location - Other	-1,057		(0,692)
Solver Location - Asia (base)	0,000		
Inverse Mills Ratio (Lambda) 2nd Stage	-7,424	***	(2,865)
Inverse Mills Ratio (Lambda) 1st Stage	8,319	**	(3,768)
	Covariates of Class 2 Membership		
Independent variables	Parameter		(S.E.)
Constant	-0,203		(0,397)
Knowledge Distance to Focal RfP	0,673	***	(0,240)

Two-tailed t -tests; * < 0.1, ** p < 0.05, *** p < 0.01

TABLE 7.6: Latent Class Model Part One

Table 7.7 shows averaged parameters for the choice-variant attributes. Except for the seeker type all attribute parameters are significantly different from zero. A Wald test for equality reveals that there are significant differences between one parameter across the two classes for seeker identity, IP disclosure, retained patenting rights and award money. Table 7.7 also shows the relative importance of attributes for the two classes with the distant class attributing more importance to financial incentives as well as publication and patenting rights.

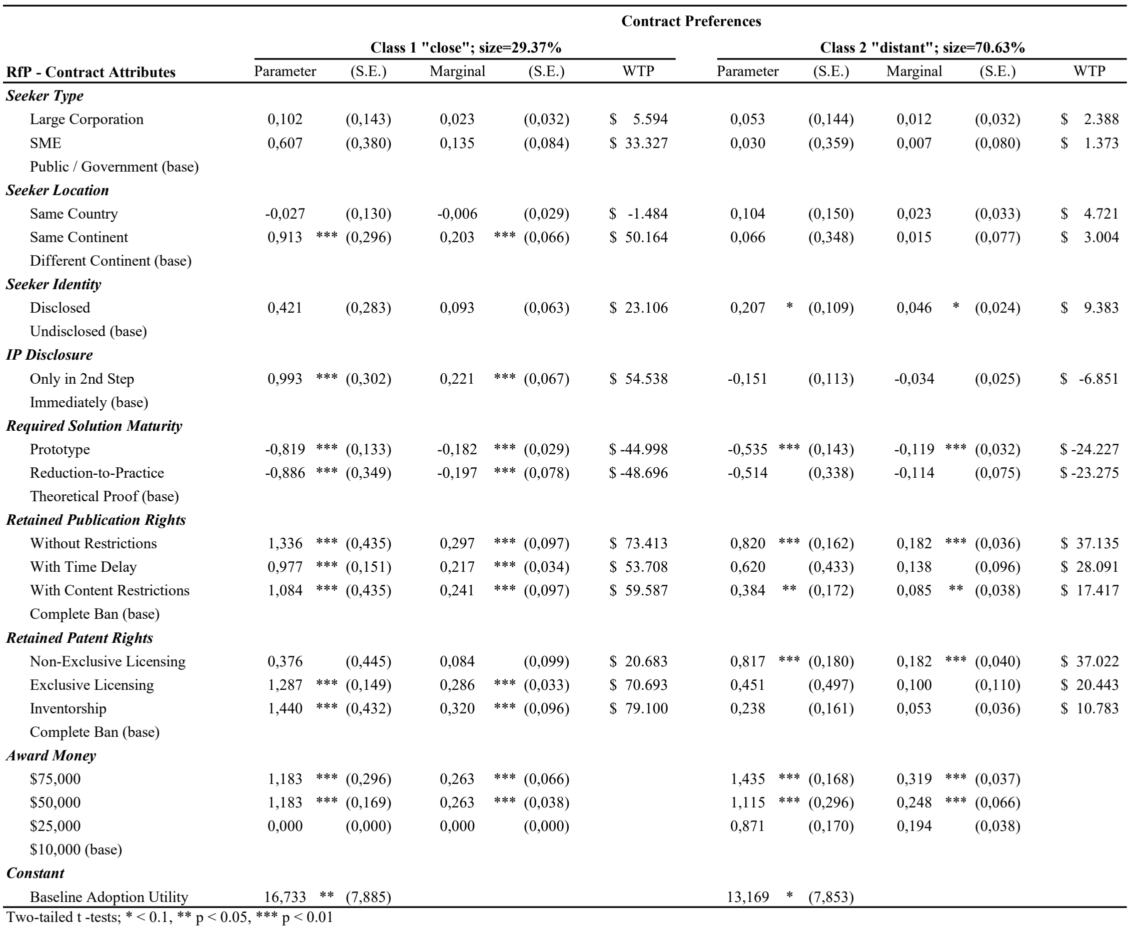
	Attribute Significance			Class Differences			Contract Preferences			
							Class 1: "close"		Class 2: "distant"	
RfP - Contract Attributes	Wald(0)	df		Wald(=)	df		Abs. Import.	Rel. Import.	Abs. Import.	Rel. Import.
Baseline Adoption Utility	188,5	2	***	188,5	1	***				
Seeker Type	3,4	4		2,7	2		0,61	8%	0,05	1%
Seeker Location	11,5	4	**	7,7	2	**	0,94	12%	0,10	3%
Seeker Identity	6,3	2	**	0,5	1		0,42	5%	0,21	5%
IP Disclosure	11,9	2	***	11,9	1	***	0,99	13%	0,15	4%
Required Solution Maturity	29,0	4	***	1,0	2		0,89	11%	0,53	13%
Retained Publication Rights	39,8	6	***	2,4	3		1,34	17%	0,82	20%
Retained Patent Rights	36,7	6	***	14,1	3	***	1,44	18%	0,82	20%
Award Money	100,1	4	***	36,8	3	***	1,18	15%	1,43	35%

Two-tailed t -tests; * < 0.1, ** p < 0.05, *** p < 0.01

TABLE 7.7: Latent Class Model Part Two

Table 7.8 shows detailed parameter estimates and standard deviations for choice attributes for both latent classes. As the parameters of the non-linear logit models cannot be easily interpreted, we also included a "willingness to pay" measure to indicate the relative importance of an attribute value relative to the financial incentive. For class one the contractual incentives / barriers are significant and affect participation in the expected way. While higher required technical maturity reduces participation likelihood, an increase in the retained publication or patenting rights or in monetary incentives increases participation likelihood. The WTP indicates the non-linear relation between levels of the categorical variables: retaining publication rights without any restriction seems to be of much higher utility than either time or content restrictions. Compared to class one the more distant class seems to be more difficult to motivate for participation with contractual settings. Several incentive levels that motivate class one solvers are not significant for class two solvers. If they are significant, the WTP is lower. Interestingly, more distant solvers exhibit a strong preference for non-exclusive patent licensing. More distant solvers also appear to prefer seekers who reveal their identity. This may indicate that uncertainty plays a role in participation likelihood for this class of solvers and that seekers can reduce this uncertainty by being more transparent.

The difference in baseline adoption utility shows that more distant solvers are less likely to approve of the contractual settings, which corresponds well to the previous results of only a few contractual parameters being useful in enticing their participation.



	Contract Preferences									
	Class 1 "close"; size=29.37%					Class 2 "distant"; size=70.63%				
RfP - Contract Attributes	Parameter	(S.E.)	Marginal	(S.E.)	WTP	Parameter	(S.E.)	Marginal	(S.E.)	WTP
Seeker Type										
Large Corporation	0,102	(0,143)	0,023	(0,032)	$ 5.594	0,053	(0,144)	0,012	(0,032)	$ 2.388
SME	0,607	(0,380)	0,135	(0,084)	$ 33.327	0,030	(0,359)	0,007	(0,080)	$ 1.373
Public / Government (base)										
Seeker Location										
Same Country	-0,027	(0,130)	-0,006	(0,029)	$ -1.484	0,104	(0,150)	0,023	(0,033)	$ 4.721
Same Continent	0,913 ***	(0,296)	0,203 ***	(0,066)	$ 50.164	0,066	(0,348)	0,015	(0,077)	$ 3.004
Different Continent (base)										
Seeker Identity										
Disclosed	0,421	(0,283)	0,093	(0,063)	$ 23.106	0,207 *	(0,109)	0,046 *	(0,024)	$ 9.383
Undisclosed (base)										
IP Disclosure										
Only in 2nd Step	0,993 ***	(0,302)	0,221 ***	(0,067)	$ 54.538	-0,151	(0,113)	-0,034	(0,025)	$ -6.851
Immediately (base)										
Required Solution Maturity										
Prototype	-0,819 ***	(0,133)	-0,182 ***	(0,029)	$ -44.998	-0,535 ***	(0,143)	-0,119 ***	(0,032)	$ -24.227
Reduction-to-Practice	-0,886 ***	(0,349)	-0,197 ***	(0,078)	$ -48.696	-0,514	(0,338)	-0,114	(0,075)	$ -23.275
Theoretical Proof (base)										
Retained Publication Rights										
Without Restrictions	1,336 ***	(0,435)	0,297 ***	(0,097)	$ 73.413	0,820 ***	(0,162)	0,182 ***	(0,036)	$ 37.135
With Time Delay	0,977 ***	(0,151)	0,217 ***	(0,034)	$ 53.708	0,620	(0,433)	0,138	(0,096)	$ 28.091
With Content Restrictions	1,084 ***	(0,435)	0,241 ***	(0,097)	$ 59.587	0,384 **	(0,172)	0,085 **	(0,038)	$ 17.417
Complete Ban (base)										
Retained Patent Rights										
Non-Exclusive Licensing	0,376	(0,445)	0,084	(0,099)	$ 20.683	0,817 ***	(0,180)	0,182 ***	(0,040)	$ 37.022
Exclusive Licensing	1,287 ***	(0,149)	0,286 ***	(0,033)	$ 70.693	0,451	(0,497)	0,100	(0,110)	$ 20.443
Inventorship	1,440 ***	(0,432)	0,320 ***	(0,096)	$ 79.100	0,238	(0,161)	0,053	(0,036)	$ 10.783
Complete Ban (base)										
Award Money										
$75,000	1,183 ***	(0,296)	0,263 ***	(0,066)		1,435 ***	(0,168)	0,319 ***	(0,037)	
$50,000	1,183 ***	(0,169)	0,263 ***	(0,038)		1,115 ***	(0,296)	0,248 ***	(0,066)	
$25,000	0,000	(0,000)	0,000	(0,000)		0,871	(0,170)	0,194	(0,038)	
$10,000 (base)										
Constant										
Baseline Adoption Utility	16,733 **	(7,885)				13,169 *	(7,853)			

Two-tailed t -tests; * < 0.1, ** p < 0.05, *** p < 0.01

TABLE 7.8: Latent Class Model Part Three

7.5 Discussion

According to Nooteboom et al. (2007) the positive effect of distant knowledge arising from its novelty is discounted by increasing demands on absorptive capacity required to assimilate distant knowledge. In the context of broadcast search platforms problems can be accurately described and a narrow solution space defined. This should enable seekers to avoid the negative effects related to distant knowledge as provided solution proposals are likely to adhere to the pre-defined solution space. In this case the novelty value of distant knowledge is merely discounted by the decreasing likelihood of distant solver′s contribution.

However, it is also possible that seekers leave the solution space open in order to attract more distant solvers. In this case absorptive capacity is likely to play a role as proposed solutions to a defined problem stemming from unfamiliar contexts still require more effort on the part of the seeker to understand and as seekers tend to ignore distant solutions when confronted with resource conflicts in the filtering stage (Piezunka and Dahlander, 2015). In this case the novelty value of distant knowledge is discounted by increasing demands on absorptive capacity as well as lower likelihood of distant solver participation. In either case the negative effect of distance on participation likelihood needs to be taken into account. Our study shows that knowledge distance conceptualized as function of RfP description and texts produced by potential solvers transformed to a vector space can be used to predict participation likelihood and reveals interactions with contractual design.

Given our research design it is important to correct for selection bias in order to correctly estimate the effect of distance on participation. We have corrected for a selection effect that occurs when inviting potential solvers to the survey and in the second step, when the solvers decide on whether to complete the survey. This selection effect offered first insights into the link between knowledge distance and participation with many solvers opting out of the survey when they were more distant to the problem.

Prior literature has shown that more distant solvers tend to contribute more valuable solutions, which is a central (implicit) benefit of the concept of broadcast search. Our findings suggest that the promise of broadcast search cannot be realized if the effect of knowledge distance on participation is not accounted for. In a latent class regression we estimate the effect of various choice-variant and invariant parameters on participation. Individual-level variables are only of limited use when attempting to determine who participates. The academic background does not appear to be a significant predictor of participation at this point. However, the selection model has shown a strong negative effect of the citation count on the decision to participate in the survey. Prior experience with broadcast search platforms affects participation likelihood, suggesting that there are either learning effects or a reduction in uncertainty. This implies that innovation intermediaries need to overcome a hurdle when inviting new solvers and should invest into their existing solver network.

Analysis of contractual parameters using a conjoint study shows that more distant solvers do not react as well to incentives. Intermediaries can only partially mitigate this effect by being more transparent or by offering higher rewards. Distant solvers appear to be more sensitive to opportunity costs as evidenced by the class two parameters for financial incentives and non-exclusive patent licensing. The distant solvers' preference for higher rewards and greater freedom with regard to contractual aspects can be explained by relatively higher anticipated costs of understanding and contributing to the problem (Haas, Criscuolo, and George, 2015).

Also, the direct benefit to a distant solver is expected to be lower: a close solver who contributes solutions to a company active in a similar context can expect to benefit from newly gained personal connections and learning effects that may apply to the solver's research. Awareness of these opportunity costs may explain distant solver's preference for higher financial rewards and relaxed contractual barriers.

For the seeking company this implies that increasing incentives will most likely attract more solvers at knowledge distance levels that are not optimal. Aside from the expense related to the incentives, additional effort would be required to filter the submissions. Prior research has shown that under these conditions seekers are likely to be overburdened and resort to discarding submissions that are more costly to evaluate due to the higher distance to their own knowledge domain (Piezunka and Dahlander, 2015). The dilemma of distant solvers being less willing to participate even though they may have the knowledge required for optimal solutions may necessitate modifications to the concept of broadcast search.

7.6 Conclusion

From a managerial perspective, the results of our paper are helpful for intermediaries and other hosts of idea contests. We show that contests can be designed more effectively (in terms of attracting more and better suited participants) if the host is aware of the trade-off implied by knowledge distance. We have validated topic modeling as tool for objectively measuring knowledge distance and preselecting survey respondents. Our results show that publication abstracts serve as a useful predictor for relatedness of researchers to certain problems. This can be useful for technology transfer offices or innovation platforms: Intermediaries may want to confront the distance dilemma by pre-selecting potential respondents with appropriate knowledge distance. Limiting the population of possible participants in this way, it may be possible to attract solvers at the optimal knowledge distance with high incentive levels, while limiting expenses and detriments related to evaluation of a flood of submissions with low probability of success due to sub-optimal knowledge distance. Furthermore, intermediaries may consider a transparent design of the intermediation process to reduce uncertainty in distant solvers.

From a theoretical perspective, we contribute to the understanding of participation motives in idea contests. Our research design that studies potential participants in idea contests allows us to distinguish between different types of participants and their respective motives and barriers to participation.

There are some limitations to our study: our topic model is based only on the abstracts of publications and the publicly available information of RFPs. A more advanced model could use full papers and more detailed descriptions of the technical problems. This might lead to a more accurate measure of distance between researchers and RFPs and thus allow for better match-making. Ideally one would compare scientific abstracts from all disciplines to all types of RFPs to assure coverage of the entire range of distance values between scientific papers and technical problems. We focused our attention on one branch of science, enabling us to download a limited number of RFPs and papers. However, with this restriction comes a possible bias in data selection as far as the possible range of knowledge distances is concerned: by excluding solvers from disciplines not related to the problem described in the RfP, the possible knowledge distance is limited with an upper limit to the possible distance. In order to keep this problem to a minimum, we focus on a general purpose technology (nanotechnology), which limits the scope of the study without infringing too much on possible distance values.

By presenting only one RFP to a potential solver, we simplify from the real situation of potential solvers having to choose from a set of RFPs. However, with regard to the effect of distance on participation this is likely to lead only to an under-estimation of the effect as an increased number of choices tends to shift solver attention to closer problems (Haas, Criscuolo, and George, 2015).

Further research may attempt to investigate possible non-linear relations between distance and participation likelihood: if a full range of possible distance values is taken into account by expanding the dataset to solvers from completely unrelated disciplines, the negative effect of distance on participation may be reinforced.

Chapter 8

Collaboration: breadth, depth and potential

The results of a study[1] on university-industry collaboration are presented in this section. Prior research has found that in a regional context, knowledge can spill over to other institutions. These spillovers also affect knowledge created at universities and hence contributes to the innovative potential of a region. The focus of this study is on collaborative research as a transfer channel between industry and academia. I.e. we try to understand how suitable cooperation between researchers employed by academia and industry is to transfer knowledge generated in academia into a more applied context. We use publication data to identify collaborative networks across German regions. We contribute to existing research by focusing on three aspects of this transfer channel: its breadth, depth and the potential for transferable knowledge.

The remainder of this chapter is structured as follows: the next sub-section introduces the topic. In further sub-chapters prior literature is discussed. We then describe the collection of data and suitable methods for analysis. We continue to report our findings and subsequently discuss their importance. Finally, this chapter concludes with a discussion of limitations and suggestions for future research.

8.1 Introduction

In 1956 Abramovitz described a "non-finding" which set in motion a new stream of research that is unabatedly popular with economists:

[1] The study is the result of joint work with Hannes Lampe (Hamburg University of Technology)

the inputs to a national economic system were insufficient to predict its outputs (Abramovitz, 1956). As a consequence, knowledge has been recognized as important resource for economies that evolve past the industrial stage. One aspect that may attract researchers to the topic is perhaps that the production of knowledge is the key competence of the academic community. Extensive prior research has refined our understanding of knowledge production and identified universities as a potential source for commercial application of knowledge (Cowan and Zinovyeva, 2013; Fagerberg and Verspagen, 2002; Ponds, Oort, and Frenken, 2010).

While universities have been found to be capable of directly translating knowledge into applications (O'Shea et al., 2005), existing research stresses the opportunities of cooperation between academia and industry due to the potential to combine the resources of both domains. Whereas universities create knowledge, industry excels at translating knowledge into successful products or services. This stream of research builds on the knowledge production framework, which estimates knowledge outputs as function of various input parameters at the regional or national level (Audretsch and Lehmann, 2005; Cowan and Zinovyeva, 2013; Ponds, Oort, and Frenken, 2010).

Two mechanisms are regarded as important for linking academic knowledge production and industrial knowledge production: spillovers that occur naturally between co-located actors and collaborations on scientific projects that include actors from both academia and industry (Maietta, 2015). We extend research on the transfer of knowledge between academia and industry by taking a closer look at scientific university-industry collaboration as well as the position of universities within the collaboration networks. We find that collaboration between academia and industry in terms of scientific co-publications positively affects regional industrial knowledge production. We add to the existing literature by showing that this effect can be divided into measures for collaboration breadth, intensity and potential. We find that collaboration breadth emerges as the most robust aspect when explaining the positive relation between collaboration and industrial knowledge outputs. We also find that collaborations enable not only direct knowledge flows from university to industry but also strengthen the effectiveness of local universities in contributing to industrial knowledge production.

8.2 Theory

As recombination of knowledge is regarded important for the economic growth of knowledge-based economies (Keupp and Gassmann, 2013 Leiponen and Helfat, 2009) significant attention has been paid to academia as one of the primary generators of knowledge. Since industry is generally considered to be more effective at translating knowledge into innovations (Robin and Schubert, 2013), a stream of research has concerned itself with the transaction of knowledge from academia to industry and the impact of such knowledge transfer on industrial knowledge outputs. A popular framework for this field of analysis is the "knowledge production function" concept according to which knowledge outputs of industry are modeled as a function of various industrial and academic input factors (Griliches, 1990; Jaffe, 1986) at regional or national levels (Anselin, Varga, and Acs, 1997).

Increasingly, characteristics of universities are taken into account to explicitly model the positive effects of academic knowledge on industrial knowledge production (Gulbrandsen and Smeby, 2005; Robin and Schubert, 2013). The relevance of academia to industrial knowledge production can be explained with the presence of spillovers (Glaeser et al., 1992): intended or unintended transfers of knowledge from universities to industry positively affect firm's abilities to innovate. Several channels have been identified for the transmission of knowledge from academia to industry: informal transfer, such as personal communication in professional or private networks (Breschi and Lissoni, 2001; Ponds, Oort, and Frenken, 2010 Singh, 2005), the transfer of knowledge through university graduates that seek employment in local industry (Almeida and Kogut, 1999; Breschi and Lissoni, 2006; Leten, Landoni, and Van Looy, 2014), university spin-offs (Zucker, Darby, and Armstrong, 1998), an indirect transfer through scientific publications accessed by industry (Leten, Landoni, and Van Looy, 2014) as well as formal collaboration as indicated by scientific publications or patents that name both a university and a company (Powell, Koput, and Smith-Doerr, 1996; Stuart, 2000).

The ease with which information can be transmitted through personal communication results in transfers that may not be intended by the transmitting organization (Marshall, 1898). These localized knowledge spillovers have been found to positively affect a region's knowledge production in the context of university-industry collaborations (Moreno, Paci, and Usai, 2005; Anselin, Varga, and Acs, 1997; Fischer and Varga, 2003; Leten, Landoni, and Van Looy, 2014).

The efficiency of information transfer decreases with distance as knowledge can be tacit (Polanyi, 1966), i.e. it is associated with high transfer costs as some aspects of the knowledge are not explicitly known by the transmitting person. Personal communication reduces these transfer costs. However, the probability that personal communication is chosen as transfer channel decreases with distance (Laursen, Reichstein, and Salter, 2011). In the special context of university-industry collaboration, the transfer may be complicated by different organizational cultures (Lissoni, 2001) increasing transaction costs and making the spatial nature of the transfer process even more relevant. Knowledge flows over longer distances can be enabled by scientific publications. Scientific publications are, as public good (Anselin, Varga, and Acs, 1997), available for any interested party and enable information exchange over greater distances (Maietta, 2015).

However, given the large number of available publications as well as the high degree of specialization inherent to advancements of scientific literature, it is to be expected that significant costs are required for firms to identify and absorb academic knowledge so that publications are less effective at transferring knowledge across the organizational boundary between academia and industry.

These transaction costs can be reduced by formal collaborations as indicated by scientific co-publications with authors from both academia and industry. A co-publication by university and industry indicates a certain degree of investment by the partners as it hints at an underlying formal collaboration (Audretsch and Lehmann, 2005). For the firm, additional transaction costs for identifying relevant literature are either not required or delegated to the partnering university. The absorption of knowledge is eased by participation of company employees in the research project. Furthermore, this type of collaboration is likely to entail a different sort of knowledge flow than alternative transfer channels. Informal communication, while offering the benefit of very low transaction costs, is limited in the depth of information that can be transmitted (Welsh et al., 2008). Graduates, while trained in the basics of their profession, do not come with the same degree of expertise in scientific work required to mitigate the aforementioned transaction costs relevant to absorbing scientific knowledge and spend most of their time at university catching up to the scientific status quo rather than expanding it. Scientific publications (i.e. those that were not part of a collaborative effort with industry) tend to be difficult to absorb for firms. In short, collaborations on scientific publications differ to other transfer channels in that they mitigate significant transaction costs and have the potential of providing more in-depth information.

Prior research on formal university-industry collaboration has used measures derived from economic geography to capture spillover effects over distances (Cowan and Zinovyeva, 2013; Ponds, Oort, and Frenken, 2010 Anselin, Varga, and Acs, 1997). We extend this approach by disentangling the effect into three mechanisms of knowledge transfer: depth, breadth and potential. Prior research on collaboration has found evidence that both the intensity of the collaboration and the number of collaborations have a positive effect on collaboration outcome (Berchicci, 2013; Bercovitz and Feldman, 2011).

