

Influence of cognitive, geographical, and collaborative proximity on  
knowledge production of Canadian nanotechnology

ELVA LUZ CRESPO NEIRA

A THESIS  
IN  
THE DEPARTMENT  
OF  
CONCORDIA INSTITUTE FOR INFORMATION SYSTEMS ENGINEERING

PRESENTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE DEGREE OF MASTER OF APPLIED SCIENCE (QUALITY SYSTEMS ENGINEERING)  
CONCORDIA UNIVERSITY  
MONTRÉAL, QUÉBEC, CANADA

APRIL 2016

© ELVA LUZ CRESPO NEIRA, 2016

CONCORDIA UNIVERSITY  
School of Graduate Studies

This is to certify that the thesis prepared

By: **Elva Luz Crespo Neira**

Entitled: **Influence of cognitive, geographical, and collaborative proximity  
on knowledge production of Canadian nanotechnology**

and submitted in partial fulfillment of the requirements for the degree of

**Master of Applied Science (Quality Systems Engineering)**

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

<u>Dr. M.Mannan</u>	Chair
<u>Dr. A. Ben Hamza</u>	CIISE Examiner
<u>Dr. I. Contreras</u>	External Examiner
<u>Dr. A. Schiffauerova</u>	Supervisor
<u>Dr. C. Beaudry</u>	Co-supervisor

Approved \_\_\_\_\_  
Chair of Department or Graduate Program Director

\_\_\_\_\_ 20 \_\_\_\_\_

Dr. J. Bentahar, Graduate Program Director,  
Concordia Institute for Information Systems Engineering  
Faculty of Engineering and Computer Science

# Abstract

Influence of cognitive, geographical, and collaborative proximity on knowledge production of Canadian nanotechnology

Elva Luz Crespo Neira

We address the question of whether or not geographical, cognitive, and collaborative proximity have an impact on citation probability in the scientific writings of Canadian nanotechnology. Even though a number of studies in the proximity literature deal with the effects of spatial distance, scientific specialization, and social network structure, to our knowledge no one has combined all three to explore the production of academic information.

We generate a feature framework based on measurements from these factors, relying on statistical and classification approaches to assess their influence on effective citations. Specifically, by means of applying binary regression models along with tree-based machine learning algorithms, we found statistical significance proving that these features have both a verifiable impact and predictive potential. Importantly, our work is the first one that we have seen combining these techniques to infer the establishment of positive citation links. Moreover, we employed inductive network analysis comprehensively to examine the co-authorship links between authors publishing in nanoscience, considering additional network metrics to the ones usually adopted in the literature.

Our findings reveal that cognitive proximity, closely followed by the collaborative aspect, are the most important elements inducing Canadian scholars to cite, with geography sometimes acting as their base. Our results enable us to reach better understanding related to the citation behavior of the nanoresearch community in Canada, making our work a valuable contribution to scholarly literature, also giving us ground to make policy recommendations.

---

## Acknowledgments

---

This thesis is the final product of a long journey that has been filled with learning and all kinds of challenges at every step of the way.

First and foremost, I would like to express my gratitude to my supervisor, Dr. Andrea Schiffauerova, for offering me the opportunity to participate in this research project in the first place, for her continuous support during my Master's study, and for her direction in the writing of this thesis.

I have been fortunate to count with guidance from Dr. Carl St-Pierre for the statistical analysis; I am thankful for his insightful comments and challenging discussions. I would like to acknowledge Dr. Ben Hamza for taking the time to listen to my questions and provide help at several stages of my work. Mukesh Kumar also deserves a special mention for suggesting the use of machine learning.

I owe my gratitude to my friend Kunwar Rattan, who supported me with his skills and extensive knowledge in computer programming, as well as with encouragement throughout the process. Without his help, it would not have been possible to conduct this research in a timely manner.

I have been blessed with an amazing group of friends, who have gone a long way providing moral support, Carolina and my roommates Lizette and Johanna most of all, I cannot thank them enough for all their sympathetic understanding. I am wholeheartedly grateful towards my family: my parents and my brothers and sister, for their unconditional motivation, for always believing in me, and for supporting me spiritually throughout the time of my studies.

Finally, *the fear of the Lord is the beginning of knowledge* (Proverbs 1:7 NIV). I give Him praise forevermore for He has been good to me, guiding me and giving me the strength to fulfill his purposes in my life.

---

## Contents

---

<b>List of Figures</b>	<b>viii</b>
<b>List of Tables</b>	<b>x</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background and Motivation . . . . .	1
1.2 Objectives . . . . .	3
1.3 Contributions . . . . .	3
1.4 Organization . . . . .	3
<b>2 Literature Review</b>	<b>4</b>
2.1 Nanotechnology research in Canada . . . . .	4
2.2 Knowledge diffusion . . . . .	5
2.3 Proximity Factors . . . . .	6
2.3.1 Defining Proximity . . . . .	6
2.3.2 Proximity Types . . . . .	7
2.3.3 Cognitive Proximity . . . . .	10
2.3.4 Geographical Proximity . . . . .	13
2.3.5 Collaborative Proximity . . . . .	17
2.3.6 Interaction between Geographical Proximity and other Dimensions . . . . .	21
2.4 Citations . . . . .	25
2.4.1 Defining Citation . . . . .	25

2.4.2	Citing Behavior . . . . .	26
2.5	Impact of Proximity on Citations . . . . .	28
2.5.1	Proximity Influence on Academic Citation . . . . .	30
2.5.2	Proximity and the Interaction between Scholarship and Innovation . . . . .	32
2.5.3	Proximity Influence on Patent Citation . . . . .	35
2.5.4	Research Gaps . . . . .	37
2.6	Research Questions . . . . .	39
<b>3</b>	<b>Methodology</b>	<b>40</b>
3.1	Research Design . . . . .	40
3.2	Data . . . . .	41
3.2.1	Instrumentation . . . . .	41
3.2.2	Setting and Participants . . . . .	41
3.2.3	Data preparation and Sampling . . . . .	42
3.3	Proximity Measuring . . . . .	47
3.3.1	Measuring Cognitive Proximity . . . . .	47
3.3.2	Measuring Geographical Proximity . . . . .	51
3.3.3	Measuring Collaborative Proximity . . . . .	59
3.4	Data Analysis . . . . .	66
3.4.1	Regression and Classification Modeling . . . . .	66
3.4.2	Hypotheses and Variables . . . . .	74
3.4.3	Applying Regression . . . . .	76
3.4.4	Applying Classification . . . . .	80
<b>4</b>	<b>Results and Discussion</b>	<b>85</b>
4.1	Results . . . . .	85
4.1.1	Regression Models . . . . .	85
4.1.2	Classification Models . . . . .	90
4.1.3	Testing the hypotheses . . . . .	94
4.2	Discussion of Results . . . . .	95
4.3	Limitations and Assumptions . . . . .	101
<b>5</b>	<b>Conclusions</b>	<b>103</b>
5.1	Conclusions . . . . .	103
5.2	Future Work . . . . .	104

<b>Bibliography</b>	<b>106</b>
<b>Appendices</b>	<b>121</b>
Appendix A	General References . . . . . 122
Appendix B	Data Structure . . . . . 125
Appendix C	Programming Scripts . . . . . 126
Appendix D	Statistics . . . . . 130
Appendix E	Collaboration Network . . . . . 133

---

## List of Figures

---

1	Nanotechnology articles per year . . . . .	43
2	Canadian authors publishing in 2010 and 2011 . . . . .	43
3	Cited articles and authors per year . . . . .	44
4	Proportion of citation links of REF authors between 2007-2010 . . . . .	46
5	Author pairs per cited year . . . . .	47
6	Distribution of cognitive distance . . . . .	50
7	Example of affiliation record in raw format . . . . .	53
8	Script querying Google Maps Geocoding API . . . . .	53
9	Google Maps response output . . . . .	54
10	Distribution of Euclidean distance . . . . .	55
11	Selection of transportation mode . . . . .	56
12	Script querying Google Maps distance API . . . . .	57
13	Distribution of Traveling Time . . . . .	58
14	Distribution of location category . . . . .	60
15	Example visualization of a network graph . . . . .	61
16	Screenshot of Sci2 software . . . . .	62
17	Network overview from Sci2 . . . . .	63
18	Distribution of Shortest Path . . . . .	65
19	Correlation Plot . . . . .	77
20	Best-performing models, as ranked by SPSS Modeler . . . . .	82
21	Modeling stream with SPSS Modeler . . . . .	84



22	Variable ranking by z-statistic in binary regressions . . . . .	86
23	Variable Importance, as ranked per classifier tree . . . . .	92
24	Shortest path node in CHAID model . . . . .	92
25	Variable importance according to Random Forest . . . . .	93
26	Variable importance according to Random Forest, reduced . . . . .	93
27	Geo-layout of nodes with positive citing link . . . . .	100
28	Network overview from Pajek . . . . .	134
29	Distribution of SNA metrics of REF author . . . . .	135
30	Distribution of control variables . . . . .	136

---

## List of Tables

---

1	Proximity dimension frameworks . . . . .	7
2	Proximity dimensions addressed in the literature . . . . .	29
3	Cognitive distance scale . . . . .	49
4	Cognitive distance into dummy variables . . . . .	50
5	Location category scale . . . . .	58
6	Location category into dummy variables . . . . .	59
7	Comparison of different decision tree algorithms . . . . .	70
8	Confusion Matrix . . . . .	81
9	Regression models summary . . . . .	87
10	Classifier performance metrics (in %) . . . . .	91
11	Cognitive fields and subfields in the data . . . . .	124
12	Table structure of paired links in MySQL database . . . . .	125
13	Descriptive statistics of independent variables . . . . .	130
14	Descriptive statistics of independent variables according to Citing . . . . .	131
15	Correlation Matrix between variables . . . . .	132

### 1.1 Background and Motivation

Knowledge and science are two concepts that are closely interrelated, so much so, that sometimes they are even used interchangeably. Science, however, entails formal processes for its creation, so we should rather call it structured knowledge.

The use of existing knowledge is vital for innovating, discovering, and generating new ideas. Learning is also essential, since it involves gaining new knowledge, as well as sharing and exchanging it. Knowledge being the raw material, its production in terms of academic research and invention is key for the development of technology and sciences.

Throughout this thesis, we will be focusing on one scientific field in particular: nanotechnology, an emerging technology based on solid particles in the size range of 1-100 nm. Government organizations reported that investment targeted at its research and evolution increased worldwide over the past years, from roughly \$432 million in 1997 to \$4.1 billion in 2005 (Yegul et al., 2008), figures which illustrate its potential for global economy. Developments and applications in this area over the past two decades have been dramatic, and will continue into the foreseeable future (EPA, 2005). Canada being one of the seven major advanced economies in the world, relevant investigations show that it is among the top countries producing nanotechnology research (Yegul et al., 2008).

Since research in general concerns the flow and expansion of knowledge, we seek to find out more about possible causes that could be behind knowledge production, driving and modeling its diffusion among the

scientific community. Moreover, we suggest that knowledge production in Canadian nanotechnology is influenced by three key proximity factors: Cognitive, geographical, and collaborative.

Cognitive proximity refers to the closeness between scientific branches. Visualizing the interactions of scientists coming from different knowledge fields or disciplines will prove helpful to get an overview of the level of interdisciplinarity, which is very important, considering the multidisciplinary nature of nanotechnology.

The geographical or spatial proximity is perhaps the easiest concept to understand from all the proposed aspects, since it clearly alludes to physical distance. With the expansion and evolution of communication technologies nowadays, it is interesting to examine if distance still has weight to determine the feasibility or facilitate academic research, since it is well known that spatial separation is no longer an obstacle for information exchange (Feldman, 2002).

With collaborative proximity, we aim to get a sense of how close scholars are to each other in terms of a co-authorship social network, as well as the importance of their position within the network. The idea is to examine the relationships between scientists and their attributes with respect to the whole network, and evaluate the cohesion level among individuals.

In the present work, we propose to test the hypotheses that these proximities have both an observable impact and a predicting influence in knowledge production, in terms of academic citation probability. We consider that finding empirical evidence for this will be of interest to the scientific society, because it will contribute to explain how proximity factors affect the resulting scholarly output. In addition, we seek to know if the importance of these elements lies within each factor separately, or if their effect is based upon an existing interaction between them.

Throughout the literature, we have a few examples of studies sharing some similarities to ours. However, they are geared towards innovation and economics, or rather focusing on other branches of science. Plus, they have been conducted either with considerably smaller databases, or inspecting proximity influence only at a macro level.

At any rate, to our knowledge, there has been no analysis concerning all three proximity aspects together in nanotechnology research, nor in any other field. Furthermore, none of these works address the establishment of citation links by means of both regression and classification modeling. Thus, we expect the results from this thesis to be a good contribution in terms of its conclusions and the proposed methodology.

## 1.2 Objectives

The main purpose of this thesis is to evaluate the importance of the aforementioned three key factors for the creation of knowledge in the area of nanotechnology in Canada:

1. Cognitive proximity,
2. Geographical proximity, and
3. Collaborative proximity.

With the results from this research, we aim to determine whether or not these aspects have an actual influence on research conducted by scholars.

## 1.3 Contributions

We consider one of the major contributions of this work to be leading to a better understanding on the production of knowledge for Canadian nanotechnology, which can provide more guidance in terms of academic research. Particularly, to have an improved grasp on the dynamics, evolution, and structure of formal research conducted in this field, focusing on the three distinctive aspects we have selected to analyze more in depth.

We expect the results of our analysis to be encouraging for nanotechnology researchers, motivating them to pursue trans-disciplinary research, that is, making use of the best research ideas and methodologies from other fields. In addition, we anticipate that observable effects from scholarly collaboration will drive institutions to promote the establishment of social connections in this academic domain.

Policy makers shall also gain result-based evidence to support fine-tuning of science and technology policies that encourage multidisciplinary research teams. We anticipate that this study will provide supportive arguments for the creation of tools that will facilitate the expansion of research fields within Canada.

## 1.4 Organization

This thesis is organized as follows: The “Literature Review” section provides an overview of the bibliography found relevant to the topic. Then follows a section that introduces the data and the methodology used in the different stages of our work.

The “Results” section presents the outcome of the various analysis techniques applied to the data, followed by a discussion on their significance and an account of the limitations found. Finally, the last section concludes by presenting a summary on the findings of this research, while also making suggestions for future research opportunities emerging from our analysis.

## 2.1 Nanotechnology research in Canada

We set the background for this thesis by taking a brief look at the development of nanotechnology, an emerging technology based on solid particles in the size range of 1-100 nm (see Yegul et al., 2008, p. 1), with special consideration in its knowledge production at the hands of Canadian scholars.

According to Delemarle et al. (2009), scientific publications in this area have substantially increased since the last decade of the 1990s (12% per year from 1998 to 2006), with nanoscience programs flourishing throughout the world. Furthermore, nanotechnology is a field where the involvement of academic research is preeminent. With universities as an important source for its knowledge generation and even innovative activities, nanotechnology appears to rely on sciences to a higher degree than other technology fields (Wang and Guan, 2011).

As mentioned previously, Canada is one of the top countries producing scientific literature in nanotechnology (Yegul et al., 2008). Moreover, Canadian nanotechnology inventors show an increasing tendency to collaborate more closely with researchers on the field (Beaudry and Schiffauerova, 2011). This scenario would then be ideal for knowledge production, given that the dynamics of such collaboration results in greater information sharing, in turn favoring its diffusion among the scientific community.

However, since this technology requires sustained investment along with favorable conditions, nanotechnology does not thrive just anywhere, nor is it spread uniformly around the globe (Delemarle et al., 2009).

The context of this study is nanotechnology, which makes for a particularly good setting to study knowledge flows and the role of proximity influences (Cunningham and Werker, 2012). Thus, we shall now take a look into the concepts of academic knowledge, aiming to better understand what elements play a major role to constitute a favorable environment for the knowledge production in nanotechnology, as measured by attributes of scholarly publications.

## 2.2 Knowledge diffusion

Knowledge is defined by Howells (2002) as a dynamic framework from which information can be stored, processed, and understood. The sharing of learned knowledge is essential for innovation to take place, going hand in hand with academic research (Moodysson and Jonsson, 2007). We could interpret this mutual exchange of knowledge and shared learning as meaning that knowledge is intrinsically a socially constructed process (Berger and Luckmann, 1991).

Extensive literature associates social networks as the pathways that channel the flow of knowledge among actors (Sorenson et al., 2006). When evaluating some properties and the structure of networks itself, it is well documented that they have a decisive influence on the spread of knowledge (Eslami et al., 2013). If we go a step further, and focus on the realm of scientific communication, it is unquestionable that individuals are connected in intellectual and social networks shaped by formal and informal channels through which knowledge flows (Marion et al., 2003).

Marion et al. (2003) also emphasize that these studies, when applied to scientific communities, have revealed their capacity for analyzing and understanding the growth and diffusion of information. Consequently, research on collaboration and knowledge flows has just recently started to be applied to areas of emerging technology (Eslami et al., 2013). This is particularly relevant, if we consider science as an “epistemic community”, which means that the sole objective of its members is the production of knowledge (Gittelman, 2007).

Considering the key role that the knowledge flow plays in a wide variety of fields (Rogers, 2010), the processes involved in its diffusion have spurred investigation since way back, with most of the work originating from the field of Sociology (e.g. Valente and Rogers, 1995; Zinkhan et al., 1992; Baber, 1992; Shapin, 1995). Moreover, the dynamics, evolution, and direction of these knowledge flows have been the aim of analysts who seek to visualize and document the patterns of the relationships involved (Marion et al., 2003), as well as discerning their influencing factors.

## 2.3 Proximity Factors

### 2.3.1 Defining Proximity

As mentioned above, social networks are the highways for knowledge diffusion. These networks are usually formed according to typical characteristics such as gender, ethnicity, age, education level, and so on, following people’s tendency to group together with individuals they have something in common. This similarity of attributes is called *homophily* by sociologists, but we shall rather use the terminology of *proximity*, since it is more common when studying scientific areas and innovation (Boschma and Frenken, 2010).

Gilly and Torre (2000) define proximity as the existing interactions between actors, following the school of thought from the French Proximity Dynamics group, formed by a collection of industrial and spatial economists. Since the early 1990s, this group took the lead in a worldwide research movement who collaborated to show the concepts of convergence and coherence under the light of new approaches in economic space (Shaw and Gilly, 2000). Their purpose was to study proximity and explain the nature of its effects within an organization setting. The book by Rallet et al. (1995) brings together the findings of this association, in which they mention that proximity is an important variable that affects human activities, whether as cause or consequence.

The interest given to the notion of proximity since then has inspired further investigation by scholars coming from multiple disciplines, since clearly this concept cannot be solely limited to the context of economics (Shaw and Gilly, 2000). Even though there is as yet little understanding of how it affects innovation over time (Boschma, 2005), proximity is considered to be an influencing factor for knowledge flows in science. Notably, former research on nanotechnology suggests that proximity plays a significant role in the understanding of collaboration and development in this area (Cunningham and Werker, 2012).

Speaking in terms of the scientific network, a certain degree of proximity is required to make the actors or agents connected to actually form the network itself. Furthermore, in the literature it is frequently argued that an increasing proximity leads to more interaction between actors, leading them in turn to learn and innovate more.

However, Boschma (2005) takes a critical stand on this, by putting to question the virtues of proximity. Boschma and Frenken (2010) call it *the proximity paradox*, which states that whereas too little proximity will prevent interactive learning and innovation from happening, too much will also be harmful for these purposes. Hence, an optimal level of proximity between agents needs to be reached and not surpassed, to avoid negative impacts on their academic and innovative performance, due to the lack of openness and flexibility (Boschma, 2005; Broekel and Meder, 2008). We shall see more details about the potential risks related to degree of proximity when we go through the concerning dimensions further on.

Proximity is a somewhat “elastic” notion (Moodysson and Jonsson, 2007), and its definition and the



number of its different dimensions vary among authors, since there are non-tangible aspects to be considered. Still, they agree that it is essential to clarify them in a framework that avoids overlap as much as possible, so the effects of each dimension can be isolated and analyzed separately.

The proximity dimensions found in the literature are discussed in the next section.

### 2.3.2 Proximity Types

Proximity has been historically associated with location, starting with von Thünen (1826), who studied its advantages in the context of urban and agricultural activities (Gilly and Torre, 2000). Nevertheless, various scholars amplify the meaning of proximity, since it clearly goes beyond a geographical connotation (Boschma, 2005). Refer to Table 1 to see a compendium of the proximity categories or dimensions that have been identified by major authors in the subject (Gilly and Torre, 2000; Zeller, 2004; Boschma, 2005; Moodysson and Jonsson, 2007). As we can see, there is, in some cases, wide conceptual variation between the categories proposed.

Gilly and Torre (2000)	Zeller (2004)	Boschma (2005)	Moodysson and Jonsson (2007)
Geographical	Spatial	Geographical	Functional
Organizational	Organizational	Organizational	
	Institutional	Institutional	
	Technological	Cognitive	
	Cultural	Social	
	Relational		Relational
	Virtual		

Table 1: Proximity dimension frameworks  
Source: Modified based on Moodysson and Jonsson (2007, p. 6)

The key contribution from the French School of Proximity Dynamics consisted in proposing for the first time that proximity covers a number of dimensions. They analyzed innovation processes from the perspective of the interface between the space economy and the industrial economy (Gilly and Torre, 2000). Among the primary elements of their research were competition conditions at local levels, innovation and technological change dimensions, and the involvement of externalities. Within the setting of economics, *externalities* (also called *spill-overs*) are unintended side effects caused by an economic activity, meaning that there is no actual payment involved for these externalities to happen.

Thus, we have the first categorization for proximity by this group, in terms of geographical and organizational proximity. Since we will go more in depth about the different approaches towards geographical proximity in a next section, we focus now on *organizational proximity*. It basically follows two ideas: similarity and adherence, referring respectively to either being part of the same relational framework, or sharing common knowledge and skills.

Conversely, the study performed by Zeller (2004) extends proximity dimensionality by adding five more categories on top of those already mentioned. His study was geared towards the pharmaceutical industry, and he specified the categories of spatial, organizational, institutional, cultural, relational, technological, and virtual proximity. In this case, geographical proximity is named *spatial proximity*, and the organizational proximity keeps one of the logics from the French Proximity Dynamics crew, the similarity aspect. Zeller gives more emphasis on the organization as a corporate unit, with its own set of rules and identity. The adherence logic is addressed in two separate dimensions: institutional and cultural proximities.

*Institutional proximity* refers to the collection of practices, laws, and rules defined by the geographical setting, that is, within a country or region. These elements are involved in the evolution of political power relations that contribute to a cultural affinity (Zeller, 2004) that influences, shapes, and constrains interactions between actors (Kirat and Lung, 1999). Having a correlation to this dimension, *cultural proximity* is based on a shared cultural background, and the consequent norms of behavior between innovative actors and researchers. It implicates as well the expected factors in social relationships, such as culture, language, and trust. However, personal relationships per se are studied by this author in the *relational proximity*, which is influenced by the cultural dimension, but focuses on the informal structures that facilitate knowledge transfer (Zeller, 2004).

*Technological proximity* is, of course, related to technology, in terms of the sharing of its knowledge bases, experiences, and infrastructure in general. It is key for innovation to occur, because people in the field that are technologically proximate, can contribute with their respective findings, developments, and know-how (Zeller, 2004). Sometimes, a *sectoral proximity* is used, which relates to the industrial distribution of innovative activity, and it is closely related to technological proximity (Maggioni and Uberti, 2009), however it does not form part of a bigger proximity categorization. The last dimension conceptualized in Zeller's framework is the *virtual proximity*, which results from the use of ICT (Information and Communications Technology).

Shortly after the categorization by Zeller (2004), Boschma (2005) introduced his model of proximity classification in five dimensions: cognitive, organizational, social, institutional, and geographical. Similarly to Zeller's relational dimension, *social proximity* deals with the ties among individuals in a social context, that is, it looks into the relations between agents in a micro-level. In the same fashion, Boschma also separates the cultural aspects of proximity, associating them to the notion of institutional proximity, which studies factors in agents at a macro-level. An institution in this context is not used as a synonym to company or firm, rather cultural elements like ethnics, beliefs, language, etc. are considered to be informal institutions, and laws and regulations would be formal institutions (Boschma, 2005).

Coming from the field of economic geography, the model by Boschma (2005) was also built on top of the work from the French research team, but differing to some extent, particularly for the organizational

dimension, which is not as broadly defined as by the French scholars. Instead, organizational proximity is delimited for relations inside or between firms, and for analytical reasons, the partaking in the same knowledge field is isolated as a new type: the cognitive dimension, which we will address in a further section.

As the most recent of these academic publications, Moodysson and Jonsson (2007) take a critical stance on previous works. Despite the French proximity framework (Gilly and Torre, 2000) being a noteworthy contribution for defining non-tangible aspects of proximity, its vague categorization would impede real applicability (Moodysson and Jonsson, 2007, p. 5). Nonetheless, they likewise take into consideration only two proximity dimensions, since according to their perspective, the models by Zeller (2004) and Boschma (2005) do not provide a consistent framework for empirical studies, due to the overlapping in their dimension categories. In spite of this assessment, they largely base their own work on these two contributions, since they were viewed as analytically sharper in defining the non-tangible aspects of proximity.

As a result, two categories were described: functional and relational proximity, in the context of biotechnology innovation projects in Sweden (Moodysson and Jonsson, 2007). Relational proximity, based on affinity and similarity, is close to the proposed organizational dimension from Gilly and Torre (2000), yet it is different in that the two logics (adherence and similarity) are seen as the same, or rather, one is the cause of the other. Finally, *functional proximity* is introduced, combining key concepts from the geographical dimension.

Sometimes it is argued that the list of dimensions can be extended, without affecting the meaning of a dimension itself (Boschma and Frenken, 2010). Such could be the case for splitting sub-factors, like religious or linguistic as their own proximity dimension. Boschma and Frenken (2010) claim that proximity dimensions are analytically orthogonal, or independent, even though many dimensions of proximity may empirically overlap and turn out to be correlated or mutually dependent (Cunningham and Werker, 2012). For instance, institutional proximity between two agents could be influenced by geographical proximity. It could imply that the cultural gap within short distances is smaller, with individuals pertaining to the same town or country.

Overall, we see that even though there are some discrepancies towards the number of proximity dimensions between scholars, the factors to contemplate are common in each of their studies. Rather, they sometimes rename, or just either separate into several, or merge diverse aspects in one dimension. For the purpose of explaining and understanding how connections are encouraged into the formation of networks, it is useful to count with a framework constituted by multiple proximity dimensions. Boschma (2005) affirms that in the context of innovation networks, proximity dimensions are substitutes rather than complements. Hence, at least one dimension is required to establish a successful connection, but if more than one is involved, it adds little to the probability of said connection to be successful or established in the first place.

In essence, proximity is mandatory to build networks that will allow knowledge diffusion resulting in learning or innovation, though not every dimension must be involved to arrange the connections. We base

on the scheme defined by Boschma (2005), and consider two categories from it: geographical and cognitive proximity. We have reviewed and put into context the frameworks into which these dimensions are situated. So, we shall now go in depth to explain the various approaches taken by several scholars when delimiting these dimensions and its implications.

### 2.3.3 Cognitive Proximity

Cognitive Proximity can be defined as the shared knowledge base and expertise of different entities, labeled sometimes as actors, and by Cunningham and Werker (2012) as agents. Given this definition for this category, one could argue that the concept is similar to Zeller's (2004) technological proximity, because it relates to the sharing of knowledge bases and experiences. However, we could say that a cognitive field is a broader concept than a technological area, which is why sometimes technological proximity has been defined as a subset of cognitive proximity.

Boschma (2005, p. 63) features the importance of cognitive proximity for enabling agents with a common background in which to communicate, absorb, comprehend, and process new information. Cohen and Levinthal (1990) start from the analogy of how an individual's prior knowledge and background is related to their cognitive basis, and coin the term *absorptive capacity* to define the ability for firms to achieve these tasks, and exploit its outcome in innovation activities. Accordingly, Nooteboom (2000) states that for learning reasons, that is, to obtain information and develop knowledge, the scope of cognition must be extended by the interaction with external partners. He also declares that comprehension at the company level can be improved when improving the company's absorptive capacity, keeping in mind that it is key to have an awareness and understanding of the current information the firm already possesses (Cohen and Levinthal, 1990). Thus, cognitive distance between entities is required to be at an optimal level in regards of their absorptive capacity, to be able to learn and realize opportunities in the others' knowledge base and expertise.

Knowledge is cumulative, localized, and tacit by nature (Antonelli, 1995). This author claims that it is because of this nature that you can generally find cognitive differences between agents. Similarly, Boschma (2005) states that in a company, the processes of knowledge creation and innovation have a high degree of tacit knowledge themselves, and are the output of cumulative and localized research. Consequently, the cognitive base of different organizations will tend to be different from each other, along with their absorptive capacity and learning potential.

Cognitive proximity between agents increases the chances for collaboration, since firms have an established tendency to look out for partners close to their own knowledge base. In addition, according to Perez and Soete (1988), in order to obtain a new technology, companies require a minimum level of knowledge, without which they would be incapable of moving across their knowledge gap. Given that there is better communication

and absorbing new information is more likely because of a better understanding produced by having close cognitive bases, being cognitively proximate ensures a more valuable outcome (Boschma, 2005; Boschma and Lambooy, 1999).

When it comes to innovation development and knowledge exchange, agents have a tendency to look for collaboration partners that are both close but at the same time complementary in their respective cognitive and technological capabilities. This is part of what Boschma and Frenken (2010) call *an evolutionary approach*, where partner selection is affected by favoring some agents over others in terms of their current knowledge base.

### **Cognitive Proximity and Interdisciplinarity**

Let us expand the idea of having agents coming from different knowledge fields or disciplines, to the concept of *interdisciplinarity*. Thus, Boschma and Frenken (2010) relate interdisciplinarity research collaboration to cognitive proximity, when visualizing the interactions of scientists from complementary fields has proven helpful to obtain an overview of interdisciplinarity in academic research outputs.

Schummer (2004) defines a scientific discipline as a category combining a cognitive and a social body, distinguished by the scientific community itself, especially, in cases where two disciplines share much of their knowledge set (e.g. biochemistry and molecular biology).

Sometimes, the term *multidisciplinarity* is used indistinctly and interchangeably with interdisciplinarity. However, there is a fundamental difference between these two terminologies, as explained by Klein (1990, pp. 56-63): When multiple disciplines are participating in the same research field, but there is no interaction between them, such field is said to be highly multidisciplinary, but not interdisciplinary. Likewise, strong interdisciplinary research (interaction between different fields) can exist without having a high degree of multidisciplinaryity.

Interdisciplinarity can sometimes be a rather vague and ambiguous concept, because the underlying cognitive and social processes can substantially differ depending on the collaborative project, the techniques used, and the social practices involved (Rafols et al., 2010). Furthermore, the author mentions that interdisciplinary research not always implies collaboration between researchers from different disciplines (Bordons et al., 2005). Plus, when it does mean collaboration, this term may refer to various practices (Laudel, 2001).

In any case, there is an increasing consensus that interdisciplinarity is characterized by knowledge integration (National Academies Committee on Science, Engineering, and Public Policy (COSEPUP), 2005), in the sense that it is a blend of concepts, tools, techniques, and the information or data itself, coming from numerous bodies of specialized knowledge. In the light of this definition, Rafols et al. (2010) note that there has been a significant increase in its adoption since the early 1990s by both scientists and policy-makers.

It is noteworthy to mention that despite the encouragement for interdisciplinarity in scientific and innovation research, this used to not be the case at all (Rafols et al., 2010), with the direct opposite being stressed instead. In this regard, Weber et al. (1946, p. 4) emphasizes that only strict specialization would ensure lifelong scientific achievements, warning that “*whoever lacks the capacity to put on blinders, so to speak, ... may as well stay away from science*”.

Nowadays, it would appear that the steps towards greater interdisciplinarity are driven by the need to meet the demands from society or the industry. Moreover, the topic of interdisciplinarity is closely related to innovation in knowledge production (Weingart, 2000, p. 30). This has pushed the scientific realm to recommend it for dealing with issues and better supporting innovation and competitiveness (Rafols et al., 2010).

Additionally, policy reports have highlighted the importance of interdisciplinarity for strategic technologies such as nanotechnology (Malsh, 1997). As a matter of fact, all funding programs in nanoscale research take a trans or interdisciplinary approach, so it is vital for the field’s development (Schummer, 2004).

### **Risks in Cognitive Proximity**

Ultimately, although having interdisciplinary agents that are cognitively proximate is beneficial, we see that when the cognitive distance is too short, this could potentially represent a hindrance for learning, knowledge production, and innovation. Boschma (2005) provides three reasons to explain why some cognitive distance should necessarily be kept among agents:

- In the first place, the construction of new knowledge is dependent to some extent on having access to complementary bodies of knowledge. Thus, agents will increase their probability for learning new information from different sources, which will, in consequence, trigger new ideas and creativity (Cohendet and Llerena, 1997).

- Secondly, having too much cognitive proximity may result in “cognitive lock-in” when agents are closed to new technologies or market possibilities. To avoid being too cognitively proximate, organizations should guarantee a diversified supply of information and be broad-minded to the outside world.

- Third, a small cognitive distance increases the risk of involuntary spill-overs, which in turn makes competing agents to become highly unwilling to share information.

In short, agents need to be separate enough so that knowledge can be combined in new ways, to avoid lock-in potential issues, and decrease in learning possibilities (Boschma, 2005). However, when there is cognitive distance between agents, the tendency for their capacity for obtaining new knowledge increases, but, it could restrict learning because of problems of communication.

Therefore, a compromise must be reached between cognitive distance, in behalf of novelty, and cognitive proximity, in behalf of absorptive capacity. Nooteboom (2000, p. 153) affirms that information is useless

if it is not new, but it is also useless if it is so new or so different that it cannot be understood. As a way to mitigate the potential risks in cognitive proximity, Maskell (2001) proposes that there must exist a geographical cluster with a common knowledge base composed of diverse, but complementary, knowledge resources.

In any case, it is noteworthy to be aware of the importance of having a balanced multidisciplinary environment for generating knowledge, due to the risks involved when balance is not achieved, as seen in the preceding section. We mentioned that a geographical cluster with a knowledge base built from diverse and complementary cognitive resources would be beneficial to attain this balance. Thus, we go into the second proximity dimension targeted in this work: geographical proximity.

### 2.3.4 Geographical Proximity

The influence of geography over learning and innovation has ever interested academics from different research fields (Boschma, 2005), making geographical proximity the most common dimension throughout proximity literature (Knoben and Oerlemans, 2006, p. 74).

Gittelman (2007) explains that learning has a collective nature, where linked individuals profit from contributing knowledge at the community level. Geography would in turn favor the transmission and sharing of knowledge, when taking into account that “*knowledge traverses corridors and streets more easily than continents and oceans*” (Feldman, 1994, p. 2).

We remarked previously on how proximity was originally attached to the notion of physical location, first researched from a metropolitan and agronomic viewpoint (von Thünen, 1826). Later on, Marshall (1890) brought new theoretical approaches for geographical proximity research within an economical and industrial context. He highlighted the importance of physical closeness between firms with his famous saying: “*the secrets of industry are in the air*”, stressing that localized spill-overs contributed to form a beneficial “*industrial atmosphere*” (Shaw and Gilly, 2000) that would boost innovation.

We first find geographical proximity as part of the French School of Proximity Dynamics framework, though the French academics did not limit the definition of this dimension of proximity with physical and natural constraints, claiming that they were not sufficient (Gilly and Torre, 2000). Instead, a social aspect was included, to allow for economic mechanisms and society factors to characterize what they define as *functional distance*, to be combined with the spatial aspect within geographical proximity. Among these influencing factors, we can find access time, affected by transport infrastructures, or the financial resources that would facilitate the use of ICTs (Information and Communication Technologies).

Likewise, we see how in the second studied categorization, Zeller (2004) states that costs and resources monitoring can decrease when having quick face-to-face interactions, possible when agents are in close-by locations. Moreover, spatially proximate actors would belong to an interpretative community, making

them prone to benefit from “noise” from their peers (in the form of suggestions, approaches, rumors, etc.) (Grabher, 1980, pp. 366-9).

Thus, geographical proximity is defined by Boschma (2005) as the spatial or physical distance between agents, either measuring it with an absolute metric of length (i.e. actual distance units, like meters), or a relative metric (e.g. travel time) (Boschma and Frenken, 2010). In their two-dimensional proximity model, Moodysson and Jonsson (2007) favor the same definition, in turn labeling it as *functional proximity*. They indicate that in society nowadays, more important than the location of actors, is the amount of effort required for their interaction, or face-to-face contact (see also Ter Wal and Boschma, 2009). This would also account for accessibility, which is limited by the time and cost variables involved in the mobility of actors. Ultimately, we see that all the proximity models introduced agree that the definition of geographical proximity should not be constrained to bare Euclidean physical distance.

In any case, Boschma (2005) remarks that it is imperative to have geographical proximity delineated in such a restrictive way, that it does not contain involvement from other dimensions, and rather keep it isolated for analytical purposes. The goal of this differentiation is being able to detect when spill-overs are influenced by geography only, without other kinds of proximity.

### **Risks in Geographical Proximity**

Nevertheless, though spatial concentration could bring significant benefits, having too much geographical proximity between agents could also be harmful for knowledge production, and innovation.

One potential risk is that agents within a region could become a too “inward-looking” community with weakened learning ability, and lose their capacity to come up with new ideas or respond to new developments (Boschma, 2005; Gittelman, 2007). This is why regions highly specialized in a certain technology have to make an effort to keep open to external flows of knowledge and solve or avoid situations of spatial lock-in. In fact, the lack of openness to the outside world is actually the factor that combined with geographical closeness, negatively affects innovation and learning (Boschma, 2005).

