MSc Thesis:

# Proximities in research collaboration within biotechnology


| | |
|---|---|
| Author: | Bart B. Lauret, BSc |
| Address: | M. v. Regteren Altenahof 1<br>1183 BT Amstelveen<br>The Netherlands |
| Student number: | 3179346 |
| E-mail: | B.B.Lauret@students.uu.nl |
| Programme: | Science and Innovation<br>Management, Utrecht University |
| ECTS: | 45 |
| Supervisor: | dr. G. J. Heimeriks |
| Second reader: | dr. F. J. van Rijnsoever |


8th February 2013

**Universiteit Utrecht**

**Abstract**

This research aims to find out the importance of proximity between research collaborators within a science-intensive sector. This research uses the biotechnological field. This is done by using regression analysis on scientometric data from Web of Science and using network analysis tools to analyse the citation impact of publications. There are five different dimensions of proximity distinguished: cognitive, organisational, social, institutional, and geographical. In addition to this, four of these five dimensions of proximity are studied on their relative importance of nearness. This paper shows that for all the dimensions of proximity between the authors of a publication under study, the organisational dimension and the cognitive dimension show a significant relationship to the quality of the research in terms of citation impact.

**Acknowledgements**

I would like to take this opportunity to acknowledge and thank those who made this work possible. First I thank my supervisor, Gaston Heimeriks. Although it has been a long journey, he constantly provided me with constructive criticism. To my second reader, Frank van Rijnsoever, for his helpful remarks during the first phase of this research I would like to say thank you as well and even more for the remarks the second time he reviewed my thesis.

Foremost I would like to express my sincerest appreciation to my family. They have supported me through all my years through university. They always stayed interested in the progress of this thesis. Although I find it hard, I would like to thank my mother in the first place. She always was there for me, believed in me, and was a good mother to me and my brothers. It goes without saying that I miss her more and more. Losing her did not make it easy to keep it together and stay focussed on something that all of a sudden seemed so futile. But finally I made it and I know she would be extremely proud of me.

To my father I would like to show my gratitude for he tried to understand and help me further with statistical questions and also for supporting me throughout my years of study.

I would also like to say thank you to Mathijs, Chris, Annelieke, Louise, Emmeline, Serena and Frank for their support and involvement in many ways. Above all, thanks for asking. Frank, thanks for your help with R.

And to Bettina, thanks for the papers that were a welcome beginning of the day. For taking care of our daughters when we needed to be elsewhere. And for asking in what way help would be appreciated.

Most importantly, to my wife, Florentine, I wish to offer my deepest thanks. I am glad I found you. Words do not do justice to all what you have done for me. We have gone through a lot different challenges of life, but together we made this possible!

Last but not least, I would like to dedicate this work to my three daughters. Feline, Berit, and Bieke, you make life worth living. Being at home with you made writing a bit more challenging from time to time, but on the other hand thank you for all the opportunities you provided to give me some fresh air on the playground to think this all through (or not). And finally I am grateful that I was able to participate so much in our daily life together, despite the constant sustained responsibility of graduating.

# Contents

# Chapter 1

# Introduction

## 1.1 Problem description

A lot of innovations in current society are derived from progress in science and technology. As such, how the progress of science evolves gets a lot of attention from innovation science. Science-based industries rely on a high number of innovations based on scientific knowledge (Pavitt, 1984).

It is said that science is in transformation in the last two decades. The concept of 'Mode 2' knowledge production tries to understand these changes (Scott et al., 1994). Where Mode 1 is academic driven, disciplinary, homogeneous, autonomous and peer-reviewed; Mode 2 knowledge is produced in the context of the application, transdisciplinary, heterogeneous, reflexive (researchers regard themselves socially accountable), and uses a novel quality control (Hessels & Lente, 2008). *"While knowledge production used to be located primarily in scientific institutions and structured by scientific principles, its locations, practices and principles are now much more heterogeneous"* (Hessels & Lente, 2008, p. 740). Moreover, Mode 2 science production transforms to more transdisciplinary collaborations between authors (Hessels & Lente, 2008).

Another model which tries to understand these changes in knowledge production is the Triple Helix model (Etzkowitz & Leydesdorff, 2000). In this model more and more collaboration is taking place between academia, industry and government, instead of disciplinary knowledge production. The complexity in the science-based sector is so high that most innovations come from inter-organisational collaboration (Powell, Koput & Smith-Doerr, 1996).

> *The Triple Helix thesis states that the university can play an enhanced role in innovation in increasingly knowledge-based societies. The underlying model is analytically different from the*

> *National Systems of Innovation (NSI) approach (see Lundvall, 1988; Freeman, 1997), which considers the firm as having the leading role in innovation. [...] The Triple Helix model focusses on the network overlay of communications and expectations that reshape the institutional arrangements among universities, industries, and governmental agencies.* (Etzkowitz & Leydesdorff, 2000, p. 109)

In figure 1.1 the relations between the different institutions are depicted. Triple helix collaborations take place in the center, where all these different kind of institutions participate.

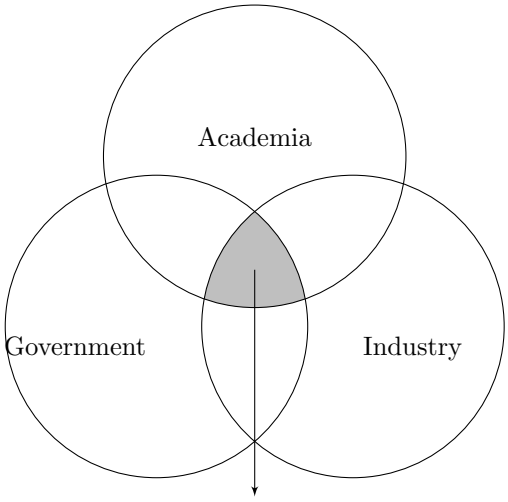



**Figure 1.1:** The Triple Helix Model of University-Industry-Government relations (Etzkowitz & Leydesdorff, 2000)

Moreover, with the growing specialisation within the scientific field and the progressive professionalisation (Cronin et al., 1998), it is becoming increasingly difficult for a researcher to possess the all necessary skills and technologies to solve problems by himself (David, 1994). This means that scientific collaboration becomes increasingly important, because the knowledge of the single researchers is combined and the whole gamut of skills necessary to conduct research is covered.

To examine if researchers collaborate between the institutional boundaries (see Triple Helix), geographical locations (see Mode 2), or across specialisations, these factors have to be examined in the context of collaboration between researchers. More specifically, these previous factors can be understood as a distance (or proximity) between these collaborators. Boschma (2005) devised a set of five dimensions of proximities between these researchers.

It is argued that geographical proximity in scientific collaboration is an important factor (Hoekman et al., 2009; Boschma, 2005; Boschma & Frenken, 2010), when looking at these interactions between collaborators. The main reason this has been argued as an important mechanism is because of face–to–face interactions, which are a prerequisite of exchange of tacit knowledge (i.e. knowledge which cannot be formulated, see e.g. Collins, 1974). But also non-spatial dimensions of proximity (cognitive, organisational, social, and institutional) play an important role (Boschma, 2005; Ponds et al., 2007; Nooteboom et al., 2007). Originally, Torre and Gilly (2000) have proposed that proximity covers a number of dimensions. This has been extended by Boschma (2005) to the following five dimensions of proximity. In this paper these dimensions of proximity are under study.

*Institutional proximity* measures if the institutions where the researchers are working for are similar or not. For collaborators to be institutional close, they should all work at the same type of organisation, for instance an university. *Organisational proximity* could be defined by the fact if researchers work for the same organisation or not. *Social proximity* has its roots in a more personal sphere and comprises of friendship or experience of repeated interaction (Boschma, 2005). As such, collaborators are socially close if they know each other personally and thus maintain personal relations. *Geographical proximity* covers the most classic view on proximity, namely the spatial nearness. Lastly, *cognitive proximity* can be described as two actors having the same technological paradigm or technological frame (see Pinch & Bijker, 1984). This means they work within the same paradigm of science, which Kuhn defines as *"universally recognized scientific achievements that, for a time, provide model problems and solutions for a community of researchers"* (Kuhn, 1996, p. 10). It is worth noting that it is possible that multiple dimensions of proximity can coexist within one scientific collaboration, and it is even probable that these dimensions correlate to one another. For instance, if collaborators work at the same university, they are often (but not necessarily so) geographically close.

To measure if these collaborations are successful, the citation impact of the publications is used which is ultimately the result of such collaborations. This is the same approach as Frenken, Ponds and Van Oort (2010) have used. This means that for this study the scientific contribution as measured in the number of citations received, is the discriminatory variable to be able to discern if these proximities between collaborators have any effect. Moreover, because publications have often more than two authors, these proximities will be aggregated per publication.

This research uses data on publications about the field of biotechnology. An important aspect in selecting this field, was that it is science-based according to the taxonomy of Pavitt (1984). Biotechnology is defined by the *UN*

*Convention on Biological Diversity* as *"any technological application that uses biological systems, living organisms, or derivatives thereof, to make or modify products or processes for specific use"* (Secretariat of the CBD, 1992, p. 3). For this research the field which served as a source of the data is secondary to the meta data about the publications, because this study looks at the interactions between researchers and the resulting publications. This research does not use the content of these publications, but rather the meta data of these papers and so the biotechnological subject could be considered as exchangeable for, for instance, nanotechnology, which is a field of study in a similar phase and has similar characteristics.

## 1.2 Research question

Against this background, a study exploring the effects of the various dimensions of proximity of collaborators in relation to the citation impact of the result of that collaboration can help to understand the dynamics within scientific progress. This leads to the following research question:

- How do cognitive, organisational, social, institutional, and geographical proximity between researchers influence the citation impact of the resultant publications in a science-based field?

