



U N I V E R S I T Y O F
T H E S S A L Y

M.Sc. IN APPLIED ECONOMICS
DEPARTMENT OF ECONOMICS

**Emergence of knowledge networks in new
fields: The case of Artificial Intelligence**

Styliani Theodorou Alevra

Advisor: Dr. Yeoryios Stamboulis, Associate Professor

Volos, 2023

Emergence of knowledge networks in new fields: The case of Artificial Intelligence

Abstract

New fields of knowledge such as Artificial Intelligence usually emerge from activities of scientists previously identified with other, pre-existing and established fields. As the number of researchers and publications in the field rise, the field becomes established and research paradigms are founded. In this study I explore how the co-authorship network in the field of Artificial Intelligence has evolved over the 51 year period from 1970 to 2020. I focus on year 2015 and study the network's structure at that time point. Finally, I probe whether and how the different dimensions of proximity affect scholars' effectiveness. Results demonstrate that co-authorship has been the dominant trend since the early 1970s. In 2015 scholars who work in China produced almost 30% of the publications. The co-authorship network is fragmented and it facilitates the entrance of new actors. The dimensions of proximity are strongly and significantly correlated, while organizational and social proximities achieve perfect negative correlation.

Keywords: Artificial Intelligence, co-authorship networks, knowledge networks, innovation networks, scientometrics, proximity.

Περίληψη

Νέα επιστημονικά πεδία όπως η Τεχνητή Νοημοσύνη συνήθως αναδύονται από τις δραστηριότητες επιστημόνων που προηγουμένως εμβάθυναν σε διαφορετικά, πιο καθιερωμένα γνωστικά αντικείμενα. Καθώς ο αριθμός των ερευνητών και των δημοσιεύσεων στο αντικείμενο αυξάνονται, το τεχνολογικό πεδίο παγιώνεται και δημιουργούνται πρότυπα συνεργασίας στην έρευνα. Στην παρούσα μελέτη διερευνώ πως εξελίχθηκε το δίκτυο γνώσης με δρώντες όσους έχουν επιστημονικές δημοσιεύσεις στον τομέα της τεχνητής νοημοσύνης σε μια περίοδο 51 ετών, από το 1970 έως το 2020. Εστιάζω στο έτος 2015 και μελετώ τη δομή του δικτύου τη συγκεκριμένη χρονική στιγμή. Τέλος, κάνω μια πρόβλεψη, για το αν και με ποιο τρόπο οι διάφορες διαστάσεις της εγγύτητας επηρεάζουν την αποτελεσματικότητα των συνεργασιών μεταξύ ερευνητών. Τα αποτελέσματα δείχνουν πως οι συνδημοσιεύσεις είναι κυρίαρχη τάση στο συγκεκριμένο κλάδο. Το δίκτυο είναι κατακερματισμένο και ευνοεί την είσοδο νέων ερευνητών. Οι διαστάσεις της εγγύτητας είναι ισχυρά και στατιστικά σημαντικά συσχετισμένες, με την κοινωνική και την οργανωσιακή εγγύτητα να επιτυγχάνουν απόλυτη αρνητική συσχέτιση.

Λέξεις κλειδιά: Τεχνητή Νοημοσύνη, δίκτυα γνώσης, δίκτυα καινοτομίας, επιστημομετρία, εγγύτητα.

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ

Δηλώνω πως η παρούσα Διπλωματική Εργασία εκπονήθηκε από εμένα την ίδια, ότι αντιπροσωπεύει τις προσωπικές μου απόψεις και ότι αναφέρω πλήρως όλες τις πηγές στις οποίες ανέτρεξα για την εκπόνηση της διπλωματικής μου εργασίας.

Βόλος, Φεβρουάριος 2023

Contents

Chapter1: Introduction	9
1.1 Aim of the Dissertation	9
1.2 Justification of the Research Topic	9
1.3 Methodological Approach	10
1.4 Structure of the Dissertation	10
Chapter 2: Co-Authorship Networks as Knowledge Networks	12
2.1 Introduction	12
2.2 Co-Authorship Networks as Knowledge Networks	12
2.2.1 <i>The Concept of Knowledge Networks</i>	12
2.2.2 <i>Knowledge Networks in Emerging Technologies</i>	14
2.2.3 <i>Co-Authorship Networks as Knowledge Networks</i>	15
2.3 The role of Proximity in Knowledge Networks	16
2.3.1 <i>Geographical Proximity</i>	16
2.3.2 <i>Organizational Proximity</i>	17
2.3.3 <i>Social Proximity</i>	17
2.3.4 <i>Cultural Proximity</i>	17
2.3.5 <i>The role of Proximity in Knowledge Networks with a high degree OF Interdisciplinarity</i>	18
2.4 The role of Network Structure in Knowledge Networks and the Social Network Analysis	18
2.4.1 <i>Social Network Analysis</i>	18
2.4.2 <i>The Role of Networks Structure in Knowledge Networks</i>	22
2.5 The case of Artificial Intelligence	23
2.6 Empirical Studies	24
2.7 Conclusions	29
Chapter 3: Methodology	30
3.1 Introduction	30
3.2 Data Sample	30
3.3 Descriptive Statistics	33
3.4 Social Network Analysis	33
3.5 The impact of proximity dimensions on author's effectiveness	34
3.5.1 <i>Geographical Proximity</i>	35
3.5.2 <i>Organizational Proximity</i>	35
3.5.3 <i>Social Proximity</i>	37
3.5.4 <i>Cultural Proximity</i>	37

3.5.5 Regressions and Multivariable Regressions Models	39
3.6 Conclusions	40
Chapter 4: Results	41
4.1 Introduction	41
4.2 Results from the Descriptive-Statistic Analysis	41
4.3 Results from the Social Network Analysis	44
4.4 Results from Spearman Correlation	48
4.5 Results from Regression and Multivariable Regression Models.....	49
5. Conclusions	53
5.1 Proposal for Future Research	54
References	56
Appendices	60

List of Tables

Table 3.1: Document Type of Publications.	31
Table 3.2: Deleted Publications.	32
Table 3.3: Data Process.	34
Table 3.4: Types of Organizations.	36
Table 3.5: Organizational Proximity.	36
Table 4.1: The characteristics of the network.	45
Table 4.2: Spearman Correlation.	48
Table 4.3 : VIF Test (Number of Authors, Average Geographical Distance, Average Organizational Distance, Average Social Distance, Average Cultural Distance)	49
Table 4.4: VIF Test (Average Geographical Distance, Average Organizational Distance, Average Social Distance, Average Cultural Distance)	50
Table 4.5: Multivariable Regression (Number of Citations~Number of Co-Authors+Average Geographical Distance+Average Organizational Distance+Average Social Distance+Average Cultural Distance)	51
Table 4.6: Multivariable Regression (Number of Citations~Average Geographical Distance+Average Organizational Distance+Average Social Distance+Average Cultural Distance)	52

List of Diagrams

Diagram 3.1: Cultural map – WVS wave 7	38
Diagram 4.1: Number of scientific publications per year.	42
Diagram 4.2: Co-authorship per number of publications.	43
Diagram 4.3: Number of Authors per country (percentage calculated on the total number of authors).	44

Diagram 4.4: Degree Distribution This chart shows the Degree Distribution. y-axis is the percentage of nodes (out of total nodes), and the x-axis is the degree.	46
---	----

List of graphs

Graph 1: Network Visualisation, with nodes colored according to their degree. 47

List of Appendices

Appendix 1 : Number of Citations~Average Geographical Distance	60
Appendix 2: Number of Citations~Average Organizational Distance	60
Appendix 3: Number of Citations~Average Social Distance	61
Appendix 4: Number of Citations~Average Cultural Distance	61

List of Abbreviations

AI Artificial Intelligence

SNA Social Network Analysis

VIF Variance Inflation Factor

Chapter1: Introduction

1.1 Aim of the Dissertation

The aim of this dissertation is to explore the nature of scientific cooperation in the technological field of Artificial Intelligence (AI). First, I analyze how the trends in scientific collaboration in AI have changed during a 51-year period, from 1970 to 2000. I consider the co-authorship network as a knowledge network. The authors of scientific papers are the actors of the network and their scientific publications are the relations between the actors. I investigate the structure of the network in 2015. Taking into consideration that the effectiveness of this kind of cooperations can be proxied by the number of citations of a publication (Zitt et al., 2000), I examine whether the dimensions of proximity, affect the effectiveness of cooperation.