With respect to the scenario of collaborations on scientific publications, the breadth is likely to have a positive effect as it increases the access to diverse sources of knowledge, which in turn enable recombination of knowledge. Hence a region that collaborates with many other regions is likely to be more effective at producing knowledge. The depth or intensity of a collaboration is likely to affect the bandwidth of the transaction channel: more frequent collaboration over the same channel is an indication for larger knowledge flows. However, frequent collaborations may also hint at transaction channels that have already been exploited (Gonzalez-Brambila, Veloso, and Krackhardt, 2013). The knowledge that can potentially be transmitted through a channel, i.e. the transfer channel's potential, is likely to depend on characteristics of the collaboration partner. Prior literature has used publication counts to indicate a university's reputation and value as collaboration partner (Butler, 2003; Leten, Landoni, and Van Looy, 2014).

An alternative measure are university R&D expenditures (Ponds, Oort, and Frenken, 2010): successful research institutions are able to acquire more funding which allows for larger staff and better equipment, enabling more and higher quality research. Other available measures for available knowledge stocks, such as publication counts or number of employees, can be subsumed and complemented (e.g. as indicator of better equipment) by R&D expenditures.

Prior research has found collaboration on scientific publications to be a viable channel for information transfer, enabling local industry to benefit from knowledge created at distant universities (Mansfield, 1998). Several mechanisms may explain this finding: information obtained through collaboration may directly translate into innovation (e.g. spin-offs resulting from university-firm cooperation). Incoming information may also be relevant to the firm's processes, making it easier to identify opportunities and absorb additional scientific knowledge (Fleming and Sorenson, 2004).

So far it has been implicitly assumed that such spillovers only exist on the direct route from local industry to distant universities. Positive effects of knowledge created at local universities have been treated as separate effect. In recent literature the network position of local universities has taken a more prominent role in explaining spillover effects. A logical next step would be to test whether scientific collaboration by local academia has a similar effect to firm-industry collaboration on regional knowledge production.

When comparing universities and industry as catalysts of scientific knowledge, some arguments can be found why either side would be (un)-suitable for the task. Private companies are considered to be effective at translating knowledge into innovations (Roessner, 1977). When abstracting from the difficulty involved in absorbing academic knowledge, local industry should emerge as the more efficient catalyst of academic knowledge. However, academic knowledge is likely to be either uncodified (experience of researchers on a given subject matter) or codified in ways that are not easy to absorb (in scientific publications).

Universities should be able to absorb scientific knowledge with lower transfer costs if that information originates from a related scientific discipline: universities usually have access to databases containing recent publications and researchers are experienced in identifying relevant articles and in efficiently extracting information. The ability of academia to advance the frontiers of science increasingly depends on their competence in networking with colleagues, as innovation in science requires a high degree of specialization (Bush and Hattery, 1956; Goffman and Warren, 1980).

Hence cooperation between specialists is an important aspect in scientific advances. However, converting information into applications has traditionally not been a focus of universities, even though studies on the entrepreneurial university and respective policy changes are changing the traditional focus and enable universities and individual researchers to commercialize their knowledge (Bramwell and Wolfe, 2008).

In summary we expect universities to have better access to knowledge from academia but to be less efficient at converting that knowledge into applications, whereas industry is more efficient at translating knowledge into applications but faces higher transaction costs when accessing scientific knowledge.

8.3 Data and methods

According to the knowledge production framework, regional knowledge outputs can be estimated as a function of regional characteristics. We follow prior literature by using patent counts as measure for regional knowledge output (Jaffe, Trajtenberg, and Henderson, 1992), focusing on the knowledge-intensive industry of nanotechnology. We collected data for 412 NUTS3 (Nomenclature des unités territoriales statistiques) regions of Germany (European Commission, 2015) to estimate regional knowledge outputs and relevant input factors. Since production of knowledge is a time-intensive process we distinguish between three time periods (Ponds, Oort, and Frenken, 2010): the knowledge production function's (KPF) output, the number of patents, is measured for the years 2008-2010. We assume that the underlying inventive effort takes place in 2007 and collect controls for that year. The third time-slice, from 2004-2007, includes collaborations that precede the inventive effort.

8.3.1 Count Data

As the collection process for patent data includes both rural regions as well as large urban centers, the resulting variable is both over-dispersed and zero-inflated. Accordingly we estimate regional knowledge output using a zero-inflated negative binomial model (Hoekman, Frenken, and Van Oort, 2009; Zuur et al., 2009) with appropriate tests to compare to non-inflated and Poisson models.

$$P_{i,t} = e^{\alpha \ln x_{i,t-1} + \beta \ln z_{i,t-1} + \gamma \ln a_{i,t-1} + \epsilon} \tag{8.1}$$

In this two-stage model excess zeroes are predicted using a logistic regression, followed by a negative binomial model to estimate the over-dispersed patent counts. A likelihood ratio test is used to compare the zero-inflated negative binomial model to a zero-inflated Poisson model, whereas the Vuong test (Vuong, 1989) is used to compare the negative binomial to the zero-inflated negative binomial model.

8.3.2 Dependent variable

As measure for regional knowledge output we employ nanotechnology patent applications filed in the years 2008-2010. Patent counts were obtained by searching the European Patent Office's statistical database (PATSTAT, 2012 edition). We use keyword-based searches on abstracts and titles (Arora et al., 2013), as well as IPC (international patenting classification) and ECLA (European classification system) classes to identify applications from the field of nanotechnology. To avoid counting one invention multiple times, we matched these applications to their corresponding patent families (Hingley and Park, 2003).

To geocode the resulting set, we used address information contained in PATSTAT as well as information from the REGPAT database (Maraut et al., 2008) and retrieved additional missing location information from Espacenet, the German patent office's online database. Addresses were then matched to their corresponding NUTS (Nomenclature des unités territoriales statistiques) codes (version of 2010), which allowed us to fractionally attribute patent counts to 412 German NUTS3 regions (i.e. to add a share of each patent application to one region according to the corresponding number of applicants/inventors from that region).

8.3.3 Independent variables

As measure for collaborative efforts we used co-publications retrieved from articles downloaded from the Web of Science in the time frame 2004-2007. We again used a key-word based search to retrieve publications on the topic of nanotechnology. Organizational names and addresses were used to geolocate publications and identify collaborations within the academic network (all article authors employed at universities) and collaborations between industry and academia (at least one author employed by a company). Publications were fractionally distributed over regions according to author organization affiliation. For industry-university collaboration, the resulting network between German regions, with aggregated co-publications between two regions as tie strength, is asymmetric as we assume that university R&D expenditures have an effect on local industry. In other words, for a given collaboration between industry in region A and academia in region B we add a directed link from B to A as transfer channel for B's academic knowledge stocks as indicated by R&D expenditures in region B.

The network for collaboration within academia is undirected, we assume that both universities benefit from knowledge stocks of their partnering university.

Our focal variable "collaboration breadth" is the degree centrality of each region within these two networks, i.e. the number of regions connected to the focal region through co-publications.

The collaboration intensity measure is the average tie strength (i.e. aggregated number of fractionally counted co-publications) of the local region to collaborating regions.

The transfer channel's potential is indicated by the knowledge stocks available at the transmitting side: university R&D expenditures of collaborating regions of the focal region are multiplied with a weight matrix constructed from the collaboration network. The weight-matrix is row-standardized in accordance with conventions in spatial analysis (Ponds, Oort, and Frenken, 2010).

8.3.4 Controls

To control for existing knowledge stocks relevant to industrial knowledge production, we use patent application counts for the year 2007, excluding nanotechnology patents. In accordance with prior literature we also control for the effect of local industrial and academic R&D expenditures on knowledge production. Data for these variables were obtained from the 'Stifterverband für Deutsche Wissenschaft' for the year 2007.

We use spatial weight matrices to control for spillovers of R&D expenditures over short distances. Due to the nature of spatial data (regions can be isolated and are not regularly shaped), we use an interpoint distance matrix based on the k-nearest neighbor metric, which gives a better indication of which regions adjoin the focal region compared to simple adjacency matrices (De Smith, Goodchild, and Longley, 2007).

We also control for general economic characteristics of regions that affect knowledge output. The ratio of employees in manufacturing relative to the number of employees in the service sector differentiates industrial regions from regions with a focus on services. The ratio of employees with tertiary education indicates the knowledge intensity of regional jobs. We also control for the average firm in terms of firms' employees. Finally, we control for regional industrial specialization using a Herfindahl index (i.e. the sum of the squared ratios) constructed from regional patent applications in 35 technology categories for the year 2007 (Jaffe and Trajtenberg, 2002).

8.4 Results

Table 8.1 shows descriptive statistics for variables of interest and controls. The number of observations is limited to 412, the number of county-level NUTS regions in Germany. Of those, not all are equally likely to have a stock of nanotechnology patents. Especially rural regions inflate the number of zeroes, indicating over-dispersion and possibly zero-inflation in the dataset. Descriptive statistics indicate certain correlation between some of the variables of interest. We calculated variance inflation factors for the independent variables with all factors at three or lower. In subsequent regressions multicollinearity issues do not appear to be an issue.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	12	(12)	(13)	(14)	(15)
(1) Nano-patents	1															
(2) Collaboration Breadth Uni-Firm	0,7533	1														
(3) Collaboration Potential Uni-Firm	0,3462	0,502	1													
(4) Collaboration Intensity Uni-Firm	0,2633	0,3153	0,5462	1												
(5) Collaboration Breadth Uni-Uni	0,5458	0,5326	0,4419	0,2361	1											
(6) Collaboration Potential Uni-Uni	0,2488	0,2866	0,3286	0,1451	0,6263	1										
(7) Collaboration Intensity Uni-Uni	0,5441	0,5548	0,4209	0,2418	0,885	0,7122	1									
(8) University R&D (ln)	0,3721	0,3603	0,3084	0,1832	0,6668	0,5244	0,6136	1								
(9) Firm R&D (ln)	0,4363	0,3846	0,4077	0,258	0,3725	0,2608	0,3627	0,368	1							
(10) Wdistance Univ. R&D (ln)	0,0396	0,0439	0,1192	0,0781	-0,0233	0,0138	0,0084	-0,0996	0,2045	1						
(11) Wdistance Firm R&D (ln)	0,1027	0,0766	0,217	0,114	0,0395	0,0701	0,0556	0,003	0,4139	0,5995	1					
(12) Manu / Service	-0,1097	-0,0693	-0,0193	-0,0351	-0,2934	-0,227	-0,2603	-0,2981	0,2126	0,1068	0,1756	1				
(12) Share of tert. educ.	0,3181	0,2156	0,298	0,2154	0,2625	0,1758	0,2575	0,2066	0,2417	0,1487	0,2278	-0,2508	1			
(13) Average firm size	0,095	0,1585	0,1023	0,0481	0,3289	0,2487	0,3003	0,4785	0,2368	-0,138	-0,1161	0,0773	-0,3069	1		
(14) Specialization index	-0,3708	-0,3208	-0,3717	-0,2388	-0,2234	-0,1091	-0,2073	-0,1834	-0,5897	-0,2851	-0,5831	-0,2596	-0,2857	0,0668	1	
(15) KB07	0,7825	0,5947	0,3992	0,2826	0,3944	0,2001	0,4165	0,2863	0,5723	0,135	0,3018	0,038	0,3516	-0,0227	-0,6505	1
Mean	1,385922	1,791262	4,142098	0,2384335	7,135922	4,07632	0,4033996	3,141576	10,08362	9,732747	11,09112	0,6371638	0,0606763	10,58634	0,617078	71,92258
SD	3,592246	4,942504	5,762463	0,5970428	15,85782	5,733922	0,7820052	4,775281	1,866488	1,964244	1,281511	0,3626118	0,0245803	2,674636	0,299276	105,9577
Min	0	0	0	0	0	0	0	0	0	0	0	0,0536842	0,0198406	5,393653	0,0427124	0
Max	47	56	13,60597	9,833333	82	13,60597	5,326707	14,26324	15,23551	12,23701	13,29929	2,636364	0,1563714	30,88058	0,9940757	1074,401

TABLE 8.1: Descriptive Statistics

Table 8.2 shows the result of several zero-inflated negative binomial regressions. A likelihood ratio test was performed for each model to compare each model to a zero-inflated Poisson regression. The test reveals that Poisson models are not suitable for our dataset. Additionally a Vuong test was conducted for each model to compare the ZINB model to a conventional negative binomial regression. Again, all models fit significantly better with the ZINB specification. As expected, some of the controls, in particular the proportion of employees with tertiary education in a region, are useful for predicting excess zeroes.

Model	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Breadth Uni-Firm		0.033***	0.028***	0.029***				0.029***
		(0,008)	(0,008)	(0,008)				(0,008)
Potential Uni-Firm			0.029***	0.038***				0.036***
			(0,011)	(0,013)				(0,013)
Intensity Uni-Firm				-0,183				-0,159
				(0,151)				(0,143)
Breadth Uni-Uni					0.009**	0.010**	0.012*	0.012*
					(0,004)	(0,004)	(0,007)	(0,007)
Potential Uni-Uni						-0,008	-0,006	-0,002
						(0,014)	(0,014)	(0,014)
Intensity Uni-Uni							-0,052	-0,176
							(0,126)	(0,130)
Controls								
University R&D (ln)	0.031*	0,022	0,021	0,02	0,015	0,018	0,017	0,014
	(0,016)	(0,016)	(0,016)	(0,016)	(0,017)	(0,018)	(0,018)	(0,017)
Firm R&D (ln)	0.182***	0.146**	0.127**	0.131**	0.161***	0.162***	0.162***	0.127**
	(0,059)	(0,058)	(0,058)	(0,058)	(0,058)	(0,058)	(0,058)	(0,057)
Wdistance Univ. R&D (ln)	0.142**	0.119**	0.113*	0.114*	0.154**	0.156***	0.159***	0.129**
	(0,058)	(0,059)	(0,058)	(0,058)	(0,060)	(0,060)	(0,060)	(0,059)
Wdistance Firm R&D (ln)	-0.227***	-0.147*	-0.142*	-0.141*	-0.191**	-0.187**	-0.190**	-0,121
	(0,080)	(0,083)	(0,081)	(0,081)	(0,081)	(0,081)	(0,080)	(0,080)
Manu/Service	-0,164	-0,17	-0,131	-0,14	0,019	0,017	0,021	-0,04
	(0,201)	(0,199)	(0,199)	(0,197)	(0,216)	(0,216)	(0,216)	(0,212)
Share of tert. educ.	12.850***	11.791***	10.630***	10.539***	11.832***	11.598***	11.581***	10.008***
	(2,392)	(2,351)	(2,338)	(2,316)	(2,407)	(2,433)	(2,428)	(2,314)
Average firm size	0.101***	0.084***	0.081***	0.080***	0.089***	0.087***	0.088***	0.076***
	(0,031)	(0,030)	(0,030)	(0,030)	(0,031)	(0,031)	(0,031)	(0,029)
Specialization index	-0.811**	-0.837***	-0.737**	-0.703**	-0.736**	-0.740**	-0.735**	-0.644**
	(0,316)	(0,313)	(0,311)	(0,312)	(0,315)	(0,314)	(0,314)	(0,309)
KB07	0.002***	0.001**	0.001***	0.001***	0.002***	0.002***	0.002***	0.001***
	(0,000)	(0,000)	(0,000)	(0,000)	(0,000)	(0,000)	(0,000)	(0,000)
Constant	-2.260**	-2.274**	-2.214*	-2.264**	-2.547**	-2.566**	-2.582**	-2.608**
	(1,152)	(1,156)	(1,142)	(1,143)	(1,143)	(1,141)	(1,137)	(1,127)
Zero-Inflated Part								
Firm R&D (ln)	-0,045	-0,145	-0,187	-0,186	-0,104	-0,108	-0,106	-0,182
	(0,361)	(0,337)	(0,353)	(0,347)	(0,339)	(0,336)	(0,335)	(0,339)
Manu/Service	2.741**	2.762**	2.980**	2.979**	2.991**	3.009**	3.018**	3.143**
	(1,368)	(1,344)	(1,442)	(1,432)	(1,383)	(1,386)	(1,388)	(1,449)
Share of tert. educ.	-46.176*	-49.135*	-56.842*	-56.900*	-49.105*	-50.681*	-50.916*	-59.855*
	(26,371)	(27,382)	(30,161)	(29,842)	(27,655)	(28,251)	(28,330)	(30,979)
Average firm size	0,198	0,162	0,177	0,173	0,17	0,165	0,165	0,157
	(0,164)	(0,160)	(0,168)	(0,167)	(0,163)	(0,163)	(0,163)	(0,166)
Specialization index	-1,302	-1,476	-1,019	-0,948	-1,026	-0,965	-0,918	-0,671
	(2,952)	(2,785)	(2,953)	(2,941)	(2,938)	(2,915)	(2,916)	(2,922)
KB07	-0.137***	-0.125***	-0.123***	-0.122***	-0.122**	-0.119**	-0.118**	-0.113***
	(0,052)	(0,046)	(0,045)	(0,044)	(0,049)	(0,048)	(0,048)	(0,042)
Constant	2,879	4,311	4,315	4,295	3,26	3,33	3,268	4,077
	(4,791)	(4,517)	(5,004)	(4,967)	(4,708)	(4,673)	(4,681)	(4,943)
Fit Statistics								
Alpna (ln)	-2.021***	-2.201***	-2.301***	-2.362***	-2.140***	-2.164***	-2.183***	-2.570***
	(0,395)	(0,397)	(0,424)	(0,442)	(0,432)	(0,440)	(0,450)	(0,538)
LL	-447,6347	-439,0784	-435,7557	-434,7072	-445,1886	-445,0272	-444,9415	-433,1528
Chi^2	232.44***	249.55***	256.2***	258.29***	237.33***	237.65***	237.83***	261.4***
LR Test (alpha=0)	12.82***	14.06***	11.89***	10.69***	10.48***	10.16***	9.02***	5.84***
Vuong	3.38***	3.52***	3.45***	3.48***	3.26***	3.25***	3.25***	3.5***
N	412	412	412	412	412	412	412	412

*Notes: 412 observation (198 zero); * $p<0.1$; ** $p<0.05$; *** $p<0.01$*

TABLE 8.2: Regression Results: Knowledge Production

Having controlled for zero inflation, we begin with a base model that incorporates local R&D expenditures by industry and academia as well as controls for spatial spillovers from adjacent regions. All

R&D measures have a positive effect on local knowledge production except for industry R&D expenditures in adjacent regions. The significant negative effect of spatially lagged industrial expenditures may indicate that urban regions with a high concentration of industry and R&D expenditures may be surrounded by regions with a higher proportion of rural areas and thus relatively less likely to benefit from spillovers. In Model 2 we add the number of regions that local industry collaborates with in terms of scientific co-publications and find a significant positive effect of this measure for collaboration breadth. We continue to add measures for collaboration potential (i.e. R&D expenditures by universities in collaborating regions) as well as collaboration intensity (models three and four). The transfer channel's potential in terms of accessible distant university knowledge has a positive effect that remains significant along with the breadth of collaborations. We cannot reject the Null-hypothesis for collaboration intensity, although the negative sign of the estimated coefficient gives some support for the idea that repeated collaboration using the same contacts may use up the potential of a transfer channel. Next (Models 5, 6 and 7) we test the corresponding measures for collaborative efforts of local academia: in this case only the collaboration breadth has a significant effect on knowledge production. Both collaboration intensity and the channels potential do not appear to have an effect, indicating that the process leading to collaboration in between universities may differ to that between industry and universities. Noticeable in Model 5 and subsequent models is that the effect of local university R&D is substituted by the university's collaboration breadth. In the full Model we compare university and industry as catalysts. The effects of collaboration breadth remains significant for both industry-university as well as university-university collaborations (albeit at a marginal level of significance in case of the latter). The potential of the industry-university transfer channel's positive effect on local knowledge production remains stable throughout all models.

8.5 Discussion

We find that from the three potential mechanisms for explaining the positive effect of collaborations, the collaboration breadth emerges as the most reliable explanatory variable. Increasing the number of knowledge sources available to industry appears to strengthen their ability to innovate by recombining knowledge. This is a strong hint that the mechanism underlying the positive effect of collaborations on local knowledge production is related to the diversity of knowledge that can be absorbed by local industry. Further research is required to differentiate between the breadth of collaborations in terms of the number of partners and the diversity of knowledge provided by these partners. When considering collaborations between universities we also find a positive effect on local knowledge production. Given the effectiveness of local knowledge spillovers over informal transfer channels, it is likely that knowledge stored at local universities will spill over to local industry. If local universities collaborate successfully with other universities, the size and quality of their own knowledge stock will increase along with their ability to contribute to local knowledge production. It is also possible that universities contribute directly to knowledge production, for example in terms of academic patents. However, it appears that this form of patenting is limited in scope and more likely to result from collaboration between university and industry (Crespi et al., 2011).

Collaboration depth does not appear to be useful for explaining knowledge transfers in either channel. Even though strong ties to a collaboration partner have been shown to be useful for successful collaboration (Tomlinson, 2010), it appears that they do not offer an advantage over weak ties. This findings aligns with the interpretation of the collaboration breadth effect as caused by the diversity of accessed knowledge: if the novelty of information is the main value arising from collaboration, then repeated collaboration with the same partners does not necessarily yield a higher return.

The transfer channel's potential does seem to play a role in the case of industry-university collaborations but not so for university-university collaborations: while we find some evidence for a positive effect of the potential in the university-university context, the effect cannot be observed in the full Model. The explanation may be different motivations for collaboration for the two transfer channels: in the university-university context, collaboration between institutions is the norm. Several interfaces such as meetings at conferences or professional networks facilitate collaboration between academics.

In the context of industry-university collaboration, more planning is likely to be required for both sides. If transaction costs represent a barrier for collaboration, it is likely that universities with good reputation are better able to signal their value to potential collaboration partners (Landry and Amara, 1998).

Another perspective would be to consider the channel's potential as indicative of the value of information that can be transmitted. So for industry-university collaboration it proves beneficial if industry is not only able to access more sources of information but also if those sources are able to provide more value. Intuitively we would expect to find that valuable information is more difficult to process and that academia would have an advantage at processing such information compared to industry. That we do not find evidence for this idea may be explained by a relatively lower efficacy of universities at translating absorbed information into applied knowledge. If universities are limited in their ability to produce knowledge and if the knowledge they can transmit to industry is limited by available local spillover channels, it is likely that only strong effects remain measurable.

When comparing the two transfer channels, the effect size is only somewhat helpful as it does not include information on the relative frequency of use for both channels. Collaborations within the academic network are much more likely than collaborations between industry and academia. As both effects are on the same order of scale, it is likely that collaborations between industry and academia are relatively more effective than a similar collaboration within academia.

8.6 Conclusion

We confirm and extend previous findings by controlling for spatial effects and explicitly modeling spillovers over longer distances using formal collaborations. We find that scientific co-publications are a viable transfer channel for academic knowledge. We add to existing literature by disentangling the positive effect of collaborations into breadth, depth and potential, finding that collaboration breadth is the strongest indicator of useful collaborations, followed by the transfer channel's potential. We do not find effects for collaborations' intensity, confirming the view that repeated use of a transfer channel tends to exploit its value.

These findings suggest that policy which provides incentives to research collaborations with strong breadth and potential may yield be more efficient. For example, research grants for pairs of universities and companies that have not cooperated in the past should be preferred over established partnerships. When selecting among new permutations those with high potential, i.e. larger knowledge stocks, should be preferred, as the size of the knowledge stocks of the collaboration partners is likely to increase the odds that the partners can contribute relevant information and thus generate new knowledge.