## 1.3 Justification

This research is embedded in innovation research using scientometric methods. Data is gathered from the biotechnological field as this field has a strong connection to technological change and is in high flux since the emergence of this field of research (Frenken et al., 2010). Biotechnological research can be seen as research in 'Mode 2', where the research is focussed on real problem-solving of practical problems through scientific means (Scott et al., 1994), instead of 'Mode 1' research, which is academic, investigator-initiated and discipline-based.

Because the geographical proximity between researchers is considered within this study, the scope of this study should be on a global scale, which means that there is no artificial boundary on which publications are included in this study based on geographic properties. Therefore, publications from around the world are included in this study and used for the analysis.

This study has both scientific and societal relevance. The scientific contribution is made by that these proximities have not yet been examined together in a citation impact analysis using advanced statistical methods. This paper

builds on a part of the methods of Frenken et al. (2010), which has also used citation impact to assess research quality. Furthermore, Boschma (2005) has defined all dimensions of proximity between collaborators under study in this paper. And this research uses the same data as the paper by Heimeriks and Boschma (2012), although the methods used and research subject differ.

The societal contribution lies within the gathering of knowledge about how to collaborate in science-intensive environments. Because a number of industries heavily rely on science to be able to innovate, the circumstances under which collaborations should take place is valuable. To be able to assist in scientific and technological progress and to be able to encourage scientific collaboration, it is instrumental to know if knowledge transfer and spillover from universities to other universities and industry or government can be facilitated through policy changes. It is also important to know for policy makers whether the proximity factors under study are able to make a significant difference. In case these factors prove to be significant, policy can be adjusted in order to create better circumstances for collaborators within research to better the progress of science and technological change.

# Chapter 2

# Theoretical Framework

## 2.1 Background

Technological innovation is important for industrial competitiveness (Dosi, 1988). The innovative process entails an intrinsically uncertain activity of search and problem-solving (Dosi, 1988). This means that this activity is based upon *"varying combinations of public and private (people-specific or firm-specific) knowledge, general scientific principles and rather idiosyncratic experience, well-articulated procedures and rather tacit competences"* (Dosi, 1988, p. 223). In other words, innovation needs a very wide gamut of factors to have a chance of success. This success being progress in technological innovation.

Dosi (1988) also introduces the concept of technological paradigm. This can be understood as the body of knowledge which guides these search and development activities. This technological paradigm is often shared by the entire community of technological and economic actors as the basis upon which one looks for technological innovations, which means improvements in process efficiency and product performances.

Dosi (1982) states that the interplay between scientific advances, economic factors, institutional variables, and unsolved difficulties on established technological paths are necessary for the emergence of new technological paradigms. In other words, this leads to the notion that scientific change combined with these other factors lie at the root of technological change itself. As such, technological change is in tandem with scientific change. Therefore, one can not exist without the other.

Leydesdorff (2000); Etzkowitz and Leydesdorff (2000) state in their Triple Helix thesis that universities can play an enhanced role in innovation in increasing knowledge-based societies. This is related to the concept of research in 'Mode 2' (where the research is focussed on real problem-solving

of practical problems through scientific means) instead of 'Mode 1' research, which is academic, investigator-initiated and discipline-based (Scott et al., 1994). Therefore, with Mode 2 research, more and more collaboration is taking place in the production of knowledge, and with a stronger innovation or application context. In other words, collaboration is increasingly taking place with different types of organisations when involved with Mode 2 knowledge production.

Thus it is getting more and more important for a scientist to be working within a network, where complementary skills exist (David, 1994). This could be collaborations that cross between academia, industry, or government (Etzkowitz & Leydesdorff, 2000). Working within a network is important because scientists are getting more specialised (Cronin et al., 1998), and it is thus harder for a scientist to have all the skills needed to progress in science on his own. This leads to the conclusion that the network of relationships between academia, industry and government is getting more important in today's society.

By claiming that technological innovation progresses through scientific advances, it is worthwhile to a look at the dynamics of scientific progress. The next sections try to give an overview of some of these aspects of scientific progress in the context of scientific publications. Section 2.2 provides the foundation used to measure the success of collaborations. Section 2.3 looks at how the interactions between the authors of the publications affect scientific progress.

## 2.2 Citation impact

To be able to understand the successfulness of a collaboration of researchers, it is worthwhile to look at the result of such a collaboration: a publication. A common measure used to determine the quality of a publication is the citation impact (Garfield, 1964). In this case this would mean the number of citations that a publication has received at the time the data was collected. As said before, this is the same approach as Frenken et al. (2010) have used. The citation impact has a number of advantages, such as the relatively ease of data gathering and the comparability in time. But through the years a number of criticisms have arisen when using citation impact as a measure of quality of a publication (MacRoberts & MacRoberts, 1996; Wouters, 1998).

- Citation patterns can differ across scientific fields or specialities. This means that this measure can not be used to compare between different fields.

- Citations do not necessarily cite influences. This could be because of informal communication or because the influence is captured in footnotes.
- Secondary sources are preferred, which means that there is a bias towards review papers. Credit is frequently not given to the discoverer, but rather to a secondary source.
- A cited publication can contain another claim than the author suggests, in other words a miscitation.
- Not all citations are captured in the Web of Science database, which has a known Anglo-Saxon bias. This means that relatively more Anglo-Saxon journals are indexed.

Because this study only studies a single field, this first criticism does not apply. The other deserve attention, but because the sample size in this study is quite large, these deviations appear as noise in the analysis (Frenken et al., 2010). Moreover, method to deal with these problems have not been developed and a better source of data has not been found.

## 2.3 Dimensions of proximity

As mentioned earlier, Boschma (2005) distinguishes five dimensions of proximity: cognitive, organisational, social, institutional, or geographical. These proximities could explain the success of knowledge transfer, which is ultimately the base of progress in science. For instance, Jaffe et al. (1993) argues that the knowledge spillovers from universities are spatially bound, where those spillovers *"exert a significant and positive effect on knowledge output as measured by patents or innovations"* (Ponds et al., 2007, p. 425). See Archibugi and Pianta (1996) for a discussion on measuring knowledge output by using scientometric data.

### 2.3.1 Cognitive proximity

To be able to identify, interpret and exploit new knowledge, absorptive capacity is needed (Cohen & Levinthal, 1990; Nooteboom et al., 2007). In other words, this means that to be able to absorb this new knowledge, it is necessary to have a certain level of understanding on the matter. The difference in understanding that exists between two collaborators could be described as cognitive distance. It is argued that successful scientific collaboration is best when the cognitive distance is neither small nor large, and thus has an inverse U-shape (Nooteboom et al., 2007). If the difference in mindset or

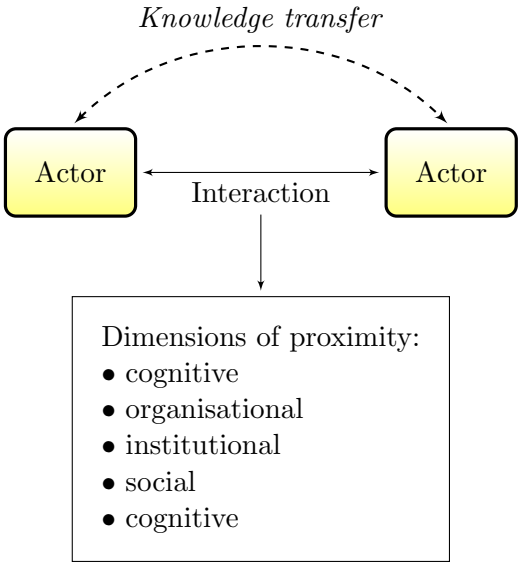



**Figure 2.1:** Dynamics between collaborators

technological frame is too small, the collaboration is not optimal, because both researchers can not effectively put in something new. On the other hand, if the difference is too large, researchers are not able to collaborate well neither due to lack of understanding of each other. This presumes that there is an optimum cognitive distance between researchers (see figure 2.2).

**Hypothesis 1.** *Collaborators with a optimal cognitive proximity (inverted U-shape relationship) produce a higher quality research.*

### 2.3.2 Organisational proximity

Organisational proximity is defined by Boschma (2005) as the extent to which relations are shared within an organisational arrangement. On the one hand, organisational proximity is assumed to be beneficial for innovation as it reduces uncertainty, because of perceived opportunism.

**Hypothesis 2.** *Collaborators with a high organisational proximity produce a higher quality research.*

But on the other hand, the Triple Helix thesis and the arguments provided by Mode 2 knowledge production hint towards a positive effect of inter-organisational collaboration. An important aspect of the science-based sector concerns the distributed nature of innovation processes. Most innovations

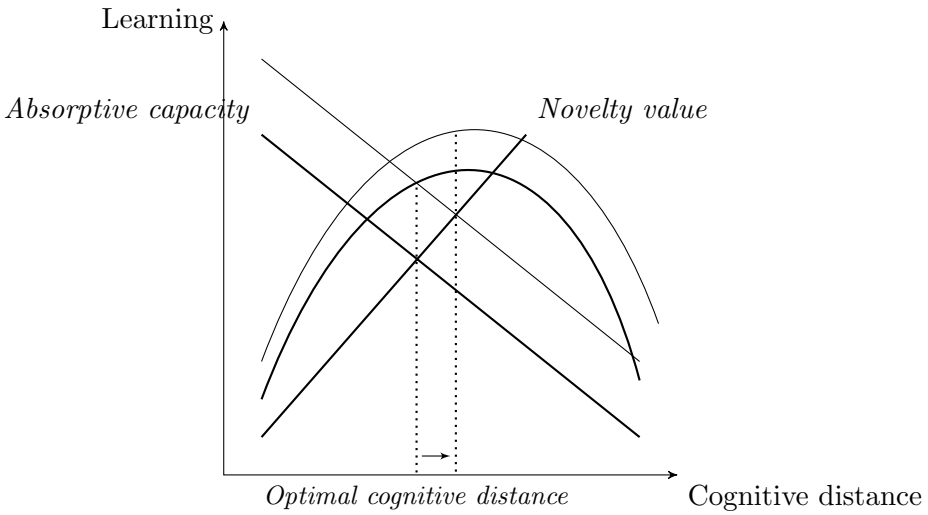



**Figure 2.2:** Optimal cognitive distance (Nooteboom et al., 2007)

come from inter-organizational collaboration and such network are also called the 'locus of innovation' in science-based industries (Powell et al., 1996; Bonaccorsi, 2008). An alternative hypothesis is used for this aspect:

**Hypothesis 3.** *Collaborators with a low organisational proximity produce a higher quality research.*

### 2.3.3 Social proximity

Social proximity measures relationship within the personal sphere of the researchers (Boschma & Frenken, 2010) and is therefore not measurable with the data available, because friendship does not express itself into the meta data available. Furthermore, it is not workable to gather this information for all agents included in the database. This dimension is therefore left out in this analysis.