Likewise, scientists could also suffer from what Merton (1973) calls *scientific parochialism*, when restricting their interactions to local colleagues would cause them to stay out of the loop of important information flows in their field. Nonetheless, the existing norms in the scientific research community would prevent these to happen, since the methods for publication and exchanges of draft papers would help scientists to keep updated with work from peers, regardless of geographic distance (Gittelman, 2007, p. 13). She goes on highlighting the importance of attending events and conferences to stimulate knowledge exchange and evaluate the possibilities to establish a collaboration relationship among distant agents with shared interests.

Yet, another risk of having short distances between agents is that it might make them fearful to share information and lose competitive advantages. The fact that companies could opt to keep the projects they



wish to patent in secret, would be a major concern, because it would mean that this knowledge would be missing from published data (Gittelman, 2007).

However, Tallman et al. (2004) claim that scholars nowadays are leaving behind their concerns on competition, and are heading to form communities of knowledge that interact with near companies. These knowledge exchanges among co-located firms would help to create a knowledge-based theory for the existence of regional geographic clusters (Maskell, 2001). Moreover, spatial proximity would enable greater transparency and lead to stronger benchmarking activities (Lublinski, 2003). In regards to potential patent projects, Gittelman (2007) explains that it would not be a potential source of bias, given that any published work has a grace period of 12 months to be patented, and thus, should not be a cause of concern.

### **Relevance of Geographical Proximity Nowadays**

Throughout the literature, several authors bring to question whether geographic proximity may still be considered a relevant factor for learning and innovation nowadays (Lublinski, 2003; Boschma, 2005; Gittelman, 2007; Dettmann and Brenner, 2010; Bouba-Olga and Ferru, 2012). It has been argued that the networks through which knowledge flows do not necessarily need to be constrained to a specific location anymore. Furthermore, the concept that geography should have any impact whatsoever on knowledge production would constitute a paradox in itself, given the intangible nature of new ideas and their potential to diffuse widely (Gittelman, 2007; Sonn and Storper, 2008). Nonetheless, there is evidence that suggests that there are strong clustering tendencies for location in some analytical sectors, such as biotechnology (Asheim and Gertler, 2005).

Several authors state that spatial proximity no longer matters because digital means are substituting more and more the need for co-location. In a global economic world as we have these days, distance is not as indispensable as much as it used to be before, back when the information and communication technologies (ICTs) as well as transportation means were not as developed (Boschma, 2005; Sonn and Storper, 2008; Bouba-Olga and Ferru, 2012). Nowadays, “*geographic separation no longer implies information deprivation*” (Feldman, 2002, p. 1), with technology making its exchange regardless of great distance a reality.

Some authors have even gone as far as dictating the death or “liberation” from the distance constraint for communications (and hence, interactive learning) (Cairncross, 2001), thanks mainly to the internet, among other factors like lower travel expenses and using English as a global language (Rallet and Torre, 1999). On the contrary, according to Frenken et al. (2010), the idea that geography is not important anymore for scientific collaborations is just a matter of popular belief. Plus, since the studies performed to test this suffer from important limitations, the death of the distance hypothesis has not been proven (Frenken et al., 2010).

On one hand, it is unquestionable that scientists can communicate well over long distances, and we have examples of this happening at high-tech centers around the world (such as Silicon Valley in the US, Shinju

Science Park in Taiwan, and Bangalore Software Park in India) (Sonn and Storper, 2008). On the other hand, Feldman (2002) explains that the impact of internet is limited due to the tacit nature of knowledge itself, and the social nature intrinsic in the innovation and learning processes.

There are two major branches in the discussion about relevance of geographical proximity: one is when considering its impact on knowledge production, and the other branch is when considering its effects on innovation. Furthermore, some authors state that innovation, a particularly knowledge-intensive economic activity, is the only exception where geography still counts at all (Bouba-Olga and Ferru, 2012). However, we cannot separate the two, as they are closely related and dependent upon each other.

It would seem that there are significant benefits of spatial proximity for technological innovation, and that they go hand in hand with specific mechanisms that facilitate rich knowledge exchange within a region. These conclusions were drawn from studies conducted using the only means of quantifiable data that would be considered the evidence of new technology: patents.

Indeed, Feldman (1994) found a positive correlation between the amount of knowledge-generating inputs and technological innovation produced within a region. Shapira and Youtie (2008) call this phenomenon the *Strong Path Dependency Hypothesis*, indicating that geographic clusters of knowledge production have a strong correlation to either previous high technology waves (like biotechnology) or to the presence of powerful institutions (such as academic or government labs). These factors would then become anchors for nanotechnology knowledge-generation (Delemarle et al., 2009).

Accordingly, some studies conclude that knowledge spill-overs from academic centers to businesses happen because spatial proximity increases the productivity of R&D (Adams and Jaffe, 1996; Sonn and Storper, 2008), Jaffe (1989) being the pioneer with his research on each US state. Therefore, the dynamics of being physically close to sources of new knowledge would account for innovative and entrepreneurial activity to cluster geographically in places such as Silicon Valley (Feldman, 2002; Gertler and Wolfe, 2006).

When speaking about the importance of geographical closeness in research collaborations, several authors in the literature point out that to establish the required partnerships, its significance or necessity is decreasing more and more (Boschma, 2005; Gilly and Torre, 2000; Bunnell and Coe, 2001; Breschi and Lissoni, 2003; Gertler and Wolfe, 2006).

In their study on European research collaborations, Bouba-Olga and Ferru (2012) hypothesize that these conclusions might have been drawn because the existing empirical studies are limited due to the collaboration indicators used. Indeed, only codified knowledge would be used to measure the knowledge spill-over process (Howells, 2002). Moreover, the indicators to measure geographical proximity, such as differentiating between national vs international collaborations, and between administrative divisions within the same country, would impact the results.

Likewise, Howells (2002) criticizes that although these studies seem to confirm that knowledge activity

is spatially constrained, the analysis of the mechanisms of knowledge transfer and sharing is greatly lacking. To this respect, Cairncross (2001) affirms that if we are reaching a point where distance is dying, it should be evident that the importance of proximity is decreasing.

However, it is hard to determine whether this importance is increasing or decreasing, due to most studies of knowledge flows using data which does not allow for historical changes versus the spatial factor. In other words, the research performed has been somewhat static, meaning that the measurements are taken at one point in time (Boschma, 2005; Sonn and Storper, 2008).

Ultimately, even though it is true that the widespread use of internet has allowed easy and quick access to information regardless of its location (Feldman, 2002), globalization does not necessarily have to signal the “death of geography” as if they were binary opposites (Morgan, 2004). Instead, once the agents involved in the processes of innovation and knowledge creation are freed from location constraints, geography could in fact become more important (Feldman, 2002; Sonn and Storper, 2008).

Globalization and localization would then be considered as complementary processes, with geography as a framework for individuals and resources in which the spill-overs associated with knowledge creation take place (Morgan, 2004). For instance, technological companies tend to locate where labor markets with pools of specialized scientists or engineers are readily available (Almazan et al., 2007; Sonn and Storper, 2008). Paradoxically, improving ICTs and the integration of a global market would result in localized technological interactions (Sonn and Storper, 2008) and a stronger localized flow of knowledge (Leamer and Storper, 2001).

Finally, the dimensions of geographical and cognitive proximity would give an overview of common environmental settings for agents. Yet, studies going in depth on these attributes are unable to fully explain the dynamics nor the fundamental characteristics of collaborative relationships (Dettmann and Brenner, 2010, p. 3).

Boschma and Frenken (2010) indicate that a dynamic proximity framework would be the basis for the geography of network formation, remarking on the growing role of networks in the study of fields with high interdisciplinarity over the last two decades. Nevertheless, they also claim that network analysis is still greatly underdeveloped in the geography of innovation and academic learning.

We have seen how the social aspect is mentioned every so often as being primordial for collaborations that generate knowledge, so we will now review the third component of our research: collaborative proximity.

### **2.3.5 Collaborative Proximity**

“*Science is done by humans*” (Heisenberg, 1969), and scientific collaboration in academic research is a complex social phenomenon (Glänzel and Schubert, 2005) that scholars have been trying to measure ever since the 1960s (de Solla Price, 1965). Indeed, nowadays academic knowledge is generally seen as a social accomplishment (Hyland, 1999). It is then evident that scientific activities are not entirely devoid of influences

from social aspects, despite the ideal definition of science as a systematic pursuit for objective truth, without any effect from personal beliefs, biases or social influence (Sarigöl et al., 2014).

*Collaborative proximity* represents the distances between scholars and their position within academic collaboration networks. However, the fact that collaboration is based upon social structure raises the question: Why do we not match this aspect to the social proximity factor defined earlier in the literature as part of proximity dimensions? As a matter of fact, we even find that the term *social proximity* has been used for this purpose in the literature (see Sorenson et al., 2006). Nevertheless, we take into consideration that there is more to the social proximity aspect than just mere acquaintanceship, because it may or may not include cultural aspects, requiring a less structured definition of the social context to which the scholars belong to. Moreover, according to Newman (2001b, p. 1), a scientific network may not be the best scenario to quantify a purely social component. This because they do not directly measure actual contact between people, though their structure would be a reflection of the society that built them. Subsequently, we rather coin the term *collaborative proximity*, adding it to the list of proximity aspects addressed in this study, by virtue of adjusting to the proper terminology in proximity frameworks.

If we consider that original ideas arising from academic publications or technological innovation are the product of collaborations between agents (either scientists or inventors) (Eslami et al., 2013, p. 1), collaboration is thence key for knowledge production. Moreover, there is an increasing importance for using natural mechanisms of cognition and information filtering for scholarly purposes through social connections (such as collaboration links). This importance is a reasonable response to the fact that available research is increasing, and that we have a limited ability to keep track of potentially relevant articles (Sarigöl et al., 2014).

However, how can we quantify collaboration? First, we must clarify that the scope of collaboration throughout this thesis refers to scientific individual collaboration or “inter-individual” (as labeled by (Katz and Martin, 1997)), as opposed to collaboration at the institution or regional level. There are many ways that show how researchers collaborate, such as co-authorship, co-partnership in projects, and co-citing publications (Jiang, 2008). Still, co-authored publication (also, *multiple-author publication* (Katz and Martin, 1997)) is typically considered as the most visible and well documented indicator for collaborative activity among scholars (Abbasi et al., 2010; Glänzel and Schubert, 2005). In scientist networks, people co-writing articles would know one another quite well (Newman, 2001b).

In such respect, Katz and Martin (1997) bring to our attention that despite the abundance of studies using this technique to investigate collaboration, there is a clear distinction between the concept of collaboration and co-authorship, and that they should not be deemed as synonyms. Among the limitations of co-authorship metrics, we find that sometimes scholars would be included as co-authoring a paper merely because of social conventions, and that in reality they could not be attributed to the research work itself, nor its credit

(Hagstrom, 1965). The opposite could also happen, when for example, two researchers have actually worked together, but decide to publish two independent field-targeted articles respectively (Katz and Martin, 1997). It could also even go to the extremes of being part of scientific fraud, when the so-called “honorary co-authors” are not even aware of their listing in a publication (LaFollette, 1992, pp. 97-101).

Another concern is how the magnitude of contribution varies during the timeline of a research project, due to the complex nature of human interaction among collaborators (Subramanyam, 1983, p. 35). Therefore, Katz and Martin (1997) insinuate that the only way in which co-authorship could truly be an accurate reflection of collaboration would be by setting a well-defined scale of “joint work intensity”, in which researchers would only be listed as co-authors if they surpassed a certain threshold. But, at the same time, they also state how impossible setting or enforcing such a criterion would be.

On the other hand, Glänzel and Schubert (2005) dismiss these caution warnings to some extent. For them, even if co-authorship were considered as a partial or approximate indicator of scientific collaboration, studying it allows us to gain valuable insight into measurable interaction between collaborative work and performance metrics of scientific communication, particularly in science-based fields such as nanotechnology (Cunningham and Werker, 2012).

Additionally, thanks to the scientific community’s strict authorship regulations, we should be able to safely assume that each co-author has indeed significantly contributed to the common research in the paper (Schummer, 2004). Furthermore, more intense collaboration would go hand in hand with increasing co-authorship between researchers, (Patel, 1973), and thus, they conclude that there is an undeniable positive correlation between the both (Glänzel and Schubert, 2005).

All considered, co-authored papers are a proxy measure adequate enough to quantify collaboration among groups of researchers. Moreover, studies and reports show that scientific collaboration (with co-authorship as metric) has intensified in all science areas, though it has not yet been proved whether this effect was caused by the formation of stable contributor teams, or if it is rather due to the temporary creation of casual links (Glänzel and Schubert, 2005). Anyhow, two scientists who have worked together at least once are more likely to later keep in touch for meaningful information exchange, and thus, it would presumably result in repeated collaboration between researchers (Agrawal et al., 2006; Beaudry and Schiffauerova, 2011).

Finally, the fundamental process of science is communication, without which, science itself cannot exist (Lievrouw, 1989). It is critical then to establish communication and this is where proximity comes to facilitate it, by bridging gaps that would otherwise keep academics from learning and exchanging vital information for the development of science.

## Collaboration and Co-authorship

We find in the literature that there are three established methods for studying scientific collaboration and endorsement (e.g. Milojevic, 2010; Ding, 2011): qualitative methods (like using surveys, interviews, or observations), bibliometric methods (using publication counting, citation counting, or co-citation analysis), and complex network methods (like shortest path, centralities, network parameters, or PageRank/HITS).

From these, there are two approaches that have the potential to explain the structure of scientific communication, which are the use of citation analysis (part of bibliometrics) and social network analysis (Marion et al., 2003). These two mechanisms have a lot in common, which is why researchers in each field tend to use similar tools when conducting their studies (Sternitzke et al., 2008; Marion et al., 2003).

A research paper is “a rhetorically sophisticated artifact” with a careful balance of both factual information and social interaction (Hyland, 1999). Based on scientific documents, citation analysis has been widely used in science research since a long time ago (e.g. Hummon and Dereian, 1989), and it is a valid way to describe relationships between scientific authors (Otte and Rousseau, 2002). Likewise, co-authorship has been traditionally studied with bibliometric analysis tools. Moreover, we find authors in our review claiming that almost every aspect of scientific collaboration can be tracked through bibliometric techniques to analyze co-authorship networks (Glänzel and Schubert, 2005).

However, bibliometric methods still have some shortfalls for co-authorship studies, because they lack ways to see how it interacts with other important processes of scientific communication, such as publication activity and citation behavior (Glänzel and Schubert, 2005). We find another limitation in that it is yet unable to examine some aspects of scientific collaboration, particularly when considering scholar research interests and social connections (Ding, 2011). This happens because in bibliometric analysis the focus is on ranking individual nodes, which causes it to overlook the relationships found between two specific nodes. It would then lack the capacity to discover scholarly communication patterns (notably for collaboration and knowledge diffusion) with finer granularity (Ding, 2011).

On the contrary, the focus in social network analysis is on the characteristics of the relationships or ties, rather than on the intrinsic characteristics of the individual members (Wetherell et al., 1994, p. 645). Remarkably, previous research shows that coauthors tend to cite each other sooner after co-publishing a paper as compared to non-coauthors (Martin et al., 2013). This strong tendency towards reciprocal citation patterns (Bethard and Jurafsky, 2010) would already give an inkling as to the influence social aspects have over scholarly citing behavior.

Furthermore, we have previously seen how scientific networks seem to be characterized by collaboration that takes place in many informal ways, (e.g. mass media and interpersonal channels like conference participation). This would then represent another impediment for tracking knowledge spread solely by the use of conventional (bibliometric) measures for spill-overs, such as cross-citation and co-publication (Murray,

2002; Rogers, 2002), without considering the social connections between researchers found when taking the collaboration network into account (Beaudry and Schiffauerova, 2011).

Nowadays, we find more and more that social network analysis is used to analyze the way scientists are interconnected (Beaudry and Schiffauerova, 2011). Social network analysis is a methodology for studying formal communication networks, as opposed to the communication ties that naturally occur in informal networks, like between colleagues, or members of an institution (Marion et al., 2003). The interest in social networks has rapidly been increasing for several years now (Borgatti, 2003; Beaudry and Schiffauerova, 2011), and its usage spans to a wide variety of fields.

Furthermore, many empirical studies that applied it for researching scientific communities have proven how powerful this kind of analysis is for understanding the growth and spread of information (Marion et al., 2003). Newman (2001b) exemplifies the high potential of information spread with the famous experiment conducted by Milgram (1967), who took a practical approach to demonstrate the *small-world hypothesis*. This experiment suggests that pairs of scientists in a population typically have a short path of intermediate acquaintances between them, even when the size of the population is very large. In other words, every researcher would “know someone who knows someone”, and thus information or knowledge diffusion would use collaboration networks as pathway.

We find some examples of this kind of analysis applied to co-authorship (e.g. Newman, 2001b, 2004; Barabási et al., 2001), where the focus is on the structure of scientific collaboration networks, taking co-authorship patterns from individuals as basis. Ultimately, collaborative proximity would be an influential factor for knowledge diffusion in the scholarly realm, even if there could be involvement from other aspects.

### 2.3.6 Interaction between Geographical Proximity and other Dimensions

Although the *definition and measurement* of proximity factors need to be strictly demarcated, this does not forcefully imply that their effects should be analyzed without involving other dimensions. For instance, the effect of the spatial dimension is not usually direct or evident, but it is rather a subtle and varied influence (Howells, 2002; Dettmann and Brenner, 2010). This is what makes it so difficult to verify the claim of whether or not geography is significant, since a methodology that completely isolates the geographical effects from other factors has not been reported yet Katz (1994).

Howells (2002) argues that geography influences all knowledge activity in five specific ways, from the perspective of the individual or agent. If we regard an individual as a “knowing self”, we can say that this entity is influenced by the various factors related to geography, such as social, cultural, and economic. Likewise, human interaction, the exchange and interpretation of information (in codified and tacit forms, also involving past experience), and learning itself (in all the multiple settings where it can take place) are all affecting the individual. And they are all closely related to geography, distance, and proximity.



Throughout the literature, we find that knowledge transfer can be associated with other proximity dimensions other than the spatial one alone. This has been accepted even by some of the strongest supporters of a regional world, such as Storper (1997).

We see examples of interactive learning taking place in settings where other dimensions, such as technological or organizational, are more relevant than the geographical aspect, like in a multinational corporation (MNC) (Zeller, 2004). This is due to the existence of information and communication technology nowadays, which in turn allows for networks to exist and be kept alive outside of spatial constraints (Rallet and Torre, 1999). Given the non-territorial definition of networks, it would consequently follow that knowledge spill-overs are not necessarily spatially bounded (Bunnell and Coe, 2001).

Hence, Boschma (2005) argues that geographical proximity should always be assessed with regard to other dimensions, given that spatial concentration alone is not an absolute generator of effective synergies (Zeller, 2004, p. 5). We could say that it rather enforces or strengthens other proximity categories, because it is a decisive factor in the channels through which knowledge is generated and transferred, which are the cultural, social, and psychological spaces (Howells, 2002). Moreover, when evaluating the prerequisites for learning to take place, geographical proximity alone is neither a necessary nor sufficient condition, despite the fact that it makes interaction and cooperation easier (Malecki and Oinas, 1999; Boschma, 2005).

In particular, it is important to consider the social aspect involved in geographical proximity, given that social interactions are a key element for all the dimensions implicated in innovation (Zeller, 2004; Tallman et al., 2004; Murray, 2002). Proof of this would be the correlation for better innovative performance when creative technological companies are located near knowledge sources when compared to that of companies placed farther away (Jaffe and Trajtenberg, 1999).

Thus, a regional cluster formed by firms tied together by the links of geographical co-location and complex social interaction (Tallman et al., 2004) will be more inclined to geographically-favored spill-overs happening around agents concentrated in space. Likewise, the larger the combined geographical and social distance between agents, the less the intensity of these positive externalities will be (Boschma, 2005). Simply put, this means that when people are situated nearer, making new contacts as well as exchanging tacit, non-verbal knowledge between the parties is easier, so much so that these informal understandings in turn would contribute to sharing technical knowledge (Katz, 1994; Tallman et al., 2004; Boschma, 2005; Dettmann and Brenner, 2010).

Either by accidental meetings or by introductions from a third common party, spatial proximity allows for initial meetings that trigger the process that defines the starting point of a collaborative relationship (Dettmann and Brenner, 2010). Storper (1997) labels the exchanges of perceptions and shared experiences originated from regular social interactions as *untraded interdependencies*. These interdependencies are formed by informal rules and codes of conduct, running in parallel with the established and formal mechanisms in



which information exchanges would typically take place (such as market procedures for licensing). Since interpersonal interaction is a requirement for untraded interdependencies, Tallman et al. (2004) state that they are more likely to be tied to geographic location, as opposed to regular economic transactions, which can spread more widely. In this respect, personal academic ties would seem to overcome geography, a behavior which might have been accentuated with the growth of Internet use (Murray, 2002).

Given that location proximity to scientific and technological knowledge would not suffice, specific transfer mechanisms need to be in place for the diffusion of externalities from where they origin to where they are ultimately implemented (Boufaden and Plunket, 2007). Since scientists are not engineers (Allen et al., 1977), these two groups have different ways to create, communicate, and draw value from knowledge (Gittelman, 2007). Thus, Boufaden and Plunket (2007) remark that identifying technological opportunities and facilitating communication require dense networks of researchers, technicians, and entrepreneurs.

We introduced previously the significance of networks as pathways for knowledge flow, particularly for knowledge externalities or spill-overs. On the one hand, it is likely that geographical proximity is a prerequisite for the very existence and sustaining of these social networks themselves, thus, making the consequent knowledge spill-overs to be spatially-bounded (Boschma, 2005).

On the other hand, being involved in the social connectedness of agents networks within a specific region is key, because these social networks would exclude outsiders, even if the agents are locally situated (Hudson, 1999). The importance of this is exemplified by an MNC trying to tap the knowledge base of a host location by setting up a local branch there (Blanc and Sierra, 1999), and failing because its members are not part of the tight networks of personal relationships through which local knowledge flows (Breschi and Lissoni, 2003; Boschma, 2005).

Dettmann and Brenner (2010) emphasize the social aspect, categorizing collaboration stages in terms of trust, which would be the outcome of personal interaction between agents, and the way the agents would evaluate reputation or prestige (Gittelman, 2007) and overall trustworthiness. In this respect, the frequency for these face-to-face contacts would be enhanced by spatial proximity, since it would be easier for the processes that help building trust between agents to take place (processes such as learning about character, respective motives, and sociocultural background) (Lublinski, 2003). Katz (1994) goes as far as to say that this communication may lead to gradually greater commitments for cooperation in a fashion similar to courtship. In any case, social networks would be the main and most productive channel for knowledge diffusion (Breschi and Lissoni, 2001), by automatically leading to cooperative behavior for its sharing and common learning processes (Asheim and Gertler, 2005).

However, not only social networks are required, or could substitute spatial nearness. Indeed, some of the other dimensions could compensate the need for geographical proximity. As a matter of fact, Boschma (2005) further hypothesizes that when spatial nearness is combined with some level of cognitive proximity,

it is a sufficient condition for interactive learning to occur. Thus, with the dynamics of knowledge creation and social networks, it is unlikely that the scientific knowledge of a specific domain would be demarcated within a single region (Gittelman, 2007). Rather, scientific communities would be geographically dispersed, with local groups being part of bigger communities that collectively respond to similar social and intellectual forces Merton (1973). Companies or academic institutions need to find ways to create proximity between their agents wherever they may be. Hence, they must either bridge the gap caused by physical dispersal, or take advantage of the benefits of concentration when it does exist (Schoenberger, 1997, p. 21).

Furthermore, there seems to be an inverse relationship between geographical and cognitive proximity, meaning that only when cognitive proximity is low can then geographical nearness become important to overcome this gap (Freel, 2003). Singh (2005) supported this hypothesis with patents data, where he found that in interdisciplinary research collaboration (when cognitive proximity is low) geographical proximity plays an important role, whereas greater spatial distances are more common when agents work in the same field (that is, cognitive proximity is high).

If we look at organizational proximity (or relational proximity), we find that there could also be an involvement of geographical proximity, in terms of the ease it brings to creating and maintaining institutional practices, such as codes of conduct, norms, and habits (Boschma, 2005). Moreover, Kraut et al. (1988) relate three dimensions: spatial, organizational, and social in terms of the probability of collaborations, proving that spatial proximity has a positive impact on it. An exemplary scenario would be agents having their offices on the same floor or in the same building, forcefully having a higher frequency of interaction, such as unintended meetings, or getting together to have lunch. At the same time, this proximity dimension could also be in many cases more important and direct than geographical proximity (Amin and Cohendet, 1999), or even act as a valid replacement (Rallet and Torre, 1999), particularly when a strong central authority (i.e. the managerial team in a MNC's headquarters) is the one coordinating all tasks and information flow (Boschma and Frenken, 2010).

Nevertheless, despite having other dimensions involved, which could maximize or even substitute the need for spatial closeness, it is essential to mention that personal contact was still required for exchanging tacit knowledge. In this respect, distance seems to be a discontinuous variable, meaning that the advantages of spatial proximity are not linear to distance (looking more like a Gaussian curve). It follows that “*to be within walking distance is very different from being slightly farther away*” (Sonn and Storper, 2008, p. 2).

Accordingly, the communication mechanisms that operate within close proximity are replaced by more formal means (like publication) beyond a certain distance where informal channels are no longer available (Adams and Jaffe, 1996). However, Boschma (2005) claims that this does not necessarily have to mean that the agents must have geographical proximity, if we refer to it as permanent co-location, but that these face-to-face contacts could be arranged by traveling.

All considered, we could say that even though we have influences of other dimensions, geography still plays a key role in learning. Particularly, when it comes by means of spill-overs happening around agents (Wallsten, 2001), allowing them to share knowledge in both formal and informal ways.

## 2.4 Citations

We have discussed by and large about the diffusion of scholarly knowledge and how proximity aspects might serve as channels for its flow. Yet, how do we track something as intangible as knowledge? Trajtenberg et al. (1997) affirm that even though knowledge flows are invisible, leaving no trail by which they may be measured and their patterns discerned, they do leave a paper footprint in the form of citations.

Citations have been extensively applied for studying knowledge diffusion across a variety of dimensions, and are considered to be valid measures for tracing out knowledge flows (Alcacer and Gittelman, 2006). We refer to de Solla Price (1965), who originally introduced the concept of scholarly research, for defining academic knowledge as a collection of highly cited papers that represent the frontiers of science. It follows that citation is key for facilitating scientific collaboration and enhancing communication within a scientific domain.

### 2.4.1 Defining Citation

Garfield (1998) labeled *citationology* as the theory and practice of citation. Citation is a basic component of an academic article, which helps authors to establish facts and communicate with others by setting the context of the knowledge addressed by the new contribution (Hyland, 1999). Moreover, in scholarly writing it is mandatory to explicitly refer to the work of others. This exigency is always enforced, given that “*new work has to be embedded in a community-generated literature to demonstrate its relevance and importance*” (Berkenkotter and Huckin, 1995, p. 3).

Formal referencing is crucial in scientific research, since citation counts are usually the raw data for evaluating scientific performance (Bornmann and Daniel, 2008), moreover, being cited is considered to be a critical goal in scholarship (Garfield, 1979). The significance of citation is highlighted by the fact that the number of citations within an article has been steadily increasing in time, which in turn has improved its value, by more focused and pertinent referencing (Hyland, 1999).

Furthermore, citations are viewed as “a complex and multidimensional phenomenon” (Bornmann and Daniel, 2008) due to authors using citations for different reasons and meanings (Garfield, 1998). Citing motivations were first categorized by Garfield et al. (1965), then by Brooks (1985), and finally by Cano (1989) and Shadish et al. (1995) (who in addition typified citation factors), by means of either the semantic content of the citing papers, or through citer surveys or interviews (Bornmann and Daniel, 2008).

However, Small (1973) criticizes that classification schemes have not been accumulative work, because each regarded their approach as unique. Indeed, useful though they may be, they suffer from methodological weaknesses that would affect their reliability and replicability for further citation categorization (Bornmann and Daniel, 2008).

Nevertheless, they all agree that there is more to the citing decision than mere content value or the need to acknowledge intellectual influences of peers; in fact, there is a wide number of non-scientific factors playing a role in this behavior (Bornmann and Daniel, 2008). These are discussed in the following section.

## 2.4.2 Citing Behavior

To cite or not to cite: what motivates scholars to reference particular works from fellow scientists?

Besides being an obligation in academic writing, we find that it is common to cite authors with a perceived degree of success (e.g. awards, Nobel laureateship, prestige, etc.) or whose publication has been deemed as a well-known “concept marker” (Case and Higgins, 2000). In other words, “famous” scholars pioneering in their field (their research is thus considered a classic reference), would likely be cited by those writers wishing to pay them homage (Garfield et al., 1965), or those believing that citing a prestigious work will promote the cognitive authority of their own paper (Case and Higgins, 2000). Remarkably, the importance of reputation is such that it causes citations to follow an approximately log-normal distribution, with notorious names roughly balanced by obscure ones, and authors of middling reputation taking up the majority (White, 2004, p. 93).

According to Garfield et al. (1965), the most frequent use of citations was what Brooks (1985) calls *professional motivations*, where they become the groundwork for the new publication by providing theoretical and practical (i.e. methodology and/or findings) content of prior authors. Sometimes, the reason for this is that preceding articles have a degree of creativity, involving unusual or innovative methods or theoretical perspectives (Shadish et al., 1995). All in all, writers use citations to “examine the products of science” (Lievrouw, 1989) and give credit to colleagues whose work they use; formal referencing would then represent intellectual or cognitive influence on scientific work (Bornmann and Daniel, 2008).

On the contrary, Radicchi et al. (2008) argue that citation does not necessarily reflect the scientific merit of the cited work (in terms of quality or relevance). Actually, it sometimes could even be superfluous, in cases where more references are required to meet compulsory citation numbers, thus becoming unnecessary (Brooks, 1985).

Moreover, although Garfield et al. (1965) recognize the need to substantiate claims by using *supportive citations* (Shadish et al., 1995), academics would not be solely motivated by the pure interest of using literature as a testimony that allows them to situate new work in the context of already accredited research (Hyland, 1999). As a matter of fact, instead of “giving credit where credit is due”, they frequently fail to

cite pertinent work, rather tending to only cite those whose views support their own (Cronin, 1982), which is why this citing motivation has been frowned upon by some authors in the literature. Not to mention that this strategy of supporting current claims (Hyland, 1999) is sometimes taken to the extreme, where citations rather portray the behavior of scholars “scouring the literature” for justification (Brooks, 1985, p. 227) and arguments that would allow them to persuade through “manipulative rhetoric” (White, 2004, p. 93).

In addition, another tool employed is the use of *negative citations* (Shadish et al., 1995), where previous works are criticized and their claims disputed, manifesting *negative homage* to the original writer (Garfield et al., 1965, p. 85). On this subject, Case and Higgins (2000) explain that citers using this tactic expect their criticism to generate a perception leading to establish them as authoritative, critical thinkers. In spite of this, Garfield et al. (1965) found that generally, scientific references were used more often for the positive purposes of citation than they were exploited as objects for refutation.

Anyhow, citing behavior would mostly suggest endorsement, authority conferral, provenance tracking, and scholarly trust (Ding, 2011). However, citation counts do not have the potential of yielding insights into the motives behind a writer’s citing behavior (Bornmann and Daniel, 2008). Indeed, since citations are sometimes made for social reasons (Shadish et al., 1995), citing behavior may be a simple indicator of more complex behaviors or social relationships (Lievrouw, 1989).

Likewise, constructing academic facts is deemed as a social process with several interactions: there is the need for acceptance based on negotiation with editors, reviewers, and also readers granting their ratification on the novel contribution (Hyland, 1999). The fact that social relations are important is made clear when, for example, new claims published by an academic whose credibility has been lost within a scientific community are pretty much ignored (Bornmann and Daniel, 2008).

Scholars seek to establish a persuasive and social framework for the approval of their arguments (Hyland, 1999), aiming to build social relationships in the scientific community through what is known as *connectional citations* (Brooks, 1985). In this regard, credit is attributed to stand for the interaction among the authors and those whom they cite (Lievrouw, 1989). By the same token, to understand the social processes of science, which is essentially communicative in nature (Lievrouw, 1989), we need to understand that citations are affected by social networks (White, 2001).

In this respect, Bornmann and Daniel (2008) take a critical stance on the two competing traditional citing behavior theories: the normative theory, which concerns the relevance of cited works, and the social constructive view.

In their findings, the latter contradicts the assumptions mentioned above about acknowledging useful background in prior publications, in turn stating that citations are influenced by personal bias and/or social pressures (e.g. by Brooks (1985): when a reference is included because the author depends on the cited writer in some way).

Among the social reasons, we find scholars citing papers authored by an influential reviewer (Shadish et al., 1995), wishing to build or maintain a professional connection to their writer (Brooks, 1985), because they want to gain the favor of editors, or colleagues (Vinkler, 1987), or simply to publicize their own or others' previous research (Brooks, 1985).

Furthermore, citation analysis reveals that clusters of research papers can be interpreted as networks of interpersonal contacts (Lievrouw, 1989), with authors mostly citing publications by people with whom they are personally acquainted (White, 2001). Additionally, scientists favor authors they have collaborated with in the past (Martin et al., 2013), resulting in a strong tendency in citation patterns (Bethard and Jurafsky, 2010). Altogether, communicative interaction exists among members of the network; though in the past, this aspect used to be neglected when describing social structures in citation (Lievrouw, 1989).

Few studies have aimed to discern the behavior that causes non-citations (Bornmann and Daniel, 2008), among which Cronin (1981) was the first to investigate the differences in writers' opinions that would lead them to question the necessity to cite. Even though there is a number of minor reasons inducing non-citing behavior, such as poor availability of the document, technical issues such as typos, and cultural aspects like language, it would seem that authors decide to make reference or not according to social reasons and specifically, acquaintanceship (Bornmann and Daniel, 2008). Moreover, Bornmann and Daniel (2008) claim that giving credit to intellectual influences is definitely not a priority for writers when it comes to choosing whom to cite.

In scientific literature, knowledge construction is in the hands of the scientific community members, so it follows that their decisions are socially grounded (Hyland, 1999). Nonetheless, citation potential varies among fields (Garfield, 1979), and representatives would be influenced by the inquiry patterns and knowledge structures of their respective academic domains (Hyland, 1999). Certainly, field variation affects the evaluation of scientific performance, causing different citation behavior according to discipline, due to, for instance, varying requirements for citation counts or unbalanced cross-discipline citations (Radicchi et al., 2008).

In conclusion, among the factors that influence citing behavior, social and cognitive reasons are involved to a great extent. In the next section, we review previous works of authors who have dealt with the effect of the targeted proximity dimensions on citation.

## 2.5 Impact of Proximity on Citations

As initially stated, our goal is to examine the influence of proximity on knowledge diffusion, as expressed by citation. Thus, in this section we include a comprehensive discussion (summarized in Table 2 of the objectives, methods, and findings of the most relevant empirical studies found in the literature addressing this topic.

## Authors studying proximity effects on citation

Author(s)	Focus	Proximity factors addressed				Variables			Data selected for analysis					
		Cog/Tech	Geo	Collab	Other	Target	Dependant Var	Network	Level	Participants	Database(s)	Location	Domain	Assesment
Baldi (1998)	Academic	✓		✓*		Citing behavior	Citation link	Social, no SNA	Micro	5,000 links		Global	Astrophysics	Logistic regression
Rafols and Meyer (2007)	Academic	✓		✓*		Crossdisciplinarity	Citation count		Micro	5 case studies		UK	Bionanotech	Descriptive stats
Ding (2011)	Academic	✓		✓		Collab and citation strength		Coauthor, citation	Micro	15,367 papers	WoS	Global	Info retrieval	Descriptive stats
Abbasi et al. (2011)	Academic			✓		Performance	Citation & pub. Count, g-index	Coauthor	Micro	2139 pub (5 uni)		USA	Sciences	Poisson multiple regression
Onel et al. (2011)	Academic			✓		Distribution fitting		Coauthor, citation	Micro	29,787 papers	WoS	Global	Nanotech	Distribution models
Wallace et al. (2012)	Academic			✓		Distributions	Citation count	Coauthor	Micro	Undefined	WoS	Global	Sciences	Distribution models
Liu et al. (2014)	Academic			✓		Citing hazard rate	Prob of citation at time t	Coauthor	Micro	16,582 pub (5 uni)	WoS	USA	Nanotech	Cox prop. hazards regression
Sarigöl et al. (2014)	Academic			✓		Scientific success	Citation count	Coauthor	Micro	108,758 pub	MSAS	Global	Comp Sci	Pairwise & rank-sum test; superv. learning
Liebeshkind et al. (1996)	Interaction			✓*	Organiz.	Interaction	Paper & patent counts		Middle	2 firms	USPTO & others	USA	Biotech	
Schummer (2004)	Interaction	✓	✓		Instit.	Interdisciplinarity	Author count per journal & field		Middle	600 papers	SCI	Global	Nanotech	
Boufaden and Plumket (2007)	Interaction	✓	✓			Interaction level	Patent application count		Middle	60 firms	EPO	Europe	Biotech	Spatial panel regression
Gittelman (2007)	Interaction			✓		Collab behavior	Paper-patent citations links		Micro	5,143 papers	USPTO, SCI	USA	Biotech	Negative binomial regression
Delemarle et al. (2009)	Interaction			✓		Scientific production	Publication count		Macro	538,000 pub	WoS	Global	Nanotech	Cluster analysis
Frenken et al. (2010)	Interaction		✓		Organiz.	Collab success	Citation count		Middle	Undefined	WoS	Netherlands	Sciences	Negative binomial regression
Wang and Guan (2011)	Interaction	✓		✓		Collab intensity	Author-inventor citation	Coauth.& Coinv.	Micro	275 patents, 1,447 articles	USPTO, SCI, IPC	China	Nanotech	Linkage network, descriptive stats
Laursen et al. (2011)	Interaction			✓		Collab uni-firm	Likelihood of collab link		Middle	8,724 firms, 99 uni	RAE	UK	Sciences	Nested logit regression
Bouhs-Olga and Ferru (2012)	Interaction			✓		Collab sci-ind	Author-inventor link		Middle	32,764 obs	CNRS, CIFRE	France	Sciences	Multinomial logit reg, sample selection
Cunningham and Werker (2012)	Interaction	✓	✓		Organiz.	Collab intensity	Coworks count		Middle	100 org, 5,050 collab	WoS	Europe	Nanotech	Negative binomial regression
Eslami et al. (2013)	Interaction			✓		Acad. prod. & tech perf.	Paper & patent counts	Coauthor	Micro	100,652 papers, 4,893 patents	USPTO, SCI	Canada	Biotech	Poisson multiple regression
Jaffe and Trajtenberg (1999)	Innovation	✓	✓		Organiz.	Citation intensity	Citation frequency and prob		Micro	50,000 patent-pairs	USPTO	Global	Sciences	Heteroskedastic probit regression
Hu and Jaffe (2003)	Innovation	✓			Organiz.	Patent citations	Citation count and frequency		Macro	59,116 patents	USPTO, NBER	Global	Sciences	Weighted nonlinear least square regression
Breschi and Lissoni (2003)	Innovation			✓		Citing behavior	Co-location link	Coinventor	Micro	3,716 links	EPO	Italy	Sciences	Odds Ratio analysis, correlation
Singh (2005)	Innovation			✓		Citing behavior	Citation link	Coinventor	Micro	2,540,991 links	USPTO	USA	Sciences	Choice-based samp. regression
Sorenson et al. (2006)	Innovation	✓	✓		Organiz.	Citing behavior	Citation link	Coinventor	Micro	72,801 links	Micro Patent, NBER	USA	Sciences	Logistic regression of rare events
Sonn and Storper (2008)	Innovation			✓	Organiz.	Patent citations	Citation count		Micro	17,761 patent-pairs	NBER	USA	Sciences	t-test
Agrawal et al. (2008)	Innovation			✓	Social	Patent citations	Inventor-patent-citation		Micro	261,888 obs	USPTO, NBER	USA/CA	Sciences	Two-way interaction means regression
Singh and Marx (2013)	Innovation			✓		Citing behavior	Citation link		Micro	8,014,434 links	USPTO	USA	Sciences	Choice-based samp. regression

\* Collaboration network not based on co-authorship.