Our study is limited by the sparsity of data that results from limiting the dataset to NUTS3 regions of Germany. Limiting the number of observations to 412 regions, of which many are rural regions that do not contribute significantly to industrial knowledge production, sets limits on the complexity of models that can be used. Several studies have investigated collaboration at the national or state (i.e. NUTS2) level which enables comparison of increasingly important international, long-distance collaboration but omits effects that are only visible over shorter ranges. Ideally further research would attempt to investigate data at the regional level over several countries. Furthermore, additional research is required on the details of collaboration breadth in the context of university-industry collaboration. Diversity of accessed knowledge is a convincing explanation for the observed effects but it is not testable given our dataset.

Chapter 9

Geographic distance and knowledge transfer

Having discussed some general characteristics of joint research in the last chapter, this chapter presents the findings of a second study[1] on the same transfer channel. In this case the focus is on the distance between actors and its effect on the transfer of knowledge. Positive effects of collaboration between academia and industry on regional knowledge production have been shown in a number of recent studies. We extend this stream of research by taking a closer look at the mechanisms underlying these effects: we consider knowledge distance and diversity as new measures for the knowledge production framework. We show that industrial regional knowledge production benefits more from collaborations with universities when academia provides access to diverse knowledge. While our findings regarding the direct effect of distance are mixed, we show that distance is positively moderated by existing regional knowledge stocks. The effect of diverse knowledge is negatively moderated by regional basic knowledge stocks. Our findings imply that characteristics of collaboration partners and receiving actor determine the success of collaboration between industry and academia.

After an introduction to the topic this chapter discusses relevant prior literature, introduces data collection and analytical methods and presents the study's results. The results and their implications for future research and policy are then discussed.

[1] This study is the result of cooperative work with Hannes Lampe (Hamburg University of Technology).

9.1 Introduction

Due to increasing global competition and significant differences in labor costs, innovation increasingly emerges as a relevant advantage of national markets. With a constant increase in the maturity of various technological domains, the space for innovation within one field is often exhausted. Hence successful innovation then requires considerable effort to merge formerly disjoint fields to expand the space for recombination. Both strategies favor division of labor in teams that span organizational boundaries (Ahuja, 2000; Morgan and Cooke, 1998).

One branch of research investigates innovation at national or regional levels using the concept of knowledge production functions (e.g. Griliches, 1990; Patel and Pavitt, 1994).

This framework has been extended with measures on collaborative efforts (Ponds, Oort, and Frenken, 2010) and analyses of characteristics of collaboration networks (Guan, Zhang, and Yan, 2015). Collaboration, particularly between firms and universities, is an increasingly important topic for industrial innovation (Powell, Koput, and Smith-Doerr, 1996). As academia produces large quantities of new knowledge, it is a potential driver of industrial knowledge production, assuming that knowledge stocks can be accessed and converted into commercial applications. Knowledge spillovers are assumed to flow from universities to firms and thus increase firms' innovative output (Maietta, 2015; D'Este and Iammarino, 2010; D'Este and Patel, 2007). In some technological domains, such as nanotechnology, advancement of basic scientific knowledge is of particular importance (Grimpe and Patuelli, 2011), hence these domains are likely to benefit from university-industry collaboration. The importance of innovation for knowledge-based economies, in combination with the innovative potential present in universities, have prompted governments to incentivize university-industry collaboration by passing legislation and funding collaborative projects (Link and Siegel, 2005) These policy measures have positive effects (Link and Scott, 2005).

However, existing research, which focusses on the structural aspects of collaboration networks, is insufficient to explain how actors gain benefits from networks (Ter Wal et al., 2016; Rodan and Galunic, 2004).

Additional information on the collaborative efforts is required to explain the mechanisms underlying successful collaborative efforts. Text-based measures have been developed to investigate the effects of diversity and distance on the economic value of innovations (Kaplan and Vakili, 2015). We create similar measures for the diversity and distance between knowledge agglomerations in order to explain the positive effect of collaboration on regional knowledge production. Our results indicate that important policy implications can be derived by extending the research on regional knowledge production with measures for network structure and knowledge content: instead of general policy measures that broadly incentivize research projects, more targeted approaches may be applied. Eventually characteristics of project partners, such as their specialization or their distance between one another, may prove useful in predicting the impact and increasing the effectiveness of the collaborative project. Furthermore, a detailed and mostly automated analysis of potential collaboration partners may enable firms to select suitable members of academia without the need of governmental incentives.

This paper, as a first step towards such policy measures, suggests new measures for the regional knowledge production framework: the diversity of the knowledge in collaborative projects, as well as the knowledge distance between the actors within a project. We find that these measures prove helpful in determining with which universities firms should collaborate to boost regional knowledge production: should they seek specialists or look to establish diverse teams? Is it important that their expertise is similar to the expertise of partners or is some distance required for effective collaboration? Answering these questions contributes to existing research on regional knowledge production in three ways: (1) we show that knowledge diversity explains the beneficial effect of collaborations on knowledge production previous literature has found, (2) we test whether distance has a direct effect on efficacy of collaborations but only find weak negative effects, (3) we find that knowledge stocks tend to counteract the direct effects of diversity and distance when they are included as interacting variable.

The remainder of this chapter is divided into four sections. The second sub-chapter reviews the literature on university-firm collaboration and collaboration network effects on knowledge production and derives our hypotheses. The third sub-chapter describes our dataset, the construction of new measures and methods chosen to test our hypotheses. Sub-chapter four presents and discusses the empirical evidence. Finally, conclusions are given and specific contributions of this chapter are discussed.

9.2 Theory and hypotheses

According to Schumpeter (1934) innovation is often the result of recombining existing knowledge (Cantner, Joel, and Schmidt, 2011). In prior literature R&D expenditures have been a useful measure for the creation of knowledge. Both internal, as well as external R&D, are relevant to innovation performance (Laursen and Salter, 2014; Freel, 2003).

Since collaboration is a means of accessing external knowledge, it is seen as a means of expanding the available re-combinatory space. A popular framework that allows to study the relation of innovation outputs and collaboration is the knowledge production function (Griliches, 1990 Patel and Pavitt, 1994). The KPF models a region's output in terms of patent counts as a function of input parameters specific to the region (Acs, Anselin, and Varga, 2002). The KPF framework can be extended to study the influence of external factors on innovation output such as knowledge spillovers between actors within a region (Grimpe and Patuelli, 2011) or of input factors from other regions (Ponds, Oort, and Frenken, 2010). This framework has also been extended with measures for collaborative efforts (Ponds, Oort, and Frenken, 2010), showing that firm-university collaboration can have an impact on regional knowledge outputs. Recently, the focus of attention has shifted to the impact of collaboration networks on knowledge production (Hölzl and Janger, 2014; Guan, Zhang, and Yan, 2015). In these studies, measures have been derived from collaboration networks to analyze the effect of collaboration on innovation output at the regional level. Next to network positions and multilevel effects of network positions (Guan, Zhang, and Yan, 2015) on knowledge production, the effects on knowledge production efficiency (Guan et al., 2016) have been evaluated. Previous research mostly analyzes collaboration in a broader setting, whereas we focus on the context of university-firm collaborations, where existing research suggests that network structure is also relevant (Casper, 2013). However, a collaboration network's structural perspective alone is insufficient to explain how actors gain benefits from networks (Ter Wal et al., 2016).

According to Ter Wal et al. (2016), the diversity of knowledge in a collaboration team determines the team's ability to innovate. The heterogeneity of a network's knowledge determines the diversity of knowledge that each network actor can access, thereby affecting their ability to recombine knowledge (Phelps, Heidl, and Wadhwa, 2012;Rodan and Galunic, 2004;Ahuja, 2000; Burt, 1992) and to determine whether information is redundant or non-redundant based on the network structure. Our article is, to our knowledge, the first to apply these concepts to the context of university-firm collaboration in a knowledge production environment.

Accessing diverse knowledge only provides benefits if actors can process the obtained information. Prior research finds that accessing similar knowledge tends to be easier (Cohen and Levinthal, 1990), so measures for the diversity of available knowledge and the distances between the knowledge stocks of actors within a network need to be considered. Our knowledge measures for knowledge distance and diversity have so far not been used in the context of regional knowledge production functions and the specific context of university-firm collaborations. We derive such measures and show that they are useful in explaining the beneficial effects of collaboration on knowledge production in the context of nanotechnology. We start by providing a brief overview of knowledge diversity and knowledge distance as mechanisms in existing research and derive hypothesis for the effect of these measures in the context of regional KPFs and how these measures are likely to interact with other variables of interest. Furthermore, we postulate hypotheses concerning the moderating role of regional knowledge stocks on the relationship between knowledge diversity / distance and regional knowledge production.

9.2.1 Knowledge diversity

We define knowledge diversity as the heterogeneity of the knowledge that can be accessed via university-firm collaboration networks. Specifically, we look at collaboration between industry in the focal region and academia in distant regions. The heterogeneity of knowledge between academic collaboration partners (excluding the focal region) determines how broad the range of accessible knowledge for the focal region is.

The access to heterogeneous knowledge in the context of the knowledge production framework should benefit a region's knowledge production process in two ways. First, the number of available discrete pieces of information increases the number of possible combinations, hence accessing more knowledge increases the chance of innovating. Since many industries have reached a certain degree of maturity, combinations that are obvious to domain experts have likely been tested, so the introduction of knowledge that differs to some degree may open up new possibilities. This view is based on prior literature which has found creative achievements to be the result of the connection of two or more disparate ideas or concepts within an individual's mind (Amabile, 1996; Fiol, 1995; Zaleznick, 1985; Wang et al., 2014). Furthermore, the Schumpeterian theory of innovation suggests that innovation consists of recombination of conceptual materials that already exist (Nelson and Winter, 2009; Shane, 2000).

In network theory, Burt (1992) makes a similar point by showing that contacts which are strongly connected are more likely to provide redundant information. Accordingly heterogeneous actors are more likely to provide non-redundant information.

Academia has been identified as source of new knowledge that can potentially be absorbed and converted into industrial knowledge output (i.e. commercial applications as indicated by patent filings). Therefore, collaboration between industry and universities may lead to particularly useful recombination as industry's abilities in commercializing knowledge and academia's specialization in creating new knowledge may complement each other.

However, an increase in the diversity of available knowledge may overburden the receiving actor's capacity to identify valuable knowledge. Hence, the costs associated with filtering potential sources of knowledge may deter actors from establishing a network with diverse partners. The necessity of engaging in collaborative efforts in order to innovate may mitigate this barrier to some extent.

In summary, we expect that a region's knowledge output increases when it has access to diverse knowledge through collaborations with scientific institutions located in other regions.

> *Hypothesis 1: Regional knowledge production is positively affected by access to diverse knowledge via collaboration with academia in distant regions.*

The effect of geographic, organizational and cultural distance on knowledge transfer has been investigated in several studies (Allen, 1977; Cummings and Teng, 2003; Hofstede, 1984; Simonin, 1999; Polanyi, 1966). Generally, increasing the distance between two actors decreases the chance of successful knowledge transfer. As a result one can observe clustering of specialized actors and a lower frequency of interaction between actors from different organizational or cultural contexts (Almeida, 1996; Uzzi, 1996). One additional type of distance that plays an important role is the distance between the knowledge stocks of two actors. This distance is referred to as knowledge distance and describes the dissimilarity of the knowledge of two actors. The literature on strategic alliances finds that such distance can interfere with learning from collaboration partners (Hamel, 1991). In the context of university-firm collaboration, greater technological proximity appears to facilitate the transfer of knowledge (Woerter, 2012).

However, it has also been recognized that knowledge distance does not only present a barrier but also an opportunity for knowledge transfer. As the adoption of knowledge grows more difficult with distance, the degree of novelty of the potentially transferable knowledge increases. Hence, the effect of knowledge distance on innovation performance is curvilinear (Nooteboom et al., 2007).

We are interested in the distance between the knowledge that can be attributed to a region's industrial sector and the knowledge of collaboration partners from academia in distant regions. In this context, access to distant knowledge should improve a region's ability to innovate as distant knowledge is more likely to be novel. However, two types of distance present a barrier to the adoption of increasingly distant knowledge. The organizational boundary between industry and academia may complicate the transfer of knowledge. The degree to which business practices, institutional heritage, and organizational culture differ between two organizations negatively impacts the odds of successful knowledge transfer (Choi and Lee, 1997; Ponds, Oort, and Frenken, 2007; Simonin, 1999).

The knowledge distance, in accordance to findings in literature on inter-firm collaboration, may increase the difficulty in absorbing knowledge (Cohen and Levinthal, 1990). If knowledge is too distant, it loses its relevance to the receiving actor and transfer becomes too costly.

In the context of knowledge transfer between academia and industry, we can expect a certain minimum distance resulting from the cultural and thematic differences between the two domains. Given a minimum distance between actors, we are thus more likely to observe the negative effect of the curvilinear relation described by Nooteboom et al. (2007) when knowledge distance between two actors increases.

> *Hypothesis 2: The distance of industry's knowledge in the focal region to the knowledge of collaborating universities negatively affects the focal region's knowledge production.*

9.2.2 The moderating role of prior knowledge

Absorptive capacity, a term introduced by Cohen and Levinthal (1990), can be understood as an actor's ability to assimilate and replicate new knowledge from external sources. Zahra and George (2002), refer to certain characteristics that are important for companies that intend to use external knowledge:

> *Acquisition refers to a firm's capability to identify and acquire externally generated knowledge that is critical to its operations.*

This implies that organizations exposed to the same external knowledge might benefit differently from it due to their dissimilar absorptive capacity. Although the acquisition of new knowledge and the underlying absorptive capacity influencing the acquisition ability is mainly analyzed on the firm level, it is also a widely accepted measure at the regional level (Miguélez and Moreno, 2015; Mukherji and Silberman, 2013; Tunzelmann, 2009).

Absorptive capacity has also been found to moderate the relation between innovation performance and network position (Tsai, 2001), so it is likely to also moderate the relation between knowledge production and university-industry collaboration. Absorptive capacity requires learning capabilities and well developed problem-solving skills. Learning capabilities are defined as the capacity to assimilate knowledge in the form of imitation. Problem-solving skills are required to create new knowledge, and thus innovation (Kim, 1998). The concept of absorptive capacity is often measured via prior knowledge stocks (Zahra and George, 2002).

For example, Griliches (1990) shows that the cumulated general knowledge stock has a positive impact on the production of new knowledge. This finding is in line with the basic assumption of the cumulative advantage model of knowledge production (Stigler, 1983; Machlup, 1984) namely “having one idea increases the likelihood of having another” (Zucker et al., 2007). Furthermore, Zucker et al. (2007) show that this argumentation also holds in the context of nanotechnology.

The validity of the cumulative advantage model is sometimes questioned. Knowledge stocks are sometimes regarded as irrelevant or even associated with a negative effect on the production of new knowledge (e.g. Kuhn, 2012): the negative effect of prior knowledge may be explained by a fixation on the status quo and lower acceptance for ideas that deviate. However, Kuhn mentions this in the context of paradigm shifts in basic science, which is different from the context of adaptation of knowledge by industry. A similar fixation on the status quo and its negative effect on R&D performance is noted in literature on the not-invented-here syndrome (Katz and Allen, 1982).

Positive effects of prior knowledge on knowledge production are a common finding in the literature on regional knowledge production. Zucker et al. (2007) show a positive effect between the size of prior knowledge stocks in all fields of science and the rate of production of new knowledge. We build upon their research by including prior knowledge stocks of various technological fields in our analysis.

We argue that knowledge stocks can be a double-edged sword in terms of their effect on knowledge production. On the one hand, in line with previous research, a region's knowledge base can have a positive effect on nanotechnology patent output: prior experience in a certain knowledge domain leads to learning effects that enable actors to more easily absorb new information (Zucker et al., 2007).

However, when combined with measures for the diversity and distance to the knowledge that is to be adopted, the effect of knowledge stocks may differ to the common findings. Access to highly diverse knowledge may be less beneficial in a region with large knowledge stocks, as such regions may not require collaboration over longer distances to connect to new sources of information: such regions may either already possess the knowledge that exists in other regions or have high confidence in their ability to generate the desired knowledge without outside help. In this case, local collaboration or informal connections may be relatively more important. Respectively, a region with low knowledge stocks is likely to benefit more from collaboration with diverse partners as it is less likely that the region already has access to diverse knowledge. We thus argue that regional basic knowledge stocks negatively affect a region's ability to exploit diverse external knowledge from collaborating with universities in other regions.

> *Hypothesis 3a: Regional knowledge stocks negatively moderate the effect of access to diverse knowledge through collaboration on the focal region's knowledge production.*

The regional knowledge stock may not only affect the relationship between knowledge heterogeneity of collaborating universities and the focal region's industry. It may also be relevant for the effect of knowledge distance between the focal region's industry and collaborating universities in other regions. Distant knowledge may be easier to process if the receiving region has large knowledge stocks and thus experience with decoding specialized information efficiently. Hence, regions with high knowledge stocks are likely to be better at recognizing the value of distant knowledge. They can use their experience to mitigate the adoption barriers that knowledge distance usually imply:

> *Hypothesis 3b: Regional knowledge stocks positively moderate the effect of distant knowledge accessible through collaboration on the focal region's knowledge production.*

Overall, we expect the two characteristics of knowledge (diversity and distance) that industry can access via collaboration with academia to have direct effects on regional knowledge production in nanotechnology. Furthermore, we expect that regional knowledge stocks, in form of patents, will have a moderating effect on the relationship between our two focal variables (knowledge diversity and distance) and regional knowledge production in the domain of nanotechnology. Figure 9.1 depicts our proposed theoretical model.

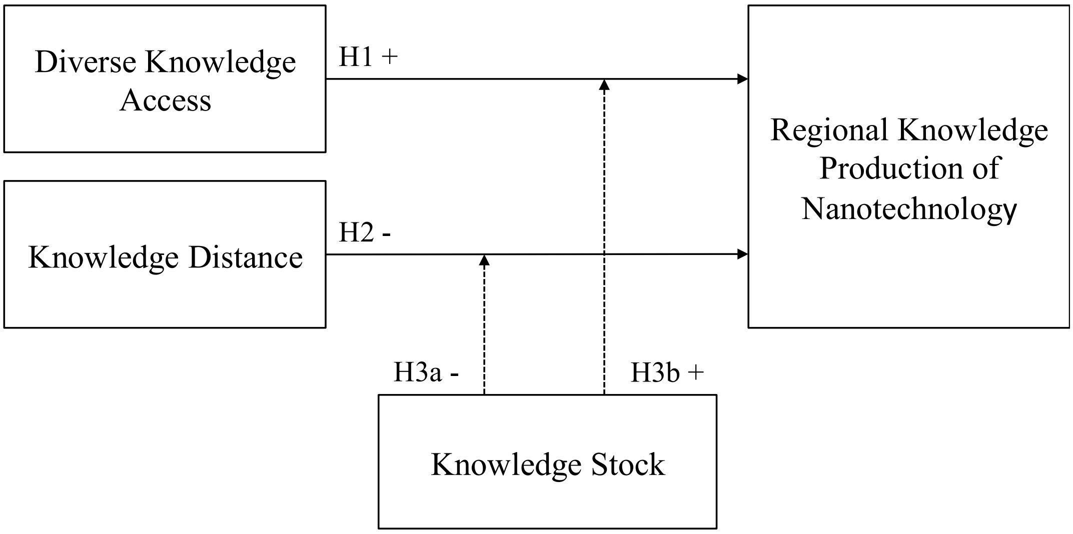


FIGURE 9.1: Conceptual Model

9.3 Methods and data

In the knowledge production framework, knowledge outputs are typically measured using patent counts (Jaffe, Trajtenberg, and Henderson, 1992) aggregated at the level of analysis. The basic KPF specification comprises firm and university R&D expenditures as inputs. As the center of analysis of this article are firm-university collaborations, we follow Ponds et al. (Ponds, Oort, and Frenken, 2007) in using scientific co-publications as indicator of formal collaboration spanning several regions (in our case the 412 NUTS level 3 regions of Germany). We follow Ponds et al. (Ponds, Oort, and Frenken, 2007) and group our data into three time periods. The first time period, t-2, is used to collect data on collaboration in a time-frame preceding the patent filing (2004-2007). The second time-frame, t-1, includes all other input and control variables except for the network-based variables (2007). Finally we collected data for our dependent variable, patent applications in nanotechnology, for the years 2008 - 2010.

9.3.1 Estimation of count data

Our output is measured by the amount of patent applications in the field of nanotechnology over a certain period of time. Thus, the dependent variable cannot assume values smaller than 0 and is initially treated as integer value. Furthermore, the distribution of our dependent variable is skewed, suggesting that Poisson regression may be appropriate. However, the dependent variable is also over-dispersed, consequently the negative binomial estimation framework was chosen for our analysis. In this case, a generalized linear model (GLM) is adopted, using the logarithm as a link function. Therefore, our model takes the form:

$$P_{i,t} = e^{\alpha \ln x_{i,t-1} + \beta \ln z_{i,t-1} + \gamma \ln a_{i,t-1} + \epsilon} \tag{9.1}$$

As we are analyzing all NUTS 3 regions in Germany, an excessive number of zero counts occurs (for example in rural regions). We correct for these structural zeroes with a zero-inflated negative binomial model (Frenken et al., 2009; Chessa et al., 2013; Hoekman, Frenken, and Van Oort, 2009). Therefore, the estimation process consists of two parts. First, a logit regression is used to explain the probability of no nanotechnology patents being filed in a region. The second part of the model is the central negative binomial model which estimates the observed patent counts.

9.3.2 Data

We use nanotechnology patent applications as dependent variable to measure knowledge production. We searched the European Patent Office's statistical database (PATSTAT, 2012 edition) for nanotechnology patents filed in Germany. Following Arora et al. (Arora et al., 2013) we use keyword-based searches on abstracts and titles, as well as corresponding IPC (international patenting classification) and ECLA (European classification system) classes to identify applications for the years 2008 to 2010. To avoid counting one invention several times, we reduce the applications to patent families, which aggregate patent applications with identical claims (i.e. modifications of applications or applications at different patent offices). To geocode the resulting set, we matched patents to the REGPAT database and retrieved additional missing location information from Espacenet, the German patent office's online database. In a final step duplicate entries for individuals and organizations were removed, which further increased the coverage of location data. Addresses were then matched to their corresponding NUTS codes (version of 2010), which allowed us to fractionally attribute patent counts to 412 German NUTS3 regions (i.e. to add a share of each patent application to one region according to the corresponding number of applicants/inventors from that region).[1]

As an indicator of collaboration between industry and academia, we use scientific co-publications in the timeframe 2004-2007. We retrieved articles on nanotechnology from the Web of Science using a

[1] As mentioned above, the distribution of patent applications is skewed and over-dispersed, which suggests that a negative binomial model is appropriate. Due to fractional counting, the dependent variable is continuous and cannot immediately be applied in a negative binomial model. We round the value to full integers. Rounding introduces some error into the variable. However, the only alternative would be to attribute a full patent to each region when a patent application has more than one applicant (which is often the case). Hence we have to choose between a rounding error and inaccuracies in the relation of patent counts between regions that result from attributing whole patent counts. Since the latter error seems more likely to distort our data, we went ahead with rounding the fractional variable.

keyword-based search resulting in 358 university-firm publications. In the resulting dataset author affiliations to universities and companies were used to geocode articles and to identify collaboration spanning 198 German NUTS3 regions. We follow prior research in using weight matrices to model knowledge transfers in the KPF (Ponds, Oort, and Frenken, 2010): in this case we assume that knowledge flows from the university to firms, enabling additional knowledge production in the firm's region. Hence our weight matrix is derived from a directed network and is asymmetric. As with the patent counts, publications are fractionally attributed to regions according to the location of co-authors. The weights arise from aggregating the fractional counts of co-publications between two regions. The collaboration networks are then used to derive more complex measures for knowledge diversity and knowledge distance.