### 2.3.4 Institutional proximity

Institutional proximity is defined as nearness in an institutional sense. For instance, university–university collaborations can be considered as institutionally nearby, as are industry–industry collaborations. It is argued that geographical proximity can possibly overcome a low institutional proximity (e.g. academia-industry collaboration) (Ponds et al., 2007). Because the agents then have a higher chance of possessing a common mindset or

technological paradigm, institutional proximity can facilitate in better communication. Boschma (2005) argues that it can even compensate for a lack of geographical proximity and vice versa can institutional proximity facilitate communication more easily over long geographical distances, as said before. Rallet and Torre (1999) has shown that tacit knowledge may be transmitted across large distances through other forms of proximity. When collaborators share a cognitive experience or are coordinated by a central authority (organisation proximity), the need for geographical proximity is shown to be rather weak.

**Hypothesis 4.** *Collaborators with a high institutional proximity produce a higher quality research.*

As with organisational proximity, the Triple Helix thesis states that inter-institutional collaborations are needed for more successful collaboration in science-intensive industries. As such, a alternative hypothesis is used to test this:

**Hypothesis 5.** *Collaborators with a low institutional proximity produce a higher quality research.*

### 2.3.5 Geographical proximity

To facilitate scientific collaboration and knowledge exchange between co-authors, geographical proximity is often cited as an important factor. This could be explained by the exchange of tacit knowledge and face-to-face contact between the collaborators (Audretsch & Feldman, 1996). Moreover, the science-base sector tends to concentrate themselves spatially (Paci & Usai, 2000). Hoekman et al. (2009) argues that in spite of globalisation, geographical proximity is an important cause of network formations. However, scholars disagree in the importance of this dimension of proximity (see Boschma, 2005). It is argued that the other dimensions of proximity could be at least as important as geographical proximity, because it could be possible to overcome a large geographical distance with one of the other dimensions of proximity.

**Hypothesis 6.** *Collaborators with a high geographical proximity produce a higher quality research.*

# Chapter 3

# Methodology

This research uses network analysis tools and quantitative analysis on the co-authorship data from an academic citation index to asses the proximities within scientific collaborations. This chapter starts with the source of the data, this is then followed by the operationalisation of the variables in this study and finally a more comprehensive foundation of the method used will be discussed.

## 3.1 Data

The data in this study has been gathered from *Web of Science*, a product offered by Thomson Scientific. This is an online academic citation index, which indexes data about publications from more than 10,000 academic journals. This database covers all major journals in the world with an Anglo-Saxon bias for the years 1988 and onwards. Web of Science allows the user to search for certain keywords or scientific disciplines and to subsequently save the results. In this case, these results have then been saved to a text format which can subsequently be parsed by the tools from Leydesdorff[1] to create a relational database. This database contains data on authors, title, keywords, journal, date of publication, citations, and affiliations, in other words: metadata on academic articles. This relational database then can be used to extract the data which will be used for the network analysis and the statistical analysis, where each publication is seen as an observation.

For this study all articles from the journals *Biotechnology and Bioengineering*, *Journal of Biotechnology* and *Biotechnology Progress* from the period 1985–2009 were selected. The number of articles from these journals are

[1] See `http://www.leydesdorff.net/software/coauth/index.htm` for more information about the tools used.

respectively 8132, 3289, and 5369 articles which sums up to 16,790 articles. These journals have an impact factor of respectively 3.946, 3.045, and 2.340[2]. These are the same jounals as Heimeriks and Boschma (2012) have selected.

## 3.2 Operationalisation

To be able to get the necessary data from the database, some operations on the dataset are necessary. This operationalisation translates the concepts from the theoretical framework into measurable variables.

### 3.2.1 Dependent variable: citation impact

The citation impact is defined as the number of citations a specific publication receives. This can measure the quality of the product of the co-authors. If a publication is central in the field of research or is well embedded within its field, the subsequent chance of more connections rises. The degree centrality is calculated using *Gephi*, a social network analysis computer program (Bastian et al., 2009).

The method to create the references network starts with the list of references per publication. This list of references contains the name of the author, initials, journal, page number, and volume per publication. The whole database is then searched using this information to find a match. As databases can contain spelling errors, or cases of ambiguous naming of for instance institutions, a fuzzy search is used. Moreover, matches are found according to the extent of similarity on all previously mentioned data fields. In other words, per publication all references are parsed and looked up in the database to find a match.

This match creates then a directional link between the publication where the link points to the publication that receives the reference. Thus, if a publication has been cited often, it means that this publication has a large citation impact and in proxy, has a high quality.

### 3.2.2 Cognitive proximity

Cognitive proximity is measured by using a co-word analysis. This is done by counting how many words co-occur in publication titles of co-authors, excluding the publication in which the co-authorship occurs (see table 3.1). To prevent that words that are often used, the so called stop words, such as for instance *the, is, at, which* and *on*, create noise in this analysis they are

[2]2011 Journal Citation Reports Science Edition

excluded. The list of stopwords which were excluded amounts to 801 words. As a result of this filter only the content-specific words remain, for only these content-specific words are interesting for the assessment of cognitive distance.

As such, this variable measures the use of certain words, which can indicate that co-authors have a high cognitive proximity or not. The number of words which co-occur is then converted into a ratio in comparison to the number of words in total that those co-authors use in their publication titles. Because publications often have more than two authors, for every co-authorship pair of a publication this number is then calculated and averaged. As such, it yields a number between 0.0 and 1.0, where a zero means that all the co-authorship pairs have no words in common and a 1.0 means that they have all words in common. This last case is a possibility, if the co-authors only have collaborated together and thus have not authored publications outside of the collaboration of the publication for which this number is calculated in the first place.

This calculating procedure is extremely laborious, as it means that for every co-authorship the entire database has to be queried for words that can co-occur.

**Table 3.1**
Cognitive distance is calculated by dividing the common words by the total unique words (excluding the publication for which it is calculated; these are greyed out in the table)

| | Publications | | | | |
|---|---|---|---|---|---|
| | From author A | | | From author B | |
| Authors | A,D,E | A,C | A,B | B,A | B,F |
| Words | $\zeta\eta$ | $\alpha\delta\epsilon$ | $\alpha\beta\gamma$ | $\alpha\beta\gamma$ | $\alpha\gamma$ |
| Result | One word in common: $\alpha$<br>Total unique words: $\alpha\gamma\delta\epsilon\zeta\eta$<br>Cognitive distance: $1/6 = 0.167$ | | | | |

### 3.2.3 Organisational proximity

Organisational proximity can be measured by comparing the organisation, that is provided in the correspondence information of a certain publication, for all the authors.

For every co-authorship in a publication it is determined if the short descrip-

tion[3] of the institution is the same across the co-authors. This will be again counted as a 0, and a 1 otherwise. In other words, in case of a 0, this means that both co-authors are linked to the same organisation.

This search is done using a fuzzy search algorithm, which means that both text fields are matched approximately, instead of exactly. This is important to still be able to get a match even when spelling mistakes are prevalent.

Subsequently, for all unique co-authorship pairs of a publication, all the pairs which work at different organisations (i.e. every pair that scores a 1) are counted, and subsequently divided by two, because else all co-author pair distances would be double counted. What this variable actually measures is organisational *distance*, where a 0 means that all co-authorship pairs work at the same organisation.

Moreover, this variable has no upper-bound, because this would be dependent on the number of co-authorships on a publication. This is done like this, because the collaborations are an individual matter, and therefore the organisational distance deserves to be summed per collaborator, instead of counting all the different organisations per publication. Furthermore, this assumes that the degree of inter-organisational collaboration is parallel to the increase of difficulty of the same collaboration. This means that for publications where all the authors work at a different organisation, the organisational distance for a publication with two authors will be 1, and for a publication with five authors it will be 10.

### 3.2.4 Institutional proximity

This variable measures if the co-authors work at the same kind of institution. Similarly to the organisational distance, for every co-authorship of a publication it is determined if they work at same sort of institution. As such, this will actually measure institutional *distance*.

This distance is calculated by looking for the phrase 'Univ'[4] in the short description of the institution. If both authors have 'Univ', it scores a 0. If they both lack 'Univ', it also scores a 0, and a 1 in all other occurrences. To also catch other institutions of higher education, the phrases 'polytech', 'inst sci', 'inst technol', 'technol inst', 'MIT', 'ETH', 'RWTH', 'Georgia Inst Technol', and 'Izmir Inst Technol' are treated in the same way. Care is taken that these phrases were not matched within a word (i.e. 'MIT' within GlaxoSmithKline or Mitsubishi).