Levels:

Micro: Author, paper, research project

Middle: Institution, firm, journal

Macro: Regional view

Table 2: Proximity dimensions addressed in the literature

### 2.5.1 Proximity Influence on Academic Citation

The works reviewed in this section concern those dealing with any proximity dimensions purely from an academic viewpoint, that is, where only citations from scientific literature are targeted.

Baldi (2013) was the first to assess scholarly impact by means of inspecting several metrics and their influence on citation probability. The goal of his research was to confirm which of the major citing behavior theories discussed above (normative or social constructivism) was the most influential for establishing an effective citation. The analysis employed almost 5,000 cases of potential citation links, extracted from a matrix formed by 100 papers publications in the domain of astrophysics. While his focus was not proximity dimension per se, his model still makes use of certain variables denoting proximity. For instance, cognitive proximity was denoted by a categorization of the research topics and subtopics of the two papers (citing/cited), revealing verifiable impact on citation probability.

In addition, other explanatory variables were included, like quality (in terms of citation counts), attributes of the article such as size (number of pages), cited author rank and prestige; interestingly, these last two failed to significantly improve the model. As for the social aspect, despite the author's claim of a network analytic approach, it did not strictly follow a collaborative definition, in that no co-authorship network nor any centrality metrics were analyzed.

Instead, social ties between authors were represented by a scale indicating whether both writers ever worked at the same institution or graduated from the same department. Moreover, the only social aspect adopted that was found to have any significance on the probability was author gender. In sum, the social constructivist hypothesis was rejected, in favor of the normative one, denoted by the cognitive closeness; although, admittedly, the need for better social and collaboration metrics was evidenced (Baldi, 1998).

Later, Rafols and Meyer (2007) assessed the cross-disciplinarity level of case studies on a bionanotechnology specialization, expressed by the citation counts according to the department affiliation of the team, a scale of the case's citation level (cited by various sources), and an existing collaboration between the authors and the outside world (again, no co-authorship network was employed).

Despite the study limitations (only 5 case studies), the authors took the result of the poor behavior of the cognitive affiliation of the team members on citation counts as potential evidence to indicate that nanotechnology is not as cross-disciplinary as in the idealistic "nano-visions" of having research teams from different disciplinary departments (Rafols and Meyer, 2007).

Furthermore, Ding (2011) inspected attributes of the co-author and paper-citation networks to verify collaboration and citation strength on papers about information retrieval from the Web of Science (WoS) database. The "strength" was denoted by a scale based on shortest path indicating the proximity between the authors in each network (the closer, the stronger the link). The top 20 authors per topic (productivity defined in terms of individual citation counts) were chosen to explore cognitive similarity between collaborators in



the co-publishing network, as well as the top 100 highly cited authors from the paper-citation network. He found that productive authors tend to directly coauthor and cite colleagues sharing the same cognitive interests (research field), while scholars with different research interests were indirectly connected to these authors. Nonetheless, the approach taken in this study was geared more to relating the two networks rather than purely assessing the effect of proximity on citations.

Contrarily, Abbasi et al. (2011) studied the effect of collaborative proximity on the performance (evaluated by the citation counts manifested in the g-index) of 2,139 publications in various science domains from the top five universities of United States. Their findings show that the research performance of scholars (g-index) is positively correlated with four Social Network Analysis (SNA) metrics: degree centrality (coauthors count), efficiency (strong co-authorship relationship to a single individual), and ties strength (defined by co-published works count), which manifested a positive significant influence, whereas eigenvector centrality displayed a negative significant effect. On the other hand, betweenness and closeness centrality metrics did not reach statistical significance, and were thus discarded as non-important measures for performance.

Even though Onel et al. (2011) and Wallace et al. (2012) addressed citation counts affected by collaborative proximity (in papers from the WoS database on nanotechnology and miscellaneous fields respectively), they followed a high-level approach. That is, they focused on averages of SNA metrics of the whole network, performing an analysis on the distributions displayed. In this regard, Onel et al. (2011) explored shortest path, degree centrality, and clustering coefficient mean values on both the co-authorship and paper-citation network; the aim of the research was fitting distribution models, so no conclusive effects of these metrics were discerned. In contrast to this approach, Wallace et al. (2012) adopted a degree centrality-based category (similar to the Erdős number), and determined that there is wide variation among fields in the propensity to cite co-authors, and co-authors of co-authors. Plus, papers citing collaborators exhibited a tendency to also cite distant collaborators (i.e. authors indirectly connected in the network).

Likewise targeting collaborative proximity, Liu et al. (2014) employed time-windows to inspect the co-authorship links between 16,582 scientific publications from five leading universities in the United States. However, a different perspective was used to assess this effect: the citation hazard rate, representing the likelihood of a paper being cited at a specific time. Among the SNA metrics adopted, degree centrality and structural holes (representing tie strength in the network) were found to be the best performing, while betweenness centrality only displayed a mild effect, and the Bonacich power (closely related to eigenvector centrality) was altogether discarded.

Abbasi et al. (2011) manifested that social network of researchers can be used to predict the future performance of scholars. By the same token, in a recent work Sarigöl et al. (2014) actually implemented predicting techniques, in addition to statistical testing and time-windows, to conclude that centrality metrics in co-authorship network, at the time of a paper publication, are indicative for future paper success (as

measured by citation counts). They focused on the largest cluster of connected authors (i.e. the giant component) in a network formed by 108,758 publications of Computer Science extracted from MSAS (the Microsoft Academic Search database), thus becoming the first large-scale analysis of the relation between citation dynamics and the researcher's position within a collaboration network.

Moreover, they combined network characteristics based on degree, betweenness, eigenvector, and k-core centrality measurements, comparing the relationship between the position shifts in time of centrality distributions versus citation success. Their findings suggest that co-authorship centrality metrics can indeed signify citation success by discovering statistical dependencies between the two. In like manner, the inverse relation was explored, hence discovering that an academic with citation success would later become more central in the collaboration network; although said author already had a good position in the network, which allegedly favored citation success to start with.

However, each centrality separately only displayed weak, if at all, correlation with citation counts. Sarigöl et al. (2014) warn that this may be due to their chosen methods to test this statistical dependency (Wilcoxon-Mann-Whitney tests) being more complex than simple correlation. Instead of, for example, having a scholar with a high number of coauthors (degree centrality) later becoming highly cited, this effect would rather be dependent on more than just single network metrics.

Finally, they attempted to predict whether a publication will be successful (highly cited), based on the position of its authors within the co-authorship network. Interestingly, by means of machine learning models, citation counts representing the success of an academic publication were in fact anticipated by the social location of its authors; such predictions achieved a precision of 60% which is impressive, considering that a random guess would only discern the same effect with a 10% precision.

## 2.5.2 Proximity and the Interaction between Scholarship and Innovation

Rather than dealing with academic citations, the works in this section deal with proximity dimensions when an interaction between science and industry exists. Citations are sometimes used to study this collaboration activity, rather than being the main target of these analyses. Mostly, products resulting from innovation (such as patents) are addressed, though in occasion, the collaborative links between scholars and inventors are inspected, and it is common to include institutional or organizational characteristics. Instead of cognitive proximity, technological distance is addressed, though they are close enough to consider findings in this dimension relevant.

For instance, Liebeskind et al. (1996) studied the exchange of scientific knowledge in biotechnology patents at the organization and individual levels, in terms of the scholarly publications authored by scientists belonging to two companies, and their patent count. Despite concluding that social networks played a critical role in organizational learning by providing firms with access to knowledge generated by academic research,

they did not adopt a structural network approach to account for collaboration proximity.

Furthermore, Schummer (2004) targeted the cognitive, geographical, and institutional dimensions by applying coauthor analysis to 600 papers from eight journals dealing with nanosciences. He developed indexes for these proximities based on categorical scales for discipline, institution, and the geographic region of the paper. As for the cognitive aspect, a publication was considered to be interdisciplinary when its coauthors belonged to more than one discipline. His findings indicate that nanotechnology research shows only an average degree of interdisciplinarity, and that it does not differ much from academic practices in sciences and engineering regarding intercontinental and interinstitutional academic collaboration.

Also, compared to both the interdisciplinarity and the interinstitutional index, the geographic collaboration index generally appears quite low; further, each geographical region seems to have its particular nanoscale research profile (e.g. Europe focuses more on physics and electrical engineering, whereas North America aims at chemistry, biomedical, and mechanical engineering). Anyhow, the number of papers authored by scientists from at least two different continents remarked on the high degree of international exchange in the field.

From the literature dealing specifically with the effect of geographical proximity, the works by Cunningham and Werker (2012) and Frenken et al. (2010) have some resemblance in that they considered a more specific categorization for location than Schummer (2004); further, they took collaborations between organizations and scientists extracted from the WoS database, and also, they both dealt with organizational proximity. These two works investigated the impact of spatial proximity on collaboration on European countries, in terms of its success (as reflected by citation counts), and intensity (shown in co-published works count) respectively.

Nevertheless, Cunningham and Werker (2012) also included a more direct depiction of geographical distance, as did Boufaden and Plunket (2007), Gittelman (2007), and Deleamarle et al. (2009), by measuring spatial distance itself, with the sole difference among them that Boufaden and Plunket (2007) converted it into a weighted matrix of distances and nearest neighbors.

Moreover, these last three authors assessed this proximity on widely different representations of science-industry collaboration: patent application counts of 60 companies in biotechnology (Boufaden and Plunket, 2007), collaboration linkage expressed by citations by papers and patents, also in biotechnology (Gittelman, 2007), and scientific production described by nanotechnology publication counts of clusters based on highly-cited cities (Deleamarle et al., 2009).

Thus, the different nuances in their findings on the importance of geographical proximity is not surprising, despite all of them having found it significant. In this regard, spatial closeness is meaningful for global knowledge production because of the high number of publications produced at spatially concentrated clusters (Deleamarle et al., 2009); plus, nearly-situated research teams tend to publish papers later cited in the firms'

patents, whereas dispersed teams publish articles that are more highly cited in scholarly works (Gittelman, 2007).

Additionally, due to also measuring technological proximity, Boufaden and Plunket (2007) found that being geographically close to patenting companies (whose employees come from an academic background) from related technology specializations does explain patent application counts.

Likewise, spatially clustered organizations augment collaboration intensity, but not as highly when the institutions are technologically proximate (Cunningham and Werker, 2012). On the other hand, according to Frenken et al. (2010), physical proximity would be more influential in some industries more than in others to achieve a successful interaction between scientific sources (university) and the industry: their results reveal that the citation impact of research collaboration is higher at the international level than at the scale of national and regional collaborations.

Moreover, in a more recent study by Bouba-Olga and Ferru (2012), traveling time (by train), along with a location category, was chosen instead of purely spatial distance to signify geographical proximity. This time, a set of 32,764 scholar-inventor links was used to denote spatial impact on collaborations between firms and scientific laboratories from various disciplines in France. Interestingly, findings from this work indicate the still present significance of geography, allegedly confirming spatial dynamics in science-industry collaboration, though its influence extent would depend on the specialization of the research team.

Similarly, Laursen et al. (2011) addressed spatial distance (by a grid-based metric and location scale as well) on the likelihood of establishing a collaboration link between 8,724 firms and 99 universities in the UK. However, they found that geographical distance plays a subtle role in shaping university-industry collaboration, given that other aspects (such as collaboration quality and the university type) revealed greater importance than spatial closeness when it comes to choose partners to collaborate with. Besides, though not directly addressing this dimension, the fact that the type of university was more influential than geography was is taken to signify a trade-off between cognitive and spatial proximity (Laursen et al., 2011, p. 24).

By the same token, Wang and Guan (2011) employed information from the USPTO (United States Patent and Trademark Office) to quantify collaboration intensity, as measured by citation links between 1,447 articles in nanotechnology research and 275 Chinese inventions (patents). The study concerned technological proximity by inspecting the different application fields of the scientific and technology products, as well as the impact of collaborative proximity.

For this purpose, three networks were built (coauthor, coinventor, and both combined) and SNA was applied to them to inspect the position of the individuals (degree centrality) and gather overall metrics of the network (nodes, edges, components, density, and diameter). Their results suggest that knowledge production and diffusion is improved by the strong interaction between science and technology discerned in nanotechnology, by means of the author/inventor positions as well as their application fields.

Finally, Eslami et al. (2013) also measured collaborative proximity by inspecting the co-authorship network of 100,652 papers and 4,893 patents in Canadian biotechnology. Said influence was assessed on their research productivity and technological performance through paper and patent counts respectively, by means of the network structural properties. From the SNA metrics analyzed, the degree and betweenness centrality revealed substantial influence on knowledge and innovation creation.

Still, we must mention that betweenness centrality only presented a significant effect due to the combination of the scientific and technological aspects in a single network; further, Eslami et al. (2013, p. 17) indicate that only when research efforts may translate into industrial applications, this metric displays its power of controlling knowledge flows in the network.

In addition, small-world properties were also found significant, particularly the clustering coefficient, which denotes a high “cliquish” network structure and would enhance knowledge creation (this result disproved previous claims as to the opposite). Altogether, their results suggest that the structure and individuals’ properties and interconnections within the collaboration network correlate to both knowledge and innovation production.

### 2.5.3 Proximity Influence on Patent Citation

In the preceding discussion, citation is sometimes used as proxy for science-technology collaboration. However, we make a clear distinction between analyzing citations from scientific publications versus citations in works produced either by the industry or academic entities (like universities), but which have an economical purpose. We now consider proximity effects in terms of citation behavior when targeted solely from the verified products of innovation: patents.

Throughout the innovation literature, we find that the impact of geography on patent citation has been the proximity dimensions most commonly measured. To begin with, Jaffe and Trajtenberg (1999) explored the citation intensity between 50,000 pairs of patents (taken from the USPTO database), by considering the effect of country location on citation frequency and probability. They affirmed that geographic localization is significant for knowledge diffusion, having found that patents with inventors in the same country are 30 to 80% more likely to cite each other than inventors from other countries.

Plus, they discovered that having pairs with the same technological class is decisive for there to be a citation link among the two, with this likelihood being 100 times greater than in patents pertaining to different classes. Similar findings by Hu and Jaffe (2003) reveal that among 59,116 patents, those being technologically proximate are preferred for citing to distant patents from another field. Finally, they discerned that knowledge diffusion (as measured by citation count and frequency) would display different macro-level patterns according to the inventor’s country of residence.

Further analyses by Breschi and Lissoni (2003), Singh (2005), and Sorenson et al. (2006) chose instead

to take a closer look at geographical distance than mere country to confirm whether they had the ability to increase the likelihood of an existing citation link between patents. Besides spatial distance, all three targeted the collaborative dimension as well by inspecting the structure of the co-inventor network.

In this regard, Breschi and Lissoni (2003) used the inventor's location category (based on intra-national boundaries) to match Italian patents from the European Patent Office (EPO) into 3,716 co-located patent links. To realize how closely related were inventors in the co-patenting network, they focused on the structural metrics of "know-who" (shortest path), "connectedness" (network component), and mobility (betweenness centrality).

Their results revealed that localization effects tended to disappear when the co-located citing/cited patents were not additionally linked by any network relationship. Thus, they claimed that geography is not a sufficient condition for knowledge diffusion, instead requiring an active participation by inventors within a network of knowledge exchanges. Moreover, while betweenness centrality and shortest path were both found to be significant, a high shortest path value (meaning a more indirect connection) would offer much less influence than the betweenness of inventors.

Differently, Singh (2005) considered the metropolitan area the inventors belong to, while Sorenson et al. (2006) quantified distance in miles (transformed); both works relied only on a shortest path scale to assess collaborative proximity. They both addressed pairings of citing and potentially cited patents from United States to measure the probability of a positive citing connection.

According to their findings, interpersonal networks expressed by collaboration ties are indeed important for knowledge diffusion by increasing the likelihood of effective citation links among inventors. Further, such probability decreases as the shortest path length increases (Singh, 2005); giving socially proximate inventors a greater advantage over distant ones for gaining knowledge (Sorenson et al., 2006).

Nonetheless, the effect of geography, as expressed by regional boundaries, decreases once interpersonal ties have been accounted for, implying that being in the same region has little to zero impact on the citing probability among inventor pairs already closely connected by network ties (Singh, 2005). Therefore, geographical proximity is regarded as a moderate influential factor on the citing likelihood between patent pairs, hence agreeing with Breschi and Lissoni (2003).

In a later analysis Singh and Marx (2013) followed a similar methodology to the one employed by Singh (2005), with the difference that spatial aspects were quantified by three metrics of geographical proximity: a geopolitical level category based on the inventor's city, a flag value for when inventors are in the same state, and by distance measurement in miles. They conclude that both country and state limits had positive independent effects on knowledge diffusion beyond the ones displayed by geographic proximity in the form of metropolitan collocation or shorter distances in the same region.

By the same token, Sonn and Storper (2008) focused on the citation counts between 17,761 patent-pairs

from the US and discovered that American inventors displayed an increasing tendency to cite local patents at three geographical levels: country, state, and metropolitan. Their results suggest that inventors choose domestic knowledge, that is, they prefer national sources, over foreign knowledge.

In addition, even though a preference over in-state knowledge was found significant, it was rather weak, meaning that inventors do not really care if their citation sources are from out-state, as long as it is domestic; plus, they would adopt knowledge from the same metropolitan area more than from outside sources.

Similarly, Agrawal et al. (2008) adopted a co-location category based on metropolitan area levels as well as Euclidean measurements to account for the spatial distances between inventor-patent-citation linkages from the US and Canada. We find it interesting that a purely social proximity aspect was considered, in terms of a co-ethnicity class that categorized inventors as Indian or non-Indian according to their names. Their findings reveal that both the spatial and social proximity dimensions increase the probability of knowledge flows between inventors.

However, only socially distant inventors would benefit from geographic closeness; spatial and social proximity thus become substitutes for being an influential factor on knowledge access. Moreover, by means of controlling for patent technological class, they conclude that being members of the same technical community helps spatially-dispersed inventors to gain access to knowledge.

Finally, it is noteworthy to mention that although measuring technological proximity was not their main purpose, those authors in this section who did not directly inspect its influence still accounted for the cognitive aspect. Indeed, most aimed to remove its effect and assess other aspects independently by using patent technological class as control variable as Agrawal et al. (2008) did.

#### **2.5.4 Research Gaps**

Although one could argue that, overall, research related to knowledge production making use of citation analysis and collaboration networks has been well documented in the multiple studies discussed above, many of them have either been conducted only with bibliometric approaches, or with a special focus on patents data, having institutions as their publication source.

It becomes evident that the literature predominantly deals with proximity concerns involved in innovative citation, or even its collaboration with scholarly communities but without really exploring their impact on citation; studies on scientific citation are, as a matter of fact, rather scarce. Further, the influence of geography has only been examined on academic citation when innovation is implicated.

Moreover, all of these works are based on either one proximity dimension or combinations of them, but never all three aspects together. Indeed, to our knowledge, there is no study dealing with the geographical, cognitive, and collaborative dimensions combined in a single analysis, let alone for exploring citation impact, on academic literature.

The only one we could find that addressed all three dimensions was Cantner et al. (2013), though we did not expand on their findings due to their focus being the behavior that would lead inventors to choose partners to establish a cooperation link with, without addressing any citation aspect.

Findings by Radicchi et al. (2008) show that citations follow a similar distribution pattern, independently of the discipline they belong to. Whereas this could suggest that the mechanisms behind citation practices are universal across fields, citation practices seem to generally differ significantly in accordance with their scientific discipline (Sarigöl et al., 2014).

As a result of this tendency, where knowledge bases would vary according to their industry (Malerba, 2005), knowledge flows would be influenced and display characteristic patterns depending on their sector of science (Gertler and Wolfe, 2006). Therefore, another important aspect to highlight is that there has been no analysis executed in Canada which solely concerns the influence of proximity aspects on scientific collaboration in nanotechnology research.

Even if Cunningham and Werker (2012) indeed targeted several types of proximity for European nanotechnology, the use of social networks was merely suggested as ground for further studies. Throughout the literature, we have examples of studies applying collaboration network analysis for scientific publications localized in Canada, but they are either outside the realm of nanotechnology, rather focusing on other branches of science (e.g. Sarigöl et al., 2014); plus, sometimes the research was more geared towards business economics (e.g. Schummer, 2004; Eslami et al., 2013), or even at a macro level (e.g. Delemarle et al., 2009).

Additionally, works dealing with SNA concerns on nanoscience knowledge production seem to be aggregated analyses that consider academic networks only from a high-level perspective (see Onel et al., 2011) or having a very limited scope (see Rafols and Meyer, 2007).

In conclusion, there are significant gaps in the literature on knowledge production and its flow specifically from a scholarly standpoint, in light of the few studies that have been performed with this approach. Moreover, we find the need to inspect knowledge-generating networks at the micro level of the academic community, exploring the attributes of its individual members. Thus, we assert the relevance of this research, seeking important insights about proximity factors and their effects concerning scientific citation.



## 2.6 Research Questions

The preceding discussion enables us to pose the questions detailed below, which motivate the research reported here.

In the field of Canadian nanotechnology,

1. Is cognitive proximity between two scholars a strong influencing factor as to increase the probability of an existing citation link among them?
2. Does geographical proximity between two researchers have impact enough to result on an increased probability for the establishment of a citation link among them?
3. Does collaborative proximity, as measured by the location of two authors within a co-authorship network, have an effect on the probability of a citation link between them?

We expect to find evidence-supported answers to our research questions throughout this thesis. We now go in detail about the data, methods, and tools used to obtain our answers in the next section.

The literature has discussed many approaches that emphasize the importance of proximity for academic research and knowledge exchange and production. Recall that the theoretical proposition is that authors who are proximate to peers within scientific communities are more likely to result in paper citations, which in turn capture their impact on knowledge spill-overs. On these grounds, we will focus throughout this work on finding out how close the studied scholars are to each other.

We shall also apply the know-how of social network analysis to the collaboration networks drawn out from articles specialized in the field of nanotechnology. According to Gay and Dousset (2005), the study of networks requires a delineation of explicit temporal and spatial boundaries. Thus, we circumscribe our research within a Canadian setting in the last 5 years.

### 3.1 Research Design

The research design is quantitative with a correlational strategy, making use of research papers by scholars publishing about nanotechnology. The purpose of the design is to measure the level of interaction between the proximity amidst referenced authors and Canadian authors, and the probability of being cited by the latter.

In our analysis, we explore several dimensions of proximity as the explanatory parameters for justifying citations, which can be relational as well as non-relational. Relational metrics correspond to a specific pair, and must be evaluated using both authors. However, we also introduce non-relational variables, which

describe attributes pertaining specifically to the cited author.

The study has been conducted in two phases. During the first phase, we develop methods to effectively measure the defined proximities. This stage also includes constructing the collaboration network of scientists, where social network analysis is performed to gather relevant network indicators. All these metrics are collected as input for the second stage of our work.

In the second phase, the association between the proximity distances and effective citation links is examined. The latter phase comprehends a quantitative method using both statistical analysis and machine learning classification, based on the data obtained from the previous phase.

## 3.2 Data

This section deals with describing in detail the tools and steps taken to select the participants in our analysis.

### 3.2.1 Instrumentation

The programming languages used throughout this project were mainly PHP and Javascript, running on a XAMPP (Apache Friends, 2015) web server. Plus, we employed statistical, modeling, and network analysis software, which are specified later on, following the various procedures in data collection and analysis. In the end, our final sample is stored in MySQL database (Oracle, 2015), with a table structure that can be found in Appendix B.

### 3.2.2 Setting and Participants

Tracking publications in academic journals has largely been viewed as the easiest and most relevant measure of scientific knowledge production (Delemarle et al., 2009; Callon et al., 1986). However, nanotechnology being an emerging and highly interdisciplinary field, scholars in the literature have raised the concern of counting with a proper data set for its study (Schummer, 2004; Delemarle et al., 2009). Among these challenges are the lack of a specific tag for nanotechnology in many traditional databases, such as the Web of Science (WoS), and the fact that words including the term “*nano*” do not exclusively refer to nanoscience publications (e.g. “*nanokelvin*”).

Fortunately, to answer our research questions we make use of a data set extracted from the SCOPUS database by another research team member (Moazami, 2012), who used specialized keywords related only to nanotechnology to select relevant academic papers, while also filtering out those with misleading terms such as “*nanosecond*”. This data set was cross-referenced with WoS (Web of Science) database to obtain further paper details, such as the scientific field of the article.

Thus, our original data set consists of 578,907 articles published in the period from the late 1900s to 2012 and written by a total of 538,780 authors (2,501,343 total authors records). We also have 2,174,607 cited articles written by 2,025,080 cited authors (12,515,392 cited authors records), leading up to 6,308,727 citation links between these articles.

Our database also includes specific details about the article, such as the location of the scholars and the article’s publication year. The data was originally provided in 6 different files in CSV format as follows, linked together by using the paper IDs and the author order:

- Papers
- Authors
- Addresses
- Cited Papers
- Cited Authors
- Cited Addresses

We propose to conduct an empirical analysis with a subset of this database by adopting complementary strategies, which we will define in a further section.

### 3.2.3 Data preparation and Sampling

We link scientists in pairs or “dyads”, which is the basic unit of analysis in social network theory (Gittelman, 2007; Jawed et al., 2015). To this respect, Butts (2008, p. 2) points out that relations need to be defined as pairs of entities, with a qualitative distinction serving to discern present vs absent relationships. Hence, we gather proximity distances between these pairs, and use a citation link as the distinction of the type of relationship, similarly to the work by Cantner et al. (2013) on patent applications. Put simply, connected pairs of scholars are assigned to the positive class, while non-connected ones are assigned to the negative class (Jawed et al., 2015).

Consequently, the statistical unit of our analysis is *CC - REF pairs* of nodes. We start by defining these types of nodes, which we shall have in our final data set, as follows:

**CC nodes:** Citing Canadian authors who published at least one *Canadian paper*.

**REF nodes:** Reference nodes, meaning authors of cited papers, who may or may not be cited by CC authors. These scientists can be from anywhere (Canadian or non-Canadian).

Throughout this research, we have established that a “*Canadian paper*” is an article with at least one Canadian scientist among its authors.

#### CC nodes

Our focus is on studying the citing behavior of Canadian scholars, and we have limited our research to only consider Canadian authors publishing in the most recent years.

You may refer to Figure 1 to see the count of Canadian publications vs the total of articles per year. As you can see, year 2012 has significantly less papers than previous years, hence, we assume that the data was extracted sometime during that year, and thus is considered as incomplete and not useful for our analysis. Consequently, we shall work with 3,981 papers published by Canadian scientists during 2010 and 2011.

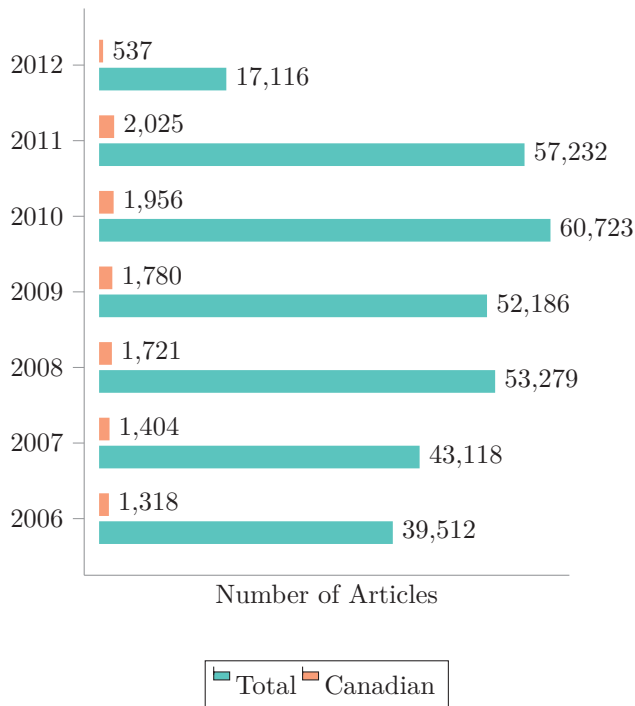


Figure 1: Nanotechnology articles per year

We use the information on author affiliation as a proxy for geographical location, and thus we are able to identify Canadian-based scientists.

The number of authors with Canadian addresses per year is shown in Figure 2 below. In total, we have 4,887 possible Canadian scholars (CC nodes).

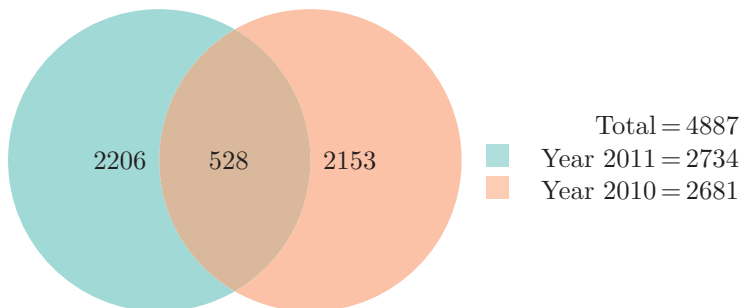


Figure 2: Canadian authors publishing in 2010 and 2011

## REF nodes

Let us now take a look at the REF side of the pairing. We have set the time window for our research to be the last 5 years. Thus, we shall evaluate the citing behavior of the CC nodes against cited authors from 3 previous years respectively. In this case, we shall pair up Canadian scientists from 2010 and 2011, against cited authors from 2007 to 2010, and dismissing year pairs like: CC 2011 - REF 2011, and CC 2010 - REF 2010, so that no pair is from the same year.

Throughout our whole database, we have 6,308,727 citation records, matched to the 2,174,607 unique cited papers for the complete time range. These cited papers are authored by 2,025,080 unique scientists (corresponding to 14,847,753 records due to repeated citations of the same paper).

As previously mentioned, our working range for cited scholars is between 2007 and 2010. Therefore, we shall consider the authors from the 395,051 cited papers published between 2007 and 2010, which gives us a total of 467,794 cited scholars (corresponding to 1,375,611 records in our database). The counts of cited articles and authors in the studied period may be found in Figure 3.

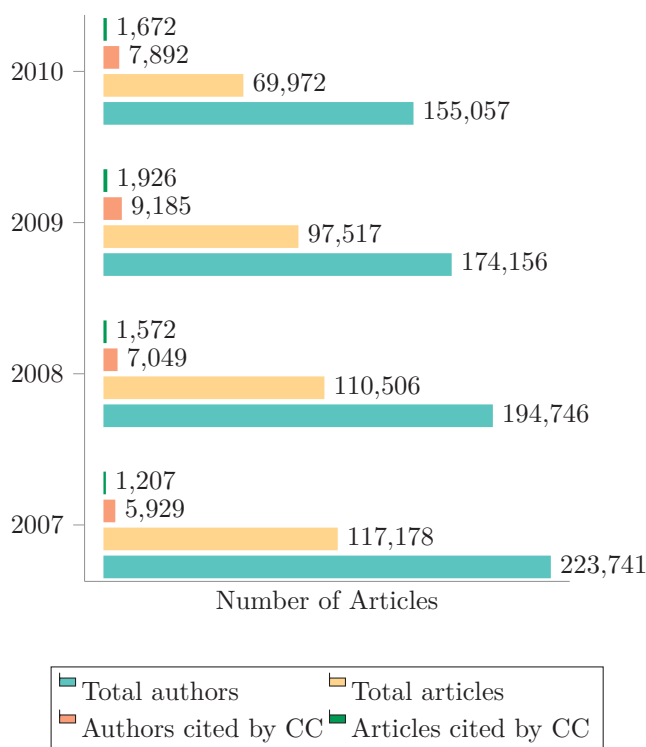


Figure 3: Cited articles and authors per year

## Sampling method

Let us assume for a moment that we are to study every possible combination between our CC nodes and the REF nodes for the selected period. With over 400K cited authors, this would give us over 2 billion pairings (2,286,109,278 to be exact). It is evident then that some sampling method is required.

Consequently, we decided to make use of R (R Core Team, 2015), a statistical computing software, to create our sample from the REF side, and then pair up combinations with our CC authors. First, we identified all the REF authors with at least one citation link to a CC author, by adding an extra column (Citing) with a boolean value.

With this identification, we were able to establish that during the period of 2007 to 2010, only 26,987 authors have at least one citation connection to a Canadian author from 2011 and 2010, while the majority (440,807 authors) does not have this citation link. The fact that the amount of cited authors with a positive citation link is very little (5.77 %) when compared to the population of all cited authors, could represent a potential issue at the moment of sampling. The proportion for cited authors with and without this citing connection is depicted in Figure 4a.

Then, we ran a trial by taking a normal random sample with a size of nearly 12% of the total population (n=54,000), which resulted in 50,765 non-linked scholars vs 3,236 of positively linked authors (94% vs 6%). While we could say that this sample is pretty representative of our REF population, it is clear that having so little number of observations with the citing connection we want to study becomes a hindrance. It follows that random sampling is not practical for our purposes because actual citation links between CC and REF authors are very rare, making meaningful estimation impossible even if we took a large sample (Singh, 2005). However, modifying data distribution by simply undersampling the majority class could negatively affect the results of later analysis, given that the distribution of the resulting data set would no longer present the same challenges as the real-world distribution (Lichtenwalter et al., 2010).

Therefore, as sampling methodology we followed the SMOTE (Synthetic Minority Over-sampling Technique) algorithm, which was developed by Chawla et al. (2002), and implemented in R through the DMwR library (Torgo, 2010). This algorithm is useful for highly imbalanced data sets, and it has been used in several fields and for various purposes, such as network intrusion and fraud detection, medical imaging intelligence (Padmaja et al., 2007; Wang et al., 2007; Zhao et al., 2009), and notably, link prediction in social networks (Munasinghe and Ichise, 2011).

SMOTE oversamples rare events by creating additional synthetic observations of that event, while at the same time undersampling the population majority that does not contain the desired effect. The definition of rare event is usually attributed to an outcome or response variable that happens less than 15% of the time in the whole population (Amunategui, 2014). With 5.77 % of positive REF scientists, it follows that we are dealing with a “rare event”.

We find similar need for a balanced sample with both cases (citing and non-citing) in the literature, particularly for the study of knowledge diffusion by making use of citation links between patents (Singh, 2005). In their case, a choice-based sampling theory called *Weighted Exogenous Sample Maximum Likelihood* (WESML) is implemented to address this requirement. Similarly, Lichtenwalter et al. (2010) overcomes imbalance by applying SMOTE, stating it is one of the best sampling strategies for highly skewed class distributions, especially for researchers dealing with link classification.

To sample our data with SMOTE, we followed the guide by Amunategui (2014) as reference to write an R script to create our sample, which appears in Appendix C.1. The SMOTE sample left us with a quite balanced observation set of 27,014 non-linked authors vs 26,987 linked authors (50.02% vs 49.98%).