Our first independent variable is collaboration network knowledge diversity (variable "knowledge diversity") for firm-university collaborations. For a focal region we define diversity as the heterogeneity of the knowledge accessible through the collaboration network, i.e. focal region's industry collaborates with academia in several other regions. We base the variable on a similar construct introduced by Ter Wal et al. (2016): the academic knowledge stocks of these distant regions are used to build the diversity measure. Using a topic model (Blei et al., 2003) we transform nanotechnology publications from each region into a vector of length 250, i.e. one publication abstract is represented by 250 topic vectors where each topic vector is a probability distribution over semantically related words. The vector representation of publications is averaged for a representation of the region's scientific knowledge which can be compared to other regions using the cosine similarity metric, i.e. the cosine of the angle between two region vectors (Tan, Steinbach, and Kumar, 2006). Knowledge diversity is then calculated as the average cosine distance between the scientific knowledge of one region and its scientific collaboration partners. The cosine-distance is defined as:

$$1 - similarity = 1 - \cos\theta = 1 - \frac{AB}{||A||\,||B||} \tag{9.2}$$

$$= 1 - \frac{\sum_{i=1}^{N} A_i B_i}{\sqrt{\sum_{i=1}^{N} A_i^2}\sqrt{\sum_{i=1}^{N} B_i^2}} \tag{9.3}$$

The second independent variable is collaboration network's knowledge distance. We measure the distance between the industrial knowledge stocks (patents) of the focal region and the academic knowledge stocks of the distant collaborating regions and take the average value for a firm's region. We again use a topic model to transform nanotechnology patents from the focal firm's region and nanotechnology publications from connected distant university regions into vectors. Individual vectors are aggregated at the region level and the regions compared using the cosine similarity measure.

We include regional prior knowledge (knowledge stock) in terms of patent applications in the year 2007 in our model for two reasons: first, as previous research has shown that regional knowledge stock has an effect on regional knowledge production (Roper and Hewitt-Dundas, 2015; Zucker et al., 2007) and second, to analyze knowledge stock's moderating effect on our focal variables.

To control for the scope of collaboration between industry and academia we include two variables in our regressions: the variable "amount of collaborations" controls for regions with industry that attract more collaboration. The number of collaboration partners controls for the breadth of knowledge that can be obtained through collaboration.

We include input variables to the production function that are commonly used in prior research, such as firm R&D expenditures and university R&D expenditures. These data are obtained from the 'Stifterverband für Deutsche Wissenschaft' for the year 2007.

We control for spatial spillovers, i.e. knowledge transfers over short distances that have been found to significantly impact knowledge production analyzed at the regional level (D'Este, Guy, and Iammarino, 2013; Leten, Landoni, and Van Looy, 2014; Grimpe and Patuelli, 2011) using spatial weight matrices (Ponds, Oort, and Frenken, 2010). The matrix contains weights based on the distance of regions to a focal region and allows to construct measures for spatial spillovers that consist of summed, weighted R&D expenditures in adjacent regions. As NUTS 3 regions differ in size and shape, we use an interpoint distance weight matrix based on a k-nearest neighbor metric instead of a simple adjacency matrix

(De Smith, Goodchild, and Longley, 2007).[2]

Further controls were built based on data from the German federal office of statistics to capture the economic structure and size of regions. The manufacturing/ service ratio of employees gives the ratio of employees in manufacturing industries in relation to those working in service. The workforce with tertiary education in percentage gives an indication of the relation of the science-based workforce. To control for firm sizes in a region, we incorporate the average firm size of a region in terms of firms' employees. Furthermore, we use a specialization index to account for regional industrial specialization. We use a Herfindahl index built from the number of patent applications in one of 35 fields of technology (classification scheme based on Jaffe and Trajtenberg, (2002) relative to the total number of filed patents.

9.4 Results

The descriptive statistics are presented in Table 9.1. Supporting prior research in the domain of nanotechnology KPF, we see that firm R&D expenditures are strongly correlated with our dependent variable. University R&D is slightly less correlated with our dependent variable, patents in nanotechnology. In accordance to findings by Zucker et al. (2007) a region's prior knowledge base is also strongly correlated with regional nanotechnology patents. The highest variance inflation factor for the independent variables (excluding interaction effects) is 3.03 and thus indicates that multicollinearity among variables is not a concern (Chatterjee and Price, 1991).

[2] Here we use row standardized weight matrices, implicitly assuming the presence of limited absorptive capacity (Ponds, Oort, and Frenken, 2010). This implies that the R&D expenditures of neighboring regions enter the focal regions knowledge production as the weighted average of R&D expenditures of neighbors. Thus, an increase in neighbors for region i results in a decrease of each neighbors' (j) spillovers towards region i.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
(1) Nano-patents	1													
(2) Knowledge diversity	0,397	1												
(3) Knowledge distance	-0,41	-0,562	1											
(4) Knowledge stock	0,783	0,4	-0,476	1										
(5) Amount of collaborations	0,713	0,327	-0,358	0,534	1									
(6) Collaboration partners	0,753	0,523	-0,519	0,595	0,732	1								
(7) Firm R&D (ln)	0,436	0,332	-0,47	0,572	0,314	0,385	1							
(8) University R&D (ln)	0,372	0,3	-0,358	0,286	0,423	0,36	0,368	1						
(9) Wdistance Firm R&D (ln)	0,103	0,09	-0,258	0,302	-0,024	0,077	0,414	0,003	1					
(10) Wdistance Univ. R&D (ln)	0,04	0,059	-0,141	0,135	-0,028	0,044	0,205	-0,1	0,6	1				
(11) Manu./serv. ratio	-0,11	-0,089	0,07	0,038	-0,18	-0,069	0,213	-0,298	0,176	0,107	1			
(12) Share of tert. educ.	0,318	0,212	-0,393	0,352	0,224	0,216	0,242	0,207	0,228	0,149	-0,251	1		
(13) Average firm size	0,095	0,112	-0,093	-0,023	0,183	0,159	0,237	0,479	-0,116	-0,138	0,077	-0,307	1	
(14) Specialization index	-0,371	-0,266	0,441	-0,651	-0,21	-0,321	-0,59	-0,183	-0,583	-0,285	-0,26	-0,286	0,067	1
Mean	1,386	0,034	0.806	71,923	4,556	1,791	10,084	3,142	11,091	9,733	0,637	0,061	10,586	0,617
SD	3,592	0,077	0.286	105,958	18,4	4,943	1,866	4,775	1,282	1,964	0,363	0,025	2,675	0,299
Min	0	0	0.24	0	0	0	0	0	0	0	0,054	0,02	5,394	0,043
Max	47	0,512	1	1074,401	277	56	15,236	14,263	13,299	12,237	2,636	0,156	30,881	0,994

TABLE 9.1: Descriptive Statistics

Table 9.2 summarizes the results from the zero inflated negative binomial regressions. The natural logarithm of the dispersion parameter is significantly different from zero in all model specifications, indicating that a negative binomial is preferable over a Poisson model. The z –value of the Vuong test (Vuong, 1989) is significant in all models and thus shows that the zero inflated specification is to be preferred over the standard negative binomial one.

Model	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Diverse knowledge		1.447***		1.257**	2.437***	0.801	2.439***
(DKA)		(0.521)		(0.543)	(0.803)	(0.560)	(0.802)
Knowledge distance			-0.454**	-0.307	-0.176	-1.257***	-1.331***
(KD)			(0.228)	(0.238)	(0.248)	(0.372)	(0.352)
Stock - divers					-0.006*		-0.009***
					(0.003)		(0.003)
Stock - distance						0.009***	0.011***
						(0.003)	(0.002)
Controls							
Knowledge stock	0.001***	0.001***	0.001***	0.001***	0.002***	0.008***	0.010***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.002)	(0.002)
Amount of	-0.006**	-0.005**	-0.005**	-0.004*	-0.003	-0.005**	-0.004
collaborations	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Collaboration	0.044***	0.040***	0.040***	0.037***	0.037***	0.039***	0.039***
partners	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Firm R&D (ln)	0.176***	0.144**	0.142**	0.128**	0.144**	0.118**	0.145***
	(0.061)	(0.059)	(0.058)	(0.057)	(0.058)	(0.057)	(0.055)
University R&D (ln)	0.017	0.016	0.017	0.016	0.014	0.017	0.016
	(0.016)	(0.016)	(0.016)	(0.015)	(0.015)	(0.015)	(0.015)
$W_{distance}$ Firm R&D	-0.189**	-0.148*	-0.186**	-0.150*	-0.153*	-0.120	-0.118
(ln)	(0.084)	(0.083)	(0.082)	(0.082)	(0.082)	(0.082)	(0.081)
$W_{distance}$ Univ. R&D	0.114**	0.108*	0.109**	0.106*	0.106*	0.072	0.059
(ln)	(0.055)	(0.056)	(0.055)	(0.056)	(0.056)	(0.055)	(0.054)
Manu./serv. ratio	-0.322*	-0.213	-0.262	-0.189	-0.268	-0.228	-0.382**
	(0.194)	(0.197)	(0.197)	(0.198)	(0.201)	(0.198)	(0.189)
Share of tert. educ.	13.182***	13.004***	11.520***	11.943***	11.238***	11.461***	10.170***
	(2.227)	(2.177)	(2.318)	(2.294)	(2.309)	(2.274)	(2.217)
Average firm size	0.089***	0.089***	0.086***	0.087***	0.078***	0.079***	0.059**
	(0.029)	(0.028)	(0.029)	(0.028)	(0.029)	(0.028)	(0.028)
Specialization index	-0.766**	-0.699**	-0.727**	-0.672**	-0.668**	-0.267	-0.170
	(0.309)	(0.306)	(0.301)	(0.301)	(0.298)	(0.317)	(0.312)
Constant	-2.106*	-2.362**	-1.806	-2.172*	-2.186*	-2.700**	-2.794***
	(1.232)	(1.183)	(1.150)	(1.142)	(1.139)	(1.148)	(1.071)
Zero-inflated part							
Knowledge stock	-0.143***	-0.141***	-0.144***	-0.142***	-0.143***	-0.147***	-0.223*
	(0.044)	(0.045)	(0.047)	(0.046)	(0.048)	(0.055)	(0.130)
University R&D (ln)	-0.277*	-0.264*	-0.236	-0.242	-0.256	-0.220	-0.007
	(0.148)	(0.159)	(0.151)	(0.156)	(0.166)	(0.207)	(0.290)
Firm R&D (ln)	0.260	0.146	0.085	0.068	0.103	0.039	0.417
	(0.465)	(0.439)	(0.384)	(0.382)	(0.413)	(0.491)	(0.943)
Manu./serv. ratio	0.977	1.335	1.507	1.605	1.509	1.895	0.177
	(1.450)	(1.532)	(1.490)	(1.502)	(1.557)	(1.916)	(2.225)
Share of tert. educ.	-10.502	-15.978	-23.315	-23.079	-25.404	-38.079	-153.585
	(26.513)	(29.700)	(28.557)	(29.590)	(32.163)	(43.936)	(152.608)
Average firm size	0.454*	0.449*	0.434*	0.439*	0.456*	0.471	0.195
	(0.246)	(0.255)	(0.252)	(0.258)	(0.272)	(0.333)	(0.416)
Specialization index	-3.091	-2.915	-2.584	-2.628	-2.552	-1.697	5.214
	(2.692)	(2.781)	(2.751)	(2.802)	(2.849)	(3.585)	(7.110)
Constant	-1.170	-0.249	0.474	0.485	0.056	-0.157	0.167
	(5.822)	(5.715)	(5.186)	(5.194)	(5.528)	(6.666)	(14.747)
Fit statistics							
Dispersion parameter	-2.747***	-2.853***	-2.671***	-2.813***	-2.900***	-2.868***	-3.084***
(ln)	(0.651)	(0.692)	(0.605)	(0.667)	(0.714)	(0.674)	(0.815)
LR chi2	256***	263.26***	259.89***	264.91***	268.50***	275.97***	283.89***
Likelihood-ratio test	3.57**	3.04**	4.25**	3.37**	2.85**	3.28**	2.07*
(alpha=0)							
Vuong-satistics	3.71***	3.67***	3.63***	3.60***	3.42***	2.83***	2.94***

Notes: 412 Observations (198 Non zero and 214 zero). * p<0.1; ** p<0.05; *** p<0.01

TABLE 9.2: ZINB Regresison Models

Model 1 is the base model, including all control variables. In line with previous findings (Zucker et al., 2007), regional knowledge stock has a significant positive effect on regional knowledge production of nanotechnology. Furthermore, we control for the scope of collaboration between industry and academia. We include the number of inter-regional collaborations between industry and academia. The effect is significantly negative in almost all models. We then include the number of regions that the focal region collaborates with as measure for the breadth of academic knowledge that is made available to industry in the focal region. The respective effect is highly significant and positive in all model specifications. Apparently the number of collaboration partners is a more robust measure for successful collaboration compared to the total number of collaborations.

The KPF's main input, namely firm R&D expenditures, is positive and significant over all models. Local university R&D is not significant in our analysis, contrary to previous research (Grimpe and Patuelli, 2011). R&D expenditures from nearby regions show significant effects with opposite signs for university and industry R&D: spatially lagged firm R&D expenditures negatively affect regional knowledge production in nanotechnology. This is surprising but might be explained by specialization effects of regions at the cost of nearby regions. University R&D expenditures from nearby regions positively affect regional knowledge, we can confirm findings from previous literature on spatial spillovers (Audretsch and Feldman, 1996, Grimpe and Patuelli, 2011).

The ratio of manufacturing to service employees is only significant in some models. The share of tertiary educated people in the total workforce as well as the average firm size of a region have significant positive effects on regional knowledge production of nanotechnology. Regional specialization negatively affects knowledge outputs, indicating that nanotechnology benefits from broader access to knowledge. This is in line with the argument that nanotechnology is a 'General Purpose Technology' with potential for application across a variety of industrial sectors (Youtie, Iacopetta, and Graham, 2008; Grimpe and Patuelli, 2011).

Model 2 includes our knowledge diversity measure to test our first hypothesis. The effect of our variable "knowledge diversity" has a significant positive effect and thus supports Hypothesis 1: diversity of the academic knowledge that a focal region's industry can access through collaboration has a positive effect on nanotechnology patent production. Model 3 tests the effect of knowledge distance, again we find a positive and significant effect. However, when testing both measures in Model 4, only the diversity measure remains significant. Thus we find only weak evidence in support of Hypothesis 2, that a focal region's industrial knowledge production is negatively affected by the knowledge distance to collaborating universities.

In Model 5 we add a term for the interaction of knowledge diversity and existing knowledge stocks. The effect is significant and negative, lending support for Hypothesis 3a: regional knowledge stocks negatively moderate the effect of access to diverse knowledge through collaboration on focal region's knowledge production in nanotechnology. Furthermore, we find a significant positive effect for the interaction of knowledge stocks and knowledge distance in Model 6. This supports Hypothesis 3b, regional knowledge stocks positively moderate the effect of distant knowledge accessible through collaboration on focal region's knowledge production. Both effects hold when combined in Model 7.

Thus we show that industry's regional knowledge production output can increase when industry gains access to diverse scientific knowledge via firm-university collaborations. Furthermore, firms' access to distant scientific knowledge has a weak negative effect on regional knowledge production in the domain of nanotechnology. This also implies that a higher similarity (a lower distance) of accessed scientific knowledge eases its adoption by industry. When analyzing the moderating effect of regional knowledge stocks on our two focal variables, we find opposing signs. On the one hand, access to diverse scientific knowledge is preferable for firms from regions with relatively low knowledge stocks. On the other hand, for firms with relatively high regional knowledge stocks it is easier to absorb distant scientific knowledge.

9.5 Conclusion

Earlier empirical research focused on collaboration network structures and their effects on regional knowledge production. In addition to the structural perspective, additional information is required to explain the mechanisms underlying the benefits actors stand to gain from networks (Ter Wal et al., 2016). We therefore extend this research with two new measures for knowledge similarity and knowledge diversity, enabling a look at the underlying mechanisms that drive the effect of collaboration on regional knowledge production.

We find that access to diverse scientific knowledge through collaboration with academia can boost a region's output in nanotechnology patents. We also find some evidence that scientific knowledge that is not too distant to the region's industrial knowledge stocks is easier to absorb.

Zucker et al. (2007) show that the knowledge base of a region has a positive effect on regional knowledge production in the field of nanotechnology. We confirm this finding and extend this research by showing a negative moderating effect of prior regional basic knowledge on the relationship between regional knowledge production in nanotechnology and access to diverse scientific knowledge via firm-university collaboration. Furthermore, regional knowledge stocks have a positive moderating effect on access to distant knowledge via firm-university collaboration.

9.5.1 Theoretical implications

Apart from having managerial implications, our findings have implications for research on innovation processes, policy research as well as network studies. Firstly, we extend research on university-firm collaborations and its conditions. Even though a broad literature analyzes the effects of university-firm collaboration on regional knowledge production, some aspects have so far been neglected. In articles which concentrate on collaboration effects, mainly collaboration network characteristics have been analyzed so far (Guan, Zhang, and Yan, 2015). We extend this research by analyzing characteristics of network actors, particularly their knowledge characteristics. We differentiate between different types of knowledge that can be accessed through collaboration. We show a positive effect of access to diverse knowledge and a negative effect of distant knowledge. Thus a firm's regional knowledge production might benefit the most from collaborations with universities that have access to a variety of knowledge that is still largely similar to the receiving actors' knowledge. This implies a trade-off between the variety and similarity of accessed knowledge as it is likely that variety cannot be increased without also ultimately affecting the similarity of knowledge. The positive effect of knowledge similarity on knowledge production (which could also be regarded as a negative effect of distant knowledge) mirrors similar findings in the literature on knowledge distance. According to Nooteboom et al. (2007), knowledge distance has a non-linear, inverted U-shape effect on performance. That is, knowledge that is either very distant or very similar tends to have a worse effect than knowledge at an intermediate distance. Since we measure knowledge distance as distance between scientific publications and industrial patents, we can expect that even results with a relatively high similarity reflect a certain distance between these dissimilar domains. This would suggest that the distance measure we use eclipses the range of extremely similar knowledge that leads to worse results according to literature.

Secondly, we extend prior research on the effects of regional knowledge stocks on the relationship between collaboration network characteristics and industrial knowledge production. Although we confirm previous findings of regional knowledge stock positively affecting regional knowledge production of nanotechnology (Zucker et al., 2007), we also find that knowledge stocks can negatively moderate the positive basic effect of collaborating universities' regional knowledge diversity on the focal region's knowledge production in nanotechnology. This suggests that access to diverse scientific knowledge is generally advantageous, but regions with a relatively low knowledge stock benefit more from these collaborations. Furthermore, industrial regional knowledge stocks positively moderate the negative basic effect of access to distant scientific knowledge. So regions with sufficient knowledge stocks may indicate higher absorptive capacity which enables firms to successfully absorb more distant knowledge.

Overall these insights into knowledge access, and especially characteristics of the accessed knowledge may trigger additional research on how the success of collaboration between industry and academia depends on characteristics of both partners.

9.5.2 Practical implications

Our findings suggest that benefits of collaboration networks do not solely depend on the collaboration network's structural perspective. In addition, firm's access to content-related characteristics of knowledge also defines the collaboration network's value. Particularly, heterogeneous knowledge as well as knowledge distance of collaboration partners defines a collaboration networks value for knowledge production. While access to diverse scientific knowledge positively affects regional knowledge production in nanotechnology, the distance to knowledge has a negative effect. Thus, similar scientific knowledge is easier to absorb. Furthermore, regional prior knowledge base negatively moderates the effect of diversity and positively moderates the effects of distant knowledge on regional knowledge production in the field of nanotechnology.

Policy makers could use these insights to foster collaborations between firms with heterogeneous universities, i.e. universities that are characterized by high diversity in their scientific knowledge stocks, for example by considering knowledge heterogeneity when assessing universities for funding or grants or by encouraging individual scientists to educate themselves in topics and domains that are not part of a faculty's portfolio. Furthermore, firms should seek collaborations with universities from regions that provide scientific knowledge that is not too distant to avoid issues with the absorption of knowledge. Therefore, policy that supports firms in identifying suitable cooperation partners that are not co-located should enable firms to improve their knowledge production processes. Policy should take into account the (dis-)similarity of the respective knowledge stocks in order to to provide intensives to cooperative efforts that are more likely to be successful. This would likely require an extensive database of past publications and suitable machine learning algorithms to automatically compare the knowledge stocks of collaboration partners. However, initially simply requesting partners to list their specialties and domains and to provide an overview over past research activities may be sufficient for a rough manual assessment.

Findings with respect to the effect of knowledge stocks show that regions where firms have relatively lower knowledge stocks are more successful in absorbing diverse scientific knowledge, which suggests a diminishing return of collaboration with partners that offer access to diverse knowledge stocks. The underlying mechanism to this observation is likely the specialization that regions with large knowledge stocks undergo. This specialization may indicate that further advances in mature technological fields require equally specialized knowledge. However, large knowledge stocks enable firms to more successfully absorb distant knowledge, which aligns with findings from research on absorptive capacity. Further research may explore which incentives (e.g. exhibitions to generate contacts and thus collaborations between selected actors or monetary incentives in the form of funding bonuses for the 'right' collaborations) could foster overall knowledge production. Concluding, policy makers should consider collaborations of firms in regions that possess a lower knowledge stock with universities from regions that have access to diverse scientific knowledge. Firms from regions with relatively high knowledge stocks may benefit more from collaborations with universities that provide access to distant knowledge. These strategies could increase overall regional knowledge production.

9.5.3 Limitations and future research

This article has a number of limitations: the data for this chapter are sampled from the area of nanotechnology; the generalization of our conclusion to other types of knowledge, especially as regards the effect of distance and diversity, may prove an interesting field for future research. One drawback of this study is the data structure. In comparison to our cross-sectional data, panel data could enable a better estimation of time-dependent effects associated with the findings of our study. The focus of this study is German regional knowledge production of nanotechnology patents, so further research is required to generalize the findings to other countries and take into account the characteristics of collaboration spanning national borders.

Chapter 10

Conclusion

The final chapter of this thesis briefly summarizes the findings of the studies discussed earlier. Advances and limitations are discussed with regard to both theory (i.e. new findings, relative to old findings and issues with the experimental setup) and methodology (implementation of new methods, their usefulness and potential issues). Finally, recommendations for future research are given.

10.1 Advances and limitations: theory of knowledge transfers

This thesis presented research into three mechanisms for the transfer of knowledge from academia to industry: academic patents, collaborative research and broadcast search. Special attention was given to the role of various types of distances in enabling or preventing successful knowledge transfer: institutional distance, which indicates a difference in organizational culture, knowledge distance, which represents the difference between information from or within various domains and geographic distance, which is simply the spatial distance between actors. Our studies reveal that these distances affect the various transfer channels differently.

10.1.1 Institutional distance

The study on academic patenting shows that organizational culture, expressed in traditions that affect the working activities of scientists, influence the propensity of scientists to engage in efforts to commercialize their research. This type of distance may vary over time, with more recent generations of researchers being less affected by traditional norms of science and thus more open to the concept of research commercialization. The study investigated to what extent researchers exhibit an attitude towards patenting and how this attitude is influenced by the incentives that are commonly used to increase academia's patent output. Our main finding indicates that the cultural distance between traditional scientific norms and the norms that are typical in the context of intellectual knowledge protection (which is mostly associated with industry) need to be taken into account when designing effective incentives. While monetary incentives are effective, fostering a culture at faculty level that encourages patenting may complement existing incentives. While we could identify an entrepreneurial attitude in scientists, a taste for patents, that is conducive to patenting activities, we did not find the negative effects that previous research was able to demonstrate when monetary incentives are applied to intrinsically motivated individuals. However, this may be a result of the experimental setup, which confronted individuals with hypothetical scenarios and did not consider motivational effects over time as is usually the case with studies that investigate crowding-out effects.