To get the total distance for the publication, for every unique co-authorship pair these calculated scores are summed up. This is similar to the organisa-

[3] *Short description* is an information field used by the Web of Science data.
[4] The short description form of *university*.

tional distance operationalisation, especially in what way this is calculated per co-authorship instead of looking at the publication as a whole.

### 3.2.5 Geographical proximity

Geographical coordinates are gathered using the *Yahoo PlaceFinder* API[5]. This webservice returns geographical coordinates from a street address. For this research the combination of the published city and country are used from the correspondence data. As the coordinates of a city and country can be established worldwide with more or less the same accuracy, more detailed geographical information is superfluous.

For every co-authorship pair of a publication, the geographical distance in kilometres between the correspondence addresses is calculated and again subsequently averaged between all co-authorship pairs. As such, it will measure geographical *distance*, where a 0 means that all the authors have the same correspondence address.

Subsequently, the distance should be calculated using the previously gathered coordinates per co-authorship, for having coordinates does not equal to having a distance. To calculate the distance between two points along a great circle of a sphere, the *haversine formula* (eq. 3.1) can be used. When applied onto the earth this formula is only an approximation as the earth is not a perfect sphere. The Earth radius $r$ varies from 6356.78 km at the poles to 6378.14 km at the equator. More importantly, the radius of curvature of a north-south line on the earth's surface is 1% greater at the poles than at the equator—so the haversine formula can't be guaranteed to be accurate more than 0.5%[6].

$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \quad (3.1)$$

where $d$ is the distance between the two points,

$r$ is the radius of the sphere (6371 km seems to be a common compromise for the earth),

$\phi_1, \phi_2$: latitude of point 1 and latitude of point 2,

$\lambda_1, \lambda_2$: longitude of point 1 and longitude of point 2

[5]See here: `http://developer.yahoo.com/geo/placefinder/` for more information.
[6]See here: `http://www.movable-type.co.uk/scripts/latlong.html`.

## 3.3 Regression analysis

All values that have been calculated in the previous operationalisation will be analysed using *Poisson regression* or *negative binomial regression*. These types of regression are appropriate, because the independent variable is a count variable. These regression analyses do not make assumptions on the distribution of the independent variable, although transforming the variable can aid in finding a relationship other than linear.

Both type of regressions use a linear combination of the independent variable to predict the natural logarithm of the dependent variable (i.e. citation impact) (see equation 3.2).

$$\ln \text{TIMES CITED} = \alpha + \beta_1 \cdot \text{COGN} + \beta_2 \cdot \text{COGN}^2 + \beta_3 \cdot \text{ORG} + \beta_4 \cdot \text{INST} + \beta_5 \cdot \text{GEO} \tag{3.2}$$

# Chapter 4

# Analysis

## 4.1 Descriptives

A descriptive analysis of the data can give insight in what analysis methods are appropriate for this data. Especially because the appropriate regression analysis method is dependent on certain aspects of the dataset.

The publication with just one author can not be analysed using the applied methods. Figure 4.1 shows the share of co-authorships through the years of publication. This shows that for almost all years under study, the most publications had more than one author.

The dataset in this research is comprised of five variables, of which one dependent variable: the number a publication has been cited, and four predictors: cognitive, organisational, institutional and geographical distance between the authors of the publications. As can be seen in figure 4.2 the degree centrality has a distribution where one article has been referenced a 152 times and where 7,120 articles have zero ties.

To be able to see if this network exhibits a scale-free degree distribution, a log-log plot of the degree distibution is shown (see figure 4.3). This plot also shows a regression line which does seem to fit well ($R^2 = 0.9281$). This means that for this network a scale-free degree distribution is demonstrated. As such, this is in agreement to the Barabási–Albert model, which is a model that is generated using a preferential attachment mechanism.

In table 4.1 an overview is given of all the variables which are calculated using the methods described in the previous chapter. For all variables but cognitive distance, there is a high occurrence of the distance '0'.

In table 4.2 the correlation matrix is presented for all the variables. This is calculated using *Spearman's rank correlation coefficient*, as this coefficient is more robust than Pearson's coefficient in the case of a non normal distribution.

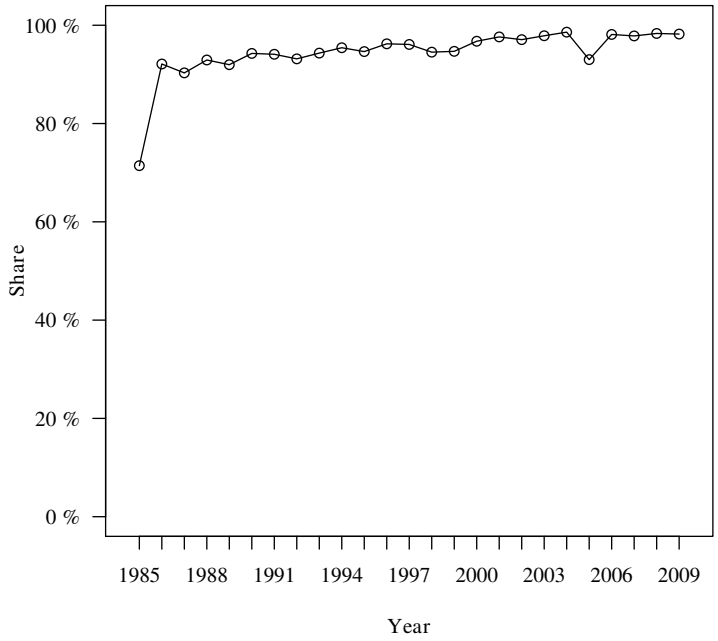



**Figure 4.1:** Share of co-authorship in biotechnology

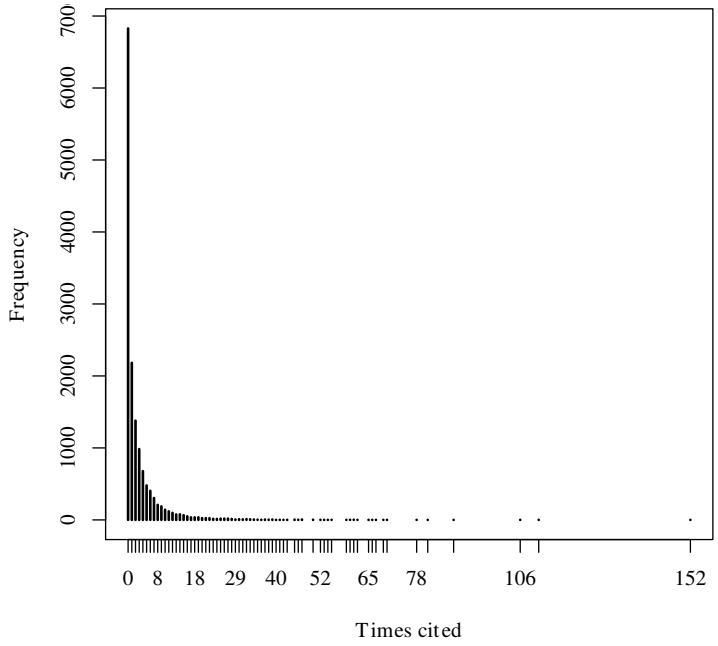



**Figure 4.2:** Distribution of node centrality

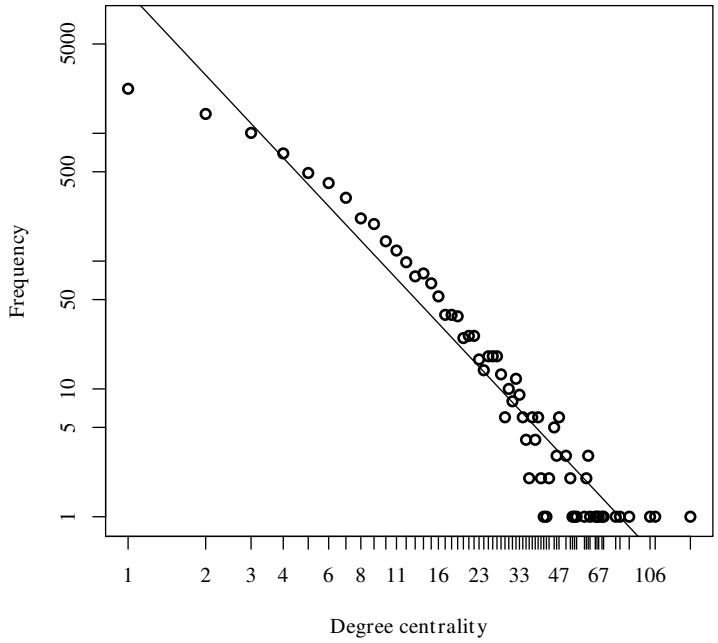



**Figure 4.3:** Log-log of distribution of node centrality

**Table 4.1**
Summary of the variables geographical, institutional, organisational, and cognitive distance; and the dependent variable centrality (i.e. connectivity)

| | Geo | Inst | Org | Cogn | Times cited |
|---|---|---|---|---|---|
| Min. | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 1st Qu. | 0.0 | 0.0 | 0.0 | 0.632 | 0.0 |
| Median | 0.0 | 0.0 | 0.0 | 0.799 | 1.0 |
| Mean | 411.3 | 0.828 | 1.684 | 0.725 | 2.71 |
| 3rd Qu. | 83.49 | 0.0 | 2.0 | 0.896 | 3.0 |
| Max. | 9795.0 | 32.0 | 70.0 | 0.997 | 152.0 |
| NA's | 32 | | | 490 | |

As previously noted, no assumptions have been made towards the distribution of the data.