1.2 Justification of the Research Topic

New fields of knowledge such as Artificial Intelligence usually emerge from activities of scientists previously studying other, pre-existing and established fields. As the number of researchers and publications in the field rise, the field and research paradigms are established.

During the last century, the dominant mode for knowledge creation in the form of scientific publications changed dramatically. Scientific research, which had been a task for solitary individuals (Acedo et al., 2006), transformed into a collaborative endeavor (Bordons et al., 2015). Scientific collaboration facilitates knowledge creation which can lead to sustainable development and economic prosperity (Abbasi et al., 2011). Moreover, scholars benefit from scientific cooperation, since it facilitates their productivity both in quantity and quality (Acedo et al., 2006; Abbasi et al., 2011). Co-authorship is a formal indicator of scientific collaboration (Acedo et al., 2006), so many researchers study collaboration by studying social networks (Abbasi et al., 2011), in which co-authors act as nodes and common publications act as edges. These are known as co-authorship networks. Co-authorship networks can provide us with information about how professional knowledge is being shared efficiently (Abbasi et al., 2012), or what kind of opportunities or constraints the actors will encounter (Borgatti et al., 2018). Moreover, we can study actors' traits such as their performance (Abbasi et al.,

2012), their behavior, or their beliefs (Borgatti et al., 2018). Finally, given that scientific cooperation occurs within rather heterogeneous networks, which include among others academia, business, and government institutions (Etzkowitz & Leydesdorff, 2000), evolving co-authorship networks may act as a proxy for other complex evolving networks, such as the World Wide Web (Barabasi et al., (2002).

Although knowledge networks (Phelps et al., 2012) and cooperation that leads to knowledge creation (Kumar, 2015) have been at the forefront of research interest, little scientific research has attempted to explain the geography of this form of cooperation (Canton et al., 2005; Van Ort et al., 2006). Here, I examine the nodes' traits in regard with several dimensions of proximity.

1.3 Methodological Approach

In this study, I derived data regarding scientific publications in the technological field of Artificial Intelligence (AI) from the Web of Science platform. First, I use descriptive statistics to analyze the transmutations on network size and co-authorship trends for the years from 1970 to 2020.

Next, I focus on year 2015 and based on graph theory, I illustrate the co-authorship network. I use Social Network Analysis (SNA) to examine the network's cohesion and modularity. Furthermore, I examine the network's geography and I identify the countries with the highest percentage of participation.

In the last part, I conduct multivariable regression analysis to probe whether and how different dimensions of proximity between co-authors affect their performance which I proxy through the number of citations.

1.4 Structure of the Dissertation

In Chapter 2, I review the theory on knowledge networks and proximity. Then I present previous empirical researches on topics relevant to the case I study.

In Chapter 3, I summarize the history of AI as a technological field, and I explain how this discipline is related to innovation. Next, I analyze the subpopulation that results from my dataset, and I explain the methodology I followed.

In Chapter 4, I present the results of descriptive statistics, the network's structure, I explain how the various dimensions of proximity characterize the network,

and I explain the results from multivariable regression and step wise regression. In Chapter 5 I summarize the most significant results and make a proposal for future research, in chapter 5.

Chapter 2: Co-Authorship Networks as Knowledge Networks

2.1 Introduction

In this chapter I describe and analyze the theoretical framework of my dissertation. I first describe the characteristics of knowledge networks and their relations to innovation. I also analyze the co-authorship networks as knowledge networks and describe their role in the evolution of emerging technologies. I present the concept of proximity and how it affects knowledge networks. I also analyze Social Network Analysis (SNA), which is the method of analyzing the relations between actors in knowledge networks. I present empirical studies relevant to my own research and conclude with the chapter's conclusions.

2.2 Co-Authorship Networks as Knowledge Networks

Drucker (1969), introduced the term “knowledge economy” to highlight how important is the role of knowledge societies on global wealth. Research suggests that the economic performance, either studied on an individual, an organizational, or a country level, is dependent on knowledge production (Blundell et al., 1999; Furman et al., 2002; Roberts, 1999).

Knowledge Networks are intertwined with innovation. This is because, as Lundvall (1992) argues, in Neo-Schumpeterian economics networks "are fundamental to knowledge production given the collective nature of innovation processes". Besides, again according to Lundvall (1992), innovations are "the result of multiple interactions between elements in an integrated system, i.e. the Innovation System". Thus, the processes of knowledge exchange and learning as well as those of the creation of new knowledge are linked to the concept of the Innovation System (Kudic, 2015). Therefore, knowledge networks are directly related to the Innovation System.

2.2.1 The Concept of Knowledge Networks

A knowledge network created for the purpose of sharing knowledge about innovation (Chaminade et al., 2016). More specifically, Phelps et al (2012) emphasizes that “a knowledge network is as a set of nodes - individuals or higher-level collectives

that serve as heterogeneously distributed repositories of knowledge and agents that search for, transmit, and create knowledge interconnected by social relationships that enable and constrain nodes' efforts to acquire, transfer, and create knowledge”.

Therefore, knowledge networks are related to the processes of creating new knowledge and to the process of interactive learning. Also, knowledge networks have to do with the ability of an organization to store and create new stocks of knowledge (Kudic, 2015). This is happening because organizations can gain access to new knowledge through internal or external knowledge channels (Malerba, 1992).

A knowledge network is made up of the actors (also known as nodes) and the relations between them. Actors of a knowledge network can be organizations (businesses, universities, research centers or institutes etc.), individuals (researchers, patent inventors, patent owners etc.), countries, regions, or states. The relations between network actors usually take the form of collaborations. In research on knowledge networks, researchers' study (Bathelt & Glückler, 2018):

- Inter-organizational relations
- Inter-personal relations

Malerba and Vonortas (2009) state that an Innovation Network has resources, and these resources are distinguished in the following dimensions:

1. The structural dimension which is determined by the location of the contacts of an actor (eg a business) in a network.
2. The relational dimension which highlights the characteristics of the relationships between the actors in the network such as reliability, trust etc.
3. The knowledge dimension which has to do with a shared knowledge base, which facilitates interaction with the network.

Malerba and Vonortas (2009) also emphasize that innovation networks have the characteristic of developing ties between actors. Thus, actors develop strong and weak ties between them. Thanks to these ties the actors of a knowledge network ensure the provision of new information and the exchange of high-quality information, complex information and tacit knowledge.

Therefore, a knowledge network is created by the collaborations or relations developed by network actors, and the purpose of the network is the flow (within the network) of information and knowledge (codified or tacit) related to innovation.

At this point I must emphasize that an innovation network can be developed through research collaboration between different organizations, or through collaborative R&D projects (Balland et al., 2019).

In fact, in the case of generating new knowledge with a high degree of complexity, inputs from other actors are required. This is happening because in the collective activity of knowledge production, actors interact and recombine existing knowledge in new ways. This is observed in the continued increase in collaborative research over time (Balland et al., 2019).

2.2.2 Knowledge Networks in Emerging Technologies

In the literature review by Rotolo et al. (2015) an emerging technology has the characteristics of “radical novelty, relatively fast growth, coherence, prominent impact, and uncertainty & ambiguity”. An emerging technology may also displace an older technology “by providing superior performance or functionality” (Li et al., 2018). But all emerging technologies are not disruptive. An emerging technology can be called disruptive when it exceeds users’ expectations and creates greater value. More specifically, a disruptive technology is related to market opportunities, competition, and adoption (Li et al., 2018). Christensen et al (2004) explain that disruptive innovations make it easier for people to do something that historically required either deep expertise or great wealth. Also, Christensen (2013) notes that a disruptive innovation also marks a major transition since a disruptive technology eventually overturns the dominant technology in the market sector. Research to date shows that knowledge networks play a catalytic role in the emergence of new or disruptive technologies (Rosenkopf, 2000; Van der Valk et al., 2011; Schmidt & Junker, 2016). In fact, Van der Valk et al (2011) emphasize that innovations are created and implemented by networks of interacting organizations and individuals. Emerging technologies require new knowledge, which is often created by recombining existing stocks of knowledge (Tsouri et al., 2022; Grillitsch et al., 2018; Boschma et al., 2012). Furthermore, the production and diffusion of knowledge affects positively the development of technologies (Iammarino & McCann, 2006; Verspagen, 2007).