Institutional distance also seemed to affect actors in the context of our second study: when researching the effectiveness of broadcast search as transfer channel we found that individuals with prior experience in similar crowd-sourcing initiatives were more likely to participate. The unfamiliarity with a concept represents a barrier that can be overcome with experience. In this case it is easy to see the similarity or overlap between institutional and knowledge distance: it appears that more knowledge about an unfamiliar institutional context builds trust.

In the context of our final studies, which mostly studied spatial distance, the institutional distance turned from a barrier into an enabler: we found several indications that the distance between academia and industry and the distance between individual organizations in these fields contributes to the effectiveness of collaboration.

While the theoretical background with regard to cultural and organizational distances is fairly well explored, our studies have shown that it is possible to quantify these concept to some degree and thus research its importance in various contexts. Still, since this type of distance exits mostly in the heads of individuals a precise measurement is not easy and represents a limitation to our studies. Our way of thinking about these distances is rather abstract and often we can only find indirect representations instead of real distance measures. It is possible that future research may leverage insights from the field of neuroscience to define better measures which should enable more detailed insights into the role of institutional distance in knowledge transfers.

10.1.2 Knowledge distance

Knowledge distance was the focus of our study on broadcast search, an alternative to the traditional transfer channels explored with data collected from real platforms and analyzed using state of the art text processing algorithms. Building on prior research regarding the optimal knowledge distance for the purpose of generating high performing innovations we research the participation likelihood as function of knowledge distance. We can thus describe a trade-off between the increase in novelty of solutions when more distant knowledge is applied and the lower participation likelihood for individuals who are unfamiliar with a subject. This result is relevant for the effectiveness of crowd-sourcing as transfer channel. A plausible interpretation of the results would be that with an increase in knowledge distance the number of possible combinations of individual knowledge items increases to the point where it is difficult for individuals to process. However, the larger pool of possible combinations also increases the likelihood that an individual will, by chance, identify a suitable application of own knowledge to a distant subject. Out study finds that scientists can be grouped into different types that react differently to incentives. The implication for knowledge transfer is to try to filter potential participants by type and knowledge in order to apply available incentives to a target group that may be more promising. Additional research is required to clarify what characteristics indicate to which group a researcher will belong. It is likely that findings with regard to institutional context will play a role, i.e. researchers trained in an environment that stresses traditional scientific norms may be opposed to crowdsourcing platforms similar to how they are less likely to patent. In this case a policy that rewards scientists for experimenting with innovation platforms may build some initial trust that is likely required before the monetary incentives take effect. Additionally, alternative incentives applied in the context of academic patenting, such as recognition of participation as of similar importance as publications may prove beneficial. We also found a strong role of knowledge distance in the context of research cooperations that span regional boundaries. Using similar text processing algorithms to those applied in the context of broadcast search, we were able to build measures for knowledge heterogeneity that explain the positive effects of research collaboration over longer distances.

10.1.3 Geographic distance

The role of spatial distance in knowledge transfers has been recognized by previous research. We contribute to this field by taking a closer look at characteristics of collaboration partners which explain the effects of spatial distance on knowledge production. In two studies we control for spatial effects and investigate several variables of collaboration between academia and industry identified through an analysis of publication networks. We find that the relation between the actors, i.e. the frequency with which they cooperate, the knowledge stocks they possess and the number of different actors they cooperate with affect regional knowledge production. Since the breadth, i.e. the number of partners, emerges as best explanatory variable we focus our attention in a second study on knowledge diversity and knowledge distance in an attempt to better understand regional knowledge production. We find that diversity, i.e. the knowledge heterogeneity of collaboration partners, seems to be conducive to successful cooperation. We use natural language processing techniques similar to those applied in previous studies to derive quantifiable distance measures which constitute the building blocks for our diversity measure. Our findings imply that policy makers may want use similar techniques when deciding over research grants in order to identify promising combinations of research partners. However, our study is limited by the sparsity of available data. Generating useful datasets is time-intensive even when limiting that analysis to one country. A broader coverage is currently difficult without sacrificing some of the regional resolutions; available datasets that cover more than one country typically have states instead of counties as regional units. In summary, we find that geographic distance is merely an abstract measure for more complex interactions that are best explained by investigating the characteristics of cooperating actors. More research is required to control for institutional factors that are easily neglected in a field that focuses on regional characteristics.

10.2 Advances and limitations: methodological

This thesis made use of some proven methodological tools such as discrete choice experiments, logistic regression and co-citation analysis. Starting with bibliometric analysis, methods from the field of computer science, in particular machine learning, were introduced. It could be shown that, in light of steadily increasing publication numbers, there is value in approaches that automatically process large amounts of scientific data. Certainly there is considerable room for improvement in this regard. While this thesis started out with established methods to find structures in scientific publications, there exist numerous clustering algorithms in the field of computer science that may yield more accurate results. Some of the short-comings of the established methods, such as mostly being applicable to publications that have received many citations, i.e. mostly older publications, could be mitigated by developing new keywords using word frequency-based metrics. A supervised classification method was demonstrated successfully; while the accuracy obtained still leaves room for improvement, the method showcases how collaborative efforts by scientists who may provide training data can quickly yield immensely useful data on the structure of a scientific discipline. The applications go beyond mere literature reviews, as the basic meta-science study discussed in Chapter 5 shows.

This thesis shows how machine learning algorithms can be applied in the field of economics. They enable researchers to use data that has hitherto not been exploited. As a result new variables can be defined and incorporated into regression analyses. The most prominent example in this thesis is the knowledge distance variable that is used to measure the distance between the knowledge stocks of two actors by applying topic modeling algorithms to patent and publications.

Since the measure has been used in most of the presented studies, it deserves some additional discussion, particularly on its validity and therewith its value. The accuracy of a topic model is often measured using the perplexity measure, which is related to a log likelihood relative to the number of words in a corpus. The perplexity depends on the data, the algorithm implementation (mostly due to different Markhov-Chain-Monte-Carlo methods being used in different implementations to estimate the posterior distribution of the model) and several algorithm parameters (most importantly, the number of topics). Several implementations and parameter combinations were tested to obtain reasonable perplexity estimates for the given data. However, perplexity is an imperfect measure when using a topic model in such a specific way. Even if it could be assumed that perplexity is sufficient, there are still subsequent transformations with impact on measure performance: distances are calculated using the cosine similarity metric. However, this metric is in itself sensitive to the number of topics specified for the model. While more topics are generally associated with lower (i.e. better) perplexity, a higher dimensionality also reduces the specificity of the cosine distance metric (just one of the set of problems commonly referred to as the "curse of dimensionality").

A possible solution to these issues appeared when conducting the study reported in chapter 7: as part of the discrete choice experiment respondents were asked to report their perceived distance to the shown RfP. While self-reported measures come with their own issues, these data points represented a sort of ground truth against which the knowledge distance measure could be tested. Due to the computational complexity involved in estimated topic models for large data sets, several models were estimated in parallel using distributed computing technologies. This enabled us to test a space of parameters and conduct the subsequent distance calculations in a reasonable time frame. Hence the effect of model input parameters on the correlation between self-reported and measured knowledge distance could be tested. The highest achieved correlation was 0.4.[1]

[1] These results came with a fairly high standard deviation; the algorithms are likely to output different results as they are usually started with random numbers. Fixing the initial random numbers allows a better comparison of models.

While this correlation is sufficient for our purposes, it also shows some of the limitations of the measure. Abstracts provide less information than full publications. One publication may be authored by several scientists and a scientist's knowledge is not adequately expressed by her publications only. Given better data (full publications, ideally single-authored and recent) should allow for considerable improvements in the measure's validity.

A less obvious but equally important advantage of the methodology chosen is that it builds skills which are applicable to data preparation and project management. Automating these aspects of research enables faster data collection, better data quality and a higher robustness towards errors.

10.3 Future research

Future research may benefit from methods for automatic literature processing described above. These methods are also likely to be useful for research in the science of sciences context. With ever increasing numbers of publications traditional methods for identification of important publications are unlikely to be suitable. Automatic text processing will enable to researchers to gain an understanding not only of a fraction of their own fields but also of adjacent fields, enabling researchers to combine disconnected knowledge sources. Furthermore, since the validation of research is likely to take on a more important role it will be useful to quickly identify studies that have the same or a similar focus.

Experiments on commercialization of academic knowledge using innovation platforms may be in the interest of platform owners as well as governmental entities: this represents an opportunity to let the free market do some of the work that publicly funded institutions such as patent offices sometimes struggle to accomplish. These experiments would represent a rich data source for validation and extension of the findings presented here.

Research into academic patenting is hampered by a lack of data regarding individuals and companies that file patents. It is probable that the application of machine learning methods to large scale datasets may mitigate this problem to some extent. Combining data from patent databases and social media may yield interesting insights.

As for regional knowledge production, future research may attempt to generalize some of the findings presented here by using more exhaustive datasets that cover regions across Europe or across continents in order to control for effects of collaborations over very long distances. Unfortunately it is difficult to obtain the corresponding data at this time due to a lack of coordination between national statistical authorities. Likewise, data quality could be improved with regard to patent and publication data: by using full texts instead of abstracts, the usefulness of distance measures may be increased.

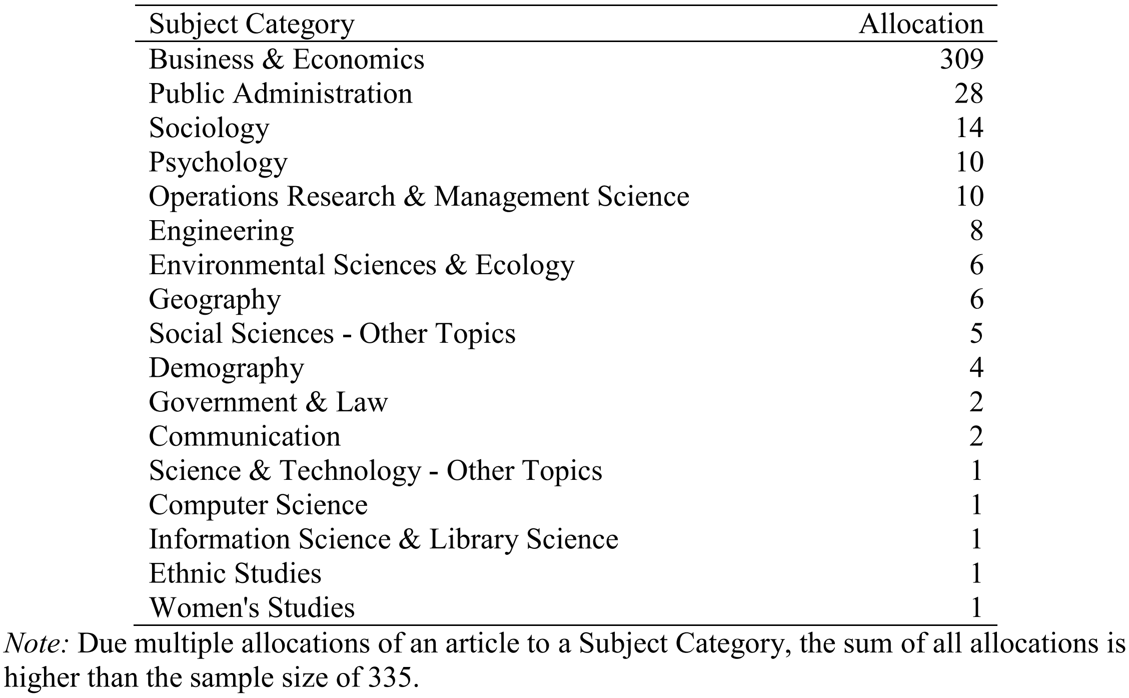
From a methodological point of view it is important to consider the limitations of complex statistical models applied to relatively small datasets. Future research will need to focus on validation of results instead of new findings that may turn out to be spurious. Advances in machine learning may be incorporated in validation studies as the methods used in this work are likely to be replaced with more precise and reliable algorithms. To complement machine learning methods more exhaustive datasets will be required.

Similar methodologies are likely to be prove useful also in adjacent research fields. A substantial part of human knowledge is codified in texts. As our ability to process large amounts of texts automatically keeps improving researchers will be able to derive new quantifiable measures in contexts where we were so far limited by relatively crude items that are often subjective or limited in scope. However, in order to leverage large amounts of text data scientists will require training in data processing and machine learning. Unfortunately, these skills are not typically taught or regarded as valuable in the field of social sciences. Also, the access to the necessary data and hardware can be expensive, which is a particular issue for faculties with limited budgets. The necessary resources are available in various multinational companies. Hence research collaboration with industry may in the future not only serve the commercialization of academic knowledge but also contribute to basic research findings by making resources and skills available to academia.

Appendix

Journal	Publications per Journal		Cumulated Publications per Journal	
	amount	percent	amount	percent
Journal of Business Venturing*	637	4.08%	637	4.08%
Small Business Economics*	517	7.40%	1154	3.31%
Entrepreneurship Theory and Practice*	305	9.35%	1459	1.96%
Entrepreneurship and Regional Development*	251	10.96%	1710	1.61%
Journal of Small Business Management*	220	12.37%	1930	1.41%
Technovation*	198	13.64%	2128	1.27%
Research Policy*	190	14.86%	2318	1.22%
International Small Business Journal*	185	16.05%	2503	1.19%
Journal of Business Ethics*	133	16.90%	2636	0.85%
International Entrepreneurship and Management Journal	125	17.70%	2761	0.80%
Journal of Business Research*	125	18.50%	2886	0.80%
Strategic Entrepreneurship Journal	116	19.25%	3002	0.74%
Forbes	115	19.98%	3117	0.74%
Harvard Business Review	111	20.69%	3228	0.71%
Strategic Management Journal*	108	21.39%	3336	0.69%
Regional Studies*	101	22.03%	3437	0.65%
African Journal of Business Management	95	22.64%	3532	0.61%
International Journal of Technology Management	95	23.25%	3627	0.61%
Journal of Management Studies*	90	23.83%	3717	0.58%
Journal of Technology Transfer	87	24.39%	3804	0.56%
Organization Studies*	83	24.92%	3887	0.53%
Organization Science*	78	25.42%	3965	0.50%
Management Decision	77	25.91%	4042	0.49%
European Planning Studies*	76	26.40%	4118	0.49%
World Development	76	26.89%	4194	0.49%
Environment and Planning C-Government and Policy	73	27.36%	4267	0.47%
Business History*	69	27.80%	4336	0.44%
International Business Review	68	28.23%	4404	0.44%
Urban Studies*	68	28.67%	4472	0.44%
Academy of Management Journal*	67	29.10%	4539	0.43%
Service Industries Journal	67	29.53%	4606	0.43%
Journal of Economic Behavior and Organization	65	29.95%	4671	0.42%
Management Science*	64	30.36%	4735	0.41%
Business History Review*	61	30.75%	4796	0.39%
International Journal of Urban and Regional Research*	61	31.14%	4857	0.39%
Journal of Management*	61	31.53%	4918	0.39%
Economic Development Quarterly*	60	31.91%	4978	0.38%
Industrial and Corporate Change	60	32.30%	5038	0.38%
Journal of World Business	60	32.68%	5098	0.38%
R and D Management	58	33.06%	5156	0.37%
Journal of International Business Studies*	57	33.42%	5213	0.37%
Journal of Product Innovation Management	57	33.79%	5270	0.37%
Journal of Evolutionary Economics*	55	34.14%	5325	0.35%

TABLE A1: Top Journals by Published Articles



Subject Category	Allocation
Business & Economics	309
Public Administration	28
Sociology	14
Psychology	10
Operations Research & Management Science	10
Engineering	8
Environmental Sciences & Ecology	6
Geography	6
Social Sciences - Other Topics	5
Demography	4
Government & Law	2
Communication	2
Science & Technology - Other Topics	1
Computer Science	1
Information Science & Library Science	1
Ethnic Studies	1
Women's Studies	1

Note: Due multiple allocations of an article to a Subject Category, the sum of all allocations is higher than the sample size of 335.

TABLE A2: WOS Subject Categories

Article	# Citations		Rank		Article	# Citations		Rank	
	Total	Per Year	Total	Per Year		Total	Per Year	Total	Per Year
Uzzi (1997)	2104	123.76	1	1	Miller and Friesen (1982)	455	14.22	21	39
Shane and Venkataraman (2000)	1578	112.71	2	2	Busenitz and Barney (1997)	447	26.29	22	22
Deshpande, Farley, and Webster (1993)	808	38.48	3	8	Sarasvathy (2001)	439	33.77	23	12
Shane (2000)	805	57.50	4	3	Stevenson and Jarillo (1990)	427	17.79	24	34
Harvey (1989)	803	32.12	5	15	Aghion and Bolton (1992)	399	18.14	25	33
Larson (1992)	728	33.09	6	13	Gimeno, Folta, Cooper, and Woo (1997)	390	22.94	26	27
Aldrich and Fiol (1994)	694	34.70	7	10	Cooper, Gimenogascon, and Woo (1994)	390	19.50	27	32
Baumol (1990)	693	28.88	8	20	Ahuja and Lampert (2001)*	379	29.15	28	19
Miller (1983)*	668	21.55	9	30	Nee (1992)	361	16.41	29	36
Oviatt and McDougall (1994)*	657	32.85	10	14	Hoang and Antoncic (2003)*	340	30.91	30	17
Stuart, Hoang, and Hybels (1999)*	623	41.53	11	7	McDougall, Shane, and Oviatt (1994)*	336	16.80	31	35
Eisenhardt and Schoonhoven (1996)	618	34.33	12	11	Lu and Beamish (2001)*	332	25.54	32	23
Evans and Jovanovic (1989)	611	24.44	13	25	Kihlstrom and Laffont (1979)**	331	9.46	33	40
Amit and Zott (2001)	587	45.15	14	6	Knight and Cavusgil (2004)*	316	31.60	34	16
Autio, Sapienza, and Almeida (2000)*	515	36.79	15	9	Cooper and Kleinschmidt (1995)	311	16.37	35	37
Peng (2003)	510	46.36	16	4	Etzkowitz, Webster, Gebhardt, and Terra (2000)	310	22.14	36	28
Davidsson and Honig (2003)	504	45.82	17	5	Shane and Stuart (2002)	304	25.33	37	24
Banerjee and Newman (1993)	484	23.05	18	26	Kaplan and Stromberg (2003)*	298	27.09	38	21
Blanchflower and Oswald (1998)*	478	29.88	19	18	Zahra and Covin (1995)	294	15.47	39	38
King and Levine (1993)	458	21.81	20	29	McDougall and Oviatt (2000)*	293	20.93	40	31

Note: * denotes articles associated to a large cluster (macro level) and ** denotes articles associated to a small cluster (meso level)

TABLE A3: Top 40 Articles

Cluster Title (total cites / # of articles / average cites per article) (year of first publication - year of latest publication)	**Description**	**tf-idf keywords** (downstream articles)
1. International entrepreneurship (6368 / 49 / 129,96) (1989 - 2008)	International entrepreneurship analyses the process of startups transforming into internationally active companies, it is regarded as a combination of the research fields of entrepreneurship and international business (McDougall and Oviatt, 2000). Relevant aspects to the growth of such companies have been identified (Autio 2000) as well as factors relating to the speed (Knight and Cavusgil, 2004) and likelihood (Oviatt and McDougall, 1994) of internationalization.	international internationalization venture firm export (945)
2. University-industry relations and entrepreneurship (4240 / 51 / 83,14) (1987 - 2009)	This large cluster investigates the relation between university and industry. For example, the role of technology transfer offices in creating startups (Siegel et al., 2003) as well as the influence of national policies (Goldfarb and Henrekson, 2003) or individual characteristics of entrepreneurs with scientific background (Murray, 2004).	university spin transfer technology academic (992)
3. Venture capital policies and financing (3053 / 31 / 98,48) (1989 - 2006)	The match-making between venture capitalists and entrepreneurs is an important factor in entrepreneurship and hence subject to intense study. Contracts may need to take into account varying motivations by investors and entrepreneurs (Aghion and Bolton, 1992) as well as moral hazard implicit to the process (Bergemann and Hege, 1998) while entrepreneurs have to optimize the relation between additional funds and shares sold (Hsu, 2004).	venture capitalist capital contract convertible (589)
4. Macroecnomic/global and regional impact of entrepreneurship (2612 / 34 / 76,82) (1987 - 2008)	The effect of entrepreneurship on economic growth is subject of research in this cluster (Reynolds et al., 2005). Related topics addressed in this cluster are differences between countries which may lead to advantages for some entrepreneurs compared to those from areas less conducive to entrepreneurship (Busenitz et al., 2000) or shifts from managed to entrepreneurial economies in developed countries (Audretsch and Thurik, 2000)	country economic regional collectivism culture (2037)
5. Entrepreneurship and liquiditiy (2202 / 8 / 275,25) (1989 - 2002)	Financial assets are an important precursor to entrepreneurial activity, whether they are the result of inheritance or investments (Blanchflower and Oswald, 1998). A related aspect is the return on investment from entrepreneurship (Hamilton, 2000).	inheritance liquidity self constraint employment (79)
6. Institutional entrepreneurship (2039 / 22 / 92,68) (1980 - 2009)	Institutional entrepreneurship is the focus of this cluster. Maguire et al. (2004) analyze institutional entrepreneurship in emerging fields, Lounsbury and Crumley (2007) develop a process model of new practice creation Beckert (1999) introduces the effect of strategic choice in this context. Furthermore, Fligstein (1997) introduce social skill in this context.	institutional actor change field agency (1922)
7. Corporate entrepreneurship (1375 / 4 / 343,75) (1983 - 1995)	Corporate Entrepreneurship is regarded as means to improve a company's long term financial performance (Zahra and Covin, 1995). Antecedents and effects are studied within this cluster.	corporate financial entrepreneurship company performance (2337)
8. Social entrepreneurship (1196 / 15 / 79,73) (2000 - 2009)	Social entrepreneurship is compared to commercial entrepreneurship (Austin et al., 2006) to determine important differences such as special performance indicators (e.g. social needs) that are not covered by commercial entrepreneurship (Mair and Marti, 2006).	social entrepreneurship civic franchise value (779)
9. Entrepreneurship and social network analysis (1102 / 6 / 183,67) (2001 - 2003)	In this cluster methods of social network analysis are applied to firm and entrepreneurial networks. E.g. theories on cohesive networks and networks with structural holes are related to firm success (Hite and Hesterly, 2001) or the evolution of personal networks through different phases of entrepreneurship is described (Greve and Salaff, 2003).	network firm social embed cohesive (1029)
10. Entrepreneurship in family firms (967 / 12 / 80,58) (2003 - 2007)	Differences between family and non-family firms are the subject of research for this cluster (Zahra et al., 2004). Family structures influence entrepreneurship (Aldrich and Cliff, 2003) and family firms exhibit special characteristics (Zahra, 2003).	family business firm involvement altruism (337)
11. Entrepreneurial education and self-efficacy (877 / 8 / 109,63) (1997 - 2006)	Self-efficacy appears to play an important role in entrepreneurship. It influences venture growth (Baum and Locke, 2004) and affects entrepreneurial learning (Zhao et al., 2005).	program student efficacy category self (68)
12. Entrepreneurial Intention (794 / 3 / 264,67) (1988 - 2000)	This cluster distinguishes between different entrepreneurial intentions (Krueger et al., 2000). Further, entrepreneurs and managers are distinguished by their self-efficacy (Chen et al., 1998).	intention efficacy student model self (189)
13. Immigration and entrepreneurship (708 / 9 / 78,67) (1985 - 1996)	This cluster, belonging to the field of sociology, analyses self-employment among immigrants. Relevant factors are, e.g., family structures (Sanders and Nee, 1996) or performance of small business entrepreneurs (Portes and Zhou, 1996). It appears to be related to the "Entrepreneurship and family firms" cluster.	immigrant korean asian export import (132)
14. Entrepreneurial opportunity detection and learning (666 / 10 / 66,6) (2005 - 2007)	This cluster explores how entrepreneurs detect business opportunities. E.g. pattern recognition, i.e. the ability to apply past experience to detect business opportunities before others, appears to be an important determinant (Baron and Ensley, 2006).	opportunity learn recognition pattern belief (274)
15. Transition economies (518 / 5 / 103,6) (2001 - 2003)	This cluster concentrates on transition countries, especially Russia. Johnson et al. (2002) analyze the effect of property right on new firms reinvestment of profits. Peng (2001) studies how entrepreneurs create wealth in transition economies.	embeddedness transition russian property hostile