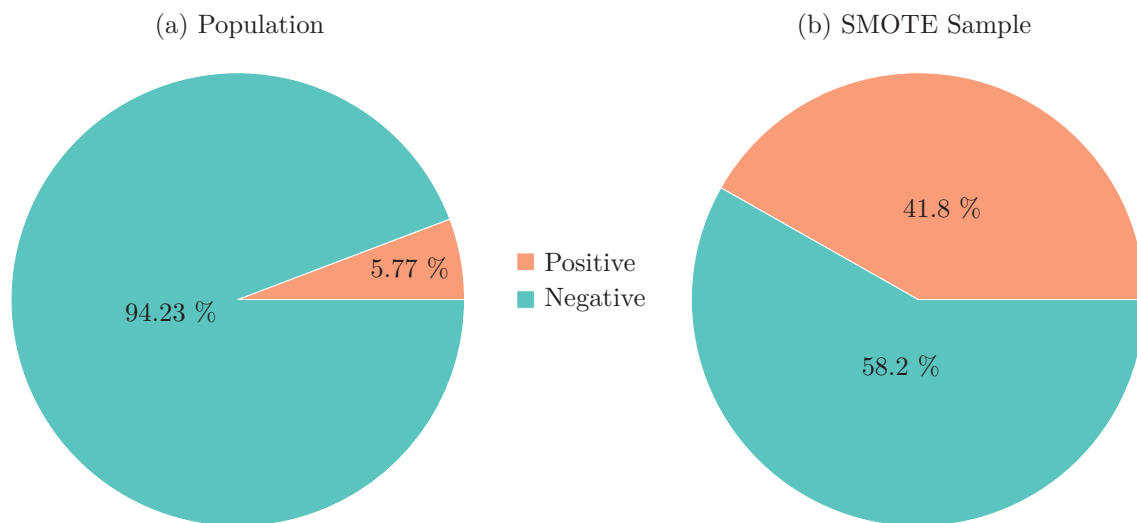


Figure 4: Proportion of citation links of REF authors between 2007-2010

As our next step, we created the combinations between our CC nodes and the REF authors from the sample. Since we have no specified path between the non-linked REF scientists and the Canadian authors, to pair these up we used an algorithm to randomize these combinations, whereas for the positively linked scholars, we followed the respective citation records. Although authors favor their own papers for citing, which has been deemed as a beneficial factor in citation counts by others (Bethard and Jurafsky, 2010), we left out all self-citing links because we consider them uninteresting for our research goals. Besides, self-citations are naturally more geographically localized (Jaffe and Trajtenberg, 1999), so including them could favorably bias the effect of the spatial dimension.

As result, there were 116,256 potential author combinations generated, however, we observed that some authors among them were missing their affiliation information, hence lacking an address that would allow measuring geographical proximity.



After discarding records with missing information, we were left with 80,091 author pairs (approximately 70% of the original sample), from which 41.8% (33,477 combinations) are positively linked, that is, there exists an actual citation connection between them, and 58.2% (46,614 combinations) have a negative, or non-citing connection (or “control” citation links). This proportion is depicted in Figure 4b.

Finally, our working sample of 80,091 pairs consists of 2,824 Canadian papers published between 2010 (59,621 pairs) and 2011 (20,470 pairs), authored by 3,747 distinct Canadian scholars (CC nodes), and 34,877 reference papers published between 2007 and 2010, authored by 47,380 distinct REF scientists. The number of pairs per reference year can be seen in Figure 5.

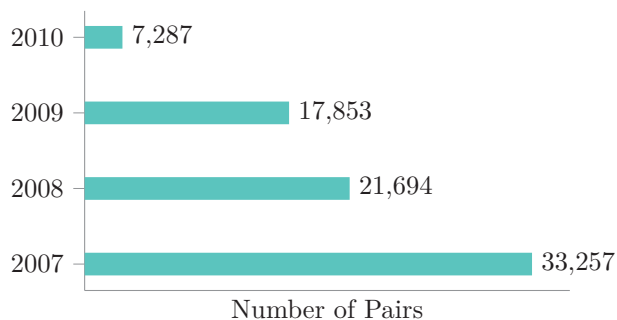


Figure 5: Author pairs per cited year

### 3.3 Proximity Measuring

The present section discusses the various methods used in the literature to measure each of the proximity factors. In addition, we introduce consequential metrics of each proximity aspect, also detailing the process followed to collect their respective data.

#### 3.3.1 Measuring Cognitive Proximity

##### Cognitive proximity measurement in the literature

We have observed in the literature that there is no predefined way to quantify interdisciplinarity (and thus, cognitive proximity) all authors can agree with (Bordons et al., 2005; Rafols et al., 2010). It would seem that this is partly due to the different definitions of interdisciplinarity each one has, thus influencing the system of disciplinary categories used (Schummer, 2004). Consequently, we see heterogeneous methods in their work, starting from the basis taken for the analysis.

On one hand, seeking a satisfactory measure for cognitive distance among fields has led some authors in the literature to come up with their own classification of disciplines (e.g. Schummer, 2004; Rafols and Meyer,

2007; Cunningham and Werker, 2012). On the other hand, we find examples adopting pre-existing field categorization systems (e.g. Jaffe and Trajtenberg, 1999; Boufaden and Plunket, 2007; Rafols and Meyer, 2010), such as the patents classification of the IPC (International Patent Classification) (WIPO, 2015). For instance, Jaffe (1989) came up with a correlation coefficient (cosine index) to measure the closeness between two companies, based on the distribution of the technology fields corresponding to their respective patents.

Likewise, Malerba et al. (1998) built up on this approach, by analyzing the frequency of co-occurrence of IPC codes assigned to individual patents. This would be an indicator of the connection strength between the knowledge bases derived from the technological areas behind those classification codes. Nonetheless, these cases are rather taken as quantifying technological proximity, since they involve technological publications regarding innovation rather than purely cognitive fields.

On the contrary, the work of Rafols et al. (2010) concerns the degree of disciplinary diversity, by employing ISI (Information Sciences Institute) subject classes and cluster analysis to categorize research topics. Still, nanoscale research being such an ambiguous field, its placement could vary among different classification systems due to the vagueness of its definition (Schummer, 2004).

Regarding the subject of study, we usually find the focus given to the interactions between patents or scientific publications, in terms of keyword or field co-occurrences, authors' affiliations, or citation links between disciplines. In the last case, the analysis would be geared towards the flow of information between the distinct disciplines of the authors' cross-disciplinary reading (Schummer, 2004).

Moreover, an alternate approach has been to classify papers according to the specialization of the journal they are published at (Katz and Hicks, 1995; Leydesdorff and Zhou, 2007). However, one drawback to this method would be that it becomes hard to distinguish multidisciplinary articles when they are published in general journals like Science or Nature (Schummer, 2004).

Furthermore, Schummer (2004) criticizes the above mentioned approaches, by stating that these factors would only investigate the cognitive aspect of interdisciplinarity in terms of information. Therefore, he proposes studying interdisciplinarity based on co-author analysis, particularly for the domain of nanoscience, due to the combination of both cognitive and social aspects found in this type of relationship. He argues that usually, the co-authors' disciplinary affiliation would correspond to their disciplinary knowledge contribution, and thus it would grant a better grasp on interdisciplinary research.

Contrarily, Rafols and Meyer (2007) warn about collaborations among diverse disciplinary affiliations not always being an accurate indicator of interdisciplinarity. In their work, they take research projects as the unit of analysis, and compare the findings from the cognitive practices of academic research (in this case, citations and references), and the social dimension (by using affiliation and researcher practical background).

In conclusion, they found that citation-based indicators better capture the generation of cross-disciplinary knowledge in nanotechnology, than through the tracking of co-authors' affiliations. In a posterior study,

the use of these two approaches to quantify disciplinary diversity was combined with network coherence analyses, thereupon obtaining a clearer picture of knowledge integration (Rafols and Meyer, 2010). Similarly, Leydesdorff and Zhou (2007) also draw conclusions about the interdisciplinarity level of nanotechnology-related journals from network indicators.

In sum, the quantification of cognitive proximity and their relation with interdisciplinarity would greatly depend upon the definition one has about these key concepts in the first place.

### Metrics and data collection of cognitive proximity

In the literature we find that the field or subject categorization by a scientific database is often adopted to measure cognitive proximity (e.g. Jaffe and Trajtenberg, 1999; Boufaden and Plunket, 2007; Rafols and Meyer, 2010; Rafols et al., 2010).

Similarly, in our data the article information that could help identifying its scientific field provided is the journal it was published in, the article’s domain, field, and subfield. These correspond to the categorization system from the Thomson Reuters WoS (Web of Science) database. We have observed that the domain is the same for all the papers, in this case: *Natural Sciences and Engineering*. The full list of the categories belonging to the data in question can be found in Table 11 in Appendix A.1.

For this proximity factor, we discern only one metric applicable in our case:

#### COGNITIVE CATEGORY

We expect to get a sense of interdisciplinarity in our citation set by creating a categorical scale of “closeness” between the cognitive fields and subfields of each author in the pairing. Refer to Table 3 for the cognitive category scale we have defined.

Value	Description
1	Same subfield (same field as well)
2	Same field (different subfields)
3	Different fields

Table 3: Cognitive distance scale

We make the assumption that the fields and subfields of an article apply to all of its authors, even though the field and subfield information in the database pertain to the paper, rather than specifically to the author. Reason for this is that we do not have particular information about the scholar’s scientific field of interest, nor their department within the respective institution. Also, we noticed that cited papers older than 1980 do not include the field and subfield attributes, instead having an “Unknown” label. Still, we disregard this lack of information as we do not work with papers from these years.

Our method for obtaining cognitive distance is pretty straightforward. First, based on the IDs of both the article and the reference (cited) paper, we fetched the cognitive field and subfield information corresponding to the publication, and stored it to our combinations table.

Next, we developed a simple program to compare the fields and subfields between the two authors, and calculate the cognitive distance according to our predefined scale. The obtained results are found in Figure 6. We can observe that the great majority of non-citing pairs come from different fields (48.15%), while the citing pairs have a more even distribution among the categories. It would seem that citing pairs display an important preference for the same subfield (category value of 1), as compared to the non-citing pairs.

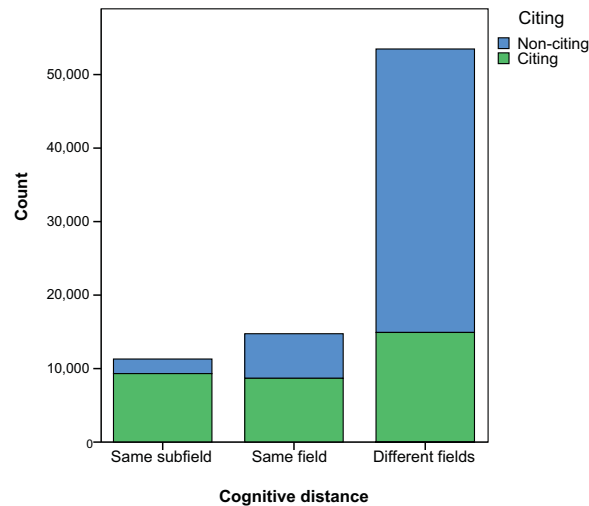


Figure 6: Distribution of cognitive distance

Finally, we split the parameter into three binary or “dummy” variables for each category (see Table 4). Reason for this was that the category values are not too numerically distant from each other, which could potentially influence our analysis. Also, it gives us the flexibility to measure each category on its own.

Dummy Variable	Value 1	Value 0
cog_1	Same subfield	Not same subfield
cog_2	Same field	Not same subfield
cog_3	Different field	Not different field

Table 4: Cognitive distance into dummy variables

### 3.3.2 Measuring Geographical Proximity

#### Geographical proximity measurement in the literature

When measuring geographical proximity, there are basically two ways to go when it comes to the data as ground for the analysis. While some studies have adopted an approach of surveys for data collection (e.g. Lublinski, 2003; Laursen et al., 2011), the majority we have seen makes use of existing formal databases, such as the Web of Science, in the case of academic publications, or databases from government patenting offices, like the USPTO (United States Patent and Trademark Office), or the EPO (European Patent Office).

We have previously made the distinction in the various definitions of what exactly is meant by geographical proximity, and how it is typically decomposed into physical or Euclidean distance and functional or travel distance.

First, we focus on the Euclidean, or a purely spatial distance. The methods used vary greatly, ranging from archaic measuring of accurate scale maps (Katz, 1994), to using specialized software applications as GIS (Geographic Information System) (e.g. Delemarle et al., 2009).

The primary concern that comes when trying to quantify distance is which address information to use, and that is directly related to the data set to be used. In the scenarios where formal scholarly databases are the data source, the data provides the affiliation information for every author, but it could happen that this address may not be the place where the study was done (because a scientist could change their affiliation, or publish using the address of the research facility headquarters) (Cunningham and Werker, 2012; Eslami et al., 2013). Nevertheless, the affiliation is forcefully the only pointer available to assess the place where the actual research was carried out. Hence, the affiliation of the authors is customarily taken as a proxy for their geographical location.

As a next step, researchers are usually required to perform *geocoding*, which refers to the translation from an address to its precise location coordinates (longitude and latitude). We have reviewed some works in which the data included zip codes of the organizational affiliations of the individual authors (e.g. Gittelman, 2007).

In some cases, there is no exact location specification and scholars are forced to find alternate ways, such as taking a general postal area for the calculations (Laursen et al., 2011). In other cases, they count with specific geographical data for every address, and can get to be more precise in their computation of distances. Some examples of the precise measurements are addresses being geocoded by using Google Earth (Cunningham and Werker, 2012), or using GIS engines such as MapPoint (Delemarle et al., 2009).

We now get to the functional aspect of measuring distance, which is associated with the transportation method and the actual time it would take to travel from point A to point B.

In spite of the arguments stating that travel time has a high correlation to Euclidean measurements of distance (Phibbs and Luft, 1995), increasing measurement precision might be necessary according to the

research being performed (Jones et al., 2010) Also, because it could be important to account for the impact of topological considerations (like mountains, rivers, etc.) and the transportation pathway networks (Jones et al., 2010). Another reason would be to avoid the potential for bias for the cases when straight line distance is not an accurate reflection of the travel time it requires (Phibbs and Luft, 1995). In any case, any Euclidean measurement is always equal to or lesser than a functional distance measurement, though the magnitude of this difference is unknown (Jones et al., 2010).

We find in the literature that some authors have taken travel time by train (Bouba-Olga and Ferru, 2012) for their functional proximity calculations. However, is it common to consider vehicles as the transportation means, in which case driving time would be the factor to use, like the case of Jones et al. (2010), in which he combines the distance traveled over a road network with Dijkstra’s algorithm (Dijkstra, 1959) to pick out the shortest path.

Moreover, it is not uncommon for studies performed in this area to include a categorical variable that provides a way of ranking distance, such as creating their own classification or scale of geographical unit (Sonn and Storper, 2008; Bouba-Olga and Ferru, 2012). Some examples are the grouping of addresses into geographic clusters (Delemarle et al., 2009), or making use of existing nomenclature standards, like EU geocode standard NUTS (Cunningham and Werker, 2012).

### **Metrics and data collection of geographical proximity**

To measure geographical proximity, we follow Cunningham and Werker (2012) and decompose it into physical proximity (referring to Euclidean or spatial distance), and functional proximity (referring to the traveling time existing between two authors).

#### **EUCLIDEAN DISTANCE**

This metric refers to the straight-line distance or metric between two points in Euclidean space, in this case, the locations of the two scholars, denoted in kilometers. It should be noted that this a measurement is in “as the crow flies” or “in a beeline” kilometers, meaning that it is the shortest distance between two geographical coordinates on a map, disregarding terrain considerations.

Note: This parameter is also referred to as “flying distance”.

To calculate spatial proximity, we followed a methodology similar to the one applied by Cunningham and Werker (2012). However, since our data set did not contain postal codes within the affiliation address, specialized geocoding software could not be used to easily convert them into coordinates (latitude and longitude), so we had to develop new methods.

To gather the required location information to calculate Euclidean distance, we first extracted the affiliation address of each author. We wrote code for this mapping, since they were situated in different tables,

linked together by the article id and the author order. The address shown in Figure 7 corresponds to the first author (order=1) of article id # 8325, as an example of an affiliation record in its raw format in the addresses table.

```
Institution: CARNEGIE-MELLON-UNIV
City: PITTSBURGH
Province: PA
Country: United States
```

Figure 7: Example of affiliation record in raw format

Next, we wrote several scripts of programming code to pass this address information to the web services offered by Google Maps, in this case, the Geocoding API (Google Inc., 2015). As stated in their developers website, geocoding is the process of converting addresses into geographic coordinates. The screenshot in Figure 8 shows our script obtaining sets of coordinates for each address, per the responses obtained from the Geocoding API.

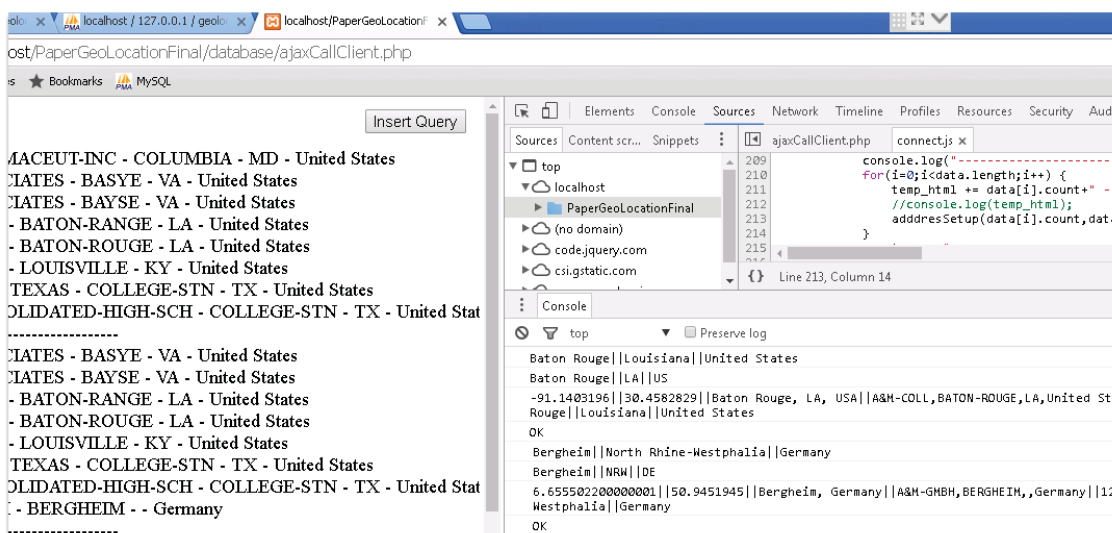


Figure 8: Script querying Google Maps Geocoding API

The response from Google Maps also allowed us to obtain complete geographical details along with the coordinates, as can be seen in the output depicted in Figure 9. We store all the returned details in a table created specifically for this purpose, which contains a unique identifier for each found address, so that there are no unnecessarily repeated records for the same location.

Nonetheless, cleanup was required for problematic addresses where, for instance, the institution name would be incomplete or acronyms would be used instead. In some cases, Google Maps indeed recognized the address, and returned the correct full institution name, along with the rest of the details. In some other

cases, it was not recognized, and it returned an empty value.

To solve this, we created several scripts to filter information and reuse geolocation data before hitting Google Maps again, since the web service limits the hits per day allowed in the free license version. The first filter code compared records with valid geocode details versus not-found addresses and associated the same location data to the missing one, provided that they had at least 60% similarity in their institution name field (string comparison) and that the city, province, and country were the same for both.

```
Address long name: Carnegie Mellon University, 5000 Forbes Avenue
Locality (city/town): Pittsburgh
Administrative Area Level 2 (province/state): Pennsylvania
Postal Code: 15213
Country: United States
Latitude: 40.4424925
Longitude: -79.94255279999999
```

Figure 9: Google Maps response output

Even though this initial filter populated the majority of the missing geocoded addresses, there were a few left that still were not found (such as empty institution names, or institution containing one or two uncommon acronym characters). For these cases, we had to use the “next best scenario” and gathered the geolocation information by using the city name instead, for which Google Maps returns the coordinates of the middle point in that city. Although for some particular cases of big cities it might not be as accurate, we consider it is still a good indicator for the spatial position, particularly when the target location is distant from the CC node.

For instance, we found a higher probability of this happening in Asian locations due to misspellings. Nevertheless, this filter should not affect much the results, since, for example, the distance from a Canadian point will not vary greatly from point A to point B in a North Korean city, yet allowing us to measure an approximate spatial separation between the two researchers. After applying these filters, we have the coordinates for all the authors in our sample.

Finally, we obtain Euclidean distance by using an implementation algorithm of the Haversine formula (GeoDataSource.com, 2015), which calculates the distance in kilometers between two points on a sphere from their longitudes and latitudes by taking into account the radius and curvature of the Earth. Refer to Figure 10 for the distribution of Euclidean distance in our sample.

## TRAVELING TIME

For the next step in our process, we collect the traveling time between scholars to have a more complete view of geographical proximity, per the literature. The time factor is collected from two different methods of



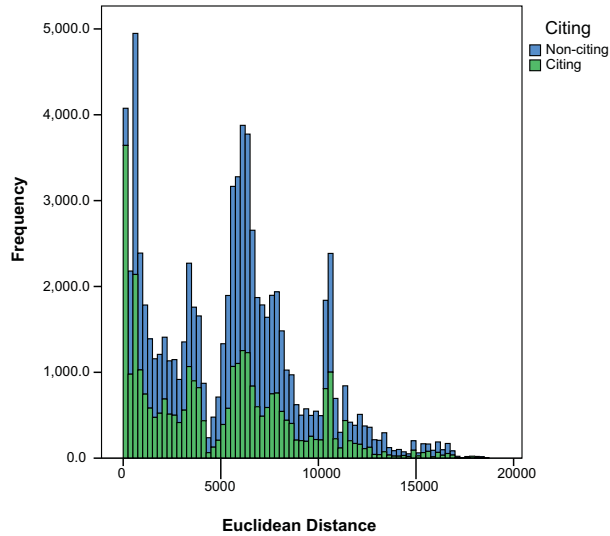


Figure 10: Distribution of Euclidean distance

transportation: plane and vehicle. Consequently, we initially divide time into two sub-metrics: Flying time and Driving time, which are later combined in a single variable: Traveling time. All time parameters are denoted in the format of hours, minutes, and seconds (hh:mm:ss).

It is important to note that driving time will only be calculated when the Euclidean distance between two authors is less or equal to 500 Km. In the literature, some authors studying the influence of spatial distance on collaboration have used the threshold of 100 miles (approx 160 Km) as maximum distance indicator for preferring travel by car over flying (Laursen et al., 2011; Bouba-Olga and Ferru, 2012). However, in Canada we find people choosing to drive for distances greater than this, for instance, from Montreal to Quebec City (around 250 Km/160 miles), or even outside their province, like from Montreal to Ontario (around 500 Km/310 miles). Therefore, we set our limit to be 500 Km as the maximum distance for which scholars would choose to drive. For all remaining distances greater than 500 Km, we make the assumption that an author would choose to mobilize to the location using an airplane as means of transportation.

To measure flying time we make use of the already calculated Euclidean distance, and as for speed, we make the airplane model assumption of a Boeing 747 aircraft with a cruising speed of 885 Km/h (Microsoft Encarta Encyclopedia, 2000). Although we acknowledge that for local flights the aircraft model might be different (resulting in other speeds), it still serves as a useful baseline for reference.

Next, we ran a programming script that uses the basic speed formula of  $time = distance/speed$  to obtain the flying time in seconds, which was later converted to time format (hh:mm:ss) and stored in our database.

Note that this time parameter is based on “as the crow flies” distance between two points, which means that it is not taking into account actual travel times or local amenities such as stop-over times, connections with different cities, delayed flights, customs waiting times, among other factors. This variable aims to

## CALCULATING TRAVEL TIME

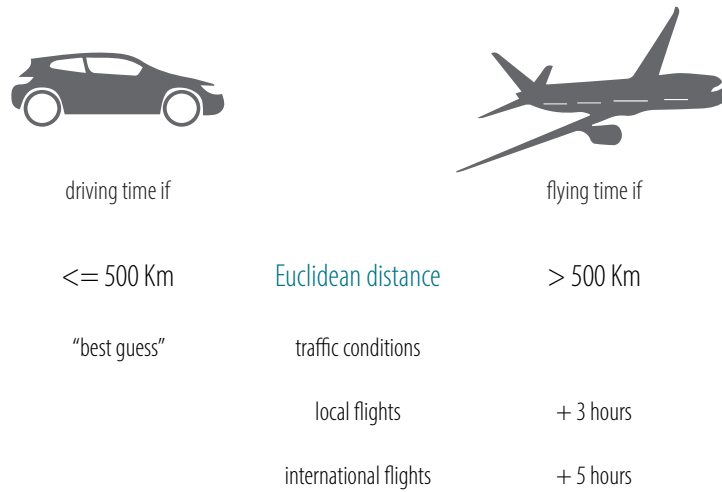


Figure 11: Selection of transportation mode

provide a reference for traveling time rather than being an exact travel measure in itself.

Nevertheless, we make some adjustments to our flying time factor, and allocate reasonable extra time to account for the time it takes to get to the airport, average waiting times before a flight, and to get from the arrival airport to the final destination (Torres et al., 2005). Thus, we apply post-processing following these considerations: for local flights, we add 3 extra hours, and for international flights, 5 extra hours. Figure 11 sums up the conditions to select the transportation method.

Recall the assumption that for Euclidean distances less or equal to 500 Km, a scholar will drive to the destination, whereas for longer distances, a scientist will prefer to travel by plane. For this reason, we calculate driving time only for all pairs meeting this distance condition. To this purpose, we wrote a programming script that queried the Google Maps Distance Matrix API web service (Google Inc., 2015) to obtain driving time.

This Google Maps web service provides travel distance and duration for a set of origins and destinations, in our case, we pass the latitude and longitude values previously collected for each author. The output returned is based on the recommended route between start and end points, as calculated by the Google Maps engine, as depicted in the screenshot in Figure 12.

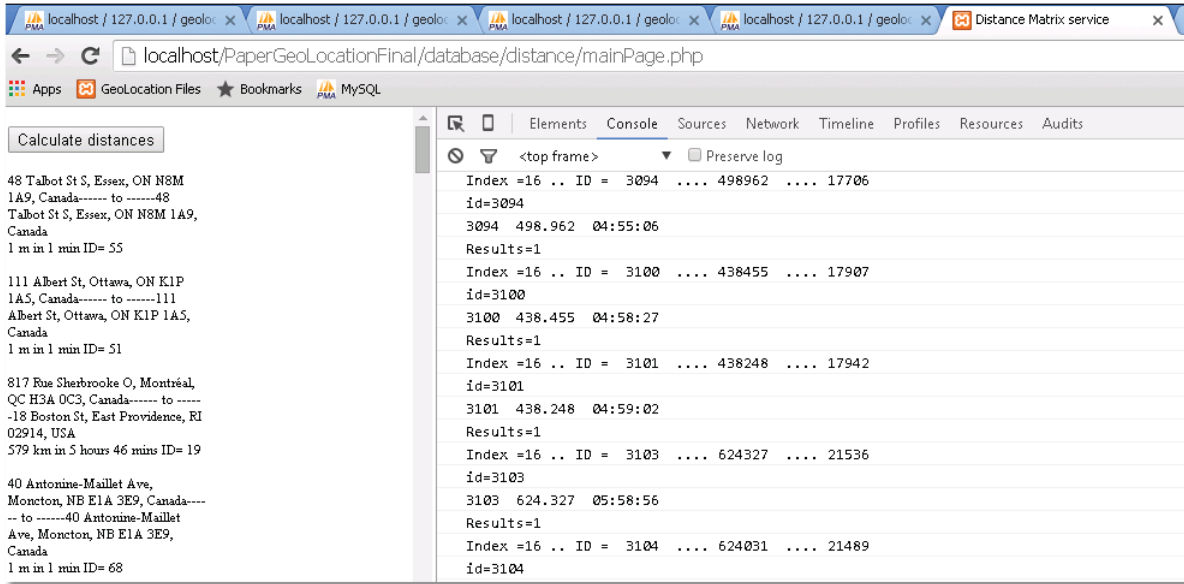


Figure 12: Script querying Google Maps distance API

Even though there are several options for mode of transportation (such as train or public transportation), we select “driving”, which indicates travel calculation using the road network. In addition, Google Maps allows for fine tuning in the travel estimations, like the influence of traffic conditions on the result, for instance, “optimistic” or “pessimistic” traffic. Regardless, we decided to use the default setting of “best guess” traffic, which is the best estimate of travel time, deeming both historical traffic conditions and live traffic.

In the end, we combine both of our traveling sub-variables into one, obtaining a consolidated metric for traveling time. Refer to Figure 13 for the distribution of traveling time measurements in the sample.

#### LOCATION CATEGORY

We find in our review of previous research that some authors include a ranking tool that serves for classifying geographical distance besides the pure spatial measurement, also taking into account geopolitical boundaries (e.g. Cunningham and Werker, 2012). Hence, we include this factor in our variable framework, and classify according to the difference or likeness in the geopolitical location between the two scientists in the dyad. Thus, our location categorization variable is rated according to the scale defined in Table 5.

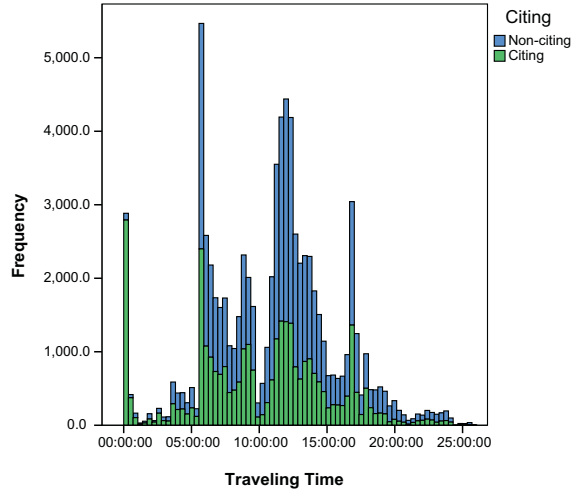


Figure 13: Distribution of Traveling Time

Value	Description
1	Same city/town
2	Same state/province
3	Same country
4	Same continent
5	Different continent

Table 5: Location category scale

As mentioned earlier, an additional benefit to finding the precise location coordinates by using Google Maps is that it also returns the address updated in a correct format. For example, we observed that some records in the data lacked the province field, or in occasions, they had obsolete geopolitical information like: Union of Soviet Socialist Republics instead of Russia, Federal Republic of Germany, or Czechoslovakia. In the last case of these, according to the specific location, Google Maps would return either Czech Republic or Slovakia as the country info, following further information in the address. In the end, this was useful for our location category metric, which depends on country, state or province, and city/town information being stored properly to be able to compare and grade according to the defined scale.

However, we still lacked continent information, since Google Maps does not provide this attribute. Hence, we wrote a short program to assign the continent according to the country, per the seven-continent model below, which is the most common.

- Europe
- Asia
- North America
- South America
- Africa
- Oceania
- Antarctica

Lastly, we created a script that compared the geopolitical fields between the two authors, and assigned the location category according to the scale mentioned in Table 5. The distribution graphs for our geography-related variables appears in Figure 14. As we did with cognitive distance, we also transform location category into five binary or “dummy” variables (see Table 6).

Dummy Variable	Value 1	Value 0
loc.1	Same city	Not same city
loc.2	Same province	Not same province
loc.3	Same country	Not same country
loc.4	Same continent	Not same continent
loc.5	Different continent	Not different continent

Table 6: Location category into dummy variables

Finally, we have all the geographical measurements for each author pair which will allow us to validate our hypothesis about spatial proximity. Now we go to the next step, and take a look into the attributes related to the social network formed by the scientists.

### 3.3.3 Measuring Collaborative Proximity

#### Collaborative proximity measurement in the literature

To obtain a picture of the collaborative proximity between researchers, we make use of the full co-authorship network between every author publishing in our working time range. The metrics used for this purpose are obtained from social network analysis applied to the co-authorship network.

#### SNA: Social Network Analysis

Social Network Analysis (SNA) is a methodology, rather than a formal theory in itself, for analyzing social networks, providing a set of statistical techniques for the study of their structure, which is why sometimes it is referred to as *structural analysis* (Wellman and Berkowitz, 1988, p. 20).

A social structure can be represented as a network consisting of a set of entities (the members of the social system, called *vertices* or *nodes*), and a set of ties showing their interconnections (usually called *edges*)

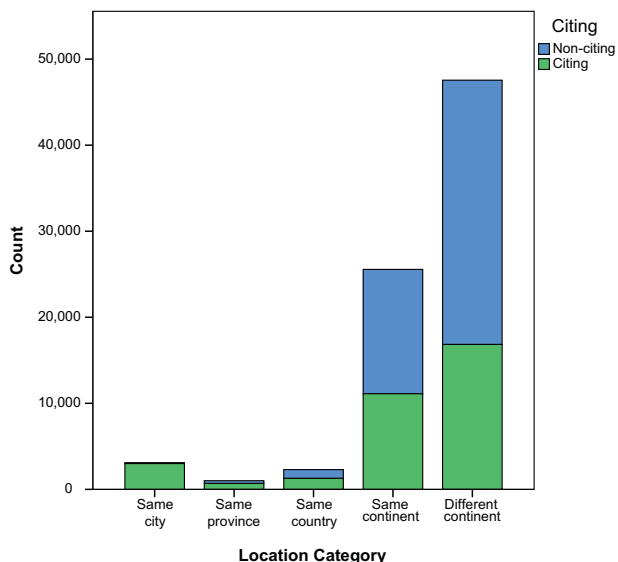


Figure 14: Distribution of location category

(Wellman and Berkowitz, 1988; Newman, 2001b; Butts, 2008), which can be visualized with a variety of graphs (Marion et al., 2003), like the one in Figure 15.

SNA originated under the influence of various fields, such as sociometrics, mathematics, and computer science (Otte and Rousseau, 2002). Based on a mathematical graph theory, SNA has become a multi-disciplinary approach with applications in many fields like sociology, information and computer sciences, geography, among others (Otte and Rousseau, 2002; Marion et al., 2003; Cunningham and Werker, 2012).

There are several parameters used to portray the characteristics of a complex network, such as calculating the network’s diameter, or various centrality measures such as eigenvector centrality and clustering coefficient. According to the literature, the most important centrality measures are considered to be degree centrality, closeness centrality, and betweenness centrality (Otte and Rousseau, 2002).

These network parameters mostly convey the way a node relates to the rest of the other nodes in the graph, while others, such as the shortest path between two vertices, may reveal features at the micro level (Ding, 2011). Moreover, SNA does not only consider people for its nodes, as is the case of Leydesdorff and Zhou (2007), where the betweenness centrality of journals dedicated to nanotechnology was used as a measure of interdisciplinarity in the field.

We remark on how network parameters may reflect on the node significance and value in terms of knowledge generation. For instance, a scholar with high level of connectivity (degree centrality) and a key position within the network, has greater potential at performing better in knowledge creation and diffusion (Liu et al., 2014). Similarly, research results by Sarigöl et al. (2014) indicate that several co-authorship centrality metrics can indicate citation success (when considering its interactions rather than a single network



Figure 15: Example visualization of a network graph  
 Source: La connaissance est un réseau (Grandjean, 2014)

measure).

Particularly, we find that SNA is highly useful to study the structural relationships within the co-authorship network of an academic knowledge base, having several examples of this in the literature (see Newman, 2001b, 2004; Barabási et al., 2001).

### Metrics and data collection of collaborative proximity

The complete database of the Canadian nanotechnology articles published in our working time range (2007 - 2010) has been employed to build the scientific collaboration network. We base the collaboration aspect on the co-authorship relationships between all the scholars participating in the period. Note that we do not build the network only with the authors present in our sample, since this would limit our view on the existing connections between scientists.

For this part of the analysis, we refer to the studied authors in the network terminology given to an individual entity, that is, as *nodes* or *vertices*. In addition, a co-authorship link would constitute a bidirectional relationship, hence its network graph would constitute an “undirected” graph, built with undirected edges (represented as simple lines without direction) as connections.

First, we had to prepare the existing data to show the co-authorship relationships for every article, given that there was one record per scholar/article. Therefore, we used a script that combined all the authors for each article in a single line separated by comma, which we exported to CSV format. The reason for this pre-processing was to conform our data to the Scopus format used by the Sci2 software, for which we only

used three of the format fields: article id, comma-separated co-authors, and year. The Science of Science (Sci2) Tool (Sci2 Team, 2009) is a software that provides tools specifically designed for the study of scientific networks, such as geospatial, topical, social networks, allowing for the analysis and visualization of scholarly data sets.

Afterwards, we loaded our data into Sci2 (refer to Figure 16), and extracted the co-authorship network, which consisted of an output file containing the lists of nodes and edges corresponding to our collaboration network from 2007-2011.

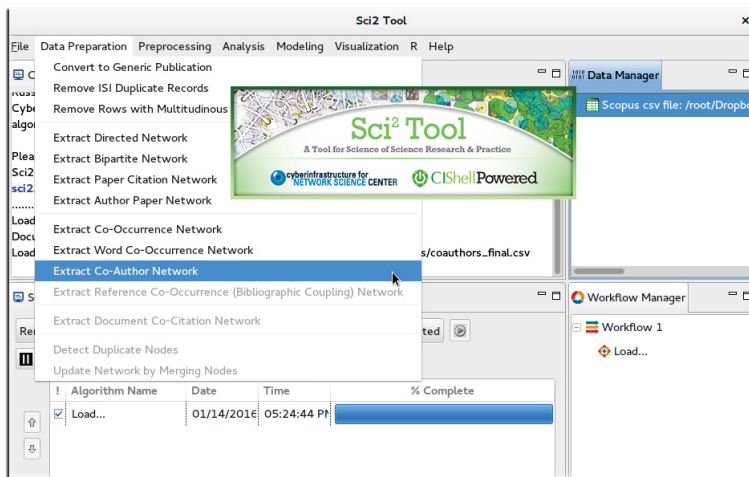


Figure 16: Screenshot of Sci2 software

Due to our network size, we had to run Sci2 on a Linux virtual machine (Fedora 23 Workstation version, 64-bit), due to the Java memory limitations when used on a Windows environment. Finally, we saved the extracted network to a CSV file, consisting of 674,113 nodes (authors) and 6,067,065 edges (co-authoring relationships).

Furthermore, the software displayed information about the extracted network in its console, among which we can observe that although there are 4,410 isolated nodes, our network is not weakly connected, with 94.87% of the nodes (639,573 in total) being part of the largest component (also called *giant component*). The complete output can be seen in Appendix E.1.