2.2.3 Co-Authorship Networks as Knowledge Networks

Co-authorship networks, i.e., the networks resulting from the co-authorship of scientific publications, are considered knowledge networks (Kumar, 2015). Abbasi & Altmann (2011) argues that scientific collaborations emerge from and mature within social networks, so they can be influenced by social networks in multiple ways. Co-authorship networks is a subset of knowledge networks where two authors are related to each other when they cooperate to publish a document. In this case of knowledge networks, the actors (nodes) in the network are the authors (researchers) of the scientific publications, and the relations of the network are the scientific collaborations that the authors develop (ie the scientific publications of the authors).

Therefore, the study of co-authorship networks, which are knowledge networks, focuses on scientific collaborations. As Sonnenwald (2007) notes: "Scientific collaborations are interactions taking place within a social context among two or more scientists that facilitates the sharing of meaning and completion of tasks with respect to a mutually shared, super-ordinated goal".

First, the study of co-authorship networks is important because it can reveal the patterns of these relations and more generally the mechanisms that shape the scientific community (Kumar, 2015). More specifically, over the last century, knowledge production in the form of scientific publications has evolved from a solitary to a collaborative endeavor (Acedo et al., 2006; Bordons et al., 2015). This phenomenon is more intense in natural sciences but has steadily been increasing in all disciplines (Acedo et al., 2006).

Co-authorship, as a demonstration of scientific collaboration, allows scholars to combine different talents, to expertise their techniques and both share and create knowledge, while expediting the research process (Bordons et al., 2015; Kumar, 2015). In the macro level it provides information about the whole network, for example it reveals how cohesive or fragmented a knowledge community is. In the micro level, it focuses on the most prestigious authors or popular “gatekeepers” (Kumar, 2015). So, scholars can be taught how to share professional knowledge in the most effective way (Abbasi & Altmann, 2011).

2.3 The role of Proximity in Knowledge Networks

In the 1990s, in France, the Proximity Dynamics group changed the way scholars would address the concept of a simple world: proximity (Torre & Gilly, 2000). Today, proximity research is conducted in different scientific fields such as geography, economics, and management. Proximity is associated with innovation and therefore there is a relationship between proximity and different types of knowledge (Boschma, 2005; Kudic, 2015).

Knoben and Oerlemans (2006) in their attempt to give a short definition of proximity, explain that proximity is “being close to something measured on a certain dimension”. Researchers from various disciplines, started studying proximity as a concept which consists of several different dimensions (Knoben & Oerlemans, 2006). According to Boschma (2005), the concept of proximity can be categorized into the following five basic dimensions:

1. Geographical Proximity
2. Organizational Proximity
3. Social Proximity
4. Cognitive Proximity
5. Institutional Proximity

Based on the five dimensions of Boschma (2005), I will analyze below the ones those I use in my research.

2.3.1 Geographical Proximity

Geographical proximity is a dimension of proximity which is related to spatial distance (Knoben & Oerlemans, 2006), and is defined as “the physical distance between actors in absolute (e.g. miles) or relative terms (e.g. travel time)” (Boschma, 2005).

Geographical proximity favors innovation and knowledge transfer (Knoben & Oerlemans, 2006). This is more evident when tacit knowledge is involved (Torre & Gilly, 2000). Moreover, despite globalization, geographical proximity favors network formation (Boschma & Frenken, 2010). Though, in some cases, if a network is characterized by too high levels of geographical proximity, it can exclude outsiders (Boschma, 2005).

2.3.2 Organizational Proximity

Organizational proximity is “the extent to which relations are shared in an organizational arrangement, either within an organization, or between organizations. It involves the rate of autonomy and control that can be exerted in organizational arrangements” (Boschma, 2005).

Organizational proximity reduces uncertainty and opportunism, so it has a positive impact on the establishment of innovation networks (Boschma & Frenken, 2010). Moreover, it is positively correlated to the external knowledge process (Boschma, 2005). However, too much organizational proximity might indicate a lack of flexibility which is necessary for innovation (Blanc & Sierra, 1999).

2.3.3 Social Proximity

Social proximity is defined in terms of “socially embedded relations between agents at the micro-level” (Boschma, 2005). These relations involve friendship, kinship, and experience through repeated reaction, which create trust.

Social proximity promotes mutual knowledge exchange and favors the formation of networks. In networks characterized by high social proximity, two nodes get related if they are introduced to each other by a third node, to whom they are both connected. This trait is known as “closure” (Boschma & Frenken, 2010). However, too much social proximity can have a negative effect, as it can lock the nodes inside a network which does not promote their own innovation (Boschma, 2005).

2.3.4 Cultural Proximity

Cultural Proximity is a subset of Institutional Proximity (Boschma, 2005). Institutional proximity involves formal and informal institutional frameworks which include laws, cultural norms, habits and values. Different countries or regions are usually characterized by different institutional frameworks (Knoben & Oerlemans, 2006). Cultural proximity refers to similarities in informal frameworks like habits and values which can be affected by cultural differences. Cultural proximity favors trust formation, communication, and interaction (Nilsson, 2019). Given that most empirical research examine institutional proximity under the prism of formal frameworks, I judge necessary to examine cultural proximity instead, so as to explore if and how the cultural similarities or differences can affect collaboration in an emerging knowledge network.

2.3.5 The role of Proximity in Knowledge Networks with a high degree of Interdisciplinarity

At this point it is important to mention the role these dimensions of proximity play in R&D networks (i.e. knowledge networks) with a high degree of interdisciplinarity.

According to Rekers and Hansen (2015), interdisciplinary research is characterized by low cognitive proximity (or big cognitive distance) because the collaborators come from different disciplines. They also note that low organizational proximity (or high organizational distance) can be expected in these knowledge networks because actors operate under different incentive structures. Thus, due to this heterogeneity of actors there is the “two-community problem” (Rekers & Hansen, 2015). According to Rekers and Hansen (2015), this problem can be overcome with high geographical and social proximity. High social proximity means frequent interaction facilitated by short geographical (or physical) distances (Rekers & Hansen, 2015).

2.4 The role of Network Structure in Knowledge Networks and the Social Network Analysis

Co-authorship networks, which as I mentioned above are knowledge networks, are studied as social networks (Kumar, 2015). Because of this, an analysis method for studying co-authorship networks is the Social Network Analysis (SNA) (Acedo et al., 2006).

With the SNA method, the structure and cohesion of the network can be studied, and it can be examined how central the position of each actor in the network is. Also, by analyzing the network structure with SNA, the openness to network expansion and how the level of network cohesion affects the flow of information and knowledge within the network can be explored. It can also be studied how the central position of the actors in the network affect the innovative performance of the network.

2.4.1 Social Network Analysis

A field of mathematics called Graph Theory is used to visualise a network, but also to study relations between network actors. Also Graph Theory for example has been used to solve the "maximal flow problem" and the "shortest path problem"

(Rosenkopf, 2000). Graph Theory is part of Social Network Analysis (SNA). SNA is an analysis method that helps to study the relations that develop between actors in an innovation network. Thus, with SNA we can investigate the structure or cohesion of an innovation network (more specifically the size and density of the network) but also the transfer of knowledge between the actors participating in the network.

2.4.1.1 Graph Theory

Graph Theory is about sets of elements (individuals, organizations, etc) and the relations between them. The elements are called "points" (nodes) and the relations "lines" (edges). From a matrix describing relations between elements, a graph of points connected by lines can be created. The presence of a relation is represented by a line (Scott, 2017).

The graphs are divided into two main categories:

- Directed Graphs
- Undirected Graphs

In the case of directed graphs the relations are directed from one factor to another. In directed graphs the line is in the form of an arrow, and the direction of the arrow represents the direction of the relation. On the contrary, in undirected graphs (as in the case of the knowledge network I am studying) each line does not have the form of an arrow, i.e. the lines are not directed (Scott, 2017).

2.4.1.2 Network Structure

Network structure includes the size of the network (or the graph that represents it), the network diameter and the concept of network coherence. The size of a network is described by the set of actors participating in the network (i.e. by the sum of the points of the graph or the nodes, and denoted by V or N) and by the set of ties or relations between them (from the sum of the lines of the graph or the connections (edges) of the network, and is denoted by E).

Thus we can describe a graph and its size as (Aggarwal, 2011):

$$\text{Network Size: } G = (V, E) \quad [2.1]$$

Where:

V = the total sum of the vertices of a graph (or the actors or nodes of the network)

E = the total sum of the lines of a graph (or the links or relations of the network).