TABLE A4: Cluster Overview

		(17)
16. Personal initiative (416 / 3 / 138,67) (1996 - 2000)	At the center of this cluster are studies that compare personal initiative in the former East and West Germany. The degree of control and complexity of work affects initiative, which can be regarded as similar to the concept of entrepreneurship (Frese et al., 1996).	initiative reactive opportunistic east planning (8)
17. Origin of entrepreneurs (412 / 5 / 82,4) (2003 - 2006)	An important antecedent for entrepreneurial behavior is the social context of the entrepreneur. Being embedded in a start-up friendly environment (Gompers et al., 2005) or in high performing academic environments with corporate links (Kenney and Goe, 2004) tends to increase the likelihood of start-ups being created.	spawn faculty science property department (43)
18. Cultural entrepreneurship in the US (365 / 3 / 121,67) (1982 - 1991)	This cluster differs from the other clusters in that it encompasses papers on media and culture rather than economics. Papers in this cluster describe the role of individuals in cultural entrepreneurship, i.e. the creation of museums or operas (Dimaggio, 1982).	boston century cultural america mediation (0)
19. Emerging economies (314 / 3 / 104,67) (2002 - 2008)	These articles analyze entrepreneurship in emerging economies (Bruton et al., 2008). Meyer and Peng (2005) concentrate on the context in Central and Eastern Europe. Specifically, three lines of theorizing have been advanced: organizational economics theories, resource-based theories and institutional theories	theory economy emerge institution eastern (698)
20. Narratives and presentation (311 / 4 / 77,75) (2007 - 2009)	This cluster analyses the relation between presenting a business model and investments. Signaling certain capabilities to potential investors is an important aspect observed in a study by Zott and Huy (2007). Short et al. (2009) points out that empirical evidence in this area is scarce which may negatively impact the application of theoretical concepts to managerial practice.	symbolic passion narrative action resource (46)
21. Business incubators (307 / 5 / 61,4) (2002 - 2005)	Co-production of business assistance in business incubators (Rice, 2002) and the effectiveness of business incubators (Colombo and Delmastro, 2002)	incubator incubation production ntbf park (37)
22. Management buyouts (259 / 4 / 64,75) (1992 - 2001)	Papers in this cluster explore the effect of leveraged buyouts on companies and describes various types of LBOs (Wright et al., 2001). It is suggested the LBOs are beneficial to a company's performance and corporate entrepreneurship and does not impact negatively on a company's RandD efforts (Zahra, 1995).	buyout company upside change performance (14)
23. Entrepreneurship in the Public Sector (253 / 3 / 84,33) (1992 - 2000)	Noticeable is that all three articles are published in Public Adminbistration Review. Articles in this cluster analyze that public entrepreneurs of the neo-managerialist persuasion pose a threat to democratic governance (Terry, 1998) as well as implications of the reinvention movement for democratic governance and its glorification of entrepreneurial management (deLeon and Denhardt, 2000).	democratic managerialism reinvention public civic (46)
24. Evans-Jovanovic entrepreneurial choice model (233 / 3 / 77,67) (1998 - 2003)	This cluster focusses on the Evans-Jovanovic entrepreneurial choice model, which states that the decision to become an entrepreneur can be modelled as optimization problem given equations for the income of wage workers and entrepreneurs.	jovanovic evan distribution credit collateral (64)
25. Habitual entrepreneurs (232 / 3 / 77,33) (1997 - 2003)	This cluster analyzes effects of entrepreneurs which were involved in more than one venture (Wrigth et al., 1997; Westhead and Wright, 1998)	habitual serial capitalist novice founder (32)
26. Entrepreneurship as a social construct (216 / 3 / 72) (2004 - 2006)	Steyaert and Katz (2004) consider Entrepreneurship as a societal rather than an economic phenomenon. Fletcher (2006) studies social constructionist thinking particularly with regard to opportunity formation processes.	myth metaphor newspaper relationally sense (8)
27. Cultural support for entrepreneurship (205 / 3 / 68,33) (2000 - 2007)	Related to the large cluster of international entrepreneurship this cluster focusses on cultural factors relevant to entrepreneurship. Individual factors shown to be important for entrepreneurial success vary by culture (Thomas and Mueller, 2000). Social concepts such as shame related to failure or social status also explains entrepreneurial action (Begley and Tan, 2001).	anglo cultural aspiration east Asian (1)
28. Venture capitalist investment decisions (198 / 3 / 66) (1992 - 1998)	The process underlying an investment decision is governed by variables attributable to the investor, such as his preference for national investments and variables describing the investment opportunity, such as the quality of the business idea. Papers in this cluster explore this relationship with methods to go beyond survey data, conjoint analyses (Muzyka et al., 1996) or policy capturing (Zacharakis and Meyer, 1998).	decision capitalist conjoint criterion venture (161)
29. Franchise I (185 / 3 / 61,67) (1988 - 1996)	This cluster concentrates on entrepreneurial franchise and starting reasons. Michael (1996) analyses decision rights and organizational form shares. Kaufmann and Dant (1996) show that capital acquisition is a relevant reason for engaging in franchising and not the assumption that franchisees manage the outlets better than company employees would if the unit were company owned.	franchise franchisee unit multi franchisor (52)
30. Women and entrepreneurship (143 / 4 / 35,75) (1991 - 2003)	Most cited articles in this cluster study differences between women and men entrepreneurs. DeMartino and Barbato (2003) explore family flexibility and wealth creation as career motivators and Caputo and Dolinsky (1998) analyze the role of financial and human capital of housholdmembers on women's choice to pursue self-employment.	woman child motivator female owner (200)
31. Gender and entrepreneurship (143 / 3 / 47,67) (2001 - 2005)	This cluster concentrates on gender diversities in Entrepreneurship. Not only differences and divisions between women business owners who are silent about gender issues and those who are not are explored (Lewis, 2006) but also formal and informal sources of business funding to illustrate how this concept impacts upon women in self-employment (Marlow and Patton, 2005).	woman gender attributional augment female (368)
32. Franchise II (118 / 4 / 29,5) (1996 - 1999)	This whole journal is published in the Journal of Business Venturing. It analyzes business-format franchising growth's in the U.S. (Lafontaine and Shaw, 1998) as well as survival patterns among franchisee and nonfranchise small firms (Bates, 1998).	franchise franchisor franchisee unit establishment (12)

TABLE A5: Cluster Overview, continued

33. Entrepreneurs vs. Managers (112 / 3 / 37,33) (1987 - 1990)	This cluster analyzes differences between Entrepreneurs and Managers in small business firms (Begley and Boyd, 1987) and in their motivational paterns (Miner, 1990).	technologically motivational manager task growth (13)
34. Alliances and Jount Ventures (111 / 3 / 37) (1994 - 1999)	Market valuation of joint ventures in terms of Joint venture characteristics and wealth gains (Park and Kim, 1997) as well as opportunistic action within research alliances (Deeds and Hill, 1999) are studies in this cluster.	joint partner alliance venture corporate (100)
35. Entrepreneurs in organizations (100 / 3 / 33,33) (1986 - 1997)	Articles located in this cluster analyze implications for organizations' structures and their Human-Resouce Mangement Practices to foster and facilitate entrepreneurship (Schuler, 1986) and effects of managers' entrepreneurial behavior on subordinates (Pearce et al., 1997).	manager subordinate behavior satisfaction corporate (89)

TABLE A6: Cluster Overview, continued

Bibliography

Abramo, Giovanni, Ciriaco Andrea D'Angelo, Flavia Di Costa, and Marco Solazzi (2011). "The role of information asymmetry in the market for university–industry research collaboration". In: *The Journal of Technology Transfer* 36.1, pp. 84–100.

Abramovitz, Moses (1956). "Resource and output trends in the United States since 1870". In: *Resource and output trends in the United States since 1870*, pp. 1–23.

Acs, Zoltan J, Luc Anselin, and Attila Varga (2002). "Patents and innovation counts as measures of regional production of new knowledge". In: *Research Policy* 31.7, pp. 1069–1085.

Afuah, A. and C. L. Tucci (2012). "Crowdsourcing As a Solution to Distant Search". In: *Academy of Management Review* 37.3, pp. 355–375.

Agarwal, Rajshree and Atsushi Ohyama (2013). "Industry or academia, basic or applied? Career choices and earnings trajectories of scientists". In: *Management Science* 59.4, pp. 950–970.

Agresti, Alan (2010). *Analysis of ordinal categorical data*. Vol. 656. John Wiley & Sons.

Ahuja, Gautam (2000). "Collaboration networks, structural holes, and innovation: A longitudinal study". In: *Administrative science quarterly* 45.3, pp. 425–455.

Ajzen, Icek (1988). *Attitudes, personality, and behavior.* Dorsey Press.

Aldridge, T Taylor and David Audretsch (2011). "The Bayh-Dole act and scientist entrepreneurship". In: *Research policy* 40.8, pp. 1058–1067.

Allen, Thomas and Gunter Henn (2007). *The organization and architecture of innovation*. Routledge.

Allen, Thomas J (1977). "Managing the flow of technology: technology transfer and the dissemination of technological information within the R and D organization". In:

Allen, Thomas J (1984). *Managing the Flow of Technology: Technology Transfer and the Dissemination of Technological Information Within the R&D Organization, volume 1 of MIT Press Books*. The MIT Press.

Almeida, Paul (1996). "Knowledge sourcing by foreign multinationals: Patent citation analysis in the US semiconductor industry". In: *Strategic management journal* 17.S2, pp. 155–165.

Almeida, Paul and Bruce Kogut (1999). "Localization of Knowledge and the Mobility of Engineers in Regional Networks". In: *Management Science* 45.7, pp. 905–917.

Amabile, Teresa M (1996). *Creativity in context: Update to" the social psychology of creativity."* Westview press.

Anselin, Luc, Attila Varga, and Zoltan Acs (1997). "Local Geographic Spillovers between University Research and High Technology Innovations". In: *Journal of Urban Economics* 42.3, pp. 422–448.

Arora, Sanjay K., Alan L. Porter, Jan Youtie, and Philip Shapira (2013). "Capturing new developments in an emerging technology: an updated search strategy for identifying nanotechnology research outputs". In: *Scientometrics* 95.1, pp. 351–370.

Arrow, Kenneth Joseph (1971). "The economic implications of learning by doing". In: *Readings in the Theory of Growth*, pp. 131–149.

Atkins, David C, Scott A Baldwin, Cheng Zheng, Robert J Gallop, and Clayton Neighbors (2013). "A tutorial on count regression and zero-altered count models for longitudinal substance use data." In: *Psychology of Addictive Behaviors* 27.1, p. 166.

Atkinson, Robert D and Stephen J Ezell (2012). *Innovation economics: the race for global advantage*. Yale University Press.

Audretsch, David B and Maryann P Feldman (1996). "R&D spillovers and the geography of innovation and production". In: *The American economic review* 86.3, pp. 630–640.

Audretsch, David B. and Erik E. Lehmann (2005). "Does the Knowledge Spillover Theory of Entrepreneurship hold for regions?" In: *Research Policy* 34.8, pp. 1191–1202.

Azoulay, Pierre, Waverly Ding, and Toby Stuart (2007). "The determinants of faculty patenting behavior: Demographics or opportunities?" In: *Journal of economic behavior & organization* 63.4, pp. 599–623.

Azoulay, Pierre, Waverly Ding, and Toby Stuart (2009). "The impact of academic patenting on the rate, quality and direction of (public) research output". In: *The Journal of Industrial Economics* 57.4, pp. 637–676.

Bagozzi, Richard P. (1984). "A prospectus for theory construction in marketing". In: *The Journal of Marketing* 48.1, pp. 11–29.

Balconi, Margherita, Valeria Lorenzi, Pier Paolo Saviotti, and Antonella Zucchella (2013). "Cognitive distance in research collaborations". In: *Università di Pavia, DEM Working Paper Series* 51.09-13, pp. 1–36.

Baldini, Nicola (2009). "Implementing Bayh–Dole-like laws: Faculty problems and their impact on university patenting activity". In: *Research Policy* 38.8, pp. 1217–1224.

Baldini, Nicola (2010). "Do royalties really foster university patenting activity? An answer from Italy". In: *Technovation* 30.2, pp. 109–116.

Baldini, Nicola, Rosa Grimaldi, and Maurizio Sobrero (2007). "To patent or not to patent? A survey of Italian inventors on motivations, incentives, and obstacles to university patenting". In: *Scientometrics* 70.2, pp. 333–354.

Bamard, Chester I (1938). *The functions of the executive*. Harvard university press.

Barga, Roger, Valentine Fontama, and Wee Hyong Tok (2014). "Introduction to Data Science". In: *Predictive Analytics with Microsoft Azure Machine Learning*, pp. 3–20.

Beaver, Donald and Richard Rosen (1978). "Studies in scientific collaboration: Part I. The professional origins of scientific co-authorship". In: *Scientometrics* 1.1, pp. 65–84.

Beaver, Donald deB (2004). "Does collaborative research have greater epistemic authority?" In: *Scientometrics* 60.3, pp. 399–408.

Bekkers, Rudi and Isabel Maria Bodas Freitas (2008). "Analysing knowledge transfer channels between universities and industry: To what degree do sectors also matter?" In: *Research policy* 37.10, pp. 1837–1853.

Benner, Mary and Joel Waldfogel (2007). *Close to You? Bias and Precision in Patent-Based Measures of Technological Proximity*. Tech. rep. w13322. National Bureau of Economic Research.

Berchicci, Luca (2013). "Towards an open R&D system: Internal R&D investment, external knowledge acquisition and innovative performance". In: *Research Policy* 42.1, pp. 117–127.

Bercovitz, Janet and Maryann Feldman (2008). "Academic entrepreneurs: Organizational change at the individual level". In: *Organization Science* 19.1, pp. 69–89.

Bercovitz, Janet and Maryann Feldman (2011). "The mechanisms of collaboration in inventive teams: Composition, social networks, and geography". In: *Research Policy* 40.1, pp. 81–93.

Bird, Steven, Ewan Klein, and Edward Loper (2009). *Natural language processing with Python*. 1st ed. O'Reilly.

Blei, David M., Andrew Y. Ng, Michael I. Jordan, and John Lafferty (2003). "Latent dirichlet allocation". In: *Journal of Machine Learning Research* 3, p. 2003.

Boudreau, Kevin J., Nicola Lacetera, and Karim R. Lakhani (2011). "Incentives and Problem Uncertainty in Innovation Contests: An Empirical Analysis". In: *Management Science* 57.5, pp. 843–863.

Boudreau, Kevin J., Eva C. Guinan, Karim R. Lakhani, and Christoph Riedl (2016). "Looking Across and Looking Beyond the Knowledge Frontier: Intellectual Distance, Novelty, and Resource Allocation in Science". In: *Management Science*. Forthcoming.

Bozeman, Barry (2000). "Technology transfer and public policy: a review of research and theory". In: *Research policy* 29.4, pp. 627–655.

Bozeman, Barry, Daniel Fay, and Catherine P Slade (2013). "Research collaboration in universities and academic entrepreneurship: the-state-of-the-art". In: *The Journal of Technology Transfer* 38.1, pp. 1–67.

Bozeman, Barry and Monica Gaughan (2007). "Impacts of grants and contracts on academic researchers' interactions with industry". In: *Research policy* 36.5, pp. 694–707.

Brabham, Daren C (2008). "Crowdsourcing as a model for problem solving an introduction and cases". In: *Convergence: the international journal of research into new media technologies* 14.1, pp. 75–90.

Bramwell, Allison and David A. Wolfe (2008). "Universities and regional economic development: The entrepreneurial University of Waterloo". In: *Research Policy* 37.8, pp. 1175–1187.

Braun, Tibor, Isabel Gómez, Aida Méndez, and Andras Schubert (1992). "International co-authorship patterns in physics and its subfields, 1981–1985". In: *Scientometrics* 24.2, pp. 181–200.

Breschi, Stefano and Francesco Lissoni (2001). "Knowledge spillovers and local innovation systems: a critical survey". In: *Industrial and corporate change* 10.4, pp. 975–1005.

Breschi, Stefano and Francesco Lissoni (2006). *Mobility of inventors and the geography of knowledge spillovers: new evidence on US data*. Università commerciale Luigi Bocconi.

Breschi, Stefano, Francesco Lissoni, and Fabio Montobbio (2005). "From publishing to patenting: Do productive scientists turn into academi inventors?" In: *Revue d'économie industrielle* 110.1, pp. 75–102.

Breschi, Stefano, Francesco Lissoni, and Fabio Montobbio (2007). "The scientific productivity of academic inventors: new evidence from Italian data". In: *Econ. Innov. New Techn.* 16.2, pp. 101–118.

Bresnahan, Timothy F and Manuel Trajtenberg (1995). "General purpose technologies 'Engines of growth'?" In: *Journal of econometrics* 65.1, pp. 83–108.

Brooks, Harvey (1993). "Research universities and the social contract for science". In: *Empowering Technology: Implementing a US Strategy*, pp. 202–234.

Burt, Ronald S (1992). *Structural holes: The social structure of competition*. Harvard university press.

Bush, George P and Lowell H Hattery (1956). "Teamwork and creativity in research". In: *Administrative Science Quarterly* 1.3, pp. 361–372.

Bush, V (1945). *As we may think Athlantic Monthly 176*.

Butler, Linda (2003). "Explaining Australia's increased share of ISI publications—the effects of a funding formula based on publication counts". In: *Research Policy* 32.1, pp. 143–155.

Calderini, Mario, Chiara Franzoni, and Andrea Vezzulli (2007). "If star scientists do not patent: The effect of productivity, basicness and impact on the decision to patent in the academic world". In: *Research Policy* 36.3, pp. 303–319.

Caloghirou, Yannis, Aggelos Tsakanikas, and Nicholas S Vonortas (2001). "University-industry cooperation in the context of the European framework programmes". In: *The Journal of Technology Transfer* 26.1-2, pp. 153–161.

Cantner, Uwe, Kristin Joel, and Tobias Schmidt (2011). "The effects of knowledge management on innovative success – An empirical analysis of German firms". In: *Research Policy* 40.10, pp. 1453–1462.

Carlson, N.R., D. Heth, and H. Miller (2007). *Psychology: The Science of Behavior*. Pearson Allyn and Bacon.

Casper, Steven (2013). "The spill-over theory reversed: The impact of regional economies on the commercialization of university science". In: *Research Policy* 42.8, pp. 1313–1324.

Cawkell, A.E. (1976). "Understanding science by analysing its literature". In: *The Information Scientist* 10.1, pp. 3–10.

CEMI's PATSTAT knowledge base. http://wiki.epfl.ch/patstat/physical. Accessed: 2014-12-15.

Chatterjee, Samprit and Bertram Price (1991). "Regression diagnostics". In: *New York*.

Che, Yeon-Koo and Ian Gale (2003). "Optimal design of research contests". In: *The American Economic Review* 93.3, pp. 646–671.

Chesbrough, Henry, Wim Vanhaverbeke, and Joel West (2006). *Open innovation: Researching a new paradigm*. Oxford University Press on Demand.

Chesbrough, Henry William (2003). *Open innovation: the new imperative for creating and profiting from technology*. Harvard Business School Press.

Chessa, Alessandro, Andrea Morescalchi, Fabio Pammolli, Orion Penner, Alexander M Petersen, and Massimo Riccaboni (2013). "Is Europe evolving toward an integrated research area?" In: *Science* 339.6120, pp. 650–651.

Chiu, Chao-Min, Eric T.G. Wang, Fu-Jong Shih, and Yi-Wen Fan (2011). "Understanding knowledge sharing in virtual communities: An integration of expectancy disconfirmation and justice theories". In: *Online Information Review* 35.1, pp. 134–153.

Choi, Chong ju and Soo Hee Lee (1997). "A knowledge-based view of cooperative interorganizational relationships". In: *Cooperative strategies: European perspectives* 2, p. 33.

Christensen, Clayton (2013). *The innovator's dilemma: when new technologies cause great firms to fail*. Harvard Business Review Press.

Clarysse, Bart, Valentina Tartari, and Ammon Salter (2011). "The impact of entrepreneurial capacity, experience and organizational support on academic entrepreneurship". In: *Research Policy* 40.8, pp. 1084–1093.

Cleveland, William S (2001). "Data science: an action plan for expanding the technical areas of the field of statistics". In: *International statistical review* 69.1, pp. 21–26.

Cohen, Wesley M and Daniel A Levinthal (1990). "Absorptive capacity: A new perspective on learning and innovation". In: *Administrative science quarterly* 35.1, pp. 128–152.

Cohen, Wesley M, Richard R Nelson, and John P Walsh (2002). "Links and impacts: the influence of public research on industrial R&D". In: *Management science* 48.1, pp. 1–23.

Cohen, William, Pradeep Ravikumar, and Stephen Fienberg (2003). "A comparison of string metrics for matching names and records". In: *Kdd workshop on data cleaning and object consolidation*. Vol. 3, pp. 73–78.

Constant, David, Lee Sproull, and Sara Kiesler (1996). "The Kindness of Strangers: The Usefulness of Electronic Weak Ties for Technical Advice". In: *Organization Science* 7.2, pp. 119–135.

Conti, Annamaria and Patrick Gaule (2011). "Is the US outperforming Europe in university technology licensing? A new perspective on the European Paradox". In: *Research Policy* 40.1, pp. 123–135.

Cooke, Philip and Loet Leydesdorff (2006). "Regional development in the knowledge-based economy: The construction of advantage". In: *The journal of technology Transfer* 31.1, pp. 5–15.

Cowan, Robin and Natalia Zinovyeva (2013). "University effects on regional innovation". In: *Research Policy* 42.3, pp. 788–800.

Crafts, Nicholas (2004). "Steam as a general purpose technology: a growth accounting perspective". In: *The Economic Journal* 114.495, pp. 338–351.

Crespi, Gustavo, Pablo D'Este, Roberto Fontana, and Aldo Geuna (2011). "The impact of academic patenting on university research and its transfer". In: *Research policy* 40.1, pp. 55–68.

Cristina, Andrea Bonaccorsi and Cristina Rossi (2004). "Altruistic individuals, selfish firms? The structure of motivation in Open Source software". In: *First Monday* 9, p. 9.

Cummings, Jeffrey L and Bing-Sheng Teng (2003). "Transferring R&D knowledge: the key factors affecting knowledge transfer success". In: *Journal of Engineering and technology management* 20.1, pp. 39–68.

Dahlander, Linus and David M. Gann (2010). "How open is innovation?" In: *Research Policy* 39.6, pp. 699–709.

Dasgupta, Partha and Paul A David (1987). "Information disclosure and the economics of science and technology". In: *Arrow and the ascent of modern economic theory*, pp. 519–542.

De Smith, Michael John, Michael F Goodchild, and Paul Longley (2007). *Geospatial analysis: a comprehensive guide to principles, techniques and software tools*. Troubador Publishing Ltd.

Deci, Edward L (1971). "Effects of externally mediated rewards on intrinsic motivation." In: *Journal of personality and Social Psychology* 18.1, pp. 105–115.

Deci, Edward L and Richard M Ryan (2000). "The" what" and" why" of goal pursuits: Human needs and the self-determination of behavior". In: *Psychological inquiry* 11.4, pp. 227–268.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman (1990). "Indexing by latent semantic analysis". In: *Journal of the American Society for Information Science* 41.6, pp. 391–407.