**Table 4.2**
Spearman's correlation matrix using pairwaise completeness with reranking for each pair

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1. Times cited | 1.00 | | | | |
| 2. Cognitive | 0.12[a] | 1.00 | | | |
| 3. Organisational | -0.09[a] | 0.00 | 1.00 | | |
| 4. Institutional | -0.07[a] | -0.02[d] | 0.71[a] | 1.00 | |
| 5. Geographical | -0.06[a] | 0.03[b] | 0.83[a] | 0.54[a] | 1.00 |

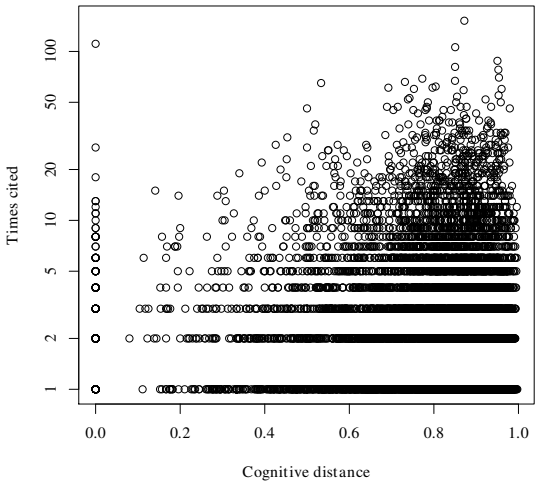
Signif. codes: 0 '[a]' 0.001 '[b]' 0.01 '[c]' 0.05 '[d]' 0.1 ' ' 1

Poisson regression tries to predict the natural logarithm of the dependent variable using a linear combination of the independent variables. In order to see how the the independent variables correlate to the dependent variable, four different scatter plots were made with the natural logarithm of the dependent variable (see fig. 4.4). To be able to establish if all independent variables measure in the same direction, cognitive proximity is displayed in this analysis as cognitive distance.

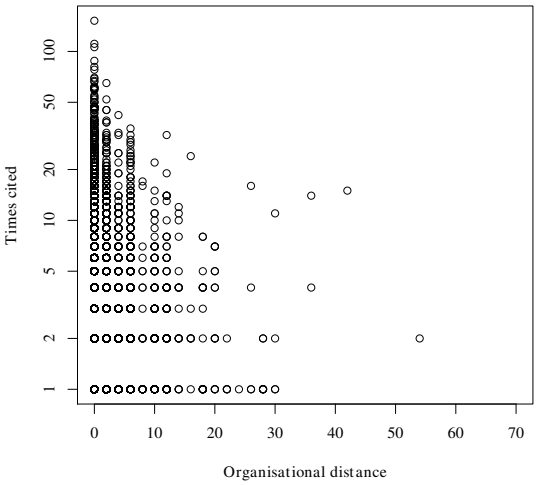
Three plots hint on a negative relationship for these predictors, where a greater distance for all proximity dimensions possibly predict a lower citation impact of the publication. Only cognitive distance hints on a positive relationship or an inverse U-shape relationship with the citation impact.

In relation to citation impact, looking at the most extreme cases can be interesting as well. When looking at the publications which received the most citations, it is striking that these 35 publications displayed in table 4.3 (0.23% of all publications) receives far more than average citations (5.5% of all citations in the dataset). More insight on the proximity dimensions can be gained by taking a closer look at the scores on these dimensions for these cases, because these publications represent a relatively large proportion of the citations. Here it is shown that in almost all the cases the organisational, institutional and geographical distance amounts to zero. This means that for almost all these top-publications, all researchers were probably working at the same organisation.

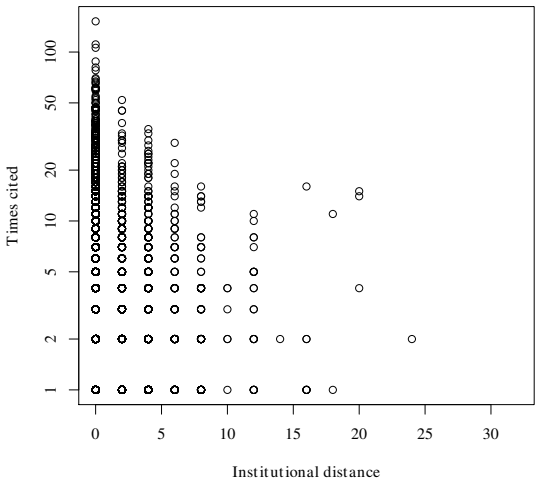
The other way around it is possible to create a list of the cases where the predictors score very high (see table 4.4). This table is sorted by the sum of all the normalised proximity dimensions. This table shows the case where the averaged and normalised predictors are highest. It is striking that almost all cases have a very low citation impact. This means that in those cases where



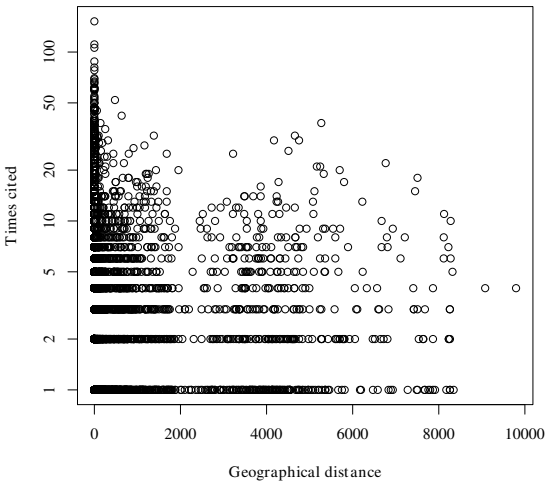

**(a)** Cognitive proximity




**(b)** Organisational proximity




**(c)** Institutional proximity




**(d)** Geographical proximity

**Figure 4.4:** Relationship between predictors and log of centrality

**Table 4.3**
Publications with largest citation impact

| Authors | Times cited | Cogn | Org | Inst | Geo |
|---|---|---|---|---|---|
| Miller, Blanch & Wilke (1988) | 152 | 0.87 | 0 | 0 | 0.00 |
| Laane, Boeren, Vos & Veeger (1987) | 111 | 0.00 | 0 | 0 | 0.00 |
| Glacken, Fleischaker & Sinskey (1986) | 106 | 0.85 | 0 | 0 | 0.00 |
| Sonnleitner & Kappeli (1986) | 88 | 0.95 | 0 | 0 | 0.00 |
| Miller, Wilke & Blanch (1989) | 81 | 0.85 | 0 | 0 | 0.00 |
| Vallino & Stephanopoulos (1993) | 78 | 0.95 | 0 | 0 | 0.00 |
| Tillet & Myerson (1987) | 70 | 0.95 | 0 | 0 | 0.00 |
| Glacken, Adema & Sinskey (1988) | 69 | 0.77 | 0 | 0 | 0.00 |
| Ozturk, Riley & Palsson (1992) | 67 | 0.85 | 0 | 0 | 0.00 |
| Ramirez & Mutharasan (1990) | 66 | 0.73 | 0 | 0 | 0.00 |
| Dandulakis, Herr & Kirwan (1994) | 65 | 0.53 | 2 | 0 | 0.00 |
| Rizzi, Baltes, Theobald & Reuss (1997) | 62 | 0.80 | 0 | 0 | 0.00 |
| Singh, Alrubeai, Gregory & Emery (1994) | 61 | 0.69 | 0 | 0 | 0.00 |
| Xie & Wang (1994) | 61 | 0.80 | 0 | 0 | 0.00 |
| Wanner & Gujer (1986) | 61 |  | 0 | 0 | 0.00 |
| Wiechert & Degraaf (1997) | 60 | 0.75 | 0 | 0 | 0.00 |
| Croughan, Sayre & Wang (1989) | 60 | 0.96 | 0 | 0 | 0.00 |
| Mazur, Fussenegger, Renner & Bailey (1998) | 59 | 0.87 | 0 | 0 | 0.00 |
| Kunas & Papoutsakis (1990) | 55 | 0.95 | 0 | 0 | 0.00 |
| Majewski & Domach (1990) | 54 | 0.85 | 0 | 0 | 0.00 |
| Ozturk & Palsson (1990) | 53 | 0.74 | 0 | 0 | 0.00 |
| Ozturk & Palsson (1991) | 52 | 0.74 | 2 | 2 | 478.29 |
| Bentley, Mirjalili, Andersen, Davis & Kompala (1990) | 52 | 0.92 | 0 | 0 | 0.00 |
| Bavarian, Fan & Chalmers (1991) | 50 | 0.96 | 0 | 0 | 0.00 |
| Murhammer & Goochee (1990) | 50 | 0.85 | 0 | 0 | 0.00 |
| Modak, Lim & Tayeb (1986) | 50 | 0.79 | 0 | 0 | 0.00 |
| Mastrangelo, Hardwick, Zou & Betenbaugh (2000) | 47 | 0.71 | 0 | 0 | 0.00 |
| Marx, Degraaf, Wiechert, Eggeling & Sahm (1996) | 47 | 0.78 | 0 | 0 | 0.00 |
| Chang & Chase (1996) | 47 | 0.86 | 0 | 0 | 0.00 |
| Ferrance, Goel & Ataai (1993) | 47 | 0.69 | 0 | 0 | 0.00 |
| Kompala, Ramkrishna, Jansen & Tsao (1986) | 47 | 0.88 | 0 | 0 | 0.00 |
| Altshuler, Dziewulski, Sowek & Belfort (1986) | 47 | 0.94 | 0 | 0 | 0.00 |
| Alrubeai & Emery (1990) | 46 | 0.50 | 0 | 0 | 0.00 |
| Petersen, Mcintire & Papoutsakis (1988) | 46 | 0.78 | 0 | 0 | 0.00 |
| Mcqueen & Bailey (1990) | 46 | 0.98 | 0 | 0 | 0.00 |

the distance is very high, the publications have not received any citations. Such an analysis is not conclusive, but only indicative.