The Network Diameter is the shortest distance between the two farthest points (or nodes) of the graph (Scott, 2017). To find if the Network Diameter is relatively long or short we need to compare it with the Average Path Length (Backstrom et al, 2012; Perez & Germon, 2016).

The network cohesion has to do with the general level of connectivity of the graph and shows how connected points are to each other (Van Der Valk et al., 2011; Laperche & Sommers, 2010; Scott, 2017; De Nooy et al., 2005). The high cohesion of a social network means that the network has many ties (De Nooy et al., 2005). Thus, cohesion indicates how sparse or dense a network is (Van Der Valk et al., 2011).

A measure that describes the cohesion of a network is the Density of Network or Graph (De Nooy et al., 2005; Scott, 2017; Laperche & Sommers, 2010). The density of a graph is the number of existing lines (i.e., connections), expressed as a percentage of the maximum possible number of lines (De Nooy et al., 2005) (Wasserman & Faust, 1994). The density of a graph describes the coherence of a network, because if the density has a relatively low value, this may mean that the network may lack mutual connections between network actors (Scott, 2017). The density of a graph is calculated by the formula (Scott, 2017; Wasserman & Faust, 1994; De Nooy et al., 2005):

$$\text{Network Density: } D = \frac{l}{n(n-1)/2} = \frac{2l}{n(n-1)} \quad [2.2]$$

Where:

n: the total sum of points of a graph

l: the total sum of a graph

$n(n-1)/2$: the maximum possible number of lines of the graph.

The maximum value that the Density of a Graph can take is 1.0 (high density) and the minimum is 0 (low density) (De Nooy, et al., 2005; Scott, 2017; Wasserman & Faust, 1994).

The density of a graph, however, depends to a large extent on the size of the network (Laperche & Sommers, 2010; Scott, 2017; De Nooy et al., 2005). This is

because the Density of a Graph is inversely related to the size of the network (De Nooy et al., 2005; Laperche & Sommers, 2010; Coulon, 2005). In this case it is better to look for the number of links of each node, that is the Degree of each point (De Nooy et al., 2005). After all, the density itself is directly related to the degree, since the higher the degrees of the points of a graph, the higher the density of the network will be (Scott, 2017; De Nooy et al., 2005). So, one more way to analyze network cohesion is Degree Distribution. More specifically the Degree Distribution is a diagram that the x-axis has the degrees in a ascending order, and the Y-axis indicates how many nodes have a given degree. This shows how many nodes have each possible degree (Golbeck, 2013).

So, when we need to compare the cohesion between networks, which vary considerably in size, then we can use the Average Degree (Laperche & Sommers, 2010; Coulon, 2005). The mathematical formula that calculates the Average Degree is (Scott, 2017; Laperche & Sommers, 2010; Coulon, 2005):

$$\text{Average Degree} = \frac{2l}{n} \quad [2.3]$$

Where:

n = the total sum of points of a graph

l = the total sum of a graph

2.4.1.3 Centrality

The Central Position of a point (node) literally means that this point is at the center of many connections, i.e. it has many direct contacts with other points (nodes). A fairly common measure of the centrality of a point or node is the degree of the point itself (Scott, 2017; Aktamov & Zhao, 2014). More specifically, if a point has a high degree then this means that it is connected to many other points or nodes, so it is better connected. In short, in this case we could say that this point (or node) has a central position in the network, or that this node "is more central" (Scott, 2017). So this kind of centrality is called Degree Centrality (Scott, 2017, Golbeck, 2013).

2.4.1.4 Fragmented nature of Social Network

Fragmented nature of a Social Network is observed when the Social Network has a very large and dense group of low degree nodes (which is called the "core") and far from the core through a chain of low-degree nodes there are few nodes with the highest degrees (these nodes are located at the “periphery” of the network or graph). Core nodes are the majority of the total number of nodes in the network (Golbeck, 2013). Often, Fragmented Social Networks have a very high Network Diameter compared to the Average Path Length (Rossoni & Hocayen-da-Silva, 2008; Bordons et al., 2013; Pauli et al., 2020).

2.4.2 The Role of Networks Structure in Knowledge Networks

The cohesion of an innovation network is a very important concept and affects its innovative performance. More specifically, when the Network Diameter is relatively long, usually is an indication that information or knowledge may spread relatively slowly within the network. This could be attributed to the fragmented nature of the network (Rossoni & Hocayen-da-Silva, 2008; Bordons et al., 2013; Pauli et al., 2020).

Also, Density has been associated with the capacity for innovation or the generation of trust (Forti et al., 2013). This can happen because in an overly sparse innovation network, the limited number of relations can limit the performance of the network. But also the very high cohesion, which has an excessively dense network of innovation, reduces the variety of knowledge and therefore the opportunities for new combinations of relations, experiences and knowledge (Van Der Valk et al., 2011). In conclusion, regarding the effect of density on scientific productivity for the case of Co-Authorship Networks, in low-density networks productivity was higher but also decreased in networks with highly redundant relations (Rossoni & Hocayen-da-Silva, 2008).

Thus, often in Co-Authorship Networks the high fragmentation may indicate low cohesion and therefore greater openness to network expansion (Rossoni & Hocayen-da-Silva, 2008; Pauli et al., 2020; Bordons et al., 2015). This expansion may mean also more possibilities for co-authoring (Pauli et al., 2020).

The concept of Centrality of the actors of an innovation network describes the interaction pattern of the network. The position occupied by a node (individual or organization) in an innovation network gives it the possibility to acquire new

experiences but also access to external information and knowledge (Antonelli, 2011; Aktamov & Zhao, 2014). Therefore, the centrality in general, but also the Degree Centrality in particular, can affect the innovative performance of the network (Antonelli, 2011).

For example, in the case of a business, its central position in an innovation network increases the chances of having access even to desirable strategic resources (such as knowledge, skills, laboratory equipment, etc.). Thus, these resources will fuel the business innovative activities by providing the external information necessary to generate new ideas (Aktamov & Zhao, 2014).

In the case of a knowledge network of individuals and more specifically in the case of a Co-Authorship Network (as in the case examined by this empirical research) the centrality has been shown to be an important predictor of a publication's impact (Abbasi et al., 2012; Hoffmann et al., 2014, Pauli et al., 2020). The fact that centrality has been positively associated with an increase in the number of citations of an article stems from the fact that the privileged position in the network allows the articles to be divulged to a larger audience (Acedo et al., 2006; Li et al., 2013; Hoffmann et al., 2014; Pauli et al., 2020).

2.5 The case of Artificial Intelligence

The term Artificial Intelligence was officially used for the first time in 1955 by John McCarthy from Dartmouth College, Marvin Minsky from Harvard, Nathaniel Rochester from IBM, and Claude Shannon from Bell Laboratories, while organizing “The Dartmouth Summer Research Project on Artificial Intelligence” (McCarthy et al., 1955). That workshop caught researchers' attention and established Artificial Intelligence as we know it today (Brooks, 1991), while McCarthy is recognized as the father of Artificial Intelligence (Ramajaran, 2014). McCarthy was the first to define Artificial Intelligence. According to him, “Artificial Intelligence is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but Artificial Intelligence does not have to confine itself to methods that are biologically observable” (McCarthy, 2007).

From the early 1970s, AI research began to intensify (McCarthy, 1984; Haenlein & Kaplan, 2019; Muthukrishnan et al., 2020). During the first decade of 21st

century, significant improvements in data storage and graphical processing units led to the birth of Deep Learning in 2006 (Muthukrishnan et al., 2020). This period can signify the third wave of Artificial Intelligence. Steady developments in computation power helped researchers overcome previous challenges, and in 2015 a great, unexpected achievement took place: a computer program managed to beat the world champion in the game Go. The field has been thriving since then.

AI promotes innovation in multiple disciplines, as it simplifies complex tasks which traditionally would require technological knowledge, time, and wealth in order to be completed (Christensen et al., 2004). AI has not only changed ordinary people's lives, but it has transformed different types of organizations. In universities we can see examples that vary from the use of virtual teaching assistants to automated scientific research conduction. In corporations AI is used in HR to scan CVs, in customer service in the form of bots, in finance in the form of robo-advisors, in retail to manage shipping, and almost every other discipline. Finally, governments use AI for multiple purposes, which vary from managing traffic lights to scheduling which restaurants to visit for health inspection based on social media information (Kaplan & Haenlein, 2019). Kalakatar et al. (2020) claim that AI is more than just an innovative discipline, as it plays a key role in innovation process.