Moreover, the Sci2 software has in-built functionality that provides direct connection to visualization software such as Gephi (Bastian et al., 2009), which displays network graphs and allows applying layout algorithms, among other functionality. However, our complete co-authorship network is too big (over 6 million edges) to obtain a good visualization.

Nonetheless, we are able to retrieve the following general information by using the *Network Analysis Toolkit (NAT)* option of Sci2, which helps us get an overview of our network, as depicted in Figure 17.

To inspect the collaboration network structure and obtain the SNA-based proximity values, we use two



```
Network Average Degree: 18.00014241
Network All Degree Centralization = 0.00935385
Watts-Strogatz Clustering Coefficient: 0.74904137
Network Clustering Coefficient (Transitivity): 0.13400960
Network Diameter: From Quinto-et (925) to Meissner-ka (314063) is distance 14
(longest shortest path)
```

Figure 17: Network overview from Sci2

popular network analysis applications, Pajek version 4.06 (Batagelj and Mrvar, 1998) and NetworkX version 1.10 (Hagberg et al., 2008), which is a library based in the Python programming language.

First, after uploading our co-authorship network file in Pajek, we ran the *Network*  $\Rightarrow$  *Info*  $\Rightarrow$  *General* option, which returned some of the general values already obtained from Sci2’s Network Analysis Toolkit (see Appendix E.1). Refer to Figure 28 in Appendix E.2 for Pajek’s output with general network information.

In terms of our network setting, the total published articles authored by each author become a node’s weight, with the total works co-authored by a pair of nodes representing the weight of the edge or connection between the pair. However, all the SNA factors are computed without considering weight parameters.

As the next step, we went in Pajek to *Network*  $\Rightarrow$  *CreateVector* in the menu, and proceeded to run the different options to obtain the network metrics, which are briefly introduced below. Most of them were calculated relatively quickly, whereas some others (betweenness and closeness centrality) took 10 days each to complete. It is noteworthy to mention that all the parameters measuring individual characteristics of a node pertain to the “target” or reference node (REF), in this case, the cited author. The only exception is the shortest path, which instead refers to pairs of scholars.

#### SHORTEST PATH

A fundamental concept in graph theory, a geodesic is the shortest path of vertices and edges that links two given nodes (Newman, 2001a). In other words, the shortest path represents how many “steps” or authors are between two researchers tied up by co-authorship links with their peers. This variable is based on the implementation of Dijkstra’s algorithm (Dijkstra, 1959) for finding the shortest path, without considering weights.

Typically, the distance between authors is short, with only a few steps in the shortest path between them. For the case when the two specific scholars have collaborated in a publication, they would be direct neighbors, and the shortest path would be 0. It is relevant to mention that there might be no path between the nodes at all, in which case we set this value to a representation of infinite (9999) (Newman, 2004).

#### DEGREE CENTRALITY

The degree centrality of a node is defined as the number of nodes that are connected to that node. In other words, this parameter is the sum of all of the node's directly connected neighbors. When an author is "central", it has often been taken as an indicator of that author's popularity in the network.

#### BETWEENNESS CENTRALITY

This centrality measures how often is a node located on the geodesic or shortest path between other nodes in the network. Since by definition betweenness centrality is normalized as the proportion of all geodesics that include the particular node, it can also be expressed as a percentage.

A node's betweenness would express its average capacity to control the flow of information in the network (Uddin et al., 2011). The idea of this parameter is that if a node with a high level of betweenness were removed, the network would fall apart into coherent clusters (Leydesdorff and Zhou, 2007). Hence, nodes acting as proxies to join different clusters would have a high betweenness value.

#### CLOSENESS CENTRALITY

This parameter is the average of the shortest path distances between a specific node and all the other nodes in a network. Hence, this value would only be available for connected nodes.

Closeness centrality would expand on the concept of of degree centrality by emphasizing how close a node is to all other nodes in the network (Uddin et al., 2011). As a result, the more central a node is, the lower its total distance from all other nodes is going to be.

#### EIGENVECTOR CENTRALITY

Eigenvector centrality computes the centrality for a node based on the centrality of its neighbors basing on the eigenvector of the largest eigenvalue of an adjacency matrix. This centrality metric relates to the influence of a node in the network by assigning relative scores to every node. The idea behind this concept is that links to nodes representing authors with high scores contribute more to the score of that particular authors than links to low-scoring authors.

Unlike degree centrality, which considers every node equally, eigenvector centrality weighs nodes according to their centralities. Since this means taking into account not only direct connections but indirect links as well, eigenvector would be a centrality measure considering the entire network pattern (Bonacich, 2007).

#### CLUSTERING COEFFICIENT

Social networks are inclined to form *cliques*, or tight groups with individual nodes highly tied together (Watts

and Strogatz, 1998). Thus, this coefficient measures the degree of the clustering tendency of a particular node in a network, by making use of the links to all the other nodes in the system.

In a co-authorship network, the clustering coefficient of a node represents the willingness of this node’s collaborators to collaborate with each other, indicating the probability that two of its collaborators wrote a paper together (Barabási et al., 2001). Roughly speaking, this metric stands for how well connected the neighborhood of the node is, thus revealing its “cliquishness”. If the node’s neighborhood is fully connected, the clustering coefficient of such node would be 1. Contrarily, a value close to 0 would mean that there are hardly any connections in that node’s neighborhood.

Moreover, networks with a high clustering coefficient and a low mean shortest path meet the criteria for being considered *small-world* networks, referring to the phenomenon in which two strangers often find that they have a friend in common (Watts and Strogatz, 1998). This feature is typically exhibited in any human social network: in any cluster of friends, each friend is also connected to other clusters, thus usually taking only a short string of acquaintances to connect any two people on earth.

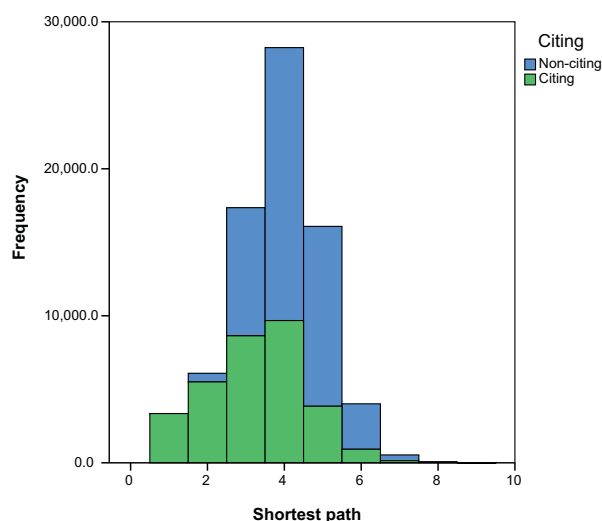


Figure 18: Distribution of Shortest Path

Figure 29 in Appendix E.3 shows the distributions for all the SNA metrics pertaining individual target nodes. In particular, for degree centrality the majority of scholars seem to have a small number of coauthors, whereas a few scientists in the network have collaborated with many or even hundreds in some cases (Newman, 2001b).

Although Pajek provides a way to get the shortest path length matrix, which calculates the geodesics for every edge in the network (option *Geodesics Matrices*), this operation is only available for small networks. Another way would be to calculate the shortest path between two nodes through manual input of the nodes identifiers. However, we need to compute the shortest path between all of the pairs in our sample, and this

large number makes it unfeasible for us to use Pajek for this purpose.

Therefore, we had to use the scripting capabilities of NetworkX to obtain the shortest path between a specific set of nodes in our sample. The Python script we created to gather the geodesics by using NetworkX can be found in Appendix C.2.

As previously stated, coauthors in the networks would have a 0 value for shortest path; however, NetworkX assigns direct neighbors a shortest path value of 1 instead. Moreover, we define in our code an alternative value of “9999” to be used as the shortest path between scientists who are not connected, to avoid getting errors when running the script. We shall discuss later how we declare these as “missing values” for our analysis. Figure 18 shows the distribution for the shortest path parameter (note that missing values have been excluded from the graph).

As the final step in the data collection concerning collaborative proximity, we imported the resulting SNA measurements to MySQL and mapped the values back to its corresponding author, keyed by the *node id*, a number assigned to each scientist by Sci2.

## 3.4 Data Analysis

In this section we start with an overview on the methods employed to analyze the various proximity factors involved in citation behavior. Then, we state the null hypotheses, summarizing as well the variables to be used in the analysis, which are based on the metrics discussed in the previous section. Finally, we implement the presented procedures with our data.

### 3.4.1 Regression and Classification Modeling

Data scientists and scholars in general dealing with quantitative research analyze data seeking to learn from it and be able to draw conclusions of various natures. The desired conclusions may range from descriptive or explanatory (like discovering occurring patterns), to predictive conclusions, such as categorical classification (Getoor, 2005).

In particular, we find that there has been a growing interest in learning from structured data, defined as data forming a graph where the nodes are objects and the edges in the graph are links or relations between objects (Getoor, 2005). Notably, citation links constitute a form of structured data that indicates relationships concerning the behavior of formal knowledge and its flow.

The Merriam-Webster Web Dictionary (2014) defines *classification* as the systematic arrangement in groups or categories according to established criteria. In machine learning and statistics, classification is the problem of identifying which class (also, category or sub-population) a new observation belongs to, on the basis of a set of data containing observations (also called instances or cases) whose class is known (Michie

et al., 1994). A class is thus defined as an outcome that must be predicted from known attributes. In short, classification is identifying group membership.

From the numerous approaches taken towards the task of classification, there are three main distinct lines of research addressing this goal: statistical analysis, machine learning, and neural networks (Michie et al., 1994). Throughout the literature, we find that two of these major branches for classification are applied in the study of citation analysis when considering its influential factors: statistical approaches and machine learning. Given the categorical nature of a citation link (i.e. citing vs non-citing), it follows that we face the problematic of its correct classification.

In terms of machine learning, classification has two distinct meanings (Michie et al., 1994; Alpaydin, 2014):

- Unsupervised learning (or clustering): When the classes or clusters are not known in advance, and they need to be construed from the data itself; and
- Supervised learning: When we know *a priori* the set of classes, and the goal is to establish a rule by which to classify a new observation into one of the existing classes.

Sometimes in statistical literature supervised learning is also be referred to as *discrimination*, where the importance resides in following classification rules drawn from given correctly classified data (Michie et al., 1994). From this point forward, whenever we refer to classification, we shall consider its definition with regard to supervised learning, i.e., when the outcome classes are already known.

## Regression

Statistical approaches rely on explicit probability models, providing observations with a probability for being in each class, rather than classifying them (Michie et al., 1994). For this reason, scientists studying the behavior of scholarly publications tend to use traditional statistical approaches and tools, such as linear discrimination analysis (LDA) or binomial logistic regression.

Particularly, statistical models based on regression techniques that account for main effects allow testing whether factors have incidence on a dependent parameter, being able to distinguish between those that do have some correlation and those constituting noise (Lemon et al., 2003). For instance, linear regression has been applied to discern future number of citations by means of authors' reputation (Castillo et al., 2007; Yan et al., 2012).

Moreover, with regression modeling we can assess the influence of statistical interactions among independent factors, while also having each factor as part of the model (Lemon et al., 2003).

Given that regression models are widely known in the academic realm, we do not explore their conceptual basis in depth, rather proceeding to discuss the other method of interest: classification in terms of machine learning.

## Classification: Machine Learning

Link prediction is an important task in network science, characterized by mining tasks performed on academic data with the goal of discovering patterns and factors to achieve such predictions. Data mining has greatly benefited from advances in methodology since the mid-1980s, with the academic community creating and increasingly improving the predictive accuracy of machine learning algorithms Breiman (2003).

A subfield of computer science, machine learning is a method of data analysis that automates analytical model building. It consists in algorithms that learn from data through their ability to independently adapt with each iteration, finding hidden insights without human intervention nor instructions of where to look (Michie et al., 1994; SAS Institute Inc., 2015). This technique is commonly used to extract useful information from large data sets, to later display it in simple and comprehensible visualizations of the relationship between the input variables (the observed attributes or features) and the responses (the pre-defined class) (Breiman, 2001; Song and Ying, 2015).

We focus on supervised learning (which is the most widely adopted machine learning method), where the learning algorithm shapes the model through a set of inputs and outputs by comparing its actual output with known correct outcomes to find errors, modifying the model accordingly (SAS Institute Inc., 2015). With prediction usually as the main goal of data analysis (Neville, 1999), supervised learning algorithms use methods like classification and regression to identify patterns that enable them to make such predictions.

Moreover, any environment naturally mapping to a network would have an equivalent mapping from link prediction applied on that network back to a key factor in the environment (Lichtenwalter et al., 2010, p. 1). Accordingly, they attempt to make the classifying expressions simple enough to be understood easily by humans, mimicking human reasoning reasonably enough as to provide useful insights into the decision process (Michie et al., 1994). Thus, they learn a (classifying) task from a series of examples by means of automatic computing procedures based on logical or binary operations, and produce reliable and repeatable decisions and results (Michie et al., 1994; SAS Institute Inc., 2015).

From their broad applicability in diverse fields, we feature their application in the study of the behavior of scholarly data, for which we find several examples. For instance, Getoor (2005) made use of the structure of binary citation links to improve classification accuracy.

Furthermore, researchers throughout the literature propose numerous approaches to predict the number of future citations and their success (based on the author's  $h$ -index<sup>1</sup>) amid scholarly work (e.g. Hirsch, 2007; Dietz et al., 2007; Acuna et al., 2012). Among these, we call attention to Dong et al. (2015) and his predictions for scientific impact and collaboration based on citation analysis and academic social networks.

While there are many algorithms for supervised learning out there, we emphasize those used to analyze factor contributions in academic data. We find that many authors in the literature make use of a logistic

---

<sup>1</sup>The  $h$ -index is an index that attempts to measure both the productivity and impact of a scholar's published works.

regression classifier (LRC) (see Getoor, 2005; Bethard and Jurafsky, 2010; Dong et al., 2015). In this case, the statistical analysis performed by logistic regression is used iteratively for training as a supervised classifier. LRC is often implemented for classification purposes due to their output consisting of linear combination of the variables with weights, which would provide insights on variable importance (Breiman, 2003)

In addition, other prominent models such as support vector machines (SVM) would also have the potential to produce accurate predictions. Yet, SVM is arguably better when applied to regression problems (i.e. when the response variable is of continuous nature) than for binary classification (Bethard and Jurafsky, 2010; Breiman, 2003).

In the same way, we reviewed models based on decision trees (DT) classifiers, which were introduced in the 1960's (see Neville, 1999; Dong et al., 2015). Decision trees are one of the most popular methods for data mining, due to their ease of use and interpretation, robustness even with missing values, and their flexibility to use both discrete (categorical) and continuous variables (Song and Ying, 2015).

Originating from decision trees, random forests (RF) algorithms (Breiman, 2001) grow an ensemble of trees (i.e. "a forest") and lets them vote for the most prominent class. Notably, RF are considered to be accurate classifiers, showing comparable or even better prediction performance than other learning methods (Breiman, 2001; Xu, 2013). They have been extensively used in classification problems (e.g. Lichtenwalter et al., 2010; Sarigöl et al., 2014; Dong et al., 2015), having led to significant improvements in classification accuracy (Breiman, 2001) thanks to their particular advantages, such as robustness against overfitting (a potential problem with decision trees).

## **Decision Trees**

Decision trees (DT) are a powerful tool for classification and prediction that operate through a series of rules based on parameter selection. These rules consist on logical (IF-THEN) expressions, and their output is displayed as limbs in the form of a tree, making up branches as they split from the root (the dependent variable) or older branches (parameters or descriptors), and finally, the unsplit nodes constitute the leaves of the tree (Neville, 1999; Tong et al., 2003). Ideally, the data in a leaf must be related to a combination between the split of the parameter value and the target measure, so that the tree represents the correct classification of data into purified groups (Neville, 1999).

The split rules would be easy to interpret intuitively as answers to questions concerning how does the association between descriptors affect the response variable, as opposed to nonlinear "black box" algorithms like neural networks (Neville, 1999). This ease of interpretation makes this method appealing because of their clear depiction of how a few inputs determine target groups (Neville, 1999; Tong et al., 2003).

Typically, DT allow missing values, treating them as a special case, which lead them to constitute an additional branch within the split rules (Michie et al., 1994). In addition, the popularity of trees would also

be due to the flexibility they provide in accepting different variable types for both target and predictors: nominal, ordinal, and continuous (Neville, 1999).

We already mentioned that trees are formed by “branching” or splitting nodes according to the input parameters, in the search of achieving purity of a specific class. Here, an ideal “purity” would be having a single class at each resulting node. The splitting criteria, namely, “impurity functions”, basically select the split that has the largest difference between the impurity of the parent node and a weighted average (Lemon et al., 2003). Some examples of impurity functions are: Gini index, entropy, and minimum error.

There are several algorithms based on the concept of decision trees, whose variations in construction and implementation have improving performance as goal (Tong et al., 2003). The tree-growing procedure is similar for all the algorithms in that they all generate trees by recursively splitting the data into smaller and smaller subcategories.

Among these, we find QUEST (Quick, Unbiased, Efficient Statistical Trees), CHAID (Chi-square-Automatic-Interaction-Detection), C5.0 (and its previous versions), and CART (or C&RT: Classification and Regression Tree). We summarize the most popular DT models by comparing their features and operating mechanism in Table 7 (see also Neville, 1999; Lemon et al., 2003; Tong et al., 2003; Song and Ying, 2015).

Methods	CART	C5.0	CHAID	QUEST
<b>Measure to select input variable</b>	Gini index; Twoing criteria	Entropy info-gain	Chi-square	Chi-square**
<b>Pruning</b>	Pre-pruning: single-pass algorithm	Pre-pruning: single-pass algorithm	Pre-pruning: Chi-square test	Post-pruning
<b>Dependent variable</b>	Categorical/Continuous	Categorical/Continuous	Categorical	Categorical
<b>Input variables</b>	Categorical/Continuous	Categorical/Continuous	Categorical/Continuous	Categorical/Continuous
<b>Split at each node</b>	Binary*	Multiple	Multiple	Binary*
<b>Missing values handling</b>	Substitute predictor fields case	Fractioning	Special category	Substitute predictor fields
<b>Stopping rules customization</b>	Yes	No	Yes	Yes

\* Split on linear combinations

\*\* Only for categorical variables

Table 7: Comparison of different decision tree algorithms  
Source: Modified based on Song and Ying (2015, p. 3)

Let us now briefly introduce the three most notable decision tree models. Based upon chi-square statistics (Lemon et al., 2003), CHAID recursively partitions data using splits that must achieve a threshold level of significance: the chi-square test of independence between the nominal target values and the branches. It terminates when no more merges or re-splits are significant.

C&RT analysis (also called CART) is a decision tree methodology first developed by Breiman et al. (1984) that examines all possible binary splits, and implements a sophisticated “pruning” mechanism to minimize misclassification error and overfitting. CART uses a generalization of the binomial variance called the Gini index, and it is considered to be one of the best decision tree methods due to its selection of most influential



predictors (Lemon et al., 2003; Loh, 2011).

A commercial improvement over C4.5 algorithm, C5.0 works by splitting the data according to the field that provides the maximum information gain (entropy) at each level (Loh, 2011). It is developed by Quinlan (2014), and for the most part, C5.0 works and performs quite similarly to CART. One notable difference between them is that CART grows the tree based on a splitting criterion iteratively applied to the data, whereas C5.0 includes the intermediate step of constructing sets of rules.

Generally, all DT methods (excepting C5.0), provide the investigator with some level of control by applying what is known as “stopping rules”. These allow the user to specify how large a tree to grow, and establish thresholds for statistical differences before declaring results as meaningful (Lemon et al., 2003).

Nevertheless, there are some weaknesses to decision trees, and the most noteworthy is that DT can be prone to overfitting, particularly when using small data sets. Overfitting is when the model fits noise by way of paying too much attention to irrelevant data, and ends up “memorizing” training data instead of learning from it.

For instance, if a decision tree were allowed to grow freely, it could classify the given data with 100% accuracy, at the expense of creating a very complex tree-structure that would most likely not perform well on new data (Michie et al., 1994). This could potentially mean a high bias to the training set, and it would ultimately limit the generalization capability of the resultant model (Song and Ying, 2015).

At any rate, Michie et al. (1994) suggest correcting or avoiding altogether the bias by using cross-validation and/or independent data samples: one to learn the classification rules and another to test it. As a result, comparing the predicted vs the true classifications on the test data would give an unbiased estimate of the error rate for the tree classifier.

Furthermore, DT may not be the best tool to quantify the impact of a single independent variable on the outcome of interest. On one hand, since decision trees estimate average effect, if the goal is to learn the impact of each parameter separately, DT should not be used as a substitute for regression techniques (Michie et al., 1994).

On the other hand, this could be seen as an advantage, given that DT already discover interactions between variables without having to specify them as a separate formula (in contrast to logistic regression). The nature of the tree would allow formation of interactions not by explicit operations on the variables, but through the tree structure itself.

Lastly, decision trees are deemed by some to be unstable, and their sensitivity to multicollinearity between independent variables has been criticized. In fact, strongly correlated inputs may result in selecting variables that improve the model statistics but do not explain the response variable (Song and Ying, 2015). However, these two issues would be addressed by means of using ensemble methods such as bagging, or random forests, instead of relying on a single tree (Breiman, 2001; Liaw and Wiener, 2002; Lichtenwalter et al., 2010).

## Random Forest

A random forest is a classifier formed by a collection of tree-structured classifiers and developed by Breiman (2001). In RF, a subset of predictors are randomly chosen at each node, and then split by using the best among them. After a large number of trees has been grown, each tree votes for the most popular class. Though the splitting strategy might seem counterintuitive, it has been proven to increase classification efficiency and performance, when compared to other classifiers such as SVM (Liaw and Wiener, 2002; Lichtenwalter et al., 2010).

RF algorithm would provide various advantages, like allowing combining categorical with numerical variables (Breiman, 2001), overfitting prevention (Lichtenwalter et al., 2010), high scalability (Sarigöl et al., 2014), and ease of setup. Indeed, in this model, only two parameters need to be set: the number of variables used at each random subset, and the number of trees in the forest. Still, they usually have low sensitivity to these values (Liaw and Wiener, 2002), which means that RF models maintain their stability throughout different parameter settings. Moreover, RF would be an efficient method of variance reduction for large data sets, because they counter increased training time for the whole forest with decreased training time for each single tree (Lichtenwalter et al., 2010).

In RF, the prediction error estimate is called *Out-Of-Bag prediction error* (OOB error), which is the equivalent to the *mean squared error* in regression, and to *misclassification error* in classification (Xu, 2013). In general, the OOB estimate of error rate is quite accurate, provided that enough trees have been generated (Breiman, 2001). Besides, RF would eliminate the need for having separate cross-validation and further error estimation, because they already have these procedures as part of their internal classification routine (Sarigöl et al., 2014).

In terms of dealing with predictor dimensionality, RF estimate the variable importance by inspecting how much the OOB error increases when that parameter is permuted while all others are left unchanged (Liaw and Wiener, 2002). In the case of statistical techniques, such as logistic regression or the Cox model, the recommended practice is to delete less important covariates to reduce the dimensionality, or the resulting model might become unstable (there would be too many competing models) (Breiman, 2003).

However, reducing dimensionality is not an easy task given that the importance of a variable might be related to its interaction with other variables (Liaw and Wiener, 2002). Indeed, RF has been known to yield accurate classifications for data with a large number of features (Sarigöl et al., 2014). In addition, deleting variables would also decrease the amount of information available for prediction (Breiman, 2001). Nevertheless, variable importance measures produced by RF are useful for model reduction by clearly identifying a few informative predictors and ignoring the other noise variables, with essentially the same prediction accuracy (Liaw and Wiener, 2002).

Finally, the main disadvantage of RF is the low interpretability due to its complex mechanism for producing a prediction. Indeed, “*trying to delve into the tangled web that generated a plurality vote from 100 trees is a Herculean task*” (Breiman, 2003, p. 10). Yet, highly easy-to-understand algorithms might not make the most accurate predictors, and simplicity would have to be sacrificed to achieve greater accuracy (Breiman, 2001; Lichtenwalter et al., 2010).

## Prediction Models vs Statistical Approaches

Arguably, the objective of statistics is to use data to obtain information about its underlying mechanism in order to understand it, and, as ultimate goal, being able to make predictions based on the conclusions drawn (Breiman, 2003). Contrarily, D. R. Cox, a prominent British statistician, admonishes in his comment on Breiman (2003, p. 18) that emphasizing successful prediction as the sole objective does not go in hand with the pure goals of statistics: interpretation and understanding.

However, Breiman (2003) criticizes that statisticians rely too heavily on assumptions and data modeling. Hence, their enthusiasm for fitting data models would be a potential issue as data becomes more complex, and it needs modifications so it can fit a model. As a result, the conclusions could turn out to be misleading in the sense that although they may pass goodness-of-fit tests and residual checks, they would be about the model’s mechanism rather than the nature of the data.

Likewise, Michie et al. (1994) bring to attention several practical problems like the removal of attributes (if they are not high contributors to the discrimination) and data transformation as well. Moreover, citation behavior may be difficult to estimate using traditional regression analysis due to their intrinsically heavy-tailed distribution of citation counts (Dong et al., 2015). In addition, it is assumed as a rule that human intervention is required to implement statistical techniques, especially in regards to variable selection and transformation, and overall structuring of the problem (Michie et al., 1994).

Alternatively, the task of citation analysis can be formulated as a classification problem (Dong et al., 2015; Jawed et al., 2015), to be solved with supervised learning techniques. Furthermore, link prediction is considered a well-known problem in field of SNA, since we intend to guess the likelihood of the occurrences of connections between nodes (Jawed et al., 2015).

Another aspect to consider is the common inadequacy of regression to discern some relationships in data in social research, and the fact that classification models like decision trees can easily get around it (Neville, 1999). Nevertheless, Michie et al. (1994) admonish practitioners of machine learning to assimilate and use statistical techniques as well, despite the ongoing debate between these two cultures. Reason for this would include, for instance, how decision trees should not be regarded as a substitute over proven statistical regression, in cases where the average effect of parameters over a response variable wishes to be discerned (Lemon et al., 2003).

### 3.4.2 Hypotheses and Variables

#### Null Hypotheses

We rephrase our research questions introduced in the previous chapter as null hypotheses, to be later assessed as statistical inferences. Importantly, we split the analysis of collaborative proximity into two sub-hypotheses.

**Null Hypothesis  $H_0$  1** *Citation probability does not increase when a Canadian author is cognitively proximate to another author.*

**Null Hypothesis  $H_0$  2** *Citation probability does not increase when a Canadian author is geographically proximate to another author.*

**Null Hypothesis  $H_0$  3a** *The probability of a Canadian author citing another does not increase due to the referenced author's position within the collaboration network.*

**Null Hypothesis  $H_0$  3b** *The probability of a Canadian author citing another does not increase if they are closely connected to each other within the collaboration network.*

#### Response Variable

We study the impact of proximity on academic knowledge production through the performance measure of an existing *citation link* in a dyad, which is a pair of authors. The citation relationship is taken from the pairings, with this link expressed as a binary value (i.e. yes/no represented by 1 and 0).

Per the literature, scientific citation probability can either be formulated as a regression problem (the traditional approach), or alternatively, as a classification problem to be solved with supervised learning techniques (see Yan et al., 2012; Jawed et al., 2015; Dong et al., 2015).

Thus, we adopt two distinct methodologies to conduct our analysis, given the boolean nature of our dependent variable:

1. Binary regression modeling
2. Machine learning-based classification modeling

Each method takes the citing behavior as the response or dependent variable, that is, if the citation link has been established or not.

#### Independent Variables

In like manner, we consider a set of independent variables based on the metrics representing collaborative proximity and general network attributes, as well as those representing cognitive and geographical proximity. Below we list the independent variables used in our analysis for testing each hypothesis. The variables are presented in the order of the hypotheses to whose validation they contribute.

- |                                      |  |
|--------------------------------------|--|
| 1. Cognitive proximity ( $H_01$ )    | (b) Betweenness centrality ( $H_03a$ ) |
| (a) Cognitive category               | (c) Closeness centrality ( $H_03a$ )   |
| 2. Geographical proximity ( $H_02$ ) | (d) Eigenvector centrality ( $H_03a$ ) |
| (a) Euclidean distance               | (e) Clustering coefficient ( $H_03a$ ) |
| (b) Traveling time                   | (f) Shortest path ( $H_03b$ )          |
| (c) Location category                | 4. Control variables                   |
| 3. Collaborative proximity           | (a) Published works                    |
| (a) Degree centrality ( $H_03a$ )    | (b) Co-authored works                  |

### Control variables

We distinguished two particular parameters that could have an important effect on citation probability. Hence, we consider them as control variables and also include them as independent parameters in our model.

#### PUBLISHED WORKS

This variable refers to the number of articles published by the target or cited author during the specified time period. Since we consider that a researcher that has published a high number of papers could be deemed prestigious in the field, we expect the scientist to be highly cited by peers regardless of their proximity measurements.

#### CO-AUTHORED WORKS

This parameter represents a simple count of the published works that have been co-authored by each pair of authors. It is likely that if two researchers have co-published scientific articles, they cite each other in their subsequent works.

The output file of the extracted network previously obtained from Sci2 also contained the tabulated information for the number of authored works for each author, and the co-authored works for every pair of scientists, since they constitute the weight per node and per edge respectively. Thus, we collected the values for our control variables, whose distributions may be found in Figure 30 in Appendix E.4.

### Data Pre-processing

Prior to running the analysis, we first had to deal with repeated pairs of the same authors, given that our statistical unit is unique CC-REF node pairs. We found that in our sample, there were 538 pairs occurring more than once, usually repeated 2 times, and 4 times at maximum. Thus, for repeated pairs, we calculated

the means for scale variables, and for categorical variables, we chose either the most frequent category, or the highest one (when there was no value occurring more often than the others).

The resulting set contains 79,524 unique CC-REF author pairs, which will serve our goal of quantifying the impact from all proximity factors on the citation probability for unique pairs of authors. Finally, we have all the required information to go into the analysis stage of our work.

### 3.4.3 Applying Regression

We use Stata (StataCorp, 2015), a data analysis and statistical software, to test the validity of our hypotheses through regression models. Importantly, we define the correct format for each variable (e.g. time format *hh:mm:ss* for traveling time) after uploading into Stata.

In addition, we distinguish “missing values” in each variable, namely rows with a 0 value in closeness centrality (isolate authors) or where no shortest path existed (9999 value), so they can be excluded from further analysis. Descriptive statistics for the independent variables can be found in Tables 13 and 14 appearing in Appendix D.

#### Correlation Analysis

We know *a priori* that some of our independent variables might be inter-associated, such as travel time and Euclidean distance, or that high collinearity may exist between some centrality metrics (Valente et al., 2008). Therefore, we first validate whether our variables are correlated or not.

For this purpose, we use the most common correlation measure: the Pearson Product-Moment Correlation Coefficient (PC), which reflects the degree of linear relationship between two variables. Conventionally, variables are considered to be uncorrelated if  $PC < 0.1$ , weakly correlated if  $0.1 < PC < 0.3$ , moderately correlated if  $0.3 < PC < 0.5$  and strongly correlated if  $0.5 < PC < 1.00$  (Schiffauerova and Beaudry, 2011). Figure 19 displays the correlation plot as a graphical representation. For the precise values, refer to Table 15 in Appendix D, which shows the correlation matrix with the coefficients for each pair of independent variables. It is noteworthy to mention that all the correlations discussed here are statistically significant, that is, they have a 2-Tailed significance value of less than or equal to 0.05.

We first proceed to explore the correlation between explanatory parameters against the dependent variable. Cognitive category and shortest path were found to have a negative moderate correlation to the citing response, whereas the other proximity metrics displayed a weak correlation. In terms of the centrality metrics, we find the weak linear correlation unsurprising due to similar findings by Sarigöl et al. (2014).

Next, we check correlations between our control variables and the rest of the independent parameters. Coauthored works was found to have a weak correlation to shortest path (evidently, co-publishing pairs would always have zero hops in their shortest path), whereas published works is strongly correlated to the

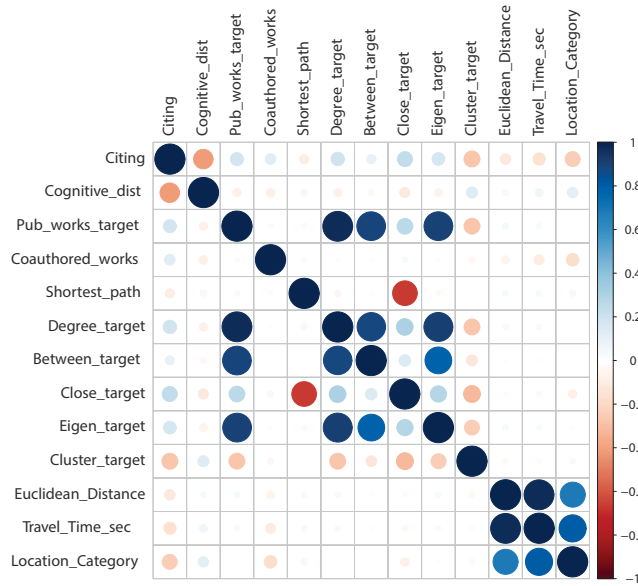


Figure 19: Correlation Plot

target author’s degree centrality, as well as to betweenness and eigenvector centrality.

Now, among the independent variables involving the social network, we predictably found a strong correlation between degree and betweenness centrality, and to eigenvector and closeness centrality. Clustering coefficient is in general weakly correlated to the rest of the SNA metrics, with the exception of a negative moderate correlation to the shortest path. Closeness centrality and shortest path have a negative moderate correlation between each other, while also being moderately correlated to degree centrality and clustering coefficient, and a positive moderate correlation to location category. Moreover, as anticipated, there is a high correlation between Euclidean distance, traveling time, and location category.

Even though the response variable is not strongly correlated to any of the studied proximity metrics, we do not take this as enough evidence to discard uncorrelated to poorly correlated variables. We consider the possibility that the correlations between response and predictors might not follow a linear relationship, which is the aim of the Pearson correlation analysis; after all, linear correlation does not necessarily mean causation.

In the end, we discard altogether the control variable of published works on account of its high correlation to degree centrality. As a matter of fact, this correlation was expected, as it is comparable to the correlation between number of collaborators and number of patents found by Schiffauerova and Beaudry (2011): a scientist’s high number of co-authors would usually be related to an increased number of publications, unless said author would choose to continue working always with the same research group for all his/her published articles. Since these two parameters have the potential to account for highly prolific authors, we cannot

control for the number of publications without removing the effect of a scientist’s degree centrality.

Finally, from the SNA metrics, we completely drop closeness centrality mostly due to being correlated with shortest path, besides correlating with the other centrality metrics as well. For the remaining centralities (degree, betweenness, and eigenvector) we decide to run the model using each one of them alternately at a time. In the same way, we shall assess the effects of the geographical metrics (Euclidean distance, traveling time, and location category) by including them in the model by turns.

## Model Setup

We implemented two regression models applicable for a binary response variable: Logistic (Logit function in Stata) and Probit.

Logit uses the logistic distribution function, whereas the probit model employs a probit link function, both having an output based on probabilities as their predicted values, hence delimited between (0,1). They are very similar in most cases, both using maximum likelihood estimation, and they are commonly used for the same kind of problems. Plus, these models are easy to interpret, enabling analysis for factor contributions and parameter settings (Dong et al., 2015). Also, they are robust to small noise in the data and are not particularly affected by mild cases of multi-collinearity (Long and Freese, 2006).

Besides, Logit/Probit models have fewer assumptions than other types of analysis in that they make no assumption on the distribution of the independent variables (Liao, 1994). This means that since they can work with skewed distributions, no normalization procedure is required for our explanatory factors. Ultimately, we decided to carry out both models as a robustness check for our analysis, relying on a 95% level of significance to test the significance of each variable and thus accept or reject the hypotheses.

Furthermore, we represent our categorical variables (i.e. cognitive and location category) with the generated dummy variables, excluding the dummy representing the last level for it to be considered as the reference category, as suggested by Le Cessie and Van Houwelingen (1991).

To start with, we need to account for the non-independence between individual dyads, since we are pairing up the CC authors multiple times with different REF authors. To this purpose, we enable the following option: `vce(cluster clustvar)`, which indicates the regression to allow for intra-group correlation. This command relaxes the usual requirement for the observations to be independent from each other so we can obtain correct standard errors (Baldi, 1998); *clustvar* specifies the association for each observation, in our case, the *source node id* identifies pairs having the same CC author.

## Interaction terms

Given that the impact of spatial distance may be better comprehended in combination with other factors, as Morgan (2004) claims, we decide to check for interaction terms. Interestingly, the interaction between



Euclidean distance and shortest path yielded significant behavior, so we shall include it in our model; we also checked for combinations between cognitive and SNA metrics, with no verifiable effect discovered.

### **Goodness-Of-Fit (GOF) testing**

The GOF tests for the model, namely, the Hosmer-Lemeshow and  $R^2$  tests, failed to reach significance in our regressions. However, it has been previously suggested that these tests might not perform well every time (as opposed to tests targeting linear models) (Le Cessie and Van Houwelingen, 1991; Hosmer et al., 1997). Accordingly, we presume that this issue might be caused by the large sample size we have. Moreover, the high number of observations increases the degrees of freedom used for the tests.

Therefore, we decide to run an experiment with a smaller random sample. We discovered that for regressions using less than 1,000 observations (we tested with a 1% random sample of about 700 observations), the fitting of our model was found to be significant. With this in mind, we proceeded to run the experiment 10 times, each time with a different 1% random sample, as way of validating our assumption.