## 4.2 Regression analysis

### 4.2.1 Poisson regression

Because the dependent variable is a count variable, a Poisson regression is first applied. The results are shown in table 4.5.

To test for the goodness of fit, a chi-squared test based on the residual deviance and degrees of freedom is used:

```
> 1-pchisq(poisson$deviance,poisson$df.residual)

[1] 0
```

This test indicates that the Poisson model does not fit the data ($p < 0.05$).

To test if there is indeed a linear relation between the predictors and the logarithm of the dependent variable, a plot of the residuals versus the fitted values is made (see figure 4.5). Here it is shown that the residuals are skewed. The line shows the moving average of all points in the plot. If a model would fit well, the distribution of the residuals should be evenly spread around the $x$-axis. This is not the case, so there should be tested for overdispersion. Overdispersion means that there exists a greater variability in the data than would be expected using the regression model used.

To test for overdispersion, the deviance divided by the residuals are compared to the chi-squared distribution to get a two-sided $p$-value:

```
> 1-pchisq(poisson$deviance/poisson$df.residual,1)

[1] 0.01844131
```

This test gives a significant result ($p < 0.05$), which means that this model is overdispersed. As a result, negative binomial regression is tried.

### 4.2.2 Negative binomial regression

The results of the negative binomial regression are given in table 4.6.

To test if the model better fits the data, the goodness of fit is estimated using a chi-squared test on the deviance and the degrees of freedom:

**Table 4.4**
Publications with highest average normalised predictors

| Authors | Times cited | Cogn | Org | Inst | Geo |
|---|---|---|---|---|---|
| Krause, Diaz, Edwards, Gartemann, Kromeke, Neuweger, Puhler, Runte, Schluter, Stoye, Szczepanowski, Tauch & Goesmann (2008) | 0 | 0.74 | 44 | 30 | 4196.44 |
| Greene, Henzl, Hosea & Darnall (2005) | 0 | | 62 | 32 | 679.23 |
| Altaner, Saake, Tenkanen, Eyzaguirre, Faulds, Biely, Viikari, Siika-Aho & Puls (2003) | 0 | 0.90 | 42 | 24 | 5652.20 |
| Collins, Clune, Meaney, O'Donoghue, Klinder, Roller, Karlsson, Bennett, O'Riordan, Dunne, O'Sullivan, Rafter, Watzl, Rechkemmer & Pool-Zobel (2005) | 0 | | 52 | 28 | 846.71 |
| Synnergren, Adak, Englund, Giesler, Noaksson, Lindahl, Nilsson, Nelson, Abbotg, Olsson & Sartipy (2008) | 0 | 0.53 | 40 | 24 | 3495.00 |
| Kuster, Hohnjec, Krajinski, El Yahyaoui, Manthey, Gouzy, Dondrup, Meyer, Kalinowski, Brechenmacher, Van Tuinen, Gianinazzi-Pearson, Puhler, Gamas & Becker (2004) | 0 | 0.81 | 50 | 28 | 373.57 |
| Anthony, Ausseil, Bechler, Benguria, Blackhall, Briarty, Cogoli, Davey, Garesse, Hager, Loddenkemper, Marchant, Marco, Marthy, Perry, Power, Schiller, Ugalde, Volkmann & Wardrop (1996) | 2 | 0.79 | 54 | 24 | 776.71 |
| Park, Lee, Lee, Park, Kim, Chung, Lee, Choi & Lee (2005) | 0 | 0.92 | 36 | 32 | 0.00 |
| Lesueur, Ingleby, Odee, Chamberlain, Wilson, Manga, Sarrailh & Pottinger (2001) | 0 | | 30 | 10 | 7225.76 |
| Jardin, Zhao, Selvaraj, Montes, Tran, Prakash & Elias (2008) | 0 | 0.83 | 28 | 18 | 4754.20 |
| Lopez-Vazquez, Song, Hooijmans, Brdjanovic, Moussa, Gijzen & Van Loosdrecht (2008) | 0 | 0.69 | 28 | 16 | 5345.32 |
| Koblizek, Maly, Masojidek, Komenda, Kucera, Giardi, Mattoo & Pilloton (2002) | 0 | 0.00 | 42 | 20 | 2030.59 |
| Radisic, Malda, Epping, Geng, Langer & Vunjak-Novakovic (2006) | 2 | 0.90 | 28 | 10 | 6807.65 |
| Yang, Bellogin, Buendia, Camacho, Chen, Cubo, Daza, Diaz, Espuny, Gutierrez, Harteveld, Li, Lyra, Madinabeitia, Medina, Miao, Ollero, Olsthoorn, Rodriguez, Santamaria, Schlaman, Spaink, Temprano, Thomas-Oates, Van Brussel, Vinardell, Xie, Yang, Zhang, Zhen, Zhou & Ruiz-Sainz (2001) | 0 | 0.97 | 30 | 16 | 4549.38 |
| Hildmann, Wegener, Riester, Hempel, Schober, Merana, Giurato, Guccione, Nielsen, Ficner & Schwienhorst (2006) | 0 | | 40 | 24 | 529.99 |
| Hermann, Kietzmann, Ivancic, Zenzmaier, Luiten, Skranc, Wubbolts, Winkler, Birner-Gruenberger, Pichler & Schwab (2008) | 0 | 0.79 | 40 | 24 | 409.60 |
| Papadimitropoulos, Mastrogiacomo, Peyrin, Molinari, Komlev, Rustichelli & Cancedda (2007) | 0 | 0.33 | 36 | 24 | 767.05 |
| Nedbal, Trtilek, Cerveny, Komarek & Pakrasi (2008) | 0 | 0.00 | 28 | 18 | 3622.43 |
| Sharm, Jani, Thungapathra, Gautam, Meena, Singh, Ghosh, Tyagi & Sharma (2008) | 0 | | 30 | 16 | 3801.51 |
| Sukyai, Rezic, Lorenz, Mueangtoom, Lorenz, Haltrich & Ludwig (2008) | 0 | 0.91 | 30 | 16 | 3770.22 |
| Sanden, Prytz, Tubulekas, Forberg, Le, Hektor, Neubauer, Pragai, Harwood, Ward, Picon, De Mattos, Postma, Farewell, Nystrom, Reeh, Pedersen & Larsson (2003) | 15 | 0.68 | 42 | 20 | 615.55 |
| Unzaga, Diaz-Ricci, Rhee, Hernandez & Schugerl (2002) | 0 | 0.88 | 12 | 6 | 8551.49 |

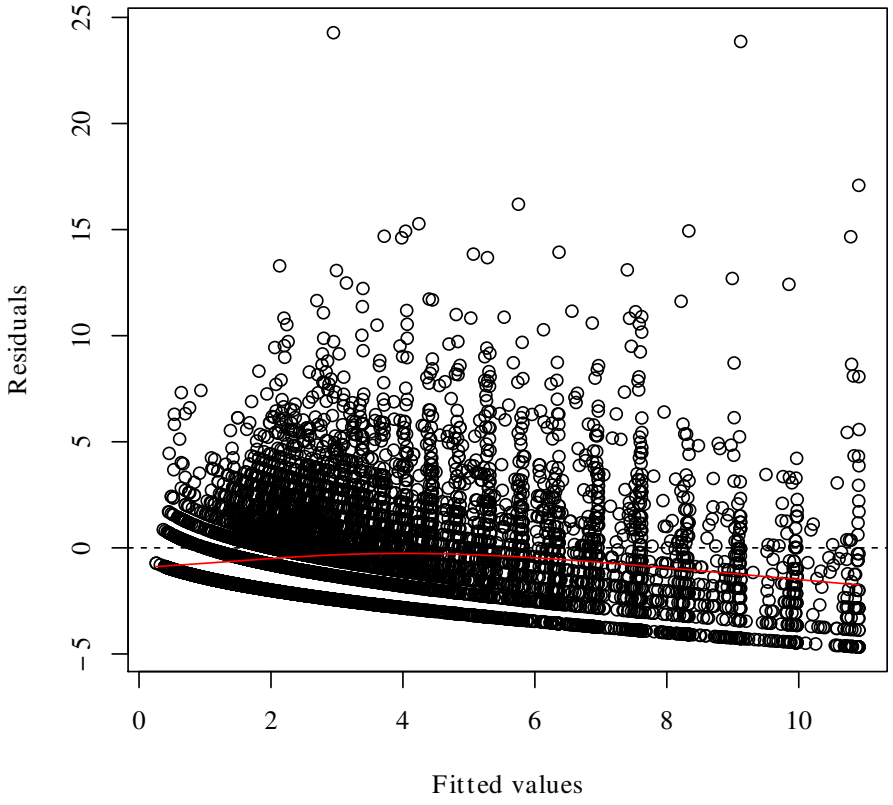



**Figure 4.5:** Residuals vs. fitted values of the Poisson regression

**Table 4.5:** Poisson regression (values are in the scale of the link function, and variables are distances)

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 179.4605 | 1.7525 | 102.40 | 0.0000 |
| Cognitive | 2.8174 | 0.1209 | 23.30 | 0.0000 |
| Cognitive$^2$ | -1.6272 | 0.0955 | -17.04 | 0.0000 |
| Organisational | -0.0236 | 0.0037 | -6.46 | 0.0000 |
| Institutional | 0.0059 | 0.0058 | 1.02 | 0.3091 |
| Geographical | 0.0000 | 0.0000 | 0.65 | 0.5132 |
| Year (control) | -0.0898 | 0.0009 | -102.26 | 0.0000 |

**Table 4.6:** Negative binomial regression (values are in the scale of the link function, and variables are distances)

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 241.2186 | 4.9428 | 48.80 | 0.0000 |
| Cognitive | 2.7295 | 0.2487 | 10.97 | 0.0000 |
| Cognitive$^2$ | -1.6730 | 0.2144 | -7.80 | 0.0000 |
| Organisational | -0.0228 | 0.0080 | -2.86 | 0.0043 |
| Institutional | 0.0216 | 0.0131 | 1.65 | 0.0985 |
| Geographical | 0.0000 | 0.0000 | 1.30 | 0.1934 |
| Year (control) | -0.1206 | 0.0025 | -48.79 | 0.0000 |

```
> 1-pchisq(negbin$deviance,negbin$df.residual)

[1] 0.253568
```

This test indicates that the model fits the data ($p > 0.05$).