2.6 Empirical Studies

From studying the literature review, I notice that several empirical researches have been conducted on co-authorship networks.

Newman (2001a) studies separately co-authorship networks formed in different disciplines, which include physics, biomedical research, and computer science. He uses data from four different databases: Medline, Stanford Public Information Retrieval System (SPIRES), Networked Computer Science Technical Reference Library (NCSTRL), and Los Alamos e-Print Archive. For each database a different co-authorship network is built, except for Los Alamos e-Print Archive, in which he identifies papers published on four different specialties within physics and builds four different networks. The research examines seven co-authorship networks in total which illustrate collaboration between co-authors in a five-year period, from 1995 to 1999. For each network he conducts statistical analysis, where he identifies the number of authors, the mean papers per author, the mean authors per paper, the average number

of collaborators per author, the size of giant component, the size of the second largest cluster, and the clustering coefficients. The result show that statistics measures follow a power-law form. In all formed networks exists a giant component. Network comparison shows the average number of collaborators per author is higher in experimental disciplines than in theoretical disciplines. Finally, the degree of network clustering is significantly lower in biomedicine than in other disciplines.

Newman (2001b) supplements the above research, by analyzing the average degrees of separation for co-authorship networks described above. The results show that the typical distance, illustrated through the shortest path, between a random pair of researchers is about 2. In addition, results demonstrate distances vary logarithmically with the number of scientists in the network.

Acedo et al. (2006) research is the first academic attempt to combine the two different approaches towards co-authorship till then, which are analysis of the reasons that promote author cooperation, and the results of this kind of cooperation, and network analysis. The research is conducted in the field of management and organizational studies. The dataset includes 11022 published documents, retrieved from the top ten ranked scientific magazines in the field. The database to retrieve the data from is Social Science Citation Index, from Web of Science.

In the first part of analysis, the researchers recognize 10176 authors, 6675 of which published only once. Of those who published only once, 5515 participated in co-authorships. Of those who published more than once, 2006 published with co-authors, and 1308 published both as co-authors and as solo authors. Logistic regressions results show that there is a non-linear, growing trend towards co-authorship, and that co-authorship leads to longer papers. As far as highly quantitative journals are concerned, they tend to include a greater number of co-authored papers compared to other journals. Scheffe test shows that the most important papers in the field are published by two co-authors who differ significantly from the other categories. Finally, the results demonstrate that co-authorship collaboration in this occurs mostly either among authors who are all highly esteemed in the field, or between novices and prominent authors, which indicate collaboration of mentoring.

In order to conduct SNA, authors created a weighted co-authorship network. In the micro level, network is characterized by low density (0.0002) and a high degree of

disconnection. In particular, there are 761 components of two authors, 280 components of three, 138 of four, 67 of five, 26 of six, and 23 of seven. The giant component accounts for 45.45% of all authors. Degree centrality has a mean value of 2.86 and a maximum value of 85. The mean neighborhood size is 2.43. The normalized betweenness centrality has a mean value of 0.017, and a maximum value of 2.53.

In the macro level, three groups are formed. The first group involves the isolated nodes, which are represented by the authors who have not participated in co-authorships and are not examined by this research. The second group involves the giant component, which is consisted of 4625 authors. The third group involves 726 authors that are characterized with a greater degree of centrality. North American authors occupy the most predominant positions within the network, but due to network's cohesion, it is implied that knowledge diffusion does not depend exclusively on them. Finally, the most efficient authors are interconnected with a high number of direct links.

Abbasi et al., (2011) used SNA, Spearman rank correlation test, and multiple regression analysis in order to explore the relation between various SNA measures and researchers' performance. They collected data on the information schools of five universities from google scholar. The formed network includes 1895 authors, 2139 publications, and 5310 co-authorships. The period covered was from 2001 to 2005. Researchers' performance is measured according to four different measures: the publication count, the citation count, the h-index, and the g-index. Spearman rank correlation test shows that the correlation coefficients between normalized degree centrality, normalized betweenness centrality, average ties strength, efficiency, and the four performance measures are high, unlike normalized closeness centrality and normalized eigenvector centrality, where no or very low correlation exists. Multiple regression analysis, after consecutive extract of not important independent variables, shows that normalized degree centrality, average ties strength, and efficiency have a positive impact on researchers' performance, while normalized eigenvector centrality has a negative impact on researchers' performance.

Abbasi et al., (2012) research which co-authorship network properties are associated with scholars' performance in the field of "Information Science and Library Science". The network build is based on data of top 9 journals of the field, published from 2000 to 2009, and are extracted from Scopus database. Sample analysis reveals a

network of 8069 authors, who participated in 12783 co-authorships, published 4837 papers, and received in total 93660 citations. The results show that a researcher's network structure correlates to their performance. In particular, Spearman correlation test shows significant strong correlation between degree centrality and scholar's performance, and between ego-betweenness and scholar's performance. Density is negatively associated with performance. Moreover, results show that researchers who maintain strong ties, tend to perform better than the ones who have more but weaker ties inside the network. Translated into my research's terms, the last finding may implies that social proximity is positively correlated to a researcher's performance.

Uddin et al. (2012) explore the co-authorship network formed in the field of steel structures, based on papers published from 1990 to 2009. Their source is Scopus. The network is composed of 5251 authors, who have published 2857 articles, and have affiliations in 1324 organizations from 83 countries. The initiative is to explore the trend and efficiency of international scientific collaboration. The network is analyzed in two different levels, which include individual node level, and network level for each year. According to their findings, the average number of authors per paper, the average number of collaborators per author, the number of collaborators who live in different countries from their partners, and the number of countries that are included in the network, are increasing gradually over time. Moreover, considering number of citation as a proxy of efficiency, they prove that the number of co-authors has an impact on published document's efficiency. Papers published by partners who live in different countries are characterized by higher efficiency compared to the ones published by authors from the same country, and this efficiency increases sharply in time. Finally, in the network created where the countries act as nodes and the common publication as edges, the degree centrality and betweenness centrality are positive correlated with the average efficiency of papers.

Liu & Xia, (2015) studied the structure and evolution of the co-authorship network formed in the field of "evolution of cooperation". The database used was Web of Science, and they examined a total of 2583 papers, published from 1961 to 2013. The structure of the network evolved gradually, initially from a "core periphery" to "chained communities" and finally to a modular structure which contains a growing cohesive core. In the cohesive core, there is a clear existence of a giant component. The giant component is examined separately, and the findings illustrate that it is highly

modular, and highly transitive, while it is characterized by a small average shortest-path-length, characteristics that classify the sub-network to small-world networks. Moreover, the giant component contains a hierarchical structure which can be classified into two layers, “grassroot” nodes, which include 90% of the nodes, and “elite” nodes, which include the remaining 10% of the nodes. Authors conclude that this structure could reflect collaboration pattern in a typical interdisciplinary field.

Rossoni and Hocayen-da-Silva (2008) study the co-authorship network in the field of Information Management in Brazil during the period 2002-2006. More specifically, they investigate the structure of the network, which consists of 359 authors-researchers. They collect data on 228 papers from the National Association of Graduate Studies and Research in Administration (ANPAD) of Brazil, and use the SNA method. From their results, they first notice that this is a co-authorship network with high fragmentation, since out of the 87 components of the network, only 11 are made up of more than 5 authors. They conclude that the fragmented nature of the network indicates a greater openness to network expansion. They also point out that fragmentation can prevent research networks from closing on redundant links.

Wanzenboeck et al (2014) study R&D collaborations in the EU. More specifically, they investigate with the SNA method the knowledge networks developed by R&D collaborations for patents (co-patent network), for scientific publications (co-authorship network) and in the European framework programs (FPs). From the results of their research they find that the nature of the co-authorship network is fragmented and this does not prevent the diffusion of knowledge in the network.

Bordons et al (2015) study co-authorship networks in the fields of Nanoscience, Pharmacology and Statistics in Spain over a three-year period (2006-2008). More specifically, they investigate the relationship between the research performance of scientists and their position in co-authorship networks. They collected their data from the Web of Science database and used the SNA method. From their results, they initially observe that the co-authorship networks of the first two fields are extremely dense, while on the contrary the Statistics network is much less coherent. They found also that the co-authorship network of the field of Statistics is highly fragmented. They conclude that this very low cohesion and high fragmentation of the Statistics network indicates a greater openness to network expansion.