Importantly, all the proximity variables remained significant throughout each run, with the exception of coauthored works. In the majority of runs, the effect of this control variable is omitted due to what is known as “separation or quasi-separation”; also called *perfect prediction*, this condition refers to a scenario in which the response variable does not change by different levels of the independent parameters (Long and Freese, 2006). In this case, having co-published at least one paper would ensure the existence of a citation link, which is the expected behavior for direct co-authors. Nevertheless, we observe that pairs with co-authored works have a very low incidence in our data (only 1,798 pairs, that is, 2.26% of the whole sample).

At any rate, there is increasing consensus that GOF statistics need not always be presented for this type of regression analysis (Menard, 2009, p. 58), due to their weaknesses (Allison, 2013).

### **Controlling for co-publishing pairs**

We noticed in the first regression run that we have a small number of *completely determined successes*. Given the condition of perfect prediction displayed by the coauthored works variable during the GOF experiment, we hypothesize that co-authoring academics will always cite each other, hence strongly affecting the model. We inspected closely the data to see if all co-publishing pairs were indeed linked by an effective citation, and found that out of the 1,798 direct co-author pairs, only three did not have a positive citing link.

Since we are measuring the impact of the co-authorship network structure on citing behavior and not collaboration itself, we decide to account for this circumstance to better observe the influence of collaborative proximity on indirectly connected scientists in the network. The number of co-publishing pairs is low enough (2.26%) as to be deemed of no consequence, but we still control for it nonetheless.

Therefore, we run the regressions excluding any pairs representing co-publishing scholars, i.e. links with

a value in the co-authored works control variable greater than 0. Despite the three non-citing pairs with this condition, we only use non-directly connected pairs in the regression (73,949), because there is a strong possibility that even if they did not cite in the period of this study, a citation between them is prone to occur later in time (or perhaps it already occurred in the years prior to our analysis data).

### 3.4.4 Applying Classification

For this stage of our analysis, we first need to identify which classifier is the best fit for our data set. We employ IBM SPSS Modeler version 17.1 and upload refined data, with properly identified data formats and missing values. Note that from this point forward, the term node may include additional connotations besides referring to an author entity in the collaboration network; in the terminology of the SPSS Modeler software, node instead refers to a particular functionality of the software, or a branching position within a decision tree.

#### Validation

Prior to getting started with classification, we decided to implement validation techniques that would improve the predicting ability of models (see Michie et al., 1994; Lichtenwalter et al., 2010). With this in mind, we adopt data partitioning and cross-validation.

First, we include a *Partition* node, which randomly splits the data according to a specified percentage in training and test portions. Typically, about two-thirds of the data is reserved for training the classifier model, while the remaining third is used for testing it. Even though not using the full sample to train the model might result in a slight loss of efficiency, this would not be a major issue when dealing with large data sets (greater than 1000) (Michie et al., 1994).

In our case, we set the percentage to be 60% for training data, which will help to construct the model. The remainder (40%) shall be used as testing partition, allowing us to see whether its classification power is good enough to predict whether a citation occurs between two authors or not, based upon the predictor variables.

Secondly, we employ 10-fold cross-validation, meaning that the training data is first divided into 10 subsets and then 10 models are trained. Each sub-model using a different subset of the data as the test partition to determine how well the model performs on new data. Once all 10 models are built, they are combined in an ensemble by taking their mean accuracy for scoring. Lastly, some models (such as decision trees) have settings for overfitting prevention, which were enabled for our processing.

## Choosing applicable models

SPSS Modeler provides a procedure called *Auto Classifier* node, which executes several algorithms in a single modeling run to find which are best fitted to successfully predict the outcome variable. Due to the binary nature of our target value, only certain models in SPSS Modeler are shown as applicable for this classifying task.

We next discuss the evaluation metrics used to appraise the results of prediction models.

## Evaluating classifiers

There are several measures used to evaluate the performance of a prediction model for binary classification throughout the literature (see Getoor, 2005; Lichtenwalter et al., 2010; Sarigöl et al., 2014; Dong et al., 2015).

First and foremost, the **Confusion Matrix** (also, classification or error matrix) summarizes the relationship between the two sources of information: actual data vs model predictions. The correct predictions, as shown in Table 8, are located on the diagonal.

		Predicted	
		Negative	Positive
Actual	Negative	TN	FP
	Positive	FN	TP

Table 8: Confusion Matrix

Considering a positive prediction as being identified, and a negative one as being rejected, we have:

TP = True Positive = correctly identified

FP = False Positive = incorrectly identified (Type I error)

TN = True Negative = correctly rejected

FN = False Negative = incorrectly rejected (Type II error)

From its results, various metrics can be calculated, in particular:

- **Overall Accuracy (AC)** is the total of correct predictions divided by the total number of cases.
- **Recall** Also known as the true positive rate or sensitivity, recall is the proportion of positive cases that were correctly identified, that is, TP divided by FN + TP.
- **Specificity** Also called the true negative rate, specificity measures the proportion of negatives that are correctly identified as such, that is, TN divided by TN + FP.
- **Precision (P)** is the proportion of the predicted positive cases that were correct, that is, TP divided by TP +FP.

- **F1-score** Also F1-measure, it combines precision and the recall in a single score. It is the harmonic mean of precision and recall:  $F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$
- **Out-of-bag (OOB) error** Term used exclusively for Random Forest, it constitutes the overall misclassification error calculated by each tree formed.
- **ROC area** is another common metric to examine the decision-making ability of classifiers, which relies instead upon the Receiver Operating Characteristic (ROC) curve (also known as area under the ROC curve). A ROC graph is a plot with the FP rate on the X axis and the TP rate on the Y axis. The top left corner is the optimal location on an ROC graph. The further the curve lies above the reference line, the more accurate the test, indicating a high TP rate and a low FP rate.

Particularly, we shall make emphasis on accuracy, precision, recall, and the F1-score, as well as in ROC area, considered as standard practice in Machine Learning to assess classifiers (following Getoor, 2005; Sarigöl et al., 2014).

### Selecting Predictors

Most importantly for our work, we need to identify which independent parameters are the best predictors among the variable set. Depending on the model, feature importance is ranked according to each algorithm's choice (Michie et al., 1994) for determining which features have more weight on the outcome. For instance, the C5.0 classifier uses the maximum information gain ratio (IGR) to select the best attributes (as does Dong et al., 2015). In SPSS Modeler, this ranking can be visualized on the **predictor importance plot** generated by each model.



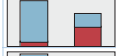

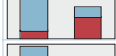











Graph	Model	Build time (sec)	Overall Accuracy (%)	No. Fields Used	Area Under Curve
	 C5 1	27	78.549	6	0.832
	 C&R Tree 1	27	77.438	6	0.804
	 CHAID 1	27	77.261	5	0.831
	 Quest 1	27	76.945	6	0.805
	 Neural Net 1	27	73.541	6	0.821
	 Logistic regression 1	27	72.728	6	0.813
	 Bayesian Network 1	27	69.694	6	0.78
	 Discriminant 1	27	69.289	5	0.754

Figure 20: Best-performing models, as ranked by SPSS Modeler

In the case of a Random Forest classifier, there are two types of variable importance measures:

- **Mean Decrease Accuracy** (MDA), also known as permutation accuracy importance. This rate is determined during the OOB error calculation phase. The more the accuracy of the random forest decreases due to the exclusion (or permutation) of a single variable, the more important that variable is deemed. Put simply, the greater the drop in prediction capability, the more significant the variable. This metric is particularly helpful for variable selection.
- **Mean Decrease Gini** (MDG). The Gini importance describes the improvement in the Gini gain splitting criterion. The Gini index (a measure of node impurity) describes the overall explanatory power of the variables. The mean decrease in Gini coefficient is a measure of how each variable contributes to the homogeneity of the nodes and leaves in the resulting RF. Each time a particular variable is used to split a node, the Gini coefficient for the child nodes is calculated and compared to that of the original node. The changes in Gini are summed for each variable and normalized at the end of the calculation. Variables that result in nodes with higher purity have a higher decrease in Gini coefficient. Simply, the Gini metric gives information about the explanatory relationship between the variables selected. In other words, predictors with high Gini values are more likely to split a branch into “pure” classes.

Finally, we shall use these measurements to determine which variables are the best contributors to the prediction task.

### Classifying with Decision Trees

From the 12 different models we set the program to test our sample data in, Figure 20 depicts the best-performing ones from the ranking of the model comparison. As a preliminary rule to pick models, we consider the overall accuracy, which refers to the percentage of observations correctly predicted by the model vs the total. It is noteworthy to mention that the models were evaluated by their performance on the testing partition.

Importantly, due to the sensitivity to correlation, we ran the models using one set of uncorrelated variables from the ones used with binary regressions: eigenvector as centrality metric, and Euclidean distance for geography measurement. In our case, decision tree algorithms turned out to be the best applicable models, outperforming other popular classifiers, like logistic regression, support vector machines, and linear discriminant analysis. Hence, we analyze the ones that did a better job at the classification problem, namely, CART tree, C5.0, and CHAID. The final modeling structure built is depicted in Figure 21.

Given the traditional concerns perceived in decision tree-based models, we applied, to the best of our knowledge, fine-tuning and validation techniques to prevent any bias and overfitting. Nevertheless, we follow the proposal by Sarigöl et al. (2014) and use a baseline predictor, and then proceed with what the literature considers to be a superior classifier: Random Forest (RF).

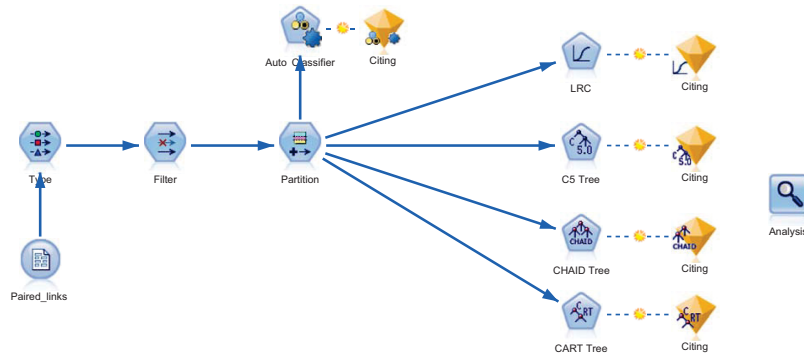


Figure 21: Modeling stream with SPSS Modeler

### Classifying with Random Forest

As previously stated, Random Forest generates an ensemble of trees as their classification mechanism. Since this model is not available by default on SPSS Modeler, we employ the *RandomForest* library (Liaw and Wiener, 2002) in R.

First, we calibrated the model settings, i.e., the number of trees to build the model with, and the number of variables to be selected per tree (*mtry* value). During the setup phase, we discovered that the algorithm has improved prediction performance (better accuracy and reduced OOB error) when using 600 trees with 6 variables per tree. Thus, we tested how the classification power varies between 100 and 600 trees. In addition, we enabled the option to omit missing values from the process, while also controlling for co-publishing dyads as in the regressions.

Given the robustness against correlation within the feature framework (Breiman, 2001; Liaw and Wiener, 2002), we decided to run the model using all the independent variables. Subsequently, we ran the RF with only 6 variables, the same ones used for regression and tree modeling, with the goal of finding out if the model is greatly affected by removing correlated predictors. For the reduced model, we set 3 variables to be used per tree (*mtry* value), also running the algorithm first for 100 and then 600 trees.

## 4.1 Results

In this section, we analyze the findings obtained from the two methods applied to predict the establishment of a citation link.

### 4.1.1 Regression Models

Two different regression models, logistic (logit) and probit, were run with varying factors accounting for the scientific domain, geography, and network centrality, as well as the spatial-social interaction, to check for statistical dependency between citing behavior and our set of explanatory variables. In total, we ran each model in 18 alternative variable combinations: 3 centralities x 3 geographic metrics x 2 interaction effects (with/without).

Variable behavior was highly similar throughout each run; we summarize the results for all 36 iterations (see Table 9) by taking the average coefficients, z-score, and robust standard error for each parameter.

Remarkably, even after controlling for any co-publishing effect, all variables in the model were found to be statistically significant at a 95% confidence level, with associated p-values of less than 0.05. The sole exception to this was the interaction term, whose importance dramatically decreased when the location category was used instead of a continuous metric.

Even if one could argue that significance was reached due to the large number of observations, the high z-score values would indicate otherwise (see Singh, 2005), thus confirming the effectiveness of our predictor

variables. In addition, we see that when actual non-significant parameters are included (as is the case of the interaction term), the model does manifest their low contribution.

### Regression performance

Moreover, our regression modeling was able to correctly classify 76.20% of the cases (average accuracy of all runs), as depicted in Table 10. Note that all predictions were calculated using 50% as a cutoff point for positive predictions. In addition to having an AUC (Area under the ROC curve) of 81.69%, we can say that the proximity factors gave the regression rather good performance.

### Variable contribution

We ranked the predictors with the z-statistic (also z-score or Wald test) as a measure of their association with the response variable (Thompson, 2009); also, because this metric does not depend on the unit of measure of the parameters, as standardized coefficients do in logistic and probit regression models. Note that, since some factors had negative effects (traveling time, Euclidean distance, shortest path, and clustering coefficient), we used the z-score's absolute value.

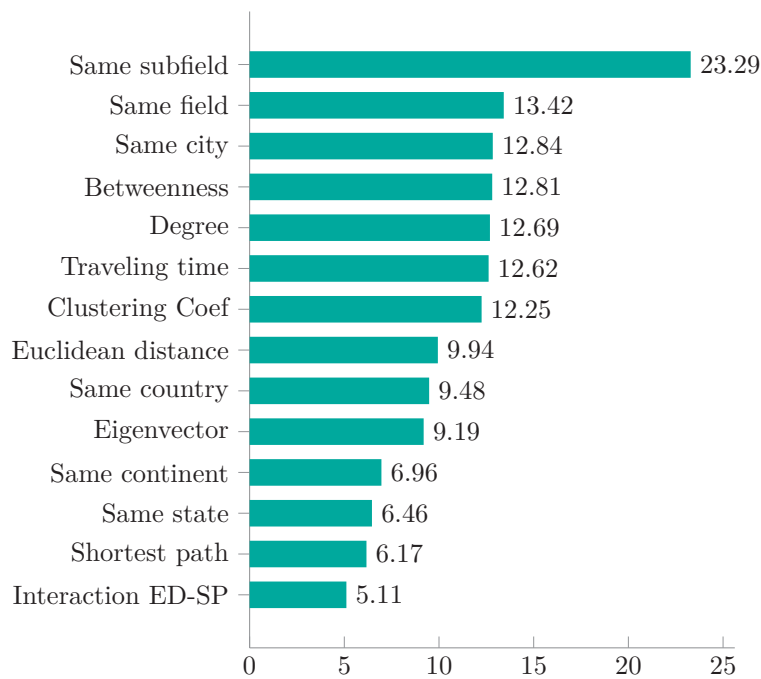


Figure 22: Variable ranking by z-statistic in binary regressions

Overall, cognitive proximity was the most influential factor in the regression, followed by collaboration, and the geographical aspect.



### Summary of 36 combinations of logit and probit regression models

		Regression models											
Variable	Significant?	Logit						Probit					
		No interaction			With interaction			No interaction			With interaction		
		Coef.	z-score	R. Std. Err.	Coef.	z-score	R. Std. Err.	Coef.	z-score	R. Std. Err.	Coef.	z-score	R. Std. Err.
Cognitive category	Yes												
cog1 - same subfield	Yes	2.3324	22.73	0.1026	2.3233	22.79	0.1019	1.3878	23.82	0.0583	1.3841	23.83	0.0581
cog2 - same field	Yes	1.3069	13.39	0.0976	1.3066	13.41	0.0974	0.7855	13.46	0.0584	0.7848	13.44	0.0584
Euclidean distance	Yes	-6.59E-05	-10.17	6.48E-06	-1.85E-04	-9.15	1.94E-05	-3.83E-05	-10.25	3.73E-06	-1.10E-04	-10.17	1.08E-05
Traveling time	Yes	-1.76E-08	-11.34	1.55E-09	-4.95E-08	-13.54	3.66E-09	-1.03E-08	-11.52	8.90E-10	-2.96E-08	-14.09	2.10E-09
Location category	Yes												
loc1 - same city/town	Yes	2.9034	12.81	0.2266	2.9012	11.91	0.2436	1.6841	13.88	0.1213	1.6875	12.77	0.1321
loc2 - same state	Yes	0.9509	6.64	0.1433	0.9483	6.22	0.1523	0.5749	6.62	0.0868	0.5786	6.34	0.0913
loc3 - same country	Yes	0.8702	10.96	0.0794	0.8680	7.97	0.1088	0.5150	10.87	0.0474	0.5182	8.09	0.0640
loc4 - same continent	Yes	0.3786	9.17	0.0413	0.3763	4.78	0.0787	0.2203	9.00	0.0245	0.2235	4.87	0.0458
Degree centrality	Yes	0.0091	12.21	0.0007	0.0093	12.15	0.0008	0.0050	13.21	0.0004	0.0051	13.21	0.0004
Betweenness centrality	Yes	1.37E+04	12.58	1.09E+03	1.38E+04	12.59	1.10E+03	6.42E+03	13.01	4.93E+02	6.48E+03	13.06	4.96E+02
Eigenvector centrality	Yes	2.61E+02	2.05	2.45E+01	2.69E+02	10.59	2.54E+01	1.37E+02	12.06	1.13E+01	1.41E+02	12.05	1.17E+01
Closeness centrality	Discarded												
Clustering coefficient	Yes	-0.9326	-10.12	0.07	-0.9402	-13.25	0.07	-0.5650	-12.80	0.04	-0.5704	-12.83	0.04
Shortest path	Yes	-0.4887	-5.32	0.09	-0.6059	-6.83	0.09	-0.2837	-5.42	0.05	-0.3547	-7.10	0.05
Inter. Euclid-Shortest	Yes*	-	-	-	3.22E-05	4.94	4.19E-06	-	-	-	1.92E-05	5.28	2.36E-06
	<b>Accuracy</b>		76.22			76.21			76.18			76.15	
	<b>Precision</b>		76.34			76.30			76.40			76.35	
	<b>Recall</b>		60.85			60.86			60.60			60.58	
<b>Performance (%)</b>	<b>F1-Measure</b>		67.72			67.71			67.59			67.55	
	<b>AUC</b>		81.70			81.72			81.68			81.70	
	<b>Run time (s)</b>		1.30			1.36			1.27			1.42	

\*non-significant at  $P > |z|$  greater than 0.5, when using location category.

Table 9: Regression models summary

#### COGNITIVE PROXIMITY

We included the scientific domain effect in the regression by using the two dummy variables (cog1 and cog2) actually representing the scenario of two cognitively close scholars. The high z-scores the cognitive parameters displayed (23.29 and 13.43 respectively) are an indication of their high contribution for establishing citation links. In fact, the magnitude of having the same subfield was the greatest factor among the explanatory variables, almost doubling the increased probability from being from the same field at every turn. Clearly, a Canadian author would be decidedly more likely to cite peers within the same cognitive subfield or field rather than outside of it.

#### GEOGRAPHICAL PROXIMITY

We successfully verified that all the representations of geographical proximity were significant throughout the regression runs. Moreover, they had comparable contributions to explaining citing behavior, either as continuous or level-based metrics.

In the categorical form of spatial distance, we observed that being in the same city/town (loc1) is decisive when it comes to establishing a link between CC and REF scholars, closely followed by being in the same country (loc3), and to a lesser degree, being within the same continent (loc4). Remarkably, the co-location term that mattered the least was when authors are in the same state or province (loc2). In any case, all the location levels remained significant even when the centrality metric varied. Altogether, we can say that our scale is a good portrayal of spatial closeness between scientists, since the average z-statistic of all four categories roughly equates the magnitude of Euclidean distance.

Additionally, both Euclidean distance and traveling time presented a negative effect on the outcome, meaning that they increase citation probability by decreasing in value. However, traveling time did seem to perform slightly better than Euclidean distance at every combination, which was a somewhat unexpected finding.

This could be explained by our composition of the traveling time variable: we defined that CC scholars located within 500 Km from a REF author would drive, otherwise electing to fly; and we also gave local flights less extra time (3 hours) than international flights. We already observed that scholar pairs in the same city, i.e. mostly driving for a few minutes, have a high citing tendency (loc1). Yet, for greater distances, the construction of traveling time resulted in the leveling of domestic co-located academics: CC-REF pairs from the same province driving for a long time, have a comparable traveling duration to those in the same country, which would have short flights, or at least, those in nearby regions in Canada. It follows that the traveling duration differences between intra-national distances were smoothed, thus falling into the same country (loc3) category, which as we know, contributed significantly to the model. Plus, since out-of-country

flights were given 5 extra hours in addition to the flying time itself, which caused a bigger duration gap between national vs internationally co-located dyads. Anyhow, the magnitudes of the Euclidean distance and traveling time effects are close enough to assume that these variables can be used interchangeably.

Regarding the behavior of the other independent variables while alternating the geography metrics, the sole difference we observed was that the interaction term became non-significant when using location category, whereas for traveling time it was not affected. This is reasonable, since traveling time acts as a close linear substitute to Euclidean distance, unlike location category, where the main effect of a continuous distance measurement is no longer present.

#### COLLABORATIVE PROXIMITY

First of all, we found significance for all the collaborative metrics included. Of course, closeness centrality is the one exception to this rule since we decided to drop it altogether from the beginning, on account of its high correlation to other parameters like shortest path, besides the other centralities. At any rate, we could have a similar perspective of its behavior by means of applying degree centrality, given that they are closely related by definition.

As with alternating spatial metrics, we noticed that the effects from degree, betweenness, and eigenvector centrality remained similar, with degree having a marginally better presence in the model than the other two. We presume that this may be due to degree being a numeric variable composed only by integers, while the values of betweenness and eigenvector centrality are not only decimals, but also the differences in their distribution are very small-scale, which is also why their standard errors went up despite remaining significant.

Surprisingly, clustering coefficient exhibited a negative effect at every turn. Recall that high clustering coefficient values refer to nodes at well-connected “neighborhoods”; hence, its negative influence could possibly mean that REF authors located in the middle of collaborator cliques are not being as highly cited. Moreover, in the iterations with eigenvector centrality, the contribution of the clustering coefficient augmented to some extent while the eigenvector variable itself reduced its effect, as opposed to their magnitudes in the presence of degree and betweenness centralities (where the z-score generally persisted).

Eigenvector centrality being the metric that better reflects a scholar’s central position in regards to the whole network structure, this behavior is coherent with the variations in clustering coefficient. In light of this, CC authors would have a citing preference towards REF nodes either with a high number of collaborators (degree) or strategically located as proxy nodes between different clusters (betweenness), but not necessarily embedded in the midst of a highly-cliquish collaborator neighborhood.

Anyhow, in the interchanging runs, the centrality metrics’ z-statistic varied among a range of 9-15 (absolute value), thus constituting themselves as primordial factors within the regression, right after the

influence of cognitive proximity.

Furthermore, shortest path reflected a negative contribution, which implies that there is indeed a certain preference for citing REF nodes that are closer to the CC authors by acquaintanceship; however, its effect is not as high. Presumably, it is because we are facing a small-world network, with almost 95% connected authors (giant component), overall clustered at 75% and with a mean of almost 4 hops between dyads. Therefore, the collaborator-chain factor, although important in the model, is not absolutely decisive to establish a citing link.

Finally, the spatial-social interaction effect created by combining Euclidean distance and shortest path was always found significant, apart from when location category was used instead, as explained above. As a matter of fact, the interaction influenced its main variables whenever included, although leaving the other parameters unchanged: it magnified the effect of shortest path by 2-3 points of z-statistic, while reducing the score of Euclidean distance, and most noticeably, traveling time. Thus, we confirm what we suspected about shortest path, in that it sometimes has the capacity to substitute for spatial proximity, with the latter serving as its proxy.

Generally speaking, we confirmed that the structure of the co-authorship network, as measured by various collaborative metrics, is meaningful to academic citation, coming next after cognitive proximity.

At any rate, disregarding what the metric selected to represent centrality and geographic aspects was, cognitive proximity still was the most important factor. Nevertheless, we conclude that all of our explanatory variables (in their main effects) kept their significance throughout all the combination runs, proving their importance towards explaining citing behavior, as well as the robustness of the models.

## 4.1.2 Classification Models

### Classification performance

Table 10 summarizes the performance metrics used to evaluate the resulting classification per model (calculated on the testing partition). We added the regression models to this table, to have a better reference for comparison.

Overall, in terms of precision, recall, and ROC area scores, the performance by all models was promising.

The logistic regression achieved an F1-score of 67.65% and a 76.20% overall accuracy, which is already good, since it means that citing links can be identified more often than not.

Likewise, the decision-tree models had similar performance. Even though they did manage to increase the overall accuracy over the regression model, the improvement is not impressive, with a few percentage points in difference as compared to logistic and probit models.

However, the predictive power was significantly boosted by applying Random Forest algorithms; this classifier clearly outperformed the baseline models (going from an F1-Measure of 71.76% of the best tree

Model	Accuracy	Precision	Recall	F1-Measure	ROC area	Run time (sec)
Logit/Probit	76.20	76.37	60.72	67.65	81.69	1.28
CHAID	77.26	67.58	74.94	71.07	83.1	2
CART	77.44	65.46	76.77	70.67	80.2	4
C5.0	78.55	66.06	78.55	71.76	83.2	8
RF red (100)	83.66	83.02	77.23	80.02	90.42	20.65
RF red (600)	83.67	83.15	77.09	80	90.54	125.55
RF (100)	86.73	85.6	82.56	84.05	93	37.35
RF (600)	86.91	85.95	82.63	84.26	93.19	215.37

Table 10: Classifier performance metrics (in %)

performer to 84.26% in the fully-featured run with 600 trees).

Regarding RF, we performed consistency checks with varying tree settings, and its output contributed to the robustness of our results. Interestingly, even the reduced RF model with 100 trees managed to outperform the best model from the decision trees. This verified the expectation of the better classifying capabilities of this model.

Ultimately, prediction accuracy was reached at the expense of increased complexity (100-600 trees would be extremely hard to inspect) and the processing time, which had a dramatic raise from just a couple of seconds to almost 4 minutes. Finally, improved performance was already accomplished with a reduced RF containing only uncorrelated variables, and though the accuracy slightly decreased, it brings the benefit of having a less complex model. We next take a look at the citing-link attributes used to achieve classification accuracy.

### Ranking predictor importance

As we can see, all tree models seemed to mostly agree on variable ranking, according to their respective predictor importance plot depicted in Figure 23. Contrarily to the regression, the predictor importance plot does not provide a clear indication of a positive or negative behavior. Notably, while generally the cognitive factor is seen by both tree and regression models as the best contributor, the perception about shortest path and spatial distance (Euclidean measurement) varies.

We focused on the importance deduction by the C5.0 model, as our top-performer tree. Its ranking was very similar to the regression one, with the exception that shortest path was given more weight in the classification job than the other SNA metrics.

Interestingly, we took a look at the CHAID model (which has the peculiarity of treating missing values as special cases in its tree-formation mechanism), and discovered that a citing link rarely occurred whenever two scientist were unconnected in the network (refer to Figure 24). Further, we inspected the data closely for this behavior and discovered that positively-linked pairs with missing shortest path predominantly had

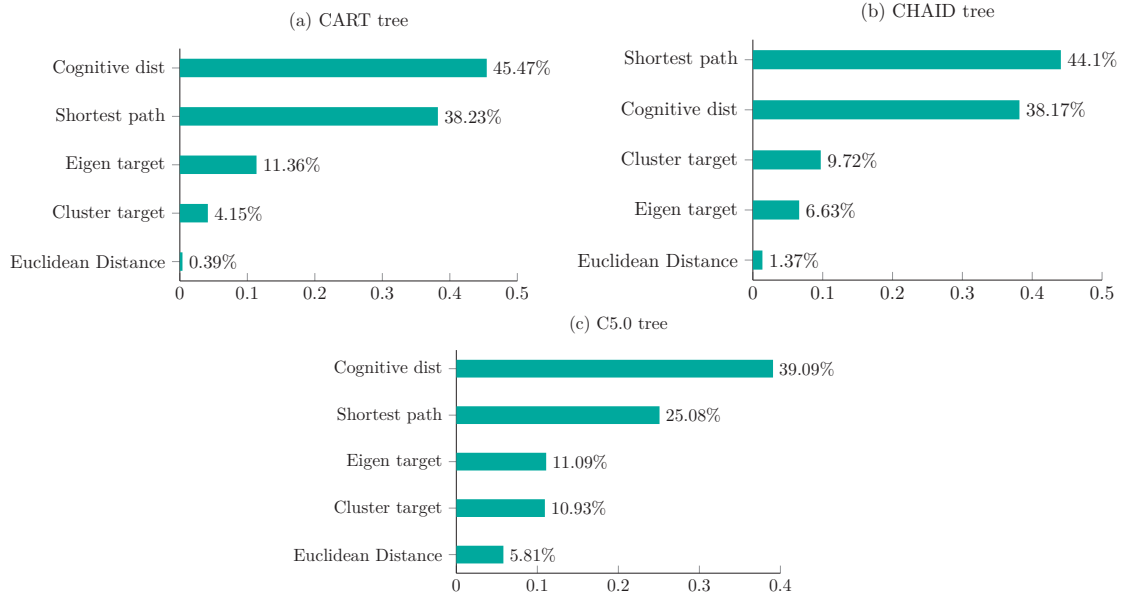


Figure 23: Variable Importance, as ranked per classifier tree

a cited author having a high degree centrality.

Eigenvector, as the metric representing centrality for tree-runs, along with clustering coefficient had comparable behavior, and each of them would have almost the same contribution to citing prediction.

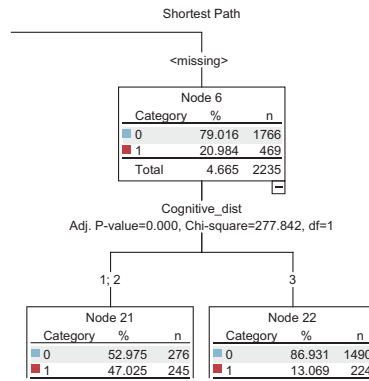


Figure 24: Shortest path node in CHAID model

Furthermore, we first inspected the findings by the Random Forest algorithm in its varying settings by looking at the Mean Decrease Accuracy (MDA). As we can see in Figure 25a, the tree-ranking for the two most influential variables held for RF as well, conveying that the model would lose the most if either cognitive distance (cog1 and cog2) or shortest path are dropped. Moreover, we see that as more trees are grown, the importance of shortest path becomes more evident for the classification.

In addition, the relevance of each centrality measure (along with the control variable of published works)

has nearly the same weight; this was expected, considering their existing correlation, and the observed behavior during the regressions. Additionally, eigenvector and degree centrality contributed slightly better to the model than betweenness, while traveling time marginally outperformed Euclidean distance, same as in the regression.

Contrarily, location category was given the lowest rating in RF; however, we presume that it might be due to the spatial levels losing importance whenever a continuous metric of geography is present. Besides, each level only accounts for a portion of the citing behavior, and correspondingly to the regressions, it is the average effect of all location categories which matches the presence of a continuous metric.

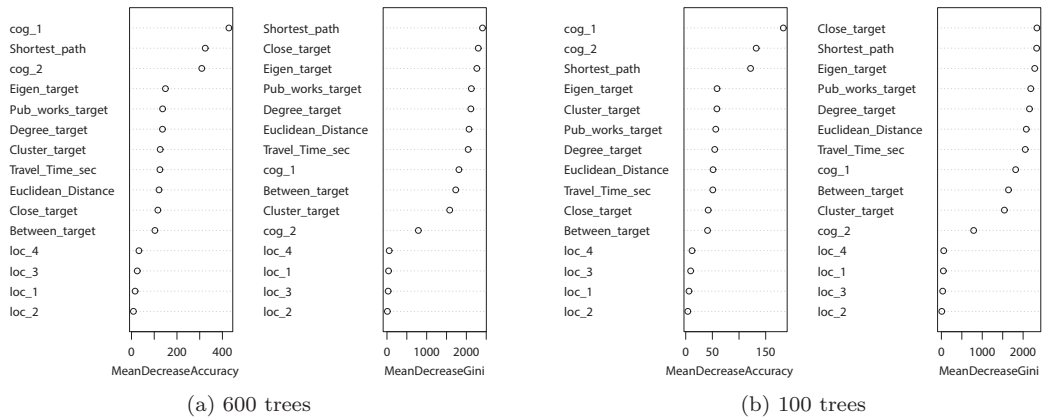


Figure 25: Variable importance according to Random Forest

In the reduced model with only 6 uncorrelated variables (see Figure 26), we can observe that their importance ranking does not seem to vary much when compared to a fully-featured model. Moreover, the sorting of predictors is pretty much identical, regardless of the number of trees generated.

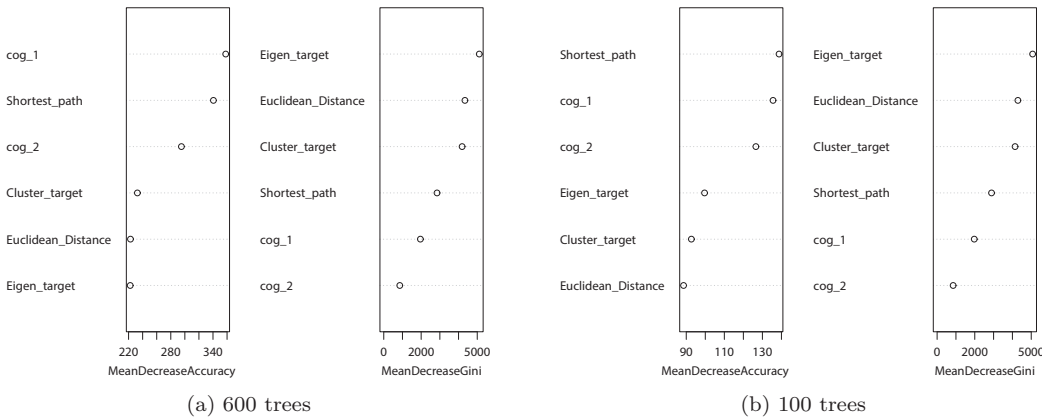


Figure 26: Variable importance according to Random Forest, reduced

With this in mind, we next inspected variable sorting with Mean Decrease Gini (MDG) categorization. Following the reduced RF, it would seem that once inside a split, the power to reach purity in classification (i.e. to know for sure whether a citation will occur or not) relies majorly on eigenvector centrality, followed by geographical distance and clustering coefficient.

Overall, the importance categorization from RF would mostly confirm the selection of independent variables for our problem and their contribution to the citing behavior of CC authors.

### 4.1.3 Testing the hypotheses

At last, we are able to assess our null hypotheses concerning the impact of proximity aspects on the probability of scholarly citation.

First of all, we tested  $H_{01}$ : *Citation probability does not increase when a Canadian author is cognitively proximate to another author.* This parameter, as represented by its two levels of same subfield (cog1) and same field (2cog), was found statistically significant in the logit and probit regression models. Moreover, it was proven to be the main contributor to the citation effect by both statistical and machine learning approaches; CC scholars will decisively cite REF authors publishing in the same subfield as themselves, and will also have a preference towards those from their own field, as opposed to academics publishing outside of it. Consequently, we reject  $H_{01}$  and instead, we assert that cognitive proximity is in fact an influential factor for citing.

Likewise, we had evidence to reject hypothesis  $H_{02}$ : *Citation probability does not increase when a Canadian author is geographically proximate to another author,* on the grounds of the significance of all the metrics representing spatial proximity. Not only, were Euclidean distance and traveling time important to increase citing likelihood, both as alternate stand-alone factors as well as a combined effect with dyad collaborative distance, but they also proved to have predicting capability. Importantly, authors co-located within the same country (loc3) and city (loc1) are very likely to be linked by citation.

Next, we examined the hypotheses concerning collaborative proximity and an academic's position within the co-authorship network. With null hypothesis  $H_{03a}$ : *The probability of a Canadian author citing another does not increase due to the referenced author's position within the collaboration network,* we attempted to verify whether or not the citation probability increased the more central a cited author was found to be.

Essentially, clustering coefficient along with the various centrality metrics (applying them one at a time) turned out to be significant, which gave us sufficient evidence to reject  $H_{03a}$ , having also displayed prediction power for the establishment of a citation link. This finding would imply that Canadian scientists are more likely to cite a REF author that has a good position within the collaboration network, in terms of: having co-published papers with a high number of people (degree), whether he is more central in the network as a whole (eigenvector), if the scholar acts as a proxy for clusters within the network (betweenness), and if the



author is situated outside of highly-cliquish neighborhoods (clustering coefficient).

Lastly, due to the significance accounted to shortest path, we rejected  $H_{03b}$ : *The probability of a Canadian author citing another does not increase if they are closely connected to each other within the collaboration network*, in preference of the alternate statement: The closer two scholars are located with respect to each other within the co-authorship network (through the chain of co-authoring peers), the higher the probability is for a positive citing link between them.

Furthermore, shortest path would interact with Euclidean distance to add up to the explanatory power of the model, while also using the latter as proxy to establish such acquaintanceship connections, and thus substitute for spatial closeness. Having collaborative closeness would thus greatly serve to anticipate a citation link between authors, with shortest path being among the best-performer variables in the citing prediction.

In sum, we determined that, regardless of the method used, the dependent variable of citation probability is successfully predicted by the chosen explanatory parameters deriving from each proximity aspect we decided to inspect, hence proving our hypotheses about the impact of proximity on citing behavior.

## 4.2 Discussion of Results

In this section we present the implications of our findings, with respect to our research questions and aforementioned works in the literature. We now present them in the order of their discovered relevance towards citation probability.

### Importance of cognitive proximity

We asked ourselves if having two scholars in the same scientific domain would increment the probability for there to be a citation between them. From our findings, it is a clear consensus that coming from close cognitive bases makes authors more prone to be linked by citation, and thus we gave a positive answer to our question. In fact, with the first place ranking by both regression and classification, it is likely the factor that influences the most at the time of choosing who to cite from.