D'Este, P., F. Guy, and S. Iammarino (2013). "Shaping the formation of university-industry research collaborations: what type of proximity does really matter?" In: *Journal of Economic Geography* 13.4, pp. 537–558.

D'Este, P. and P. Patel (2007). "University–industry linkages in the UK: What are the factors underlying the variety of interactions with industry?" In: *Research Policy* 36.9, pp. 1295–1313.

D'Este, Pablo and Simona Iammarino (2010). "The spatial profile of university-business research partnerships: The spatial profile of u-b research partnerships". In: *Papers in Regional Science* 89.2, pp. 335–350.

D'este, Pablo and Markus Perkmann (2011). "Why do academics engage with industry? The entrepreneurial university and individual motivations". In: *The Journal of Technology Transfer* 36.3, pp. 316–339.

Diener, K and FT Piller (2010). *The Market for Open Innovation: Increasing the efficiency and effectiveness of the innovation process*. RWTH Aachen University.

Dietz, James S and Barry Bozeman (2005). "Academic careers, patents, and productivity: industry experience as scientific and technical human capital". In: *Research policy* 34.3, pp. 349–367.

DiMasi, Joseph A (2002). "The value of improving the productivity of the drug development process". In: *Pharmacoeconomics* 20.3, pp. 1–10.

Ding, Waverly W, Fiona Murray, and Toby E Stuart (2006). "Gender differences in patenting in the academic life sciences". In: *Science* 313.5787, pp. 665–667.

Donaldson, Ken, Vicki Stone, CL Tran, Wolfgang Kreyling, and Paul JA Borm (2004). "Nanotoxicology". In: *Occupational and environmental medicine* 61.9, pp. 727–728.

Eisenberg, Rebecca S (1996). "Public research and private development: patents and technology transfer in government-sponsored research". In: *Virginia Law Review* 82.8, pp. 1663–1727.

Etzkowitz, Henry (1988). "The making of an entrepreneurial university: the traffic among MIT, industry, and the military, 1860–1960". In: *Science, Technology and the Military*, pp. 515–540.

Etzkowitz, Henry (2004). "The evolution of the entrepreneurial university". In: *International Journal of Technology and Globalisation* 1.1, pp. 64–77.

Etzkowitz, Henry (2008). *The triple helix: university-industry-government innovation in action*. Routledge.

Etzkowitz, Henry and Loet Leydesdorff (2000). "The dynamics of innovation: from National Systems and "Mode 2" to a Triple Helix of university–industry–government relations". In: *Research policy* 29.2, pp. 109–123.

European Commission, ed. (2015). *Regions in the European Union: nomenclature of territorial units for statistics ; NUTS 2013/EU-28*. OCLC: 935926395. Publ. of the Europ. Union.

Fagerberg, Jan and Bart Verspagen (2002). "Technology-gaps, innovation-diffusion and transformation: an evolutionary interpretation". In: *Research Policy* 31.8-9, pp. 1291–1304.

Feller, Joseph (2005). *Perspectives on free and open source software*. MIT Press.

Feynman, Richard P (1960). "There's plenty of room at the bottom". In: *Engineering and science* 23.5, pp. 22–36.

Fiol, C Marlene (1995). "Thought worlds colliding: The role of contradiction in corporate innovation processes". In: *Entrepreneurship: Theory and Practice* 19.3, pp. 71–91.

Fischer, Manfred M. and Attila Varga (2003). "Spatial knowledge spillovers and university research: Evidence from Austria". In: *The Annals of Regional Science* 37.2, pp. 303–322.

Fischer, Timo and Joachim Henkel (2013). "Complements and substitutes in profiting from innovation—A choice experimental approach". In: *Research Policy* 42.2, pp. 326–339.

Fleming, Lee and Olav Sorenson (2004). "Science as a map in technological search". In: *Strategic Management Journal* 25.89, pp. 909–928.

Florida, Richard (1999). *Engine or lnfrastructure? The University Role in Economic Development*. Tech. rep. Carnegie Mellon University.

Forbes, Daniel P, Patricia S Borchert, Mary E Zellmer-Bruhn, and Harry J Sapienza (2006). "Entrepreneurial team formation: An exploration of new member addition". In: *Entrepreneurship Theory and Practice* 30.2, pp. 225–248.

Fox, C. R. and A. Tversky (1995). "Ambiguity Aversion and Comparative Ignorance". In: *The Quarterly Journal of Economics* 110.3, pp. 585–603.

Franke, N., C. Lettl, S. Roiser, and P. Tuertscher (2014). ""Does God Play Dice?" - Randomness vs. Deterministic Explanations of Crowdsourcing Success". In: *Academy of Management Proceedings* 2014.1, pp. 15164–15164.

Franke, Nikolaus, Peter Keinz, and Katharina Klausberger (2013). ""Does This Sound Like a Fair Deal?": Antecedents and Consequences of Fairness Expectations in the Individual's Decision to Participate in Firm Innovation". In: *Organization Science* 24.5, pp. 1495–1516.

Franke, Nikolaus, Marc Gruber, Dietmar Harhoff, and Joachim Henkel (2008). "Venture Capitalists' Evaluations of Start-Up Teams: Trade-Offs, Knock-Out Criteria, and the Impact of VC Experience". In: *Entrepreneurship Theory and Practice* 32.3, pp. 459–483.

Franzoni, Chiara and Giuseppe Scellato (2010). “The grace period in international patent law and its effect on the timing of disclosure”. In: *Research policy* 39.2, pp. 200–213.

Freel, Mark S (2003). “Sectoral patterns of small firm innovation, networking and proximity”. In: *Research policy* 32.5, pp. 751–770.

Frenken, Koen, Jarno Hoekman, Suzanne Kok, Roderik Ponds, Frank van Oort, and Joep van Vliet (2009). “Death of distance in science? A gravity approach to research collaboration”. In: *Innovation networks*, pp. 43–57.

Frey, Bruno S (2010). “Geld oder anerkennung? zur oekonomik der auszeichnungen”. In: *Perspektiven der Wirtschaftspolitik* 11.1, pp. 1–15.

Frey, Bruno S (2012). “Crowding out and crowding in of intrinsic preferences”. In: *Reflexive government and global public goods*, pp. 75–83.

Frey, Bruno S and Reto Jegen (2001). “Motivation crowding theory”. In: *Journal of economic surveys* 15.5, pp. 589–611.

Garfield, Eugene (1955). “Citation indexes for science: A new dimension in documentation through association of ideas”. In: *Science* 122.3159, pp. 108–111.

Garfield, Eugene, Morton V. Malin, and Hmy Small (1983). *Citation Data as Science Indicators*.

Gartner, William B (1990). “What are we talking about when we talk about entrepreneurship?” In: *Journal of Business venturing* 5.1, pp. 15–28.

Gilsing, Victor, Bart Nooteboom, Wim Vanhaverbeke, Geert Duysters, and Ad van den Oord (2008). “Network embeddedness and the exploration of novel technologies: Technological distance, betweenness centrality and density”. In: *Research policy* 37.10, pp. 1717–1731.

Giuri, Paola, Myriam Mariani, Stefano Brusoni, Gustavo Crespi, Dominique Francoz, Alfonso Gambardella, Walter Garcia-Fontes, Aldo Geuna, Raul Gonzales, Dietmar Harhoff, et al. (2007). “Inventors and invention processes in Europe: Results from the PatVal-EU survey”. In: *Research policy* 36.8, pp. 1107–1127.

Glaeser, Edward L., Hedi D. Kallal, Jose A. Scheinkman, and Andrei Shleifer (1991). *Growth in cities*. Tech. rep. National Bureau of Economic Research.

Glaeser, Edward L., Hedi D. Kallal, José A. Scheinkman, and Andrei Shleifer (1992). “Growth in Cities”. In: *Journal of Political Economy* 100.6, pp. 1126–1152.

Glenna, Leland L, Rick Welsh, David Ervin, William B Lacy, and Dina Biscotti (2011). “Commercial science, scientists’ values, and university biotechnology research agendas”. In: *Research Policy* 40.7, pp. 957–968.

Gmür, Markus (2003). “Co-citation analysis and the search for invisible colleges: A methodological evaluation”. In: *Scientometrics* 57.1, pp. 27–57.

Goffman, William and Kenneth S Warren (1980). *Scientific information systems and the principle of selectivity*. Praeger Publishers.

Gogotsi, Yury and Volker Presser (2013). *Carbon nanomaterials*. CRC Press.

Göktepe-Hulten, Devrim and Prashanth Mahagaonkar (2010). “Inventing and patenting activities of scientists: in the expectation of money or reputation?” In: *The Journal of Technology Transfer* 35.4, pp. 401–423.

Goldberg, Amir, Michael T Hannan, and Balázs Kovács (2016). “1. Title: What Does It Mean to Span Cultural Boundaries? Variety and Atypicality in Cultural Consumption”. In: *American Sociological Review* 81.2.

Gonzalez-Brambila, Claudia N., Francisco M. Veloso, and David Krackhardt (2013). “The impact of network embeddedness on research output”. In: *Research Policy* 42.9, pp. 1555–1567.

Gordon, Michael (1980). “A critical reassessment of inferred relations between multiple authorship, scientific collaboration, the production of papers and their acceptance for publication”. In: *Scientometrics* 2.3, pp. 193–201.

Grant, Robert M. and Charles Baden-Fuller (1995). "A knowledge-based theory of inter-firm collaboration." In: *Academy of Management Proceedings*. Vol. 1995. 1. Academy of Management, pp. 17–21.

Grant, Robert M. and Charles Baden-Fuller (2004). "A Knowledge Accessing Theory of Strategic Alliances". In: *Journal of Management Studies* 41.1, pp. 61–84.

Green, Paul E, Abba M Krieger, and Yoram Wind (2001). "Thirty years of conjoint analysis: Reflections and prospects". In: *Interfaces* 31.3_supplement, S56–S73.

Greene, William H (2003). *Econometric analysis*. Pearson Education India.

Griliches, Zvi (1979). "Issues in assessing the contribution of research and development to productivity growth". In: *The bell journal of economics* 10.1, pp. 92–116.

Griliches, Zvi (1990). *Patent statistics as economic indicators: a survey*. Tech. rep. National Bureau of Economic Research.

Grimaldi, Rosa, Martin Kenney, Donald S Siegel, and Mike Wright (2011). "30 years after Bayh–Dole: Reassessing academic entrepreneurship". In: *Research Policy* 40.8, pp. 1045–1057.

Grimpe, Christoph and Roberto Patuelli (2011). "Regional knowledge production in nanomaterials: a spatial filtering approach". In: *The Annals of Regional Science* 46.3, pp. 519–541.

Guan, Jian Cheng, Kai Rui Zuo, Kai Hua Chen, and Richard C.M. Yam (2016). "Does country-level R&D efficiency benefit from the collaboration network structure?" In: *Research Policy* 45.4, pp. 770–784.

Guan, Jiancheng, Jingjing Zhang, and Yan Yan (2015). "The impact of multilevel networks on innovation". In: *Research Policy* 44.3, pp. 545–559.

Gulbrandsen, Magnus and Jens-Christian Smeby (2005). "Industry funding and university professors' research performance". In: *Research Policy* 34.6, pp. 932–950.

Haas, Martine. R., Paoloa Criscuolo, and Gerard George (2015). "Which Problems to Solve? Online Knowledge Sharing and Attention Allocation in Organizations". In: *Academy of Management Journal* 58.3, pp. 680–711.

Haeussler, Carolin and Jeannette A. Colyvas (2011). "Breaking the ivory tower: academic entrepreneurship in the life sciences in UK and Germany". In: *Research Policy* 40.1, pp. 41–54.

Hagstrom, Warren O. (1965). *The scientific community*. Basic books.

Hamel, Gary (1991). "Competition for competence and interpartner learning within international strategic alliances". In: *Strategic management journal* 12.S1, pp. 83–103.

Hanel, Petr and Marc St-Pierre (2006). "Industry–university collaboration by Canadian manufacturing firms". In: *The Journal of Technology Transfer* 31.4, pp. 485–499.

Harhoff, Dietmar and Karin Hoisl (2007). "Institutionalized incentives for ingenuity—patent value and the German Employees' Inventions Act". In: *Research Policy* 36.8, pp. 1143–1162.

Harrison, David A and Katherine J Klein (2007). "What's the difference? Diversity constructs as separation, variety, or disparity in organizations". In: *Academy of management review* 32.4, pp. 1199–1228.

Hartig, Juliane (2011). *Learning and innovation @ a distance an empirical investigation into the benefits and liabilities of different forms of distance on interactive learning and novelty creation in german biotechnology SMEs*. Gabler Verlag.

Heckman, James J. (1979). "Sample Selection Bias as a Specification Error". In: *Econometrica* 47.1, pp. 153–161.

Hess, Stephane, Moshe Ben-Akiva, Dinesh Gopinath, and Joan Walker (2011). *Advantages of latent class over continuous mixture of logit models*.

Tech. rep. Institute for Transport Studies, University of Leeds. Working paper.

Hingley, Peter and Walter G Park (2003). "Patent family data and statistics at the European Patent Office". In: *WIPO-OED workshop on statistics in the patent field, Geneva*.

Hölzl, Werner and Jürgen Janger (2014). "Distance to the frontier and the perception of innovation barriers across European countries". In: *Research Policy* 43.4, pp. 707–725.

Hoekman, Jarno, Koen Frenken, and Frank Van Oort (2009). "The geography of collaborative knowledge production in Europe". In: *The Annals of Regional Science* 43.3, pp. 721–738.

Hofmann, Thomas (1999). "Probabilistic latent semantic indexing". In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pp. 50–57.

Hofstede, Geert (1984). *Culture's consequences: International differences in work-related values*. Sage.

Holmstrom, Bengt and Paul Milgrom (1994). "The firm as an incentive system". In: *The American Economic Review*, pp. 972–991.

Hornyak, Gabor L, Joydeep Dutta, Harry F Tibbals, and Anil Rao (2008). *Introduction to nanoscience*. CRC press.

Hounshell, David A. (1996). "The evolution of industrial research in the United States". In: *Engines of innovation: US industrial research at the end of an era* 13, pp. 51–56.

Howe, Jeff (2006). "The rise of crowdsourcing". In: *Wired magazine* 14.6, pp. 1–4.

Howe, Jeff (2016). *Crowdsourcing: A definition*. http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html. Blog.

Huber, Joel and Klaus Zwerina (1996). "The importance of utility balance in efficient choice designs". In: *Journal of Marketing research* 33.3, pp. 307–317.

Ikeda, Kaname (2009). "ITER on the road to fusion energy". In: *Nuclear Fusion* 50.1, pp. 1–10.

Ireland, R. Duane, Christopher R. Reutzel, and Justin W. Webb (2005). "Entrepreneurship research in AMJ: what has been published, and what might the future hold?" In: *Academy of Management Journal* 48.4, pp. 556–564.

Jaccard, Paul (1901). *Distribution de la Flore Alpine: dans le Bassin des dranses et dans quelques régions voisines*. Rouge.

Jacobs, Jane (1970). *The economy of cities.* Jonathan Cape.

Jaffe, Adam B. (1986). *Technological opportunity and spillovers of R&D: evidence from firms' patents, profits and market value*. National Bureau of Economic Research.

Jaffe, Adam B. (1989). "Real effects of academic research". In: *The American Economic Review* 79.5, pp. 957–970.

Jaffe, Adam B. and Manuel Trajtenberg (2002). *Patents, citations, and innovations: A window on the knowledge economy*. MIT Press.

Jaffe, Adam B., Manuel Trajtenberg, and Rebecca Henderson (1992). *Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations*. Working Paper 3993. National Bureau of Economic Research.

Jaro, Matthew A. (1989). "Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida". In: *Journal of the American Statistical Association* 84.406, pp. 414–420.

Jensen, Richard A., Jerry G. Thursby, and Marie C. Thursby (2003). "Disclosure and licensing of University inventions:'The best we can do with the s** t we get to work with'". In: *International Journal of Industrial Organization* 21.9, pp. 1271–1300.

Jeppesen, Lars Bo and Karim R. Lakhani (2010). "Marginality and problem-solving effectiveness in broadcast search". In: *Organization science* 21.5, pp. 1016–1033.

Kaplan, Sarah and Keyvan Vakili (2015). "The double-edged sword of recombination in breakthrough innovation: The Double-Edged Sword of Recombination". In: *Strategic Management Journal* 36.10, pp. 1435–1457.

Katz, J. Sylvan (1994). "Geographical proximity and scientific collaboration". In: *Scientometrics* 31.1, pp. 31–43.

Katz, J. Sylvan and Ben R. Martin (1997). "What is research collaboration?" In: *Research policy* 26.1, pp. 1–18.

Katz, Ralph and Thomas J. Allen (1982). "Investigating the Not Invented Here (NIH) syndrome: A look at the performance, tenure, and communication patterns of 50 R & D Project Groups". In: *R&D Management* 12.1, pp. 7–20.

Kenney, Martin and Donald Patton (2009). "Reconsidering the Bayh-Dole Act and the current university invention ownership model". In: *Research Policy* 38.9, pp. 1407–1422.

Kenney, Martin and Donald Patton (2011). "Does inventor ownership encourage university research-derived entrepreneurship? A six university comparison". In: *Research Policy* 40.8, pp. 1100–1112.

Kessels, Roselinde, Peter Goos, and Martina Vandebroek (2006). "A comparison of criteria to design efficient choice experiments". In: *Journal of Marketing Research* 43.3, pp. 409–419.

Keupp, Marcus Matthias and Oliver Gassmann (2013). "Resource constraints as triggers of radical innovation: Longitudinal evidence from the manufacturing sector". In: *Research Policy* 42.8, pp. 1457–1468.

Kilger, Christian and Kurt Bartenbach (2002). "New rules for German professors". In: *Science* 298.5596, pp. 1173–1175.

Kim, Linsu (1998). "Crisis construction and organizational learning: Capability building in catching-up at Hyundai Motor". In: *Organization science* 9.4, pp. 506–521.

King, Robert G and Ross Levine (1993). "Finance, entrepreneurship and growth". In: *Journal of Monetary economics* 32.3, pp. 513–542.

Kleer, Robin (2010). "Government R&D subsidies as a signal for private investors". In: *Research Policy* 39.10, pp. 1361–1374.

Koblank, Peter (2012). *Die Goering Speer-Verordnung. Arbeitnehmererfindungsrecht im Dritten Reich*. Deutsche Nationalbibliothek.

Kotha, Reddi Rayalu, Gerard George, and Kannan Srikanth (2012). "Bridging the Mutual Knowledge Gap: Coordination and the Commercialization of University Science". In: *SSRN Electronic Journal* 56.2, pp. 498–524.

Kuhn, Thomas S (2012). *The structure of scientific revolutions*. University of Chicago press.

Lacetera, Nicola and Lorenzo Zirulia (2008). "Knowledge spillovers, competition, and taste for science in a model of R&D incentive provision". In: *Universita'di Bologna, Working Paper*.

Lach, Saul and Mark Schankerman (2008). "Incentives and invention in universities". In: *The RAND Journal of Economics* 39.2, pp. 403–433.

LaFollette, Marcel Chotkowski (1992). *Stealing into print: fraud, plagiarism, and misconduct in scientific publishing*. Univ of California Press.

Lakhani, Karim R., Lars Bo Jeppesen, Peter Andreas Lohse, and Jill A. Panetta (2007). *The value of openess in scientific problem solving*. Division of Research, Harvard Business School.

Lam, Alice (2011). "What motivates academic scientists to engage in research commercialization:'Gold','ribbon'or 'puzzle'?" In: *Research policy* 40.10, pp. 1354–1368.

Lampe, Hannes W. and Dennis Hilgers (2015). "Trajectories of efficiency measurement: A bibliometric analysis of DEA and SFA". In: *European Journal of Operational Research* 240.1, pp. 1–21.

Landry, Réjean and Nabil Amara (1998). "The impact of transaction costs on the institutional structuration of collaborative academic research". In: *Research Policy* 27.9, pp. 901–913.

Lane, Peter J. and Michael Lubatkin (1998). "Relative absorptive capacity and interorganizational learning". In: *Strategic Management Journal* 19.5, pp. 461–477.

Laursen, Keld, Toke Reichstein, and Ammon Salter (2011). "Exploring the Effect of Geographical Proximity and University Quality on University–Industry Collaboration in the United Kingdom". In: *Regional Studies* 45.4, pp. 507–523.

Laursen, Keld and Ammon Salter (2006). "Open for innovation: the role of openness in explaining innovation performance among UK manufacturing firms". In: *Strategic management journal* 27.2, pp. 131–150.

Laursen, Keld and Ammon J. Salter (2014). "The paradox of openness: Appropriability, external search and collaboration". In: *Research Policy* 43.5, pp. 867–878.

Lawani, Stephen M. (1986). "Some bibliometric correlates of quality in scientific research". In: *Scientometrics* 9.1-2, pp. 13–25.

Lazarsfeld, Paul Felix, Neil W. Henry, and Theodore Wilbur Anderson (1968). *Latent structure analysis*. Houghton Mifflin Boston.

Lee, Sooho and Barry Bozeman (2005). "The impact of research collaboration on scientific productivity". In: *Social studies of science* 35.5, pp. 673–702.

Lee, You-Na, John P Walsh, and Jian Wang (2015). "Creativity in scientific teams: Unpacking novelty and impact". In: *Research Policy* 44.3, pp. 684–697.

Leiponen, Aija and Constance E. Helfat (2009). "Innovation objectives, knowledge sources, and the benefits of breadth". In: *Strategic Management Journal*, pp. 224–236.

Leptien, Christopher (1995). "Incentives for employed inventors: an empirical analysis with special emphasis on the German law for employee's inventions". In: *R&D Management* 25.2, pp. 213–225.

Leten, Bart, Paolo Landoni, and Bart Van Looy (2014). "Science or graduates: How do firms benefit from the proximity of universities?" In: *Research Policy* 43.8, pp. 1398–1412.

Liao, Chien Hsiang (2011). "How to improve research quality? Examining the impacts of collaboration intensity and member diversity in collaboration networks". In: *Scientometrics* 86.3, pp. 747–761.

Link, Albert N. and John T. Scott (2005). "Universities as partners in U.S. research joint ventures". In: *Research Policy* 34.3, pp. 385–393.

Link, Albert N. and Donald S. Siegel (2005). "University-based technology initiatives: Quantitative and qualitative evidence". In: *Research Policy* 34.3, pp. 253–257.

Lissoni, Francesco (2001). "Knowledge codification and the geography of innovation: the case of Brescia mechanical cluster". In: *Research Policy* 30.9, pp. 1479–1500.

Lissoni, Francesco, Patrick Llerena, Maureen McKelvey, and Bulat Sanditov (2008). "Academic patenting in Europe: new evidence from the KEINS database". In: *Research Evaluation* 17.2, pp. 87–102.

Lotka, Alfred James (1926). "The frequency distribution of scientific productivity." In: *Journal of Washington Academy Sciences* 16.12, pp. 317–323.

Louviere, Jordan J, Terry N Flynn, and Richard T Carson (2010). "Discrete choice experiments are not conjoint analysis". In: *Journal of Choice Modelling* 3.3, pp. 57–72.

Machlup, Fritz (1984). *Knowledge, Its Creation, Distribution, and Economic Significance: The branches of learning*. Princeton university press.

Macho-Stadler, Inés and David Pérez-Castrillo (2010). “Incentives in university technology transfers”. In: *International Journal of Industrial Organization* 28.4, pp. 362–367.

Maietta, Ornella Wanda (2015). “Determinants of university–firm R&D collaboration and its impact on innovation: A perspective from a low-tech industry”. In: *Research Policy* 44.7, pp. 1341–1359.

Mansfield, Edwin (1998). “Academic research and industrial innovation: An update of empirical findings”. In: *Research Policy* 26.7-8, pp. 773–776.