Pauli et al (2020) study the co-authorship network of Brazilian postgraduate programs in the field of management. They collect their data from CAPES and use the SNA method to examine the network's structure. Their results show that the nature of the network is fragmented. Thus, they conclude that the network's high fragmentation is an indicator of low cohesion, which may facilitate network's expansion. They also emphasize that this expansion may mean more possibilities for co-authoring.

2.7 Conclusions

Studying previous bibliography, I conclude that co-authorship networks may be considered knowledge networks. Therefore, network structure and proximity dimensions play an important role in the evolution and performance of knowledge networks.

Also, I notice that several empirical studies have been conducted to investigate the influence of network structure on co-authorship networks. There are also empirical studies which investigate how the production of scientific publications evolves in a number of technological fields (Newman, 2001a).

There are several empirical studies examining how proximity affects knowledge networks that explore inter-organizational collaborations. But from the review of the literature, I ascertain that it has not been investigated how the different dimensions of proximity affect both the performance (citations) and the evolution of co-authorship networks, specifically.

Chapter 3: Methodology

3.1 Introduction

For this study, I downloaded data regarding scientific publications on AI from Web of Science, for the 51 year period from 1970 to 2020. In this chapter I describe and analyze my sample and the methods I use. More specifically, I analyze:

- The technological field of Artificial Intelligence.
- My data sample.
- The method of Descriptive Statistics.
- The method of Social Network Analysis.
- How, I calculate the dimensions of Proximity.
- The Multi-Variable Regression Model.

3.2 Data Sample

This study is based on data about scientific publications on the field of Artificial Intelligence that I have gathered from Web of Science of Clarivate. The key words my search was based on are “Artificial Intelligence”, “Machine Learning”, “Learning Algorithm”, “Deep Learning”, and “Pattern Recognition”. The initial data consisted of 539,333 documents which have been published from 1970 to 2020. Because our data were downloaded manually in a group of 1,000 documents each time, I need to mention that the gathering period lasted from 16 February 2022 to 14 March 2022, and there might be slight differences in the number of citations that I could not notice during this period. The data we used for each document include the document’s title, a unique WOS id, the authors’ full names and addresses, the language in which the document was written, the affiliations, and the times each document had been cited in all databases.

My first step is to clean the data according to variable “document type”, to keep those that would help me evaluate the cooperation between the (co)authors. For this purpose, I keep a total of 405,926 documents, which include 342,163 articles, 339 articles – book chapters, 301 articles – data papers, 44,984 articles – proceedings papers, and 18,478 reviews. The items cleaned according to their document type, are labeled as it follows: 1 abstract of published item, 2 art exhibit review, 869 article –

early access, 2 article – early access – retracted publication, 1 article – proceedings paper – retracted publication, 1 article – publication with expression of concern, 62 article – retracted publication, 6 article – withdrawn publication, 16 bibliography, 66 biographical item, 861 book review, 3 book review – early access, 822 correction, 82 correction – addition, 1 correction – book chapter, 10 correction – early access, 2 database review, 51 discussion, 7,088 editorial material, 4 editorial material – book chapter, 4 editorial material – early access, 2 expression of concern, 1 fiction – creative prose, 6 film review, 3 hardware review, 19 item about an individual, 4 item withdrawal, 1,682 letter, 6 letter – early access, 8,316 meeting abstract, 264 news item, 774 note, 1 poetry, 36 reprint, 32 retraction, 420 review – book chapter, 32 review - early access, 3 review – retracted publication, and 22 software review.

Table 3.1: Document Type of Publications.

Final Sample According to "Document Type"	
Document Type	Count of Documents
Article	342163
Article - Book Chapter	301
Article - Data Paper	44984
Review	18478
TOTAL	405926

Source: Own processing

From the raw data, I extracted the unique WOS id and the authors' full names. I used the names to count the number of co-authors per document. The dataset that occurred is used in the first part of the study. For the second part (SNA) and the third part (multiple regression analysis) I focus on year 2015. The dataset for this year after the first cleaning consisted of 16008 publications.

Table 3.2: Deleted Publications.

Deleted Items According to "Document Type"	
Document Type	Count of Documents
Abstract of Published item	1
Art Exhibit Review	2
Article - Early Access	869
Article - Early Access - Retracted Publication	1
Article - Proceedings Paper - Retracted Publication	1
Article - Retracted Publication	62
Article - Withdrawn Publication	6
Bibliography	16
Bibliographical Item	66
Book Review	861
Book Review - Early Access	3
Correction	822
Correction - Addition	82
Correction - Book Chapter	1
Correction - Early Access	10
Database Review	2
Discussion	51
Editorial Material	7088
Editorial Material - Book Chapter	4
Editorial Material - Early Access	4
Expression of Concern	2
Fiction - Creative Prose	1
Film Review	6
Hardware Review	3
Item about an Individual	19
Item Withdrawal	4
Letter	1682
Letter - Early Access	6
Meeting Abstract	8316
News Item	264
Note	774
Poetry	1
Reprint	36
Retraction	32
Review - Book Chapter	420
Review - Early Access	32
Review - Retracted Publication	3
Software Review	22
TOTAL	21575

Source: Own processing

Since in recent years there have been noticed published items that include an extremely big number of co-authors, I have kept the documents that were created by up to 8 co-authors and represent 94 per cent of all published items for year 2015. This is a necessary step, as the existence of a large number of co-authors indicates an incidence of co-authorship and cannot give us information about relationships (Acedo et al., 2006). Then I had to exclude the publications which did not contain information for every co-author, a phenomenon that would become more and more intense as the number of co-authors increased. This procedure gave us a dataset of 13173 publications. Later, I had to take one final cleaning step, which I describe in detail at chapter 3.6.2. In order to measure cultural distance, I had to exclude 24 countries from the dataset. This reduced the sample by 64 publications, leaving us a final sample of 13109 publications (81,9% of all publications in the database for the year 2015).

3.3 Descriptive Statistics

In the first part of the research, I use descriptive statistics to explore the trends of international collaboration. Similar studies have been published in the past (Newman, 2001a; Acedo et al., 2006; Udin et al., 2012). I conduct an empirical analysis based on data of 51 years to probe how the social network has emerged and grown. For each year I examine the transmutations in the number of authors who publish in the field of Artificial Intelligence, the number of published documents, and the transmutations in the number of co-authors per article.

Next, for the year 2015, which is the year I focus on at the second and third part of the study, I examine the volume of published documents per country.

3.4 Social Network Analysis

In the second part, following the previous bibliography (Newman, 2001a; Newman, 2001b; Newman, 2001c; Acedo et al., 2006; Liu & Xia, 2015), I analyze the structure of the knowledge network for the 2015 with the SNA method. I investigate the level of network cohesion, and whether this leads to closing in on redundant links. More specifically, using Gephi software I create a network visualization and study the density, diameter, modularity, and degree distribution of the network.

3.5 The impact of proximity dimensions on author's effectiveness

In order to explore the degree that different aspects of proximity affect the effectiveness of the published items, I conduct Spearman Correlation Test and Multiple Regression Analysis, a method used to explore the relationship between co-authorship network properties and authors' performance (Abbasi et al., 2011; Abbasi et al., 2012).

The raw dataset had to be processed. The first step was to extract only the information that was necessary for the research. This includes for each document the unique WOS id, the citations count, the language in which the document was written, the authors' full names, and the full addresses. From the authors' names I counted the number of (co)authors that participated in each publication. Then, each address was processed in Power Query in order to extract for each author the name of the organization they work for, and the city and country they live in. In those cases where an author would declare multiple addresses, I kept the address that matched their co-authors, provided I had that option. In the rest of the cases I kept the first address each author had declared. The process is summarized in the following table.

Table 3.3: Data Process.