Moreover, our results align with those by Ding (2011), who indicated that productive authors prefer peers sharing their same research field to both cite and coauthor with, by analyzing academic citations in the field of information retrieval. In turn, we confirmed that despite nanotechnology commonly being depicted as a highly multidisciplinary field, cognitive similarity still matters. In fact, Schummer (2004) had previously stated that nanotechnology displays a merely average multidisciplinary level in its scientific citation patterns, not differing much from other scientific bases.

We also compare the impact of the cognitive proximity category to findings from innovation literature,

where Hu and Jaffe (2003) similarly indicated that technological proximity is important for knowledge flows, as evidenced by increasing patent counts. Likewise, Jaffe and Trajtenberg (1999) had claimed that technologically proximate inventors are preferred for citing, with Cunningham and Werker (2012) affirming the same behavior for patenting firms.

In sum, the fact that cognitive interests are important for citing behavior, be it academic or innovative, remains unchallenged. More specifically, nanoscience citation appears to follow the classic tendency regarding information exchange that collaboration partners also display (Boschma and Frenken, 2010): knowledge base weighs heavily on peer-referencing selection .

However, previous research indicates that interdisciplinarity is primordial in the development of nanotechnology (Malsh, 1997) for contributing to the academic society in terms of integrating knowledge from assorted domains. Also, since articles with an interdisciplinary background are allegedly more successful due to having increased value for the scholarly community (Katz and Martin, 1997), interdisciplinarity is an important factor for knowledge production. Besides, high levels of it in research are closely related to innovation (Weingart, 2000; Rafols et al., 2010), and, on a practical note, it has been brought up that programs funding nanoscale research usually take interdisciplinary approaches (Schummer, 2004).

In short, Canadian nano-scientists would have to balance the need of being cognitively proximate, and try to aim for the “idealistic nano-visions” (Rafols and Meyer, 2007) of producing more interdisciplinary knowledge, in light of its weighty benefits for knowledge production and innovation. This could be achieved by seeking information from complementary cognitive domains, thus avoiding the potential issues coming from cognitive lock-in Boschma (2005) warns about.

### **Importance of collaborative proximity**

For our next research question, we hypothesized about the effect of collaboration links between authors on the chance of being cited, for which we employed additional metrics to the ones usually adopted in inductive network analysis.

Typically, authors in the literature find it sufficient to focus on the effects of just one or two variables, like degree centrality (e.g. Wang and Guan, 2011) and/or shortest path. Instead, we preferred to examine the impact of this proximity on citations by constructing a comprehensive framework which accounted for various collaborative metrics.

While the literature seems to agree with our results in that having a high degree centrality is favorable for citation purposes (either academic or innovative) (e.g. Wallace et al., 2012; Eslami et al., 2013), there are wide discrepancies regarding the other parameters.

In this regard, Liu et al. (2014) argued that while degree is the best-performing centrality metric, betweenness centrality only evidenced a mild effect on scholarly citation. Similarly, further analyses suggested

that solely by combining academic and innovation networks of collaboration (i.e. when scientific research translates into practical applications), is the power of betweenness centrality to control knowledge flows revealed (Eslami et al., 2013).

Furthermore, in the study by Abbasi et al. (2011), both betweenness and closeness centralities did not even reach statistical significance and were thus discarded. As for eigenvector centrality, whereas Liu et al. (2014) deemed it as non-significant, Abbasi et al. (2011) claimed that it had a negative effect on scholarly citation.

Contrarily, our results prove otherwise, with the exception of closeness centrality, which was dropped from our statistical modeling. In this respect, although not evidenced by regression, the performance of closeness centrality was granted a comparable importance to all other centrality metrics in the prediction RF models.

As a matter of fact, the positional features of authors in the co-authorship network, as characterized by the three centrality parameters of degree, betweenness, and eigenvector, were not only fairly equal (with eigenvector slightly less so) to influence the probability of citation, they also had similar predicting capabilities. Indeed, a Nanotechnology research paper is more likely to be cited if its authors are better located in the collaboration network.

Moreover, the behavior exhibited by the clustering coefficient of individual scholars, a factor that has been largely disregarded in previous works, contradicted the claim that being located in highly-cliquish neighborhoods enhances the possibility to receive citations from peers Eslami et al. (2013).

Additionally, in spite of the significance displayed by betweenness centrality in his study, Breschi and Lissoni (2003) argued that shortest path was a more critical factor for resulting in citations, at least for patent literature. This conclusion would go more in hand with the behavior of shortest path in our analysis of citing prediction by using machine learning algorithms, where geodesic distance outperformed centrality metrics in the classification task. Yet, this is not a regular centrality metric, but rather a precise value for every pair of scientists, for which reason we formulated a separate hypothesis to test this aspect.

At any rate, although the impact of shortest path was not the highest among the factors in the regression, our results from the statistical analysis displayed strong support for the general agreement that citing likelihood decreases as shortest path increases (Singh, 2005; Sorenson et al., 2006). This tendency revealed in our analysis contrasts with Wallace et al. (2012), who claimed that apart from direct self-citations, they did not find a strong preference to cite authors close in the co-authorship network, referring to number of hops in the shortest path.

In addition, we must mention that we found only one case attempting to examine the impact of proximity on citation by means of prediction models. Sarigöl et al. (2014) were able to predict citation success only with SNA metrics, achieving what they considered to be high precision (60% accuracy). We suspect that

our prediction accuracy has been much improved (with 76.5% accuracy at minimum to 86.91% maximum) by the other proximities, particularly cognitive distance, which turned out to be the most influential one.

Nevertheless, we partially agree with their findings, in that they determined the effect of most centrality measures by themselves to be almost negligible, having in turn to employ combined effects. Moreover, we presume that the differences may be because, in addition to their chosen statistical methodology, their focus is predicting citation success rather than the establishment of a single citing link, which, although related, is not quite the same.

In our case, the fact that the Random Forest algorithm gave notable weight to the SNA metrics, along with their statistical significance in the regressions, gives us evidence to claim collaboration is decisive when it comes to pick academic references. Particularly, the effects of the centrality metrics along with clustering coefficient show the influence of how central authors (taking the whole co-authorship network into consideration) are preferred for citing.

Ultimately, we affirm that collaborative proximity, as measured by co-authorship, is relevant to the practice of citation, revealing the impact of social structure on the diffusion of scientific knowledge.

### **Importance of geographical proximity**

Finally, we examine the most controversial factor from our research questions: geographical proximity.

It has been argued in the literature (Bouba-Olga and Ferru, 2012) that geography only matters when innovation, a knowledge-intensive economic activity, is involved. Bouba-Olga and Ferru (2012) proceeded to support this claim by means of analyzing the science-industry collaboration network of patent literature, being the only one that ever attempted to account for the traveling aspect of the geographical proximity definition in Boschma's (2005) proximity framework, besides spatial levels.

In our work, we refute this claim, since each of our geographical metrics were found significant by the regressions and deemed important by the classifiers, manifesting a valid statistical dependence between spatial considerations and citation probability. Thus, our work constitutes, to our knowledge, the first one that addresses scholarly citation without any involvement from innovation literature or sources.

Moreover, we remark on the uniqueness of having considered the impact of several measures of geographical proximities on citation (as opposed to one or two, which is the common case in the literature). We also highlight the novelty of our construction of the traveling time variable, which contemplated two different means of transportation according to distance conditions, and which required a high-level of precision in its measurement. Fortunately, we could achieve this precision by means of the developed methods which allowed us to benefit from Google Maps and its geographic services, also allowing us to gather the actual affiliation location of the authors, with the corresponding correct location categories.

Anyhow, our findings do seem to agree with conclusions from patent analyses, which assessed different

scales of location category and their influence to establish a citing link (Breschi and Lissoni, 2003; Singh, 2005), in that geography is not a sufficient condition for knowledge diffusion, usually requiring an interaction from other aspects, like collaborative or cognitive proximity, as well. Other studies manifest the same for collaboration intensity as represented by patent citation count and frequency, although only one level (country) was considered (Jaffe and Trajtenberg, 1999; Hu and Jaffe, 2003).

Furthermore, we discovered that from various location categories, being co-located in the same country and city is more important for citing, which was a similar case (only country was proven) for patents from the US Sonn and Storper (2008).

Overall, all three of our spatial metrics revealed that they are not inconsequential for studying citing behavior in a scholarly domain. However, given their perceived importance per different models, along with the valid interaction discovered with shortest path, we take a conservative approach towards the conclusions on its effect.

Essentially, it would be the combination between distance and other aspects that have a greater impact on citation probability. Accordingly, we align with empirical studies presented on innovation works, as well as with researchers in the conceptual proximity literature (such as Feldman, 2002; Morgan, 2004; Sonn and Storper, 2008) that remark on the importance of geographical distance as complementary for other factors.

To support our finding about geographical proximity, we take a look at the citation network, as represented by the precise locations<sup>1</sup> of positively linked scientists in Figure 27. We employed SNA attributes as filters to better adjust the visualization, in our case, the greater the node's in-degree (meaning an author being highly cited by others), the bigger the node shape is, and the darker its color.

We can see a high density of incoming citations in North America, all over Europe, and in Asia near China, while citations are sparser in other places. Thus, from the perspective of Canadian scientists, although spatial co-location may be important, it would not make much difference of where the cited author is located, as compared to the richness of source in terms of scientific field. Lastly, despite Canada's place among leading countries in nanotechnology, it would still be surpassed by conglomerate areas of knowledge with a major concentration in the field, located in the US, Europe, and Asia.

Moreover, cognitive proximity would overall be more influential than geography, at least for Canadian scientists dealing with nanoscience research. Thus, spatial distance would act more like an underlying framework for scientific collaboration and the expansion of knowledge, as it has the potential to boost the development of other, more influential, proximities.

Furthermore, the spatial impact would be better evidenced in the interaction between scholarly productivity and innovation (Gittelman, 2007; Almazan et al., 2007; Sonn and Storper, 2008), rather than in the mechanics of academic knowledge alone. Nonetheless, it is important to keep in mind that, however high its

---

<sup>1</sup>We used the GeoLayout plugin with Mercator projection in Gephi. The background map was manually added for approximate visualization purposes.

interplay with other attributes may be, geographical distance, still remained a critical actor in the evolution of science.



Figure 27: Geo-layout of nodes with positive citing link

Last but not least, many studies have been conducted with positive citation links, usually, in terms of citation counts, whereas we chose to use both negative and positive linkages, and more importantly, because having defined our problem as a classification task, we were able to employ a prediction approach as well.

Thus, we remark on the use of machine learning models to study the behavior of the proximity factors and discovering how strong a predictor they can be for citation probability. Research combining citation analysis and machine learning usually comes from other knowledge bases, such as physics (Lichtenwalter et al., 2010) and computer science (Bethard and Jurafsky, 2010; Dong et al., 2015; Getoor, 2005). Notably, we did not find any of them aimed at academic papers on nanotechnology nor addressing any combined proximity concerns. For example, in the study by Ibáñez et al. (2009), tokens from abstract and keywords were used to predict future citation patterns in terms of current citations.

In the reviewed literature, the closest examples we could find was citing prediction (in the field of computer science) on the basis of collaboration network through supervised learning (Sarigöl et al., 2014) and unsupervised learning (Jawed et al., 2015). Likewise, Dong et al. (2015) also makes use of SNA metrics, by anticipating scientific impact in terms of the author's degree within the co-authorship network.

Generally, all similar works in the domain of nanotechnology scholarly data have been conducted with statistical-only approaches (e.g. Cunningham and Werker, 2012; Liu et al., 2014; Singh, 2005), or with

distribution models comparison (e.g. Onel et al., 2011). And so, we supplement to the know-how of scholarly citation analysis by using both approaches in our investigation: statistical approaches and machine learning.

Admittedly, the most difficult task was to determine precisely which among these contributors is the most important. Nevertheless, our goal was to identify if the factors have influence or not by taking a three-fold proximity perspective, rather than exacting which among them is the most important.

Finally, it is satisfactory to become, to our knowledge, the first analysis which examines the impact of three proximities: cognitive, geographical, and collaborative, on citations at the same time. Plus, we were able to confirm, by means of statistical evidence along with machine learning, that certain hypotheses and patterns of proximity influence from innovative citation also hold for academic citation, a feat which had not been previously undertaken.

### 4.3 Limitations and Assumptions

Throughout our research we were exposed to some limitations, the first of which is related to the sample selection. We had to narrow down our analysis to a few years, and restrict the selected authors within the period to come up with a manageable sample size. Had we taken the entire population for our working time range, our sample would have been constituted by millions of pairs due to all the possible author combinations.

Clearly, although we would have liked to extend the period, we did not have the computational resources required to handle the task, such as the gathering of collaboration metrics. In particular, given the dependency on third-party applications for the geographical measurements, working with a greater data set would have been exceedingly time-consuming.

However, we question whether a bigger sample would have been altogether beneficial given that the case still would be that of a highly imbalanced data set. And, as we had established, the chosen methodology does not hold well in such scenario. Moreover, we were able to overcome this potential issue by means of the sampling technique implemented.

Regarding our statistical analysis, we based our statistical inferences on the applied regression model in spite of the poor performance on the Goodness-of-Fit tests. We presume that this limitation is due to the large number of observations included in our sample. In this respect, we expect that the experiment performed apropos helps elucidate this point, and provides good evidence for the validity of our conclusions.

Concerning the collaboration aspect, our research is circumscribed within a 5-year window, implying the very real possibility that two authors may have co-authored a paper outside this time range. As consequence, in our network they appear as disconnected, affecting the SNA metrics and shortest path, likely leading to isolated clusters. Surely, extending the collaboration network by more years would derive in more relationships between the authors and a closer-node network. Anyway, the big majority of authors were found to be

part of the giant component of the network, so this limitation did not turn out to be highly detrimental to our work.

In addition, there is a chance that successful scholars (in terms of published articles) may have stopped publishing in year one of our focus period, due to passing away or simply having retired, just to name a few possible reasons. In such respect, the high degree that would account for receiving citations from Canadian researchers would be missing, whereas if we considered previous years, such author would be more central. Be that as it may, our insights about the collaboration factor are based on a glance “as-is”, which certainly could be improved upon by using time-windows and/or analyzing more years.

Another limitation was the lack of address information in the database for some scientists. This resulted in the weeding out of a portion of our observations, because otherwise not every case would have had the geographic attributes required for our analysis. We suspect that the referenced scientific databases do not include affiliation details for all the possible co-authors in an article. Hence, although the proportion of discarded pairs was low, we advise caution to this potential bias in the interpretation of our results related to spatial distance.

For cognitive proximity, we encountered the constraint of counting with field and subfield information pertaining to the paper rather than specifically to the author. Neither did we have specifics about the scholar’s scientific field of interest, nor the department within their respective institution (which could have provided an inkling to their scientific specialty).

Thus, we made the assumption that the fields and subfields of an article apply to its writers. Regrettably, this would provide a somewhat limited and not entirely correct view on cognitive impact, particularly for papers resulting from a highly multidisciplinary collaboration team. Nevertheless, we consider it is still a good indication (and the results support our reasoning) for knowing the field of interest the author is publishing in, which could bring them to cite published works coming from similar fields. Besides, it has been argued in the literature that the discipline affiliation of coauthors corresponds to the discipline of their knowledge contribution (Rafols and Meyer, 2007; Schummer, 2004, p. 438),

Additionally, we considered papers with at least one Canadian author, without accounting for the fact that although one of the authors may have a Canadian address, the research may have been majorly conducted elsewhere. This restriction would be better examined through co-authoring pairs rather than citation dyads.

Finally, we followed the literature and restricted our study to the domain of nanotechnology. As consequence, any collateral effects due to different citation patterns from different disciplines were dismissed, although at the expense of generalization. However, in this limitation, we also find an opportunity for future research.



## 5.1 Conclusions

This thesis explores the network of nanotechnology scientists in Canada and examines the relationship with the scholars they reference in their publications. We posed the questions of whether they are more likely to cite authors who are close to them in terms of: (a) cognitive proximity, (b) geographical proximity, and (c) collaborative proximity.

Our study concerned a quantitative research that implemented mixed methods to establish the validity of our claims, and so, we validated our findings by means of a traditional statistical approach as well as through the more contemporary methods of machine learning. Consequently, we found as answer to research question (a) that Canadian authors do show a statistical strong tendency to reference within their own knowledge base. Moreover, this seems to be the distinct most influential factor among those inspected.

For research question (b) our results established that there is a significant association between the spatial separation between scientists and whether they choose to cite or not. Yet, given the revealed interactions with other factors arising from the results, we take a moderate stance in the affirmation that geography can single-handedly cause citations.

Finally, regarding research question (c) according to the results, the analyzed co-authorship network proved that the collaborative aspect in regards to increasing citation probability is significant. In addition, an author's chances of getting cited would increase either by having a central position within the network, or by being connected through collaborating peers.

Our findings related to geographical proximity support previous works coming from innovative literature that shows spatial separation acts in hand with the other elements as a framework for knowledge diffusion. Moreover, we took a critical approach at what the implications of nanoscale researchers having a marked preference for citing within their own cognitive domain could be for their research and its contribution to science and technology. Furthermore, our findings in the collaboration aspect challenged the idea that social centrality metrics have no predicting power by themselves in regards to citation, proving otherwise. Besides, our results agreed with other authors in the canon about the significance of being well-connected in the network.

Our work contributes to the literature by being a relational way of analyzing distinct effects on knowledge creation, as measured by citation. Specifically, it augments on the understanding of nanotechnology information flow, given the few studies targeting this domain from a purely academic standpoint. Also, it adds to the scholarly community know-how by inspecting the predicting power of proximity factors and structural properties of the collaboration network on scientific knowledge output.

Furthermore, these findings would constitute an attention call for the government and policy-makers, to find ways to stimulate interdisciplinarity on nanoresearch. While existing policies may be beneficial towards encouraging the international collaboration of scientific community, our results could be of interest to determine that much can be done to improve other aspects, such as the already important social impact.

In conclusion, either by regression or classification, we can explain and even predict the citation behavior of nanoscience researchers, by inspecting the factors of cognitive specialization, geography, and collaboration. We are thus satisfied after having been able to determine through a diversified methodology that these attributes indeed contribute greatly to knowledge diffusion in Canadian nanotechnology.

## 5.2 Future Work

As a result of our study, further research might well be conducted on the suggested lines of investigation:

**Replicate the analysis for innovative citation.** It would be interesting to examine whether our model holds for Canadian innovators as well and if the same factors have incidence on patent creation, and compare the results to see if any difference is found between both knowledge-generating sources. Another research avenue would be to combine patent and academic literature on nanoscience, to obtain better understanding of the effects of networks and proximities on innovation and the development of technological advances.

**Expand the proximity framework** to investigate more proximity types, such as organizational or social. We wonder, for example, about the importance of cultural (such as ethnicity, see Agrawal et al. (2008)) and language factors for forming scientific collaboration ties. Future work could inspect their interactions with our studied parameters, though some of them may be harder to quantify.

Moreover, composite metrics could be developed to measure collaboration proximity in terms of SNA centralities, such as forming combinations among them, or with values from a directed citation network. In addition, the concept of cognitive proximity could be expanded to include specialization (topic clusters), given the wide variety of applications nanotechnology has. We could thus find by how much our model can improve its classification power for predicting citations by including these or other aspects. Besides, the same aspects of cognitive domain and geography could be inspected within the collaboration network, to see if the same proximities influence whom scientists choose to coauthor with.

**Implement time-windows** , and widen the temporal coverage of the collaboration network. By adding a time-based frame for anticipating citation links, we might identify trends and expose other complex behaviors in citing practices. Likewise, associating the social networks of collaboration and citation within time-frames might reveal practices that increase citation probability, such as authors preferring to cite authors they have cited before.

**Compare to other scientific databases** to validate if our conclusions remain for other scholarly collections. Although Scopus has a wide coverage from journals across the globe, this would eliminate any potential bias caused by choosing literature published only in the English language.

**Examine the reverse citation effect.** Our findings reveal who it is that Canadian scientists are citing, based on their proximity and attributes of the reference author. But, it would also be interesting to study the opposite effect, that is, who among the academic nanotechnology community is citing Canadian scholars and why.

**Examine other fields** to see if our models are applicable for bodies of knowledge from other disciplines. We could learn if they adhere to the observed patterns or if they exhibit different tendencies on the ranking of variable importance and their interactions. Finding affinities in their behavior would enable us to make generalized conclusions about scientific citation.

Additionally, it would provide valuable insights to determine how similar or dissimilar diverse areas of science can be from one another. For instance, which fields rely more heavily on geographical proximity, or which academic community has a greater dependency on their social peers.

Along these lines, we set this thesis as groundwork opening up an ample research agenda that would contribute to better understanding of the underlying mechanisms in the academic community of nanotechnology, and potentially in others as well.

---

## Bibliography

---

- Abbasi, A., Altmann, J., and Hwang, J. Evaluating scholars based on their academic collaboration activities: two indices, the rc-index and the cc-index, for quantifying collaboration activities of researchers and scientific communities. *Scientometrics*, 83(1):1–13, 2010.
- Abbasi, A., Altmann, J., and Hossain, L. Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures. *Journal of Informetrics*, 5(4):594–607, 2011.
- Acuna, D. E., Allesina, S., and Kording, K. P. Future impact: Predicting scientific success. *Nature*, 489(7415):201–202, 2012.
- Adams, J. D. and Jaffe, A. B. Bounding the effects of R&D: An investigation using matched establishment-firm data. Technical report, National Bureau of Economic Research, 1996.
- Agrawal, A., Cockburn, I., and McHale, J. Gone but not forgotten: knowledge flows, labor mobility, and enduring social relationships. *Journal of Economic Geography*, 6(5):571–591, 2006.
- Agrawal, A., Kapur, D., and McHale, J. How do spatial and social proximity influence knowledge flows? evidence from patent data. *Journal of Urban Economics*, 64(2):258–269, 2008.
- Alcacer, J. and Gittelman, M. Patent citations as a measure of knowledge flows: The influence of examiner citations. *The Review of Economics and Statistics*, 88(4):774–779, 2006.
- Allen, T. J. et al. Managing the flow of technology: Technology transfer and the dissemination of technological information with the R&D organization. *Cambridge (US)*, 1977.

- Allison, P. Why i don't trust the hosmer-lemeshow test for logistic regression. <http://statisticalhorizons.com/hosmer-lemeshow>, 2013. Accessed: 2015-12-15.
- Almazan, A., De Motta, A., and Titman, S. Firm location and the creation and utilization of human capital. *The Review of Economic Studies*, 74(4):1305–1327, 2007.
- Alpaydin, E. *Introduction to machine learning*. The MIT Press, 2014.
- Amin, A. and Cohendet, P. Learning and adaptation in decentralised business networks. *Environment and Planning D: Society and Space*, 17:87–104, 1999.
- Amunategui, M. Smote - supersampling rare events in r. <http://amunategui.github.io/smote/>, 2014. Accessed: 2015-11-06.
- Antonelli, C. *The economics of localized technological change and industrial dynamics*, volume 3. Springer Science & Business Media, 1995.
- Apache Friends. MySQL database, 2015. URL <https://www.apachefriends.org/>.
- Asheim, B. and Gertler, M. The geography of innovation. *The Oxford handbook of innovation*, pages 291–317, 2005.
- Baber, Z. Sociology of scientific knowledge. *Theory and Society*, 21(1):105–119, 1992.
- Baldi, S. Normative versus social constructivist processes in the allocation of citations: A network-analytic model. *American Sociological Review*, pages 829–846, 1998.
- Barabási, A.-L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., and Vicsek, T. Evolution of the social network of scientific collaborations. *arXiv preprint cond-mat/0104162*, 2001.
- Bastian, M., Heymann, S., Jacomy, M., et al. Gephi: An open source software for exploring and manipulating networks. *ICWSM*, 8:361–362, 2009. URL <http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154>.
- Batagelj, V. and Mrvar, A. Pajek-program for large network analysis. *Connections*, 21(2):47–57, 1998.
- Beaudry, C. and Schiffauerova, A. Impacts of collaboration and network indicators on patent quality: The case of canadian nanotechnology innovation. *European Management Journal*, 29(5):362–376, 2011.
- Berger, P. L. and Luckmann, T. *The social construction of reality: A treatise in the sociology of knowledge*. Penguin UK, 1991.
- Berkenkotter, C. and Huckin, T. N. *Genre knowledge in disciplinary communication: Cognition/culture/power*. Lawrence Erlbaum Associates, Inc, 1995.

- Bethard, S. and Jurafsky, D. Who should i cite: learning literature search models from citation behavior. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 609–618. ACM, 2010.
- Blanc, H. and Sierra, C. The internationalisation of R&D by multinationals: a trade-off between external and internal proximity. *Cambridge Journal of Economics*, 23(2):187–206, 1999.
- Bonacich, P. Some unique properties of eigenvector centrality. *Social Networks*, 29(4):555–564, 2007.
- Bordons, M., Morillo, F., and Gómez, I. Analysis of cross-disciplinary research through bibliometric tools. In *Handbook of quantitative science and technology research*, pages 437–456. Springer, 2005.
- Borgatti, S. P. The state of organizational social network research today. *Department of Organization Studies, Boston University, manuscript*, 2003.
- Bornmann, L. and Daniel, H.-D. What do citation counts measure? a review of studies on citing behavior. *Journal of Documentation*, 64(1):45–80, 2008.
- Boschma, R. Proximity and innovation: a critical assessment. *Regional studies*, 39(1):61–74, 2005.
- Boschma, R. and Frenken, K. The spatial evolution of innovation networks. a proximity perspective. *The handbook of evolutionary economic geography*, pages 120–135, 2010.
- Boschma, R. A. and Lambooy, J. G. Evolutionary economics and economic geography. *Journal of evolutionary economics*, 9(4):411–429, 1999.
- Bouba-Olga, O. and Ferru, M. Does geographical proximity still matter? *HAL FR*, 2012.
- Boufaden, N. and Plunket, A. Proximity and innovation: Do biotechnology firms located in the paris region benefit from localized technological externalities? *Annales d’Economie et de Statistique*, pages 197–220, 2007.
- Breiman, L. Random forests. *Machine learning*, 45(1):5–32, 2001.
- Breiman, L. Statistical modeling: The two cultures. *Quality control and applied statistics*, 48(1):81–82, 2003.
- Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. *Classification and regression trees*. CRC press, 1984.
- Breschi, S. and Lissoni, F. Knowledge spillovers and local innovation systems: a critical survey. *Industrial and corporate change*, 10(4):975–1005, 2001.
- Breschi, S. and Lissoni, F. Mobility and social networks: Localised knowledge spillovers revisited. *Milan: University Bocconi, CESPRI Working Paper*, 142, 2003.

- Broekel, T. and Meder, A. The bright and dark side of cooperation for regional innovation performance. Technical report, Jena economic research papers, 2008.
- Brooks, T. A. Private acts and public objects: an investigation of citer motivations. *Journal of the American Society for Information Science*, 36(4):223–229, 1985.
- Bunnell, T. G. and Coe, N. M. Spaces and scales of innovation. *Progress in Human geography*, 25(4):569–589, 2001.
- Butts, C. T. Social network analysis: A methodological introduction. *Asian Journal of Social Psychology*, 11(1):13–41, 2008.
- Cairncross, F. *The death of distance: How the communications revolution is changing our lives*. Harvard Business Press, 2001.
- Callon, M., Law, J., and Rip, A. Mapping the dynamics of science and technology. *Book*, 1986.
- Cano, V. Citation behavior: Classification, utility, and location. *Journal of the American Society for Information Science*, 40(4):284, 1989.
- Cantner, U., Hinzmann, S., and Wolf, T. The coevolution of innovative ties and technological proximity. 2013.
- Case, D. O. and Higgins, G. M. How can we investigate citation behavior? a study of reasons for citing literature in communication. *Journal of the American Society for Information Science*, 51(7):635–645, 2000.
- Castillo, C., Donato, D., and Gionis, A. Estimating number of citations using author reputation. In *String processing and information retrieval*, pages 107–117. Springer, 2007.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, pages 321–357, 2002.
- Cohen, W. M. and Levinthal, D. A. Absorptive capacity: a new perspective on learning and innovation. *Administrative science quarterly*, pages 128–152, 1990.
- Cohendet, P. and Llerena, P. Learning, technical change and public policy: how to create and exploit diversity. *Systems of Innovation, Technologies, Institutions and Organizations*, London: Pinter, 1997.
- Cronin, B. Agreement and divergence on referencing practice. *Journal of Information Science*, 3(1):27–33, 1981.

- Cronin, B. Norms and functions in citation: The view of journal editors and referees in psychology. *Social Science Information Studies*, 2(2):65–77, 1982.
- Cunningham, S. W. and Werker, C. Proximity and collaboration in european nanotechnology. *Papers in Regional Science*, 91(4):723–742, 2012.
- de Solla Price, D. J. Networks of scientific papers. *Science*, 149(3683):510–515, 1965.
- Delemarle, A., Kahane, B., Villard, L., and Laredo, P. Geography of knowledge production in nanotechnologies: a flat world with many hills and mountains. *Nanotech. L. & Bus.*, 6:103, 2009.
- Dettmann, A. and Brenner, T. Proximity is a social process: a conceptual modification. In *DRUID Winter Conference*, 2010.
- Dietz, L., Bickel, S., and Scheffer, T. Unsupervised prediction of citation influences. In *Proceedings of the 24th international conference on Machine learning*, pages 233–240. ACM, 2007.
- Dijkstra, E. W. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271, 1959.
- Ding, Y. Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of informetrics*, 5(1):187–203, 2011.
- Dong, Y., Johnson, R. A., and Chawla, N. V. Will this paper increase your h-index?: Scientific impact prediction. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 149–158. ACM, 2015.
- EPA. Nanotechnology, biotechnology, and information technology: Implications for future science at EPA. A Workshop of the United States Environmental Protection Agency (EPA) Science Advisory Board. EPA-SAB-WKS-05-001. [http://yosemite.epa.gov/sab/sabproduct.nsf/3C26721F6B9C6E7A852570B30077D1B5/\\$File/Nanotech+Biotech+and+Info+Tech+EPA-SAB-WKS-05-001+with+Appendices+A-K.pdf](http://yosemite.epa.gov/sab/sabproduct.nsf/3C26721F6B9C6E7A852570B30077D1B5/$File/Nanotech+Biotech+and+Info+Tech+EPA-SAB-WKS-05-001+with+Appendices+A-K.pdf), 2005. Accessed: 2014-11-06.
- Eslami, H., Ebadi, A., and Schiffauerova, A. Effect of collaboration network structure on knowledge creation and technological performance: the case of biotechnology in canada. *Scientometrics*, 97(1):99–119, 2013.
- Feldman, M. P. *The geography of innovation*, volume 2. Springer, 1994.
- Feldman, M. P. The internet revolution and the geography of innovation. *International Social Science Journal*, 54(171):47–56, 2002.



- Freel, M. S. Sectoral patterns of small firm innovation, networking and proximity. *Research policy*, 32(5): 751–770, 2003.
- Frenken, K., Ponds, R., and Van Oort, F. The citation impact of research collaboration in science-based industries: A spatial-institutional analysis. *Papers in regional science*, 89(2):351–271, 2010.
- Garfield, E. Is citation analysis a legitimate evaluation tool? *Scientometrics*, 1(4):359–375, 1979.
- Garfield, E. Random thoughts on citationology its theory and practice. *Scientometrics*, 43(1):69–76, 1998.
- Garfield, E. et al. Can citation indexing be automated. In *Statistical association methods for mechanized documentation, symposium proceedings*, volume 1, pages 189–92. National Bureau of Standards, Miscellaneous Publication 269, Washington, DC, 1965.
- Gay, B. and Dousset, B. Innovation and network structural dynamics: Study of the alliance network of a major sector of the biotechnology industry. *Research policy*, 34(10):1457–1475, 2005.
- GeoDataSource.com. Distance calculation in php. <https://www.geodatasource.com/developers/php>, 2015. Accessed: 2015-03-12.
- Gertler, M. S. and Wolfe, D. A. Spaces of knowledge flows: Clusters in a global context. *Clusters and regional development: Critical reflections and explorations*, pages 218–35, 2006.
- Getoor, L. Link-based classification. In *Advanced methods for knowledge discovery from complex data*, pages 189–207. Springer, 2005.
- Gilly, J.-P. and Torre, A. Proximity relations. elements for an analytical framework, 2000.
- Gittelman, M. Does geography matter for science-based firms? epistemic communities and the geography of research and patenting in biotechnology. *Organization Science*, 18(4):724–741, 2007.
- Glänzel, W. and Schubert, A. Analysing scientific networks through co-authorship. In *Handbook of quantitative science and technology research*, pages 257–276. Springer, 2005.
- Google Inc. Google maps web service apis, 2015. URL <https://developers.google.com/maps/web-services/>.
- Grabher, G. Ecologies of creativity: the village, the group, and the heterarchic organisation of the british advertising industry. *Advertising Age*, 1980.
- Grandjean, M. La connaissance est un réseau. *Les Cahiers du numérique*, 10(3):37–54, 2014.

- Hagberg, A. A., Schult, D. A., and Swart, P. J. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA, USA, Aug. 2008.
- Hagstrom, W. The scientific community basic books. *New York*, 1965.
- Heisenberg, W. Der teil und das ganze gespräche im umkreis der atomphysik. *Piper, München*, 1969.
- Hirsch, J. E. Does the h index have predictive power? *Proceedings of the National Academy of Sciences*, 104(49):19193–19198, 2007.
- Hosmer, D. W., Hosmer, T., Le Cessie, S., Lemeshow, S., et al. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in medicine*, 16(9):965–980, 1997.
- Howells, J. R. Tacit knowledge, innovation and economic geography. *Urban studies*, 39(5-6):871–884, 2002.
- Hu, A. G. and Jaffe, A. B. Patent citations and international knowledge flow: the cases of korea and taiwan. *International journal of industrial organization*, 21(6):849–880, 2003.
- Hudson, R. “the learning economy, the learning firm and the learning region” a sympathetic critique of the limits to learning. *European Urban and Regional Studies*, 6(1):59–72, 1999.
- Hummon, N. P. and Dereian, P. Connectivity in a citation network: The development of dna theory. *Social Networks*, 11(1):39–63, 1989.
- Hyland, K. Academic attribution: Citation and the construction of disciplinary knowledge. *Applied linguistics*, 20(3):341–367, 1999.
- Ibáñez, A., Larrañaga, P., and Bielza, C. Predicting citation count of bioinformatics papers within four years of publication. *Bioinformatics*, 25(24):3303–3309, 2009.
- Jaffe, A. B. Real effects of academic research. *The American Economic Review*, pages 957–970, 1989.
- Jaffe, A. B. and Trajtenberg, M. International knowledge flows: evidence from patent citations. *Economics of Innovation and New Technology*, 8(1-2):105–136, 1999.
- Jawed, M., Kaya, M., and Alhajj, R. Time frame based link prediction in directed citation networks. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 1162–1168. ACM, 2015.
- Jiang, Y. Locating active actors in the scientific collaboration communities based on interaction topology analyses. *Scientometrics*, 74(3):471–482, 2008.

- Jones, S. G., Ashby, A. J., Momin, S. R., and Naidoo, A. Spatial implications associated with using euclidean distance measurements and geographic centroid imputation in health care research. *Health services research*, 45(1):316–327, 2010.
- Katz, J. S. Geographical proximity and scientific collaboration. *Scientometrics*, 31(1):31–43, 1994.
- Katz, J. S. and Hicks, D. The classification of interdisciplinary journals: a new approach. In *Proceedings of the fifth international conference of the international society for scientometrics and informetrics. Learned Information, Melford*. Citeseer, 1995.
- Katz, J. S. and Martin, B. R. What is research collaboration? *Research policy*, 26(1):1–18, 1997.
- Kirat, T. and Lung, Y. Innovation and proximity territories as loci of collective learning processes. *European urban and regional studies*, 6(1):27–38, 1999.
- Klein, J. T. *Interdisciplinarity: History, theory, and practice*. Wayne State University Press, 1990.
- Knoben, J. and Oerlemans, L. A. Proximity and inter-organizational collaboration: A literature review. *International Journal of Management Reviews*, 8(2):71–89, 2006.
- Kraut, R., Egidio, C., and Galegher, J. Patterns of contact and communication in scientific research collaboration. In *Proceedings of the 1988 ACM conference on Computer-supported cooperative work*, pages 1–12. ACM, 1988.
- LaFollette, M. C. *Stealing into print: fraud, plagiarism, and misconduct in scientific publishing*. Univ of California Press, 1992.
- Laudel, G. Collaboration, creativity and rewards: why and how scientists collaborate. *International Journal of Technology Management*, 22(7):762–781, 2001.
- Laursen, K., Reichstein, T., and Salter, A. Exploring the effect of geographical proximity and university quality on university–industry collaboration in the united kingdom. *Regional studies*, 45(4):507–523, 2011.
- Le Cessie, S. and Van Houwelingen, J. A goodness-of-fit test for binary regression models, based on smoothing methods. *Biometrics*, pages 1267–1282, 1991.
- Leamer, E. E. and Storper, M. The economic geography of the internet age. Technical report, National Bureau of Economic Research, 2001.
- Lemon, S. C., Roy, J., Clark, M. A., Friedmann, P. D., and Rakowski, W. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Annals of behavioral medicine*, 26(3):172–181, 2003.