The plot of the residuals is given in figure 4.6. This shows that the variability of the residuals is lower than with the Poisson regression.

Firstly, to compare if the negative binomial regression does give a better fit, the *Akaike information criterion* can be used to compare different models (Burnham & Anderson, 2004). It seems that the negative binomial model has better support (Poisson: AIC = 75579.48; Negative binomial: AIC = 42763.74).

Secondly, the deviance is used as an approximate goodness of fit (here a value of 1 is expected):

```
> poisson$deviance/poisson$df.residual
```

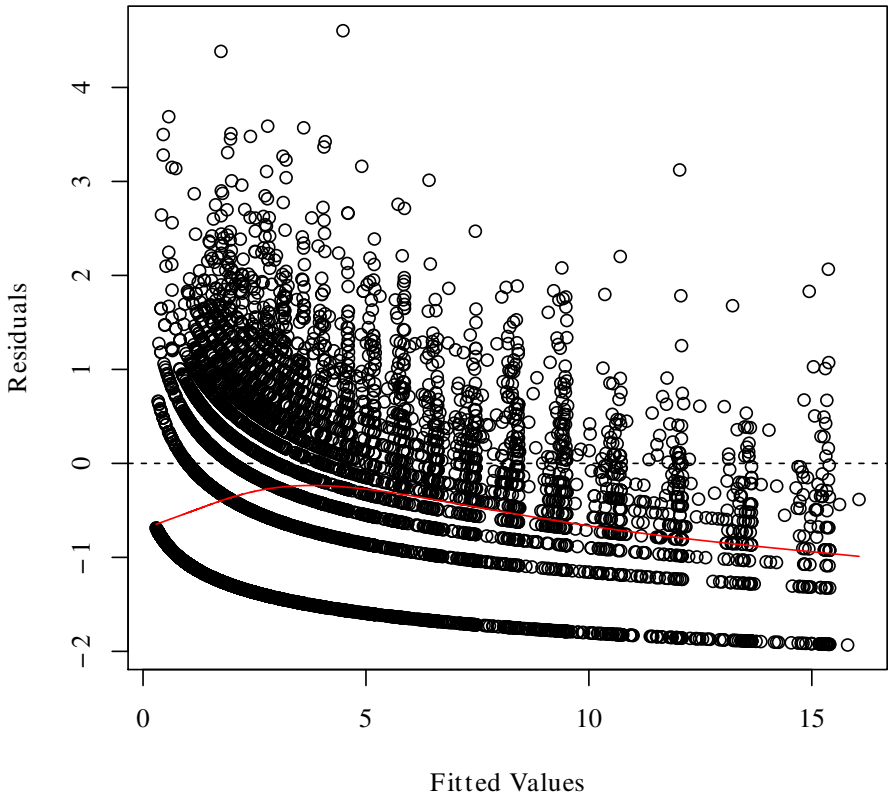



**Figure 4.6:** Residuals vs. fitted values of the negative binomial regression

```
[1] 5.553733
```

```
> negbin$deviance/negbin$df.residual
```

```
[1] 1.009249
```

This means that the negative binomial regression is a good model to use on this dataset.

## 4.3 Interpretation

The form of the model equation for negative binomial regression is the same as that for Poisson regression (see equation 3.2). The log of the outcome is predicted with a linear combination of the predictors:

$$\ln \text{TIMES CITED} = \alpha + \beta_1 \cdot \text{COGN} + \beta_2 \cdot \text{COGN}^2 + \beta_3 \cdot \text{ORG} + \beta_4 \cdot \text{INST} + \beta_5 \cdot \text{GEO} \tag{4.1}$$

To get results in the same unit as the dependent variable, 4.1 should be exponentiated. When exponentiated we get:

$$\begin{aligned} \text{TIMES CITED} &= e^{(\alpha + \beta_1 \cdot \text{COGN} + \beta_2 \cdot \text{COGN}^2 + \beta_3 \cdot \text{ORG} + \beta_4 \cdot \text{INST} + \beta_5 \cdot \text{GEO})} \\ &= e^{\alpha} \cdot e^{\beta_1 \cdot \text{COGN}} \cdot e^{\beta_2 \cdot \text{COGN}^2} \cdot e^{\beta_3 \cdot \text{ORG}} \cdot e^{\beta_4 \cdot \text{INST}} \cdot e^{\beta_5 \cdot \text{GEO}} \end{aligned} \tag{4.2}$$

For instance, if we then examine the case where the variable INST is one higher with all else equal, we get:

$$\begin{aligned} \text{TIMES CITED} &= e^{\alpha} \cdot e^{\beta_1 \cdot \text{COGN}} \cdot e^{\beta_2 \cdot \text{COGN}^2} \cdot e^{\beta_3 \cdot \text{ORG}} \cdot e^{\beta_4 \cdot (\text{INST}+1)} \cdot e^{\beta_5 \cdot \text{GEO}} \\ &= e^{\alpha} \cdot e^{\beta_1 \cdot \text{COGN}} \cdot e^{\beta_2 \cdot \text{COGN}^2} \cdot e^{\beta_3 \cdot \text{ORG}} \cdot e^{\beta_4 \cdot \text{INST}} \cdot e^{\beta_4} \cdot e^{\beta_5 \cdot \text{GEO}} \end{aligned}$$

What this means is that the coefficients have an additive effect in the $ln$(TIMES CITED) scale and a multiplicative effect in the TIMES CITED scale. So, when the confidence intervals are exponentiated we get estimates on how many times centrality changes for every unit increase of the variable (see table 4.7).

Table 4.7 should be interpreted as follows: for every collaborating author pair of a certain publication which do not work at the same type of institution, the citation impact increases with 2.2% ($= 1 - 1.0219 \cdot 100\%$). Similarly, for every pair which does not work at the same organisation, citation impact

**Table 4.7:** Exponentiated confidence intervals

| | Estimate | 2.5 % | 97.5 % |
|---|---|---|---|
| Cognitive | 15.33 | 9.50 | 24.67 |
| Cognitive$^2$ | 0.19 | 0.12 | 0.29 |
| Organisational | 0.98 | 0.96 | 0.99 |
| Institutional | 1.02 | 1.00 | 1.05 |
| Geographical | 1.00 | 1.00 | 1.00 |
| Year (control) | 0.89 | 0.88 | 0.89 |

decreases with 2.3% ($= 1 - 0.97748 \cdot 100\%$). For every 1000 kilometres extra in the average distance between all collaborators, the citation impact does not change ($= 1 - 1.00002 \cdot 1000 \cdot 100\%$). Because the case of cognitive distance has also a squared term, it is not possible to give a single value.

When looking at the plots of the dimensions of proximity (of all researchers of a certain publication) against the citation impact of that publication, three plots hint on a negative relationship for these predictors, where a greater distance for all proximity dimensions possibly predict a lower centrality of the publication. Only cognitive distance hints on a positive relationship with centrality. This is backed by the quantitative analysis: for every collaborating author pair of a certain publication which do not work at the same type of institution, it is expected that centrality is 6.6% lower. Similarly, for every pair which does not work at the same organisation, it is expected that centrality is 4.0% lower. For every 1000 kilometres extra in the average distance between all collaborators, it is expected that centrality is 2.9% lower. In the case of cognitive distance which already measures a ratio, it means that for every percentage point of increase of cognitive distance, it is expected that citation impact is 2.5% higher. These numbers are filled into equation 4.2 which results in equation 4.3.

$$\begin{aligned}\text{TIMES CITED} =& 241.22 \cdot 1.53^{\text{COGN}} \cdot 1.88^{\text{COGN}^2} \cdot 9.77^{\text{ORG}} \\ & \cdot 1.02^{\text{INST}} \cdot 1.00^{\text{GEO}/1000}\end{aligned} \tag{4.3}$$

# Chapter 5

# Discussion

This research has shown a significant relationship between certain dimensions of proximities between researchers and the quality of the research these authors published. The organisational proximity should be high to expect a publication with a high citation impact. For the cognitive dimension an inverse-U shape relationship is demonstrated. The other dimensions under study did not show a significant relationship to the citation impact.

But as with every research, there were certain shortcomings unavoidable. The most apparent is that it was not possible to include the dimension of social proximity. If and in what amount this factor influences the success of publications was unfortunately not determinable, for it was not possible to gather the right information for all the authors from the dataset. But it seems logical that this dimension of proximity could lead to very interesting results. As Rallet and Torre (1999) have already mentioned, it could be that dimensions such as geographical proximity or organisational proximity are proxies of social proximity. This would mean that these dimensions correlate heavily and if social proximity would be included in an analysis, the other dimension lose their possible relationship to citation impact.