Data Extracted for each document	Contribution to Model Creation
Unique WOS id	Distinguish between the documents
Citation Count	Dependent Variable
Authors' Full Names	• Calculation of the independent variable "Number of (co)authors" • Calculation of the independent variables regarding social proximity •
Addresses	• Derived Organization to calculate the independent variable regarding organizational proximity • Derived City to calculate the independent variable regarding geographical proximity • Derived Country to calculate the independent variable regarding cultural proximity

Source: Self Processing

3.5.1 Geographical Proximity

To define geographical proximity between co-authors, I measure their geographical distance. I isolate the cities from the authors' addresses and then produced latitude and longitude with Microsoft Office Power Query. Then I calculate geographical distance according to the type:

$$GeoDist = ACOS(COS(RADIANS(90 - Latitude_1)) * COS(RADIANS(90 - Latitude_2)) + SIN(RADIANS(90 - Latitude_1)) * SIN(RADIANS(90 - Latitude_2))) * COS(RADIANS(Longtitude_1 - Longtitude_2))) * 3671 \quad [3.1]$$

Geographical Distance (GeoDist) is a numeric variable, which takes values from 0 km for the closest partners, to 19947.58 km for the most distant ones. For each published document present in my research, I calculate the maximum geographical distance (MaxGeoDist), the minimum geographical distance (MinGeoDist), and the average geographical distance according to the type:

$$AvGeoDist = \Sigma(GeoDist_{ij})/n \quad [3.2]$$

Where n is the number of pairs formed between the (co)authors, and for a total of k (co)authors can be calculated through the combinations of second class of k elements according to the type:

$$n = \frac{k!}{2!(k-2)!} \quad [3.3]$$

3.5.2 Organizational Proximity

In my dataset, I identified 11 different types of organizations that participate in the knowledge network. These include universities, colleges, academies, military academies, research institutions, research centers, hospitals, government organizations, international organizations, non-profit organizations, and business.

In order to simplify the procedure, and given that organizational distance between some of these types is negligible, I group organization types in 5 clusters, according to the following table:

Table 3.4: Types of Organizations.

Universities	Research Centers & Institutions	Government	Business	International	Non - Profit
Universities	Research Centers	Hospitals			
Colleges	Research Institutions				
Military Academies	Academies				

Source: Own Processing

Taking into consideration the fact that cooperation might take place either within the same organization, or in different organizations, which often are classified in different types, I create a numeric variable $OrgDist_{ij}$ to measure organizational distance between partners. $DistOrg_{ij}$ takes values from 0 for the closest (co)authors, to 5 for the most distant ones.

First, I examine whether the pair of partners belongs to the same organization. In this case, organizational proximity is the highest possible and takes the value 0.

If the cooperation takes place between co-authors who belong in different organizations, I measure organizational proximity based on the following table:

Table 3.5: Organizational Proximity.

Στήλη1	Universities	Research Centers & Institutions	Government	Business	International Organizations	NonProfit Organizations
Universities	1	2	3	5	4	3
Research Centers & Institutions		1	3	5	4	3
Government			1	5	3	4
Business				1	5	5
International Organizations					1	4
Non-Profit Organizations						1

Source: Own Processing

In my multiple regression model, for each published document I calculate the average organizational distance (*AvDistOrg*), which is calculated by the type:

$$AvDistOrg = \frac{\sum DistOrg_{ij}}{n} \quad [3.4]$$

Where n is the number of (co)authors pairs formed in each document.

3.5.3 Social Proximity

For social proximity, I follow the proxy used by several researchers (Brokel & Boschma, 2012; Balland et al., 2013; Heringa et al., 2016), which is based on the number of previous cooperation. I define a numeric variable *DistSoc_{ij}* to measure the social distance between each pair of partners, where *DisSoc_{ij}* equals to the number of times the pair i,j has participated in a scientific publication during the year 2015. *DistSoc_{ij}* is a numeric variable, which takes values from 0 for a pair with the lowest social proximity, to 9 for the pair with the highest one.

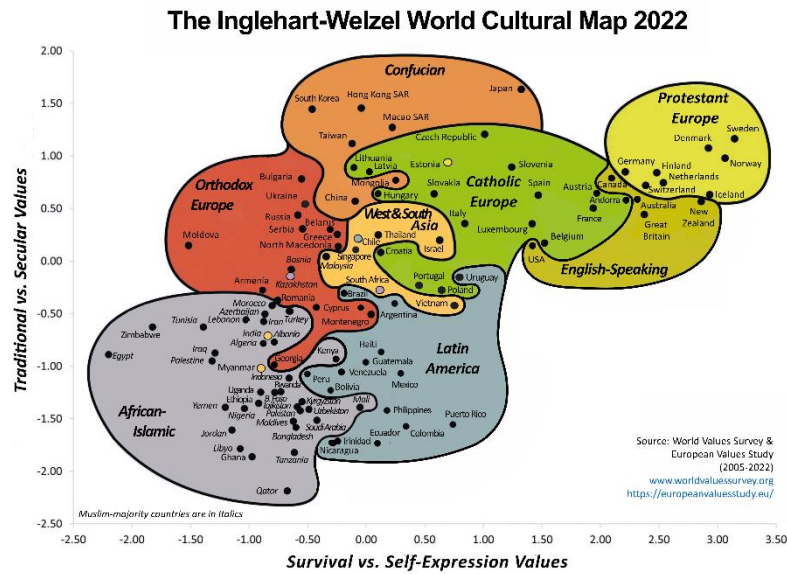
In my model the variable affected by *DistSoc_{ij}* is average social distance (*AvDistSoc*).

3.5.4 Cultural Proximity

For cultural proximity I use Culture Map - Wave 7 data from the World Value Survey (WVS) (<http://www.worldvaluessurvey.org/>). I chose to use the WVS data because they are applied in cross-cultural analysis research, which connects the Human Development with the culture (Inglehart & Welzel, 2005; Welzel et al., 2003).

These data cover 111 countries. The Cultural Map of the WVS is a diagram where the x-axis takes the values of the "Survival vs Self-Expression Values" index and the y-axis takes the values of the "Traditional vs Secular Values" index (Diagram 3.). On the one hand the "Survival vs Self-Expression Values" index focuses on "survival values" (related to "economic and physical security") and on the other hand it emphasizes "self-expression values" ("subjective well-being, self-expression and quality of life") (Inglehart & Welzel, 2005). The "Traditional vs Secular Values" index highlights the importance of values such as religion, parent-child bonds, respect for authority and traditional family values (Inglehart & Welzel, 2005).

Diagram 3.1: Cultural map – WVS wave 7



Source: World Value Survey (WVS) (<http://www.worldvaluessurvey.org/>, accessed 2/1/2023)

Since the nodes of the network, we are investigating are the authors of the publications, then I consider that each author gets the same values (from the two indexes we mentioned above) as the country of his affiliation. So, to measure the cultural distance I calculate the vector distance between of the countries.

$$\text{Cultural Distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad [3.5]$$

x_2 is the value of the 2nd Author from the "Survival vs Self-Expression Values" index.

x_1 is the value of the 1st Author from the "Survival vs Self-Expression Values" index

y_2 is the value of the 2nd Author from the "Traditional vs Secular Values" index

y_1 is the value of the 1st Author from the "Traditional vs Secular Values" index

I remove from my data (the data from Web of Science) those publications that have authors from 24 countries for which there were no Culture Map - Wave 7 values (of the WVS) (Barbados, Bermuda, Botswana, Cambodia, Costa Rica, Cuba, Dem. Rep. Congo, Fiji, Grenada, Guinea, Kuwait, Laos, Madagascar, Malta, Mauritius,

Mozambique, Nepal, North Korea, Oman, Rep Congo, Sri Lanka, Sudan, Suriname, U. Arab Emirates).

At this point, I must note that a weakness of my research is the fact that we took into account the affiliation address of each author (ie their work address) and not the country of origin. Thus, we took into account the culture of the country where each author works rather than the culture of their country of origin.

3.5.5 Regressions and Multivariable Regressions Models

In order to explore the degree that different aspects of proximity affect the effectiveness of the published items, I focus on year 2015. I consider the four independent variables regarding the dimensions of proximity, as I have defined them in the previous units. All independent variables would get values between 0 and 10, except for geographical distance, which had to be normalized in order to return values that would range from 0 to 5. For my estimations, I considered one additional variable, the number of co-authors per publication, which gets values between 2 and 8.

The dependent variable is the number of citations of each published item. The number of citations is a common proxy to measure authors' performance (Zitt et al., 2000; Acedo et al., 2006; Abbasi et al., 2011), acting as a reflection of the quality of the research (Acedo et al., 2006) and offering the researcher high appraisal in the academic community (Abbasi et al., 2011). The number of citations varied from 0 to 25306, so I normalized it as well, to return values that range from 0 to 5.