Maraut, Stéphane, Hélène Dernis, Colin Webb, Vincenzo Spiezia, Dominique Guellec, et al. (2008). *The OECD REGPAT Database: A Presentation*. Tech. rep. OECD Publishing.

Maritz, Alex, Axel Koch, and Marlen Schmidt (2016). “The Role of Entrepreneurship Education Programs in National Systems Of Entrepreneurship and Entrepreneurship Ecosystems”. In: *International Journal of Organizational Innovation (Online)* 8.4, p. 7.

Markman, Gideon D, Peter T Gianiodis, Phillip H Phan, and David B Balkin (2005). “Innovation speed: Transferring university technology to market”. In: *Research Policy* 34.7, pp. 1058–1075.

Marshall, Alfred (1898). *Principles of economics. Vol. 1*. Macmillan And Co., Limited; London.

Martin, Ben R and John Irvine (1983). “Assessing basic research: some partial indicators of scientific progress in radio astronomy”. In: *Research policy* 12.2, pp. 61–90.

McCallum, Andrew Kachites (2002). “MALLET: A Machine Learning for Language Toolkit”.

McFadden, Daniel (1973). “Conditional logit analysis of qualitative choice behavior”. In: *Frontiers in Econometrics*, pp. 105–142.

McLure Wasko, M. and Samer Faraj (2000). ““It is what one does”: why people participate and help others in electronic communities of practice”. In: *The Journal of Strategic Information Systems* 9.2-3, pp. 155–173.

Merton, Robert K (1973). *The sociology of science: Theoretical and empirical investigations*. University of Chicago press.

Miguélez, Ernest and Rosina Moreno (2015). “Knowledge flows and the absorptive capacity of regions”. In: *Research Policy* 44.4, pp. 833–848.

Moed, Henk F (2010). “Measuring contextual citation impact of scientific journals”. In: *Journal of Informetrics* 4.3, pp. 265–277.

Mohanty, Madhu S. (2001). “Testing for the specification of the wage equation: double selection approach or single selection approach”. In: *Applied Economics Letters* 8.8, pp. 525–529.

Monsen, Erik, Holger Patzelt, and Todd Saxton (2010). “Beyond Simple Utility: Incentive Design and Trade-Offs for Corporate Employee-Entrepreneurs”. In: *Entrepreneurship Theory and Practice* 34.1, pp. 105–130.

Moreno, Rosina, Raffaele Paci, and Stefano Usai (2005). “Spatial spillovers and innovation activity in European regions”. In: *Environment and Planning A* 37, pp. 1793–1812.

Morgan, Kevin and Philip Cooke (1998). *The associational economy: firms, regions, and innovation*.

Mowery, David C., Joanne Oxley, and Brian Silverman (1998). “Technological overlap and interfirm cooperation: implications for the resource-based view of the firm”. In: *Research Policy* 27.5, pp. 507–523.

Mowery, David C and Bhaven N Sampat (2005). “The Bayh-Dole act of 1980 and university-industry technology transfer: a model for other OECD governments?” In: *Essays in honor of Edwin Mansfield*, pp. 233–245.

Mowery, David C, Richard R Nelson, Bhaven N Sampat, and Arvids A Ziedonis (2001). “The growth of patenting and licensing by US universities: an assessment of the effects of the Bayh–Dole act of 1980”. In: *Research policy* 30.1, pp. 99–119.

Mukherji, Nivedita and Jonathan Silberman (2013). “Absorptive capacity, knowledge flows, and innovation in US metropolitan areas”. In: *Journal of Regional Science* 53.3, pp. 392–417.

Neckermann, Susanne, Reto Cueni, and Bruno S Frey (2009). *What is an award worth? An econometric assessment of the impact of awards on employee performance*. Tech. rep. CESifo working paper series.

Nelson, Richard R and Sidney G Winter (2009). *An evolutionary theory of economic change*. Harvard University Press.

Netzer, Oded, Olivier Toubia, Eric T Bradlow, Ely Dahan, Theodoros Evgeniou, Fred M Feinberg, Eleanor M Feit, Sam K Hui, Joseph Johnson, John C Liechty, et al. (2008). “Beyond conjoint analysis: Advances in preference measurement”. In: *Marketing Letters* 19.3-4, pp. 337–354.

Nielsen, Bo Bernhard and Sabina Nielsen (2009). “Learning and Innovation in International Strategic Alliances: An Empirical Test of the Role of Trust and Tacitness”. In: *Journal of Management Studies* 46.6, pp. 1031–1056.

Nilsson, Anna S, Annika Rickne, and Lars Bengtsson (2010). “Transfer of academic research: uncovering the grey zone”. In: *The Journal of Technology Transfer* 35.6, pp. 617–636.

Noble, David F (1979). *America by design: Science, technology, and the rise of corporate capitalism*. Oxford University Press, USA.

Nooteboom, Bart, Wim Van Haverbeke, Geert Duysters, Victor Gilsing, and Ad Van den Oord (2007). “Optimal cognitive distance and absorptive capacity”. In: *Research policy* 36.7, pp. 1016–1034.

OECD (2013). “Knowledge transfer channels and the commercialisation of public research”. In: *Commercialising Public Research*, pp. 17–23.

O’Shea, Rory P., Thomas J. Allen, Arnaud Chevalier, and Frank Roche (2005). “Entrepreneurial orientation, technology transfer and spinoff performance of U.S. universities”. In: *Research Policy* 34.7, pp. 994–1009.

Owen-Smith, Jason (2003). “From separate systems to a hybrid order: Accumulative advantage across public and private science at research one universities”. In: *Research Policy* 32.6, pp. 1081–1104.

Owen-Smith, Jason and Walter W Powell (2001). “To patent or not: Faculty decisions and institutional success at technology transfer”. In: *The Journal of Technology Transfer* 26.1-2, pp. 99–114.

Patel, Parimal and Keith Pavitt (1994). “National innovation systems: why they are important, and how they might be measured and compared”. In: *Economics of innovation and new technology* 3.1, pp. 77–95.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Perkmann, Markus, Zella King, and Stephen Pavelin (2011). “Engaging excellence? Effects of faculty quality on university engagement with industry”. In: *Research Policy* 40.4, pp. 539–552.

Perkmann, Markus and Kathryn Walsh (2009). “The two faces of collaboration: impacts of university-industry relations on public research”. In: *Industrial and Corporate Change* 18.6, pp. 1033–1065.

Phelps, Corey, Ralph Heidl, and Anu Wadhwa (2012). “Knowledge, networks, and knowledge networks a review and research agenda”. In: *Journal of Management* 38.4, pp. 1115–1166.

Piatetsky-Shapiro, Gregory (2012). "Big Data Hype (And Reality)". In: *Harvard Business Review*.

Pierre, Levy (1997). *Collective Intelligence: Mankind's Emerging World in Cyberspace*. Perseus Books.

Pieters, Rik GM (1988). "Attitude-behavior relationships". In: *Handbook of economic psychology*, pp. 144–204.

Piezunka, Henning and Linus Dahlander (2015). "Distant Search, Narrow Attention: How Crowding Alters Organizations' Filtering of Suggestions in Crowdsourcing". In: *Academy of Management Journal* 58.3, pp. 856–880.

Poetz, Marion K and Martin Schreier (2012). "The value of crowdsourcing: can users really compete with professionals in generating new product ideas?" In: *Journal of Product Innovation Management* 29.2, pp. 245–256.

Poland, Craig A, Rodger Duffin, Ian Kinloch, Andrew Maynard, William AH Wallace, Anthony Seaton, Vicki Stone, Simon Brown, William MacNee, and Ken Donaldson (2008). "Carbon nanotubes introduced into the abdominal cavity of mice show asbestos-like pathogenicity in a pilot study". In: *Nature nanotechnology* 3.7, pp. 423–428.

Polanyi, Michael (1966). *The Tacit Dimension*. University of Chicago Press.

Ponds, Roderik, Frank van Oort, and Koen Frenken (2007). "The geographical and institutional proximity of research collaboration". In: *Papers in Regional Science* 86.3, pp. 423–443.

Ponds, Roderik, Frank van Oort, and Koen Frenken (2010). "Innovation, spillovers and university-industry collaboration: an extended knowledge production function approach". In: *Journal of Economic Geography* 10.2, pp. 231–255.

Porter, Alan L, Jan Youtie, Philip Shapira, and David J Schoeneck (2008). "Refining search terms for nanotechnology". In: *Journal of nanoparticle research* 10.5, pp. 715–728.

Postigo, Hector (2003). "From Pong to Planet Quake: Post-industrial transitions from leisure to work". In: *Information Communication & Society* 6.4, pp. 593–607.

Powell, Walter W, Kenneth W Koput, and Laurel Smith-Doerr (1996). "Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology". In: *Administrative science quarterly*, pp. 116–145.

Rehurek, Radim and Petr Sojka (2010). "Software Framework for Topic Modelling with Large Corpora". In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50.

Roach, Michael and Henry Sauermann (2010). "A taste for science? PhD scientists' academic orientation and self-selection into research careers in industry". In: *Research Policy* 39.3, pp. 422–434.

Robertson, Stephen (2004). "Understanding inverse document frequency: on theoretical arguments for IDF". In: *Journal of documentation* 60.5, pp. 503–520.

Robin, Stéphane and Torben Schubert (2013). "Cooperation with public research institutions and success in innovation: Evidence from France and Germany". In: *Research Policy* 42.1, pp. 149–166.

Robinson, Richard D (1988). *The international transfer of technology: theory, issues, and practice*. Ballinger Publishing Company.

Rodan, Simon and Charles Galunic (2004). "More than network structure: how knowledge heterogeneity influences managerial performance and innovativeness". In: *Strategic Management Journal* 25.6, pp. 541–562.

Roessner, J. D. (1977). "Incentives to Innovate in Public and Private Organizations". In: *Administration & Society* 9.3, pp. 341–365.

Rogers, Everett M (2010). *Diffusion of innovations*. Simon and Schuster.

Romer, Paul (1989). *Endogenous technological change*. Tech. rep. National Bureau of Economic Research.

Roper, Stephen and Nola Hewitt-Dundas (2015). "Knowledge stocks, knowledge flows and innovation: Evidence from matched patents and innovation panel data". In: *Research Policy* 44.7, pp. 1327–1340.

Rose, J. and M. Bliemer (2009). *NGENE 1.0 User Manual & Reference guide*. Choice Metrics.

Rosenberg, Nathan (2004). "Innovation and economic growth". In: *Innovation and Economic Growth*.

Rosenberg, Nathan and Richard R Nelson (1994). "American universities and technical advance in industry". In: *Research policy* 23.3, pp. 323–348.

Rosenkopf, Lori and Paul Almeida (2003). "Overcoming Local Search Through Alliances and Mobility". In: *Management Science* 49.6, pp. 751–766.

Rosenkopf, Lori and Atul Nerkar (2001). "Beyond local search: boundary-spanning, exploration, and impact in the optical disk industry". In: *Strategic Management Journal* 22.4, pp. 287–306.

Rothaermel, Frank T, Shanti D Agung, and Lin Jiang (2007). "University entrepreneurship: a taxonomy of the literature". In: *Industrial and corporate change* 16.4, pp. 691–791.

Rothaermel, Frank T. and Warren Boeker (2008). "Old technology meets new technology: complementarities, similarities, and alliance formation". In: *Strategic Management Journal* 29.1, pp. 47–77.

Sahal, Devendra (1981). "Alternative conceptions of technology". In: *Research policy* 10.1, pp. 2–24.

Salton, Gerard and Michael J McGill (1986). *Introduction to modern information retrieval*. McGraw-Hill, Inc.

Samuelson, William and Richard Zeckhauser (1988). "Status quo bias in decision making". In: *Journal of Risk and Uncertainty* 1.1, pp. 7–59.

Sandmo, Agnar (2011). *Economics evolving: A history of economic thought*. Princeton University Press.

Saragossi, Sarina and Bruno van Pottelsberghe de la Potterie (2003). "What patent data reveal about universities: the case of Belgium". In: *The Journal of Technology Transfer* 28.1, pp. 47–51.

Sauermann, Henry and Michael Roach (2012). "Taste for science, taste for commercialization, and hybrid scientists". In: *DRUID Proceedings*. DRUID.

Sauermann, Henry and Paula E Stephan (2010). *Twins or strangers? Differences and similarities between academic and industrial science*. Tech. rep. NBER working paper.

Schildt, Henri A, Shaker A Zahra, and Antti Sillanpää (2006). "Scholarly communities in entrepreneurship research: a co-citation analysis". In: *Entrepreneurship Theory and Practice* 30.3, pp. 399–415.

Schumpeter, Joseph A (2013). *Capitalism, socialism and democracy*. Routledge.

Schumpeter, Joseph Alois (1934). *The theory of economic development: An inquiry into profits, capital, credit, interest, and the business cycle*. Vol. 55. Transaction publishers.

Segaran, Toby (2007). *Programming collective intelligence: building smart web 2.0 applications*. " O'Reilly Media, Inc."

Shane, Scott (2000). "Prior knowledge and the discovery of entrepreneurial opportunities". In: *Organization science* 11.4, pp. 448–469.

Shane, Scott and Sankaran Venkataraman (2000). "The promise of entrepreneurship as a field of research". In: *Academy of management review* 25.1, pp. 217–226.

Siegel, Donald S., David Waldman, and Albert Link (2003). "Assessing the impact of organizational practices on the relative productivity of university technology transfer offices: an exploratory study". In: *Research policy* 32.1, pp. 27–48.

Simon, Herbert A (1991). "Bounded rationality and organizational learning". In: *Organization science* 2.1, pp. 125–134.

Simonin, Bernard L (1999). "Ambiguity and the process of knowledge transfer in strategic alliances". In: *Strategic management journal* 20.7, pp. 595–623.

Singh, Jasjit (2005). "Collaborative Networks as Determinants of Knowledge Diffusion Patterns". In: *Management Science* 51.5, pp. 756–770.

Slaughter, Sheila and Larry L Leslie (1997). *Academic capitalism: Politics, policies, and the entrepreneurial university.* ERIC.

Small, Henry (1973). "Co-citation in the scientific literature: A new measure of the relationship between two documents". In: *Journal of the American Society for information Science* 24.4, pp. 265–269.

Small, Henry and Edwin Greenlee (1980). "Citation context analysis of a co-citation cluster: Recombinant-DNA". In: *Scientometrics* 2.4, pp. 277–301.

Smith, Mapheus (1958). "The trend toward multiple authorship in psychology." In: *American psychologist* 13.10, pp. 596–599.

So, Anthony D, Bhaven N Sampat, Arti K Rai, Robert Cook-Deegan, Jerome H Reichman, Robert Weissman, and Amy Kapczynski (2008). "Is Bayh-Dole good for developing countries? Lessons from the US experience". In: *PLoS Biology* 6.10, pp. 2078–2084.

Solla Price, Derek J de and Donald Beaver (1966). "Collaboration in an invisible college." In: *American psychologist* 21.11, pp. 1011–1018.

Solla Price, Derek John de, Derek John de Solla Price, Derek John de Solla Price, and Derek John de Solla Price (1986). *Little science, big science... and beyond*. Columbia University Press New York.

Solow, Robert M (1957). "Technical change and the aggregate production function". In: *The review of Economics and Statistics* 39.3, pp. 312–320.

Spradlin, D. (2012). "Are you solving the right problem? Asking the right questions is crucial." In: *Harvard Business Review* 90.9, pp. 84–101.

Stephan, Paula E and Sharon G Levin (1992). *Striking the mother lode in science: The importance of age, place, and time*. Oxford University Press.

Stern, Scott (2004). "Do scientists pay to be scientists?" In: *Management science* 50.6, pp. 835–853.

Stigler, George J. (1983). "Nobel Lecture: The Process and Progress of Economics". In: *Journal of Political Economy* 91.4, pp. 529–545.

Straus, Joseph (2000). *Expert opinion on the introduction of a grace period in the European patent law: Submitted upon request of the European Patent Organisation*. Max Planck Inst. for Foreign, Internat. Patent, Copyright, and Competition Law.

Stuart, Toby E (2000). "Interorganizational alliances and the performance of firms: A study of growth and innovation rates in a high-technology industry". In: *Strategic management journal* 21.8, pp. 791–811.

Stuart, Toby E. and Joel M. Podolny (2007). "Local search and the evolution of technological capabilities". In: *Strategic Management Journal* 17.S1, pp. 21–38.

Surowiecki, James (2004). *The wisdom of crowds: why the many are smarter than the few and how collective wisdom shapes business, economics, society and nations*. Anchor.

Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar (2006). *Introduction to data mining*. Pearson Education India.

Tarasconi, Gialuca and Byeongwoo Kang (2015). "PATSTAT revisited". In:

Team, Sci (2009). *Science of Science (Sci2) Tool. Indiana University and SciTech Strategies*.

Ter Wal, Anne LJ, Oliver Alexy, Jörn Block, and Philipp G Sandner (2016). "The Best of Both Worlds The Benefits of Open-specialized and Closed-diverse Syndication Networks for New Ventures' Success". In: *Administrative Science Quarterly* 61.3, pp. 393–432.

Terwiesch, Christian and Yi Xu (2008). “Innovation contests, open innovation, and multiagent problem solving”. In: *Management science* 54.9, pp. 1529–1543.

Thomas, James B., Stephanie Watts Sussman, and John C. Henderson (2001). “Understanding “Strategic Learning”: Linking Organizational Learning, Knowledge Management, and Sensemaking”. In: *Organization Science* 12.3, pp. 331–345.

Thursby, Jerry G and Marie C Thursby (2011). “Has the Bayh-Dole act compromised basic research?” In: *Research Policy* 40.8, pp. 1077–1083.

Tobler, Waldo R (1970). “A computer movie simulating urban growth in the Detroit region”. In: *Economic geography* 46.sup1, pp. 234–240.

Tomlinson, Philip R. (2010). “Co-operative ties and innovation: Some new evidence for UK manufacturing”. In: *Research Policy* 39.6, pp. 762–775.

Train, Kenneth E. (2009). *Discrete choice methods with simulation*. Cambridge University Press.

Tsai, Wenpin (2001). “Knowledge transfer in intraorganizational networks: Effects of network position and absorptive capacity on business unit innovation and performance”. In: *Academy of management journal* 44.5, pp. 996–1004.

Tunzelmann, Nick von (2009). “Regional capabilities and industrial regeneration”. In: *Technological Change and Mature Industrial Regions: Firms, Knowledge and Policy*, pp. 11–28.

Uzzi, Brian (1996). “The sources and consequences of embeddedness for the economic performance of organizations: The network effect”. In: *American sociological review* 61.4, pp. 674–698.

Uzzi, Brian, Satyam Mukherjee, Michael Stringer, and Ben Jones (2013). “Atypical combinations and scientific impact”. In: *Science* 342.6157, pp. 468–472.

Valiquette, Claude AM, Pierre Valois, Raymond Desharnais, and Gaston Godin (1988). “An item-analytic investigation of the Fishbein and Ajzen multiplicative scale: The problem of a simultaneous negative evaluation of belief and outcome”. In: *Psychological reports* 63.3, pp. 723–728.

Van Looy, Bart, Paolo Landoni, Julie Callaert, Bruno Van Pottelsberghe, Eleftherios Sapsalis, and Koenraad Debackere (2011). “Entrepreneurial effectiveness of European universities: An empirical assessment of antecedents and trade-offs”. In: *Research Policy* 40.4, pp. 553–564.

Van Noorden, Richard et al. (2015). “Interdisciplinary research by the numbers”. In: *Nature* 525.7569, pp. 306–307.

Von Hippel, Eric (1994). ““Sticky information” and the locus of problem solving: implications for innovation”. In: *Management science* 40.4, pp. 429–439.

Von Hippel, Eric (2005). “Democratizing innovation: The evolving phenomenon of user innovation”. In: *Journal für Betriebswirtschaft* 55.1, pp. 63–78.

Vuong, Quang H. (1989). “Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses”. In: *Econometrica* 57.2, pp. 307–333.

Walter, Thomas, Christoph Ihl, René Mauer, and Malte Brettel (2013). “Grace, gold, or glory? Exploring incentives for invention disclosure in the university context”. In: *The Journal of Technology Transfer*, pp. 1–35.

Wang, Chunlei, Simon Rodan, Mark Fruin, and Xiaoyan Xu (2014). “Knowledge networks, collaboration networks, and exploratory innovation”. In: *Academy of Management Journal* 57.2, pp. 484–514.

Wang, Jian (2014). “Unpacking the Matthew effect in citations”. In: *Journal of Informetrics* 8.2, pp. 329–339.

Welsh, Rick, Leland Glenna, William Lacy, and Dina Biscotti (2008). "Close enough but not too far: Assessing the effects of university–industry research relationships and the rise of academic capitalism". In: *Research Policy* 37.10, pp. 1854–1864.

Wilhelm, Sven Karl-Heinrich (2010). "Evidence on individual decision making in the context of patenting activities in German academia". PhD thesis. RWTH Aachen University.

Will, Birgit E, Roland Kirstein, et al. (2002). *Effiziente Vergütung von Arbeitnehmererfindungen. Eine ökonomische Analyse einer deutschen Gesetzesreform.* Tech. rep. Saarland University, CSLE-Center for the Study of Law and Economics.

Williams, Katherine Y and Charles A O'Reilly III (1998). "A review of 40 years of research". In: *Res Organ Behav* 20, pp. 77–140.

Winkler, William E. (1999). *The State of Record Linkage and Current Research Problems*. Tech. rep. Statistical Research Division, U.S. Census Bureau.

Woerter, Martin (2012). "Technology proximity between firms and universities and technology transfer". In: *The Journal of Technology Transfer* 37.6, pp. 828–866.

Wuyts, Stefan, Massimo G. Colombo, Shantanu Dutta, and Bart Nooteboom (2005). "Empirical tests of optimal cognitive distance". In: *Journal of Economic Behavior & Organization* 58.2, pp. 277–302.

Youtie, Jan, Maurizio Iacopetta, and Stuart Graham (2008). "Assessing the nature of nanotechnology: can we uncover an emerging general purpose technology?" In: *The Journal of Technology Transfer* 33.3, pp. 315–329.

Zahra, Shaker A and Gerard George (2002). "Absorptive capacity: A review, reconceptualization, and extension". In: *Academy of management review* 27.2, pp. 185–203.

Zaleznick, Abraham (1985). *Organizational reality and psychological necessity in creativity and innovation.*

Zhao, Liming and Arnold Reisman (1992). "Toward meta research on technology transfer". In: *IEEE Transactions on engineering management* 39.1, pp. 13–21.

Zucker, Lynne G and Michael R Darby (2006). *Movement of star scientists and engineers and high-tech firm entry*. Tech. rep. National Bureau of Economic Research.

Zucker, Lynne G., Michael R. Darby, and Jeff Armstrong (1998). "Geographically Localized Knowledge: Spillovers Or Markets?" In: *Economic Inquiry* 36.1, pp. 65–86.

Zucker, Lynne G., Michael R. Darby, and Marilynn B. Brewer (1998). "Intellectual human capital and the birth of US biotechnology enterprises." In: *American Economics Review* 88.1, pp. 290–306.

Zucker, Lynne G., Michael R. Darby, Jonathan Furner, Robert C. Liu, and Hongyan Ma (2007). "Minerva unbound: Knowledge stocks, knowledge flows and new knowledge production". In: *Research Policy* 36.6, pp. 850–863.

Zuniga, P, D Guellec, H Dernis, M Khan, T Okazaki, and C Webb (2009). *OECD Patent Statistics Manual*. OECD.

Zuur, Alain F, Elena N Ieno, Neil J Walker, Anatoly A Saveliev, and Graham M Smith (2009). "Zero-truncated and zero-inflated models for count data". In: *Mixed effects models and extensions in ecology with R*, pp. 261–293.