- Leydesdorff, L. and Zhou, P. Nanotechnology as a field of science: Its delineation in terms of journals and patents. *Scientometrics*, 70(3):693–713, 2007.
- Liao, T. F. *Interpreting probability models: Logit, probit, and other generalized linear models*. Number 101. Sage, 1994.
- Liaw, A. and Wiener, M. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002. URL <http://CRAN.R-project.org/doc/Rnews/>.
- Lichtenwalter, R. N., Lussier, J. T., and Chawla, N. V. New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 243–252. ACM, 2010.
- Liebesskind, J. P., Oliver, A. L., Zucker, L., and Brewer, M. Social networks, learning, and flexibility: Sourcing scientific knowledge in new biotechnology firms. *Organization science*, 7(4):428–443, 1996.
- Lievrouw, L. A. The invisible college reconsidered bibliometrics and the development of scientific communication theory. *Communication Research*, 16(5):615–628, 1989.
- Liu, X., Jiang, S., Chen, H., Larson, C. A., and Roco, M. C. Nanotechnology knowledge diffusion: measuring the impact of the research networking and a strategy for improvement. *Journal of Nanoparticle Research*, 16(9):1–15, 2014.
- Loh, W.-Y. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):14–23, 2011.
- Long, J. S. and Freese, J. *Regression models for categorical dependent variables using Stata*. Stata press, 2006.
- Lublinski, A. E. Does geographic proximity matter? evidence from clustered and non-clustered aeronautic firms in germany. *Regional Studies*, 37(5):453–467, 2003.
- Maggioni, M. A. and Uberti, T. E. Knowledge networks across europe: which distance matters? *The Annals of Regional Science*, 43(3):691–720, 2009.
- Malecki, E. J. and Oinas, P. *Making connections: technological learning and regional economic change*. Ashgate Publishing Company, 1999.
- Malerba, F. Sectoral systems of innovation: a framework for linking innovation to the knowledge base, structure and dynamics of sectors. *Economics of Innovation and New Technology*, 14(1-2):63–82, 2005.
- Malerba, F., Breschi, S., and Lissoni, F. Knowledge proximity and technological diversification, 1998.

- Malsh, I. The importance of interdisciplinary approaches: The case of nanotechnology. *IPTS Report*, 13, 1997.
- Marion, L. S., Garfield, E., Hargens, L. L., Lievrouw, L. A., White, H. D., and Wilson, C. S. Social network analysis and citation network analysis: Complementary approaches to the study of scientific communication. sponsored by sig met. *Proceedings of the American Society for Information Science and Technology*, 40(1):486–487, 2003.
- Marshall, A. *Principles of economics*. Royal Economic Society/Macmillan, London, 1890.
- Martin, T., Ball, B., Karrer, B., and Newman, M. Coauthorship and citation patterns in the physical review. *Physical Review E*, 88(1):012814, 2013.
- Maskell, P. Towards a knowledge-based theory of the geographical cluster. *Industrial and corporate change*, 10(4):921–943, 2001.
- Menard, S. *Logistic Regression: From Introductory to Advanced Concepts and Applications: From Introductory to Advanced Concepts and Applications*. Sage Publications, 2009.
- Merriam-Webster Web Dictionary. Classification. (n.d.). <http://www.merriam-webster.com/dictionary/classification>, 2014. Accessed: 2014-03-25.
- Merton, R. C. An intertemporal capital asset pricing model. *Econometrica: Journal of the Econometric Society*, pages 867–887, 1973.
- Michie, D., Spiegelhalter, D. J., Taylor, C. C., and Campbell, J., editors. *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, Upper Saddle River, NJ, USA, 1994. ISBN 0-13-106360-X.
- Microsoft Encarta Encyclopedia. 747 jumbo jet, 2000. URL <http://hypertextbook.com/facts/2002/JobyJosekutty.shtml>. CDROM. Accessed: 2015-03-16.
- Milgram, S. The small world problem. *Psychology today*, 2(1):60–67, 1967.
- Milojevic, S. Modes of collaboration in modern science: Beyond power laws and preferential attachment. *Journal of the American Society for Information Science and Technology*, 61(7):1410–1423, 2010.
- Moazami, A. *A Network Perspective of Nanotechnology Innovation: A Comparison of Quebec, Canada and the United States*. PhD thesis, Concordia University Montreal, Quebec, Canada, 2012.
- Moodysson, J. and Jonsson, O. Knowledge collaboration and proximity the spatial organization of biotech innovation projects. *European urban and regional studies*, 14(2):115–131, 2007.

- Morgan, K. The exaggerated death of geography: learning, proximity and territorial innovation systems. *Journal of economic geography*, 4(1):3–21, 2004.
- Munasinghe, L. and Ichise, R. Time aware index for link prediction in social networks. In *Data Warehousing and Knowledge Discovery*, pages 342–353. Springer, 2011.
- Murray, F. Innovation as co-evolution of scientific and technological networks: exploring tissue engineering. *Research Policy*, 31(8):1389–1403, 2002.
- National Academies Committee on Science, Engineering, and Public Policy (COSEPUP). Facilitating interdisciplinary research. 2005.
- Neville, P. G. Decision trees for predictive modeling. *SAS Institute Inc*, 1999.
- Newman, M. E. Who is the best connected scientist? a study of scientific coauthorship networks. *Phys. Rev. E*, 64(016131), 2001a.
- Newman, M. E. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, 2001b.
- Newman, M. E. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the national academy of sciences*, 101(suppl 1):5200–5205, 2004.
- Nooteboom, B. *Learning and innovation in organizations and economies*. Oxford University Press, 2000.
- Onel, S., Zeid, A., and Kamarthi, S. The structure and analysis of nanotechnology co-author and citation networks. *Scientometrics*, 89(1):119–138, 2011.
- Oracle. Mysql database, 2015. URL <https://www.mysql.com/>.
- Otte, E. and Rousseau, R. Social network analysis: a powerful strategy, also for the information sciences. *Journal of information Science*, 28(6):441–453, 2002.
- Padmaja, T. M., Dhulipalla, N., Krishna, P. R., Bapi, R. S., and Laha, A. An unbalanced data classification model using hybrid sampling technique for fraud detection. In *Pattern Recognition and Machine Intelligence*, pages 341–348. Springer, 2007.
- Patel, N. Collaboration in professional growth of american sociology. *SOCIAL SCIENCE INFORMATION SUR LES SCIENCES SOCIALES*, 12(6):77–92, 1973.
- Perez, C. and Soete, L. Catching up in technology: entry barriers and windows of opportunity. *Technical Change and Economic Theory*, pages 458–479, 1988.

- Phibbs, C. S. and Luft, H. S. Correlation of travel time on roads versus straight line distance. *Medical Care Research and Review*, 52(4):532–542, 1995.
- Quinlan, J. R. *C4. 5: programs for machine learning*. Elsevier, 2014.
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL <https://www.R-project.org/>.
- Radicchi, F., Fortunato, S., and Castellano, C. Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, 105(45):17268–17272, 2008.
- Rafols, I. and Meyer, M. How cross-disciplinary is bionanotechnology? explorations in the specialty of molecular motors. *Scientometrics*, 70(3):633–650, 2007.
- Rafols, I. and Meyer, M. Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics*, 82(2):263–287, 2010.
- Rafols, I., Park, J., and Meyer, M. Hybrid nanomaterial research: Is it really interdisciplinary. *The Supramolecular Chemistry of Organic-Inorganic Hybrid Materials*, 2010.
- Rallet, A. and Torre, A. Is geographical proximity necessary in the innovation networks in the era of global economy? *GeoJournal*, 49(4):373–380, 1999.
- Rallet, A., Torre, A., and Antonelli, C. *Economie industrielle et économie spatiale*. Economica, 1995.
- Rogers, E. M. Diffusion of preventive innovations. *Addictive behaviors*, 27(6):989–993, 2002.
- Rogers, E. M. *Diffusion of innovations*. Simon and Schuster, 2010.
- Sarigöl, E., Pfitzner, R., Scholtes, I., Garas, A., and Schweitzer, F. Predicting scientific success based on coauthorship networks. *EPJ Data Science*, 3(1):1–16, 2014.
- SAS Institute Inc. Machine learning: What it is & why it matters. [http://www.sas.com/en\\_us/insights/analytics/machine-learning.html](http://www.sas.com/en_us/insights/analytics/machine-learning.html), 2015. Accessed: 2015-09-22.
- Schiffauerova, A. and Beaudry, C. Star scientists and their positions in the canadian biotechnology network. *Economics of Innovation and New Technology*, 20(4):343–366, 2011.
- Schoenberger, E. J. *The cultural crisis of the firm*. Blackwell, 1997.
- Schummer, J. Multidisciplinarity, interdisciplinarity, and patterns of research collaboration in nanoscience and nanotechnology. *Scientometrics*, 59(3):425–465, 2004.
- Sci2 Team. Science of science (sci2) tool, 2009. URL <https://sci2.cns.iu.edu>.

- Shadish, W. R., Tolliver, D., Gray, M., and Gupta, S. K. S. Author judgements about works they cite: three studies from psychology journals. *Social Studies of Science*, 25(3):477–498, 1995.
- Shapin, S. Here and everywhere: Sociology of scientific knowledge. *Annual review of sociology*, pages 289–321, 1995.
- Shapira, P. and Youtie, J. Emergence of nanodistricts in the united states path dependency or new opportunities? *Economic Development Quarterly*, 22(3):187–199, 2008.
- Shaw, A. T. and Gilly, J.-P. On the analytical dimension of proximity dynamics. *Regional studies*, 34(2):169–180, 2000.
- Singh, J. Collaborative networks as determinants of knowledge diffusion patterns. *Management science*, 51(5):756–770, 2005.
- Singh, J. and Marx, M. Geographic constraints on knowledge spillovers: political borders vs. spatial proximity. *Management Science*, 59(9):2056–2078, 2013.
- Small, H. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science*, 24(4):265–269, 1973.
- Song, Y.-y. and Ying, L. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2):130, 2015.
- Sonn, J. W. and Storper, M. The increasing importance of geographical proximity in knowledge production: an analysis of us patent citations, 1975-1997. *Environment and Planning A*, 40(5):1020, 2008.
- Sorenson, O., Rivkin, J. W., and Fleming, L. Complexity, networks and knowledge flow. *Research Policy*, 35(7):994–1017, 2006.
- StataCorp. Stata statistical software: Release 14.1, 2015.
- Sternitzke, C., Bartkowski, A., and Schramm, R. Visualizing patent statistics by means of social network analysis tools. *World Patent Information*, 30(2):115–131, 2008.
- Storper, M. *The regional world: territorial development in a global economy*. Guilford Press, 1997.
- Subramanyam, K. Bibliometric studies of research collaboration: A review. *Journal of information Science*, 6(1):33–38, 1983.
- Tallman, S., Jenkins, M., Henry, N., and Pinch, S. Knowledge, clusters, and competitive advantage. *Academy of management review*, 29(2):258–271, 2004.



- Ter Wal, A. L. and Boschma, R. A. Applying social network analysis in economic geography: framing some key analytic issues. *The Annals of Regional Science*, 43(3):739–756, 2009.
- Thompson, D. Ranking predictors in logistic regression. *Paper D10-2009, Assurant Health, West Michigan*, 2009.
- Tong, W., Hong, H., Fang, H., Xie, Q., and Perkins, R. Decision forest: combining the predictions of multiple independent decision tree models. *Journal of Chemical Information and Computer Sciences*, 43(2):525–531, 2003.
- Torgo, L. *Data Mining with R, learning with case studies*. Chapman and Hall/CRC, 2010. URL <http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR>.
- Torres, E., Dominguez, J., Valdes, L., and Aza, R. Passenger waiting time in an airport and expenditure carried out in the commercial area. *Journal of Air Transport Management*, 11(6):363–367, 2005.
- Trajtenberg, M., Henderson, R., and Jaffe, A. University versus corporate patents: A window on the basicness of invention. *Economics of Innovation and new technology*, 5(1):19–50, 1997.
- Uddin, S., Hossain, L., Abbasi, A., and Rasmussen, K. Trend and efficiency analysis of co-authorship network. *Scientometrics*, 90(2):687–699, 2011.
- Valente, T. W. and Rogers, E. M. The origins and development of the diffusion of innovations paradigm as an example of scientific growth. *Science communication*, 16(3):242–273, 1995.
- Valente, T. W., Coronges, K., Lakon, C., and Costenbader, E. How correlated are network centrality measures? *Connections (Toronto, Ont.)*, 28(1):16, 2008.
- Vinkler, P. A quasi-quantitative citation model. *Scientometrics*, 12(1-2):47–72, 1987.
- von Thünen, J. H. Der isolierte staat in beziehung auf nationalökonomie und landwirtschaft. *Gustav Fischer, Stuttgart (reprinted 1966)*, 1826.
- Wallace, M. L., Larivière, V., and Gingras, Y. A small world of citations? the influence of collaboration networks on citation practices. *PloS one*, 7(3):e33339, 2012.
- Wallsten, S. J. An empirical test of geographic knowledge spillovers using geographic information systems and firm-level data. *Regional Science and Urban Economics*, 31(5):571–599, 2001.
- Wang, G. and Guan, J. Measuring science–technology interactions using patent citations and author-inventor links: an exploration analysis from chinese nanotechnology. *Journal of Nanoparticle Research*, 13(12):6245–6262, 2011.

- Wang, J., Xu, M., Wang, H., and Zhang, J. Classification of imbalanced data by using the smote algorithm and locally linear embedding. In *Signal Processing, 2006 8th International Conference on*, volume 3. IEEE, 2007.
- Watts, D. J. and Strogatz, S. H. Collective dynamics of "small-world" networks. *nature*, 393(6684): 440–442, 1998.
- Weber, M., Gerth, H. H., and Mills, W. C. Science as a vocation. *From Max Weber*, page 129–156, 1946.
- Weingart, P. Interdisciplinarity: The paradoxical discourse. *Practising interdisciplinarity*, pages 25–41, 2000.
- Wellman, B. and Berkowitz, S. D. *Social structures: A network approach*, volume 2. CUP Archive, 1988.
- Wetherell, C., Plakans, A., and Wellman, B. Social networks, kinship, and community in eastern europe. *Journal of Interdisciplinary History*, pages 639–663, 1994.
- White, H. Reward, persuasion, and the sokal hoax: A study in citation identities. *Scientometrics*, 60(1): 93–120, 2004.
- White, H. D. Authors as citers over time. *Journal of the American Society for Information Science and Technology*, 52(2):87–108, 2001.
- WIPO, W. International patent classification (ipc). <http://www.wipo.int/classifications/ipc/en/>, 2015. Accessed: 2015-12-10.
- Xu, R. *Improvements to random forest methodology*. PhD thesis, Iowa State University, 2013.
- Yan, R., Huang, C., Tang, J., Zhang, Y., and Li, X. To better stand on the shoulder of giants. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, pages 51–60. ACM, 2012.
- Yegul, M. F., Yavuz, M., and Guild, P. Nanotechnology: Canada's position in scientific publications and patents. In *Management of Engineering & Technology, 2008. PICMET 2008. Portland International Conference on*, pages 704–713. IEEE, 2008.
- Zeller, C. North atlantic innovative relations of swiss pharmaceuticals and the proximities with regional biotech arenas. *Economic Geography*, 80(1):83–111, 2004.
- Zhao, Y.-A., Chen, J.-J., and Mu, X.-F. Application of unbalanced data approach in high-speed network intrusion detection. *Journal of Computer Applications*, 7:015, 2009.
- Zinkhan, G. M., Roth, M. S., and Saxton, M. J. Knowledge development and scientific status in consumer-behavior research: a social exchange perspective. *Journal of Consumer Research*, pages 282–291, 1992.

# Appendices

---

 General References
 

---

## A.1 Cognitive Fields

Field	Subfields	
Arts	Fine Arts & Architecture	Performing Arts
Biology	Agricult & Food Science	General Biology
	Botany	General Zoology
	Dairy & Animal Science	Marine Biology & Hydrobiology
	Ecology	Miscellaneous Biology
	Entomology	Miscellaneous Zoology
Biomedical Research	Anatomy & Morphology	Microbiology
	Biochemistry & Molecular Biology	Microscopy
	Biomedical Engineering	Misc Biomedical Research
	Biophysics	Nutrition & Dietetic
	Cell Biology Cytology & Histology	Parasitology
	Embryology	Physiology
	General Biomedical Research	Virology
	Genetics & Heredity	

---

Chemistry	Analytical Chemistry	Organic Chemistry	
	Applied Chemistry	Physical Chemistry	
	General Chemistry	Polymers	
	Inorganic & Nuclear Chemistry		
Clinical Medicine	Addictive Diseases	Nephrology	
	Allergy	Neurology & Neurosurgery	
	Anesthesiology	Obstetrics & Gynecology	
	Arthritis & Rheumatology	Ophthalmology	
	Cancer	Orthopedics	
	Cardiovascular System	Otorhinolaryngology	
	Dentistry	Pathology	
	Dermatology & Venereal Disease	Pediatrics	
	Endocrinology	Pharmacology	
	Env & Occupational Health	Pharmacy	
	Fertility	Psychiatry	
	Gastroenterology	Radiology & Nuclear Medicine	
	General & Internal Medicine	Respiratory System	
	Geriatrics	Surgery	
	Hematology	Tropical Medicine	
	Immunology	Urology	
	Misc Clinical Medicine	Veterinary Medicine	
	Earth and Space	Astronomy & Astrophysics	Geology
		Earth & planetary Science	Meteorology & Atmosph Science
Environmental Science		Oceanography & Limnology	
Engineering and Technology	Aerospace Technology	Materials Science	
	Chemical Engineering	Mechanical Engineering	
	Civil Engineering	Metals & Metallurgy	
	Computers	Misc Engineering & Technology	
	Electrical Eng & Electronics	Nuclear Technology	
	General Engineering	Operations Research	
	Industrial Engineering		

Health	Geriatrics & Gerontology	Rehabilitation
	Health Policy & Services	Social Sciences, Biomedical
	Nursing	Social Studies of Medicine
	Public Health	Speech Lang Pat & Audiology
Humanities	History	Miscellaneous Humanities
	Language & Linguistics	Philosophy
	Literature	Religion
Mathematics	Applied Mathematics	Miscellaneous Mathematics
	General Mathematics	Probability & Statistics
Physics	Acoustics	Miscellaneous Physics
	Applied Physics	Nuclear & Particle Physics
	Chemical Physics	Optics
	Fluids & Plasmas	Solid State Physics
	General Physics	
Professional Fields	Communication	Management
	Education	Misc Professional Field
	Inf Science & Library Science	Social Work
	Law	
Psychology	Behav Science & Comp Psychology	Human Factors
	Clinical Psychology	Miscellaneous Psychology
	Developmental & Child Psychology	Psychoanalysis
	Experimental Psychology	Social Psychology
	General Psychology	
Social Sciences	Anthropology and Archaeology	International Relations
	Area Studies	Misc Social Sciences
	Criminology	Planning & Urban Studies
	Demography	Political Science and Public Admin
	Economics	Science studies
	General Social Sciences	Sociology
	Geography	

Table 11: Cognitive fields and subfields in the data

## Data Structure

Field	Column	Data Type	Field	Column	Data Type
serial (Primary)		int(11)	Paper_ID		int(11)
Count		tinyint(4)	P_order		tinyint(4)
Citing		tinyint(1)	P_Author		varchar(100)
Cognitive_dist		tinyint(4)	Source_node		int(11)
Coauthored_works		int(11)	P_year		smallint(6)
Shortest_path		int(11)	Source_lat		double
Degree_target		int(11)	Source_lng		double
Between_target		double	Target_lat		double
Close_target		double	Target_lng		double
Eigen_target		double	P_Field		varchar(50)
Cluster_target		float	P_Subfield		varchar(50)
Flying_Distance		float	Cited_paper_ID		int(11)
Flying_Time		time	C_order		tinyint(4)
Driving_Distance		float	C_Author		varchar(100)
Driving_Time		time	Target_node		int(11)
Travel_Time		time	C_year		smallint(6)
Location_Category		tinyint(4)	C_Field		varchar(50)
DistanceID		int(11)	C_Subfield		varchar(50)

Table 12: Table structure of paired links in MySQL database

## C.1 SMOTE sampling with R

```
#Load required library
library(DMwR)

#Load population file in R
mydata <- read.csv('C:\\Users\\Elvita\\Downloads\\cited_auth_population.csv', header=T)

#Check field names in loaded data and first 10 rows
colnames(mydata)
head(mydata,10)

#Summary of values in Citing and Year columns
table(mydata$YOP)
table(mydata$Citing)

#Check the balance of positive versus negative outcomes
prop.table(table(mydata$Citing))
```



```

#Setting seed for sampling
set.seed(1234)

#Setting sample size
n <- 54000

#We randomly split the data set into 2 equal portions
mysample <- createDataPartition(mydata$Citing, p = .50, list = FALSE, times = 1)

mysample <- mydata[ splitIndex,]
mysamplettest <- mydata[-splitIndex,]

#Check if the balance is still approx the same in the partition
prop.table(table(trainSplit$Citing))
prop.table(table(testSplit$Citing))

#Preparing for SMOTE
trainSplit$Citing <- as.factor(trainSplit$Citing)
trainSplit <- SMOTE(Citing ~ ., trainSplit, perc.over = 100, perc.under=200)
trainSplit$Citing <- as.numeric(trainSplit$Citing)

#Check for proportion of results
table(trainSplit$Citing)
prop.table(table(trainSplit$Citing))

#Export Smoted sample to a CSV file
write.csv(file='C:\\Users\\Elvita\\Desktop\\SMOTEDsample.csv', x=trainSplit, row.names=
FALSE)

```

## C.2 Shortest Path with NetworkX and Python

```
#Call required libraries
print "Importing libraries..."
import networkx as nx
import csv
import numpy as np

#Import network in Pajek format .net
myG=nx.read_pajek("MyNetwork_0711_onlylabel.net")
print "Finished importing Network Pajek file"

#Simplify graph into networkx format
G=nx.Graph(myG)
print "Finished converting to Networkx format"

#Network info
print "Nodes found: ",G.number_of_nodes()
print "Edges found: ",G.number_of_edges()

#Reading file and storing to array
with open('paired_nodes.csv','rb') as csvfile:
    reader = csv.reader(csvfile, delimiter = ', ', quoting=csv.QUOTE_MINIMAL)#, quotechar
    = ',')
    data = [data for data in reader]
paired_nodes = np.asarray(data)
paired_nodes.astype(int)
print "Finished reading paired nodes file"

#Add extra column in array to store shortest path value
paired_nodes = np.append(paired_nodes,np.zeros([len(paired_nodes),1],dtype=np.int),1)
print "Just appended new column to paired nodes array"

#Get shortest path for every pair of nodes
for index in range(len(paired_nodes)):
```

```

    try:
        shortest=nx.shortest_path_length(G,paired_nodes[index,0],paired_nodes[
index,1])
        #print shortest
        paired_nodes[index,2] = shortest
    except nx.NetworkXNoPath:
        #print '99999' #Value to print when no path is found
        paired_nodes[index,2] = 99999
print "Finished calculating shortest path for paired nodes"

#Store results to csv file
f = open('shortest_path_results.csv','w')

for item in paired_nodes:
    f.write(','.join(map(str,item)))
    f.write('\n')
f.close()

print "Done writing file with results, bye!"

```

# APPENDIX D

## Statistics

This appendix contains several tables corresponding to statistical information and analysis run on our data set.

### General Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
Cognitive_dist	79524	1	3	2.53	.730	.533
Pub_works_target	79524	1	2144	11.68	47.625	2268.110
Coauthored_works	79524	0	35	.05	.505	.255
Shortest_path	75747	1	9	3.82	1.193	1.423
Degree_target	79524	0	5864	42.41	149.239	22272.410
Between_target	79524	0.00000000000	0.01900220600	0.00002727572	0.00030925763	.000
Close_target	79135	0.00000296686	0.33010975100	0.21090911701	0.04550922486	.002
Eigen_target	79135	0.00000000000	0.11385647029	0.00069788694	0.00352368293	.000
Cluster_target	79524	0.0000000	1.0000000	.565966332	.3532162066	.125
Euclidean_Distance	79524	0.00	18558.70	5454.02	3739.50	13983846.36
Travel_Time	79524	0:00:00	25:58:13	10:51:33	4:44:26	291248913.233
Location_Category	79524	1	5	4.43	.920	.847
Valid N (listwise)	75747					

Table 13: Descriptive statistics of independent variables

### Descriptive Statistics by Citing

Citing	Cognitive_dist	Pub_works_target	Coauth_works	Short_path	Degree_target	Between_target	Close_target	Eigen_target	Cluster_target	Euc_Dist	TravelTime	Location_Cat
N	46614	46614	46614	43645	46614	46614	46295	46295	46614	46614	46614	46614
Mean	2.78	4.27	.00	4.23	17.96	.00000354675237	.20138126783372	.00017329566513	.648629674	5830.452011768210	11:30:04	4.62
Minimum	1	1	0	1	0	0.000000000000	.000002966862	.000000000000	0.0000000	0.0000000000	0:00:00	1
Maximum	3	278	2	9	933	.000734096715	.295036900000	.028098652457	1.0000000	18558.7000000000	25:58:13	5
Std. Deviation	.504	7.817	.011	.932	30.973	#####	#####	#####	.3394362847	3591.2464480542000	4:12:02	.591
Variance	.254	61.099	.000	.869	959.349	.000	.002	.000	.115	12897051.051	228694951.113	.349
Skewness	-2.317	9.636	146.898	.328	7.204	17.608	-2.835	13.654	-.328	.423	.265	-1.745
N	32910	32910	32910	32102	32910	32910	32840	32840	32910	32910	32910	32910
Mean	2.17	22.18	.13	3.27	77.03	.00006088561782	.22434065712152	.00143741047147	.448881282	4920.848858600670	9:57:00	4.16
Minimum	1	1	0	1	0	0.000000000000	.000002966862	.000000000000	0.0000000	0.0000000000	0:00:00	1
Maximum	3	2144	35	9	5864	.019002206000	.330109751000	.113856470294	1.0000000	18304.5000000000	25:40:59	5
Std. Deviation	.841	72.154	.778	1.284	224.534	#####	#####	#####	.3389639044	3877.9683138033500	5:16:55	1.195
Variance	.707	5206.188	.606	1.648	50415.721	.000	.002	.000	.115	15038638.243	361581743.702	1.428
Skewness	-.329	12.128	13.782	.035	10.739	22.985	-2.477	8.653	.549	.525	-.083	-1.660
N	79524	79524	79524	75747	79524	79524	79135	79135	79524	79524	79524	79524
Mean	2.53	11.68	.05	3.82	42.41	.00002727571547	.21090911700553	.00069788693626	.565966332	5454.024269567710	10:51:33	4.43
Minimum	1	1	0	1	0	0.000000000000	.000002966862	.000000000000	0.0000000	0.0000000000	0:00:00	1
Maximum	3	2144	35	9	5864	.019002206000	.330109751000	.113856470294	1.0000000	18558.7000000000	25:58:13	5
Std. Deviation	.730	47.625	.505	1.193	149.239	#####	#####	#####	.3532162066	3739.4981421347100	4:44:26	.920
Variance	.533	2268.110	.255	1.423	22272.410	.000	.002	.000	.125	13983846.355	291248913.233	.847
Skewness	-1.200	18.101	21.364	-.233	15.815	35.519	-2.588	12.985	.024	.434	-.049	-2.243

Table 14: Descriptive statistics of independent variables according to Citing

## Correlation Matrix

		Citing	Cognitive_dist	Pub_works_target	Coauthored_works	Shortest_path	Degree_target	Between_target	Close_target	Eigen_target	Cluster_target	Euclidean_Distance	Travel_Time	Location_Category
Citing	Pearson Correlation	1	-.415**	.185**	.123**	-.397**	.195**	.091**	.249**	.177**	-.279**	-.120**	-.161**	-.245**
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	N	79524	79524	79524	79524	75747	79524	79524	79135	79135	79524	79524	79524	79524
Cognitive_dist	Pearson Correlation	-.415**	1	-.077**	-.082**	.217**	-.078**	-.038**	-.110**	-.066**	.134**	.037**	.064**	.118**
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	N	79524	79524	79524	79524	75747	79524	79524	79135	79135	79524	79524	79524	79524
Pub_works_target	Pearson Correlation	.185**	-.077**	1	.022**	-.177**	.973**	.900**	.279**	.905**	-.275**	.040**	.035**	-.002
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	.496
	N	79524	79524	79524	79524	75747	79524	79524	79135	79135	79524	79524	79524	79524
Coauthored_works	Pearson Correlation	.123**	-.082**	.022**	1	-.251**	.015**	.013**	.038**	.017**	-.041**	-.067**	-.106**	-.174**
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	N	79524	79524	79524	79524	75747	79524	79524	79135	79135	79524	79524	79524	79524
Shortest_path	Pearson Correlation	-.397**	.217**	-.177**	-.251**	1	-.189**	-.087**	-.542**	-.184**	.323**	.145**	.214**	.364**
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	N	75747	75747	75747	75747	75747	75747	75747	75747	75747	75747	75747	75747	75747
Degree_target	Pearson Correlation	.195**	-.078**	.973**	.015**	-.189**	1	.887**	.310**	.919**	-.276**	.035**	.031**	-.002
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	.532
	N	79524	79524	79524	79524	75747	79524	79524	79135	79135	79524	79524	79524	79524
Between_target	Pearson Correlation	.091**	-.038**	.900**	.013**	-.087**	.887**	1	.151**	.782**	-.125**	.016**	.015**	-.002
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	.662
	N	79524	79524	79524	79524	75747	79524	79524	79135	79135	79524	79524	79524	79524
Close_target	Pearson Correlation	.249**	-.110**	.279**	.038**	-.542**	.310**	.151**	1	.299**	-.366**	-.027**	-.034**	-.084**
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	N	79135	79135	79135	79135	75747	79135	79135	79135	79135	79135	79135	79135	79135
Eigen_target	Pearson Correlation	.177**	-.066**	.905**	.017**	-.184**	.919**	.782**	.299**	1	-.249**	.035**	.029**	-.016**
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	N	79135	79135	79135	79135	75747	79135	79135	79135	79135	79135	79135	79135	79135
Cluster_target	Pearson Correlation	-.279**	.134**	-.275**	-.041**	.323**	-.276**	-.125**	-.366**	-.249**	1	-.034**	-.020**	.034**
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	N	79524	79524	79524	79524	75747	79524	79524	79135	79135	79524	79524	79524	79524
Euclidean_Distance	Pearson Correlation	-.120**	.037**	.040**	-.067**	.145**	.035**	.016**	-.027**	.035**	-.034**	1	.978**	.687**
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	N	79524	79524	79524	79524	75747	79524	79524	79135	79135	79524	79524	79524	79524
Travel_Time	Pearson Correlation	-.161**	.064**	.035**	-.106**	.214**	.031**	.015**	-.034**	.029**	-.020**	.978**	1	.805**
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	N	79524	79524	79524	79524	75747	79524	79524	79135	79135	79524	79524	79524	79524
Location_Category	Pearson Correlation	-.245**	.118**	-.002	-.174**	.364**	-.002	-.002	-.084**	-.016**	.034**	.687**	.805**	1
	Sig. (2-tailed)	0.000	0.000	.496	0.000	0.000	.532	.662	0.000	0.000	0.000	0.000	0.000	0.000
	N	79524	79524	79524	79524	75747	79524	79524	79135	79135	79524	79524	79524	79524

\*\* . Correlation is significant at the 0.01 level (2-tailed).

Table 15: Correlation Matrix between variables

## E.1 Sci2 Output

This graph claims to be undirected.

Nodes: 674113

Isolated nodes: 4410

Node attributes present: label, number\_of\_authored\_works, id

Edges: 6067065

No self loops were discovered.

No parallel edges were discovered.

Edge attributes:

Did not detect any non-numeric attributes.

Numeric attributes:

	min	max	mean
number_...	1	166	1.37194
weight	1	166	1.37194

This network seems to be valued.

Average degree: 18.00014241

This graph is not weakly connected.

There are 13488 weakly connected components. (4410 isolates)

The largest connected component consists of 639573 nodes

Density1 [loops allowed] = 0.00002670

Density (disregarding weights): 0

Additional Densities by Numeric Attribute

## E.2 Pajek Output

Number of vertices (n): 674113

	Arcs	Edges
Total number of lines	0	6067065
Number of loops	0	0
Number of multiple lines	0	0

Density1 [loops allowed] = 0.00002670

Density2 [no loops allowed] = 0.00002670

Average Degree = 18.00014241

Figure 28: Network overview from Pajek



### E.3 Distribution of network metrics

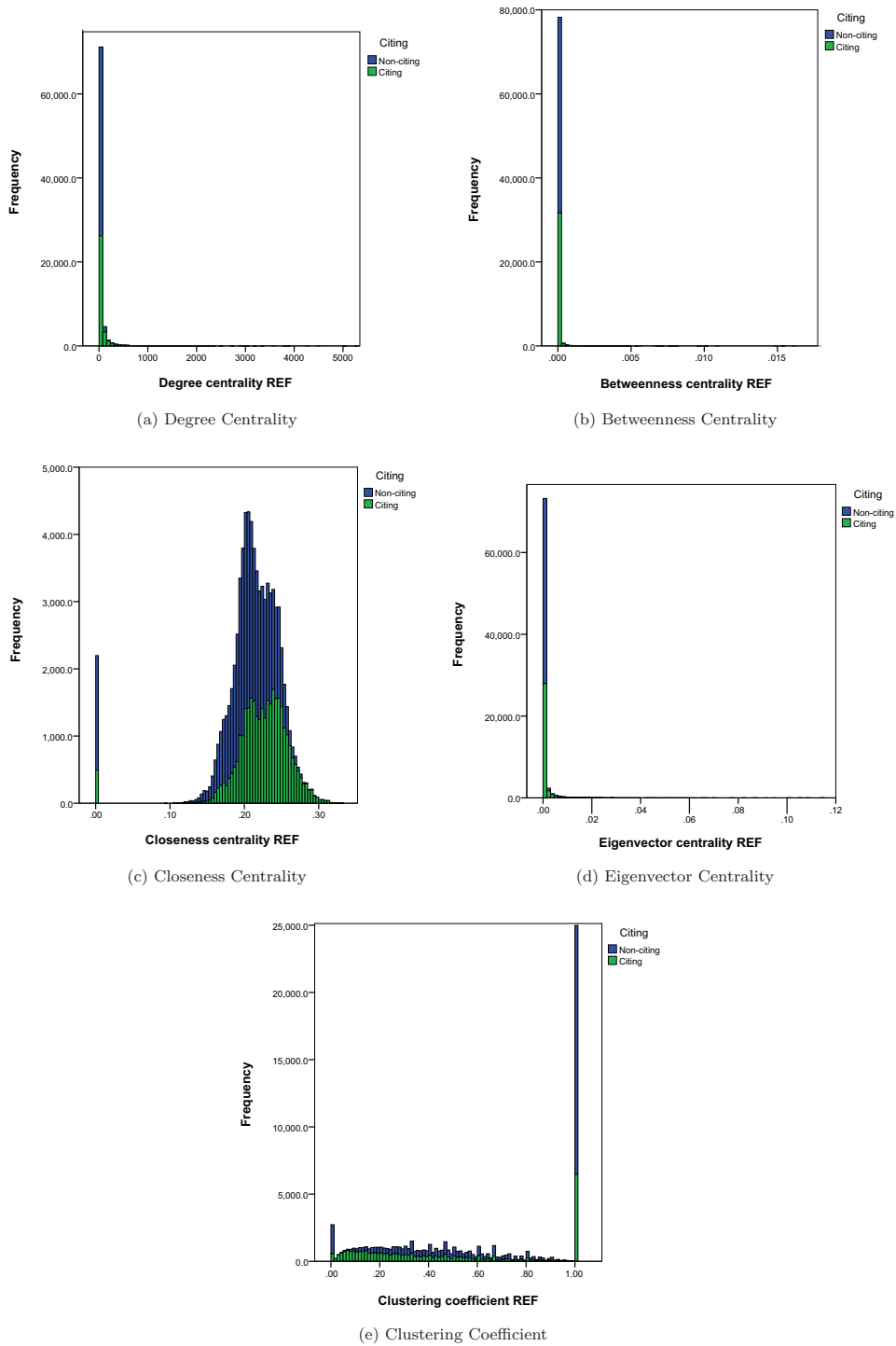
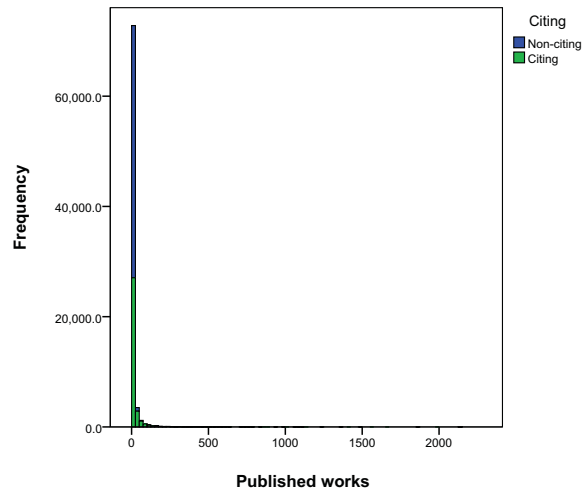
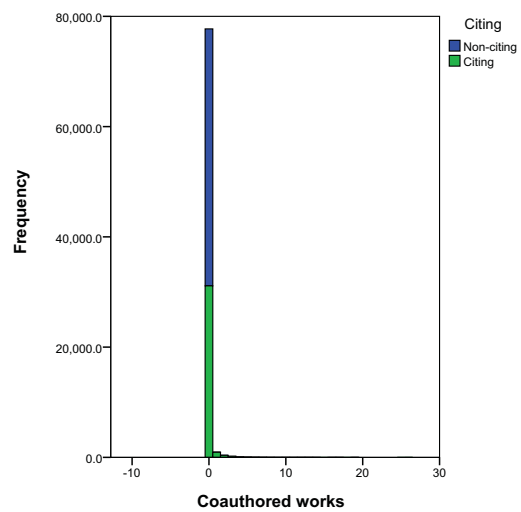


Figure 29: Distribution of SNA metrics of REF author

## E.4 Distribution of control variables



(a) Published works



(b) Coauthored works

Figure 30: Distribution of control variables