The operationalisation of data from a data source such as *Web of Science* has proven to be difficult. Because this meta data is very concise, the dimension of cognitive distance was hard to operationalise. It would have been much easier if it would have been possible to extract a cognitive proximity idea more easily in a way from full-text sources. As full-text was unfortunately not available in this case, the title had to be used, but as already said, this actually proved to be too sparse information.

The dataset had a lot of cases where these cases score zero on some of the dimensions. This made the regression analysis probably a bit weaker, although this is not tested by using a zero-truncated regression design, as this would have thrown away a lot of important data and no theoretical

reason to choose such an analysis has been found. Such regression analyses need a variable which explains such a large number of zeros, which have not been found.

Also, the effect of the correlation of the different proximities was not taken into account. It is for some dimensions not probable that one dimension scores high and the other low. Most organisations have locations that are geographically nearby, rather than that they are spread out around the globe. If both authors work at the same organisation (or even the same location), they are institutionally nearby as well.

As noted previously, the difficulties when using citation impact in such an analysis, could be the reason that some variables did not show a significant relation. When the dependent variable has too much variation because of the reasons described in section 2.2, a regression analysis as used in this paper gets weaker.

Unfortunately the effect of time has not been analysed. Because citation patterns change through time, the used models had to be corrected for time. This also means that the effect of time is precluded from analysis. As such, the change in citation patterns has not been analysed.

Based on this research, I have two recommendations to both the field of scientometrics and innovation studies. First and foremost, this analysis should be carried out on other fields of science and/or other journals as well, to be able to verify its means. Secondly, a case study on the effect of social proximity would be advisable, for all the reasons written above.

# Chapter 6

# Conclusion

This paper has explored the relationship between proximities between research collaborators and the quality of the work they deliver. To explore these effects the question *How do cognitive, organisational, social, institutional, and geographical proximity between researchers influence the citation impact of the resultant publications in a science-based field?* has been tried to answer.

The quantitative regression analysis has shown that there exists a inverse U-shape relationship between cognitive distance (or proximity) and the quality of research. This means that hypothesis 1, *Collaborators with a optimal cognitive proximity (inverted U-shape relationship) produce a higher quality research*, is supported by this research.

Hypothesis 2, *"Collaborators with a high organisational proximity produce a higher quality research"*, is supported by the regression analysis. A significant relation has been found between organisational proximity and the quality of research. Researchers seem to only collaborate effectively and successfully while they are working at the same organisation. There seems to have a positive effect on the citations received of a publication when organisational proximity are high. But hypothesis 3, *"Collaborators with a low organisational proximity produce a higher quality research"*, can be rejected. There has been no evidence found that in this science-intensive sector, inter-organisational collaboration leads to successful research.

Hypothesis 4, *"Collaborators with a high institutional proximity produce a higher quality research"*, hypothesis 5, *"Collaborators with a low institutional proximity produce a higher quality research"*, and hypothesis 6, *"Collaborators with a high geographical proximity produce a higher quality research"*, have been rejected by the regression analysis. No significant relationship has been found in this analysis. This also means that it is not possible to discriminate between hypotheses 4 and 5.

What this means in terms of the theoretical foundation which lies at the base of this paper, is that there is no evidence of the Triple Helix model or Mode 2 knowledge creation found specifically, because you would then expect the inter-organisational ties to be stronger and more successful. Moreover, a significant relationship between collaboration within a type of institution and the quality of the research has been found in the science-intensive sector under study, which lies perpendicular to the model of the Tripe Helix. On the other hand, the proposition that universities can play an enhanced role in a knowledge based society is not undermined.

What this research has added to the whole body of knowledge on proximities in research collaboration, is the conclusion that face–to–face contact remains an important factor in the creation of knowledge through organisational proximity. Furthermore, the principle of an optimal cognitive distance has been confirmed.

Finally, for policy it is important to know that if good progress within science is desirable, scientists probably should be brought together for 'interdisciplinary' collaboration and should be coordinated by a central authority. The other dimensions of proximity have not been confirmed and should therefore not be central in policy change. It should be possible for a government to be the central authority, which is also what the Triple Helix thesis proposes. If multidisciplinary collaborators are brought together in this way, this would in turn create a organisational proximity between collaborators, which is demonstrated in this research.

# References


Archibugi, D. & Pianta, M. (1996). Measuring technological change through patents and innovation surveys. *Technovation*, *16*(9), 451 - 468.

Audretsch, D. B. & Feldman, M. P. (1996). R&d spillovers and the geography of innovation and production. *The American Economic Review*, *86*(3), 630–640.

Bastian, M., Heymann, S. & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. In *International AAAI Conference on Weblogs and Social Media*.

Bonaccorsi, A. (2008). Search regimes and the industrial dynamics of science. *Minerva*, *46*, 285–315.

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

Boschma, R. & Frenken, K. (2010). The spatial evolution of innovation networks. A proximity perspective. In R. Boschma & R. Martin (Eds.), *The Handbook Of Evolutionary Economic Geography* (p. 120 - 135). Northampton, MA: Edward Elgar Publishing.

Burnham, K. P. & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*(2), 261-304.

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

Collins, H. M. (1974). The TEA set: Tacit knowledge and scientific networks. *Science Studies*, *4*(2), 165 - 185.

Cronin, B., Snyder, H. W., Rosenbaum, H., Martinson, A. & Callahan, E. (1998). Invoked on the web. *Journal of the American Society for Information Science*, *49*(14), 1319 - 1328.

David, P. A. (1994). Why are institutions the 'carriers of history'?: Path dependence and the evolution of conventions, organizations and institutions. *Structural Change and Economic Dynamics*, *5*(2), 205 - 220.

Dosi, G. (1982). Technological paradigms and technological trajectories: A suggested interpretation of the determinants and directions of technical

change. *Research Policy*, *11*(3), 147 - 162.

Dosi, G. (1988). The nature of the innovative process. In G. Dosi, C. Freeman, R. Nelson, G. Silverberg & L. Soete (Eds.), *Technical Change and Economic Theory* (p. 221 - 238). London and NY: Pinter.

Etzkowitz, H. & Leydesdorff, L. (2000). The dynamics of innovation: from National Systems and "Mode 2" to a Triple Helix of university–industry–government relations. *Research Policy*, *29*(2), 109 - 123.

Freeman, C. (1997). The diversity of national research systems. In R. Barré, M. Gibbons, J. Maddox, B. Martin & P. Papon (Eds.), *Science in Tomorrow's Europe* (p. 5 - 31). Paris: Economica International.

Frenken, K., Ponds, R. & Van Oort, F. (2010). The citation impact of research collaboration in science-based industries: A spatial-institutional analysis. *Papers in Regional Science*, *89*(2), 351–271.

Garfield, E. (1964). Science citation index: A new dimension in indexing. *Science*, *144*(3619), 649–654.

Heimeriks, G. & Boschma, R. (2012). *The path-and place-dependent nature of scientific knowledge production in biotech 1986-2008* (Papers in Evolutionary Economic Geography). Utrecht University.

Hessels, L. K. & Lente, H. van. (2008). Re-thinking new knowledge production: A literature review and a research agenda. *Research Policy*, *37*(4), 740 - 760.

Hoekman, J., Frenken, K. & van Oort, F. (2009). The geography of collaborative knowledge production in europe. *The Annals of Regional Science*, *43*, 721 - 738.

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

Kuhn, T. S. (1996). *The structure of scientific revolutions* (Vol. 2). Chicago, IL: University of Chicago press.

Leydesdorff, L. (2000). The triple helix: an evolutionary model of innovations. *Research Policy*, *29*(2), 243 - 255.

Lundvall, B. (1988). Innovation as an interactive process: from user–producer interaction to the national system of innovation. In G. Dosi, C. Freeman, R. Nelson, G. Silverberg & L. Soete (Eds.), *Technical Change and Economic Theory* (p. 349 - 369). London and NY: Pinter.

MacRoberts, M. & MacRoberts, B. (1996). Problems of citation analysis. *Scientometrics*, *36*(3), 435–444.

Nooteboom, B., Van Haverbeke, W., Duysters, G., Gilsing, V. & van den Oord, A. (2007). Optimal cognitive distance and absorptive capacity. *Research Policy*, *36*(7), 1016 - 1034.

Paci, R. & Usai, S. (2000). Technological enclaves and industrial districts: An analysis of the regional distribution of innovative activity in europe. *Regional Studies*, *34*(2), 97–114.

Pavitt, K. (1984). Sectoral patterns of technical change: Towards a taxonomy and a theory. *Research Policy*, *13*(6), 343 - 373.

Pinch, T. J. & Bijker, W. E. (1984). The Social Construction of Facts and Artefacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other. *Social Studies of Science*, *14*(3), 399 - 441.

Ponds, R., van Oort, F. & Frenken, K. (2007). The geographical and institutional proximity of research collaboration. *Papers in Regional Science*, *86*(3), 423 - 444.

Powell, W. W., Koput, K. W. & Smith-Doerr, L. (1996). Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology. *Administrative Science Quarterly*, *41*(1), 116 - 145.

Rallet, A. & Torre, A. (1999). Is geographical proximity necessary in the innovation networks in the era of global economy? *GeoJournal*, *49*, 373–380.

Scott, P., Gibbons, M., Nowotny, H., Limoges, C., Trow, M. & Schwartzman, S. (1994). *The new production of knowledge: The dynamics of science and research in contemporary societies*. Sage Publications Limited.

Secretariat of the CBD. (1992). *Convention on biological diversity.* Available from `http://www.cbd.int/convention/text/`

Torre, A. & Gilly, J.-P. (2000). On the analytical dimension of proximity dynamics. *Regional Studies*, *34*(2), 169–180.

Wouters, P. (1998). The signs of science. *Scientometrics*, *41*, 225–241.