My next step is to conduct Spearman Correlation Test, to investigate the strength of association between the variables. Then I create two different multivariable regression models. In the first I consider all the independent variables:

$$\begin{aligned} CitNum \sim \alpha + \beta_1 AuthNum + \beta_2 AvGeoDist + \beta_3 AvOrgDist + \beta_4 AvSocDist \\ + \beta_5 AvCultDist \end{aligned} \quad [3.6]$$

In the second model, I consider only the independent variables regarding proximities:

$$\begin{aligned} CitNum \sim \alpha + \beta_1 AvGeoDist + \beta_2 AvOrgDist + \beta_3 AvSocDist \\ + \beta_4 AvCultDist \end{aligned} \quad [3.7]$$

Where:

CitNum = Citations Number

AuthNum = Number of co-authors

AvGeoDist = Average Geographical Distance

AvOrgDist = Average Organizational Distance

AvSocDist = Average Social Distance

AvCulDist = Average Cultural Distance

For each model, I calculate the Variance Inflation Factor (VIF), to check the models for the incidence of multicollinearity. Then I check the relation of each independent variable to the number of citations, conducting regression analysis.

$$CitNum \sim a + \beta AvGeoDist \quad [3.8]$$

$$CitNum \sim a + \beta AvOrgDist \quad [3.9]$$

$$CitNum \sim a + \beta AvSocDist \quad [3.10]$$

$$CitNum \sim a + \beta AvCultDist \quad [3.11]$$

Chapter 4: Results

4.1 Introduction

In this chapter I present the study's results. First, I describe the transmutations on network size and on co-authorship trends. Then I analyze the network's structure in 2015, I present the countries with the highest participation in the network, and I examine how different dimensions of proximity characterize relations in the studied network. Finally, I present the results of the multivariable regression analysis, explain the limitations and try to approach a more effective model.

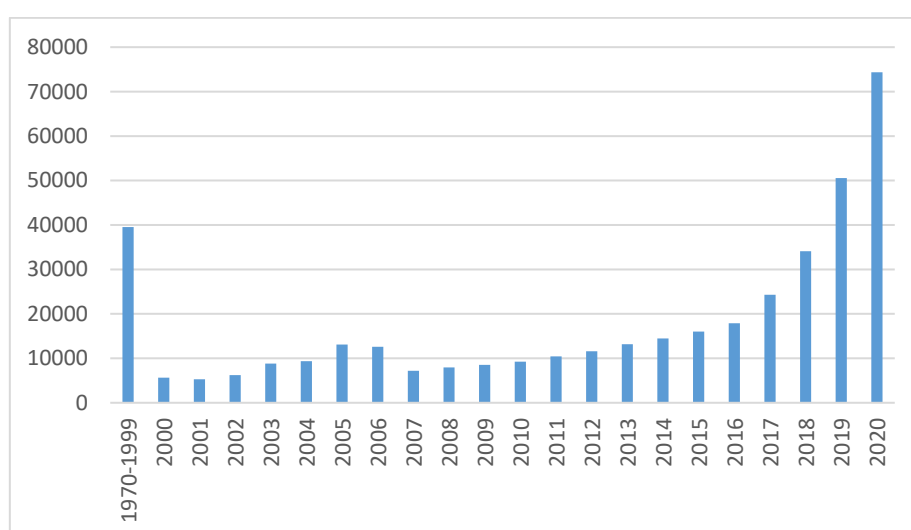
4.2 Results from the Descriptive-Statistic Analysis

In Diagram 4.1 I notice that the number of scientific publications is extremely low in the period 1970-1999. This can be explained since the scientific field emerged in 1955 and for the first 50 years researchers had to deal with serious limitations regarding computational power. Expectations could not be met, the discipline was judged regarding not only its effectiveness and the morality of the concept, but funding was also limited, and scholars would choose to focus on different disciplines.

In the 2000s society was less superstitious and more open to the idea of AI. This combined with fact that during that decade took place serious scientific discoveries that improved computational power led to an increased number of publications in the technological field. The diagram illustrates a small transmutation from 2003 to 2007, reaching a local peak in 2005. This possibly explains how the increased interest in the field led to the birth of deep thinking in 2006. The establishment of the new concept might have made the researchers reluctant, and we notice a considerable drop in the number of publications in 2007. From 2007 to 2017 we notice a steady increase, as deep thinking is getting established in scientific community.

The production of scientific publications starts to increase rapidly from 2017 (24,289 publications) to reach 74,344 publications in 2020. This can be explained by the fact that AI has transformed into a multidisciplinary scientific field (Laato et al., 2020). Moreover, improvements in computational power are rapid, and people have realized that AI is part of our daily lives. This has caught researchers' interest, and we notice a sharp increase in the number of publications, which is believed to continue in the following years.

Diagram 4.1: Number of scientific publications per year.



Source: Own processing

Diagram 4.2 indicates that the tendency for co-authorship has been increasing since 1988, and the highest percentages are noted in the period 1997-2020 (in 2019 and 2020 the percentage of co-authorship is the highest (95%)). The reason for this increase in co-authorships from 1997 onwards may also be the widespread use of the internet.

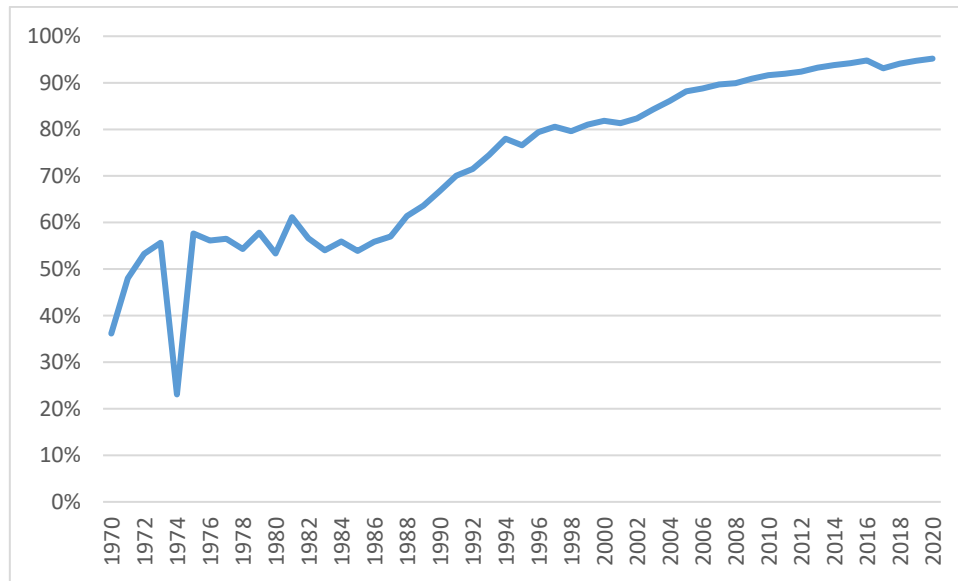
The results confirm previous studies which indicate that co-authorship the dominant mode in scientific publications (Acedo et al., 2006; Abbasi et al., 2011; Bordons et al., 2015). Abbasi et al. (2011) claim that scientific publications were considered a solitude endeavor during the first part of the 20th century. The diagram shows that this norm was still evident in the field of AI in the early 1970s. Though this fact had changed till 1972, when we notice more co-authored than single-authored publications. Since then, co-authorship has remained the dominant trend, despite the fluctuations from 1976 to 1988. From 1994 to 1996 co-authorship as a trend experiences a considerable increase, and since then it has kept increasing slightly, from 80% to 95%.

The diagram illustrates a dramatic drop in year 1974. This is most possibly caused by an abnormality in my dataset. The average number of publications per year in the 1970s is close to 100, but in 1974 we only have 26 publications. This number is too low to lead to conclusions.

The trend for co-authorship is followed by a trend for multiple co-authors. The median of co-authors per publication has steadily increased from 1 in the early 1970s, to 4 in the late 2010s. In the early 1990s, publications with extreme values in the number of co-authors per publication appear for the first time. Since then, such values have been evident almost every year.

If we examine Diagram 4.1 and Diagram 4.2 at the same time, we can conclude that from 1970 to 2020, publications on AI have increased, and so has the trend for co-authorship in the discipline. This can possibly be related to results of previous researches, which indicate that co-authorship facilitates authors' productivity in quantity (Acedo et al., 2006; Abbasi et al, 2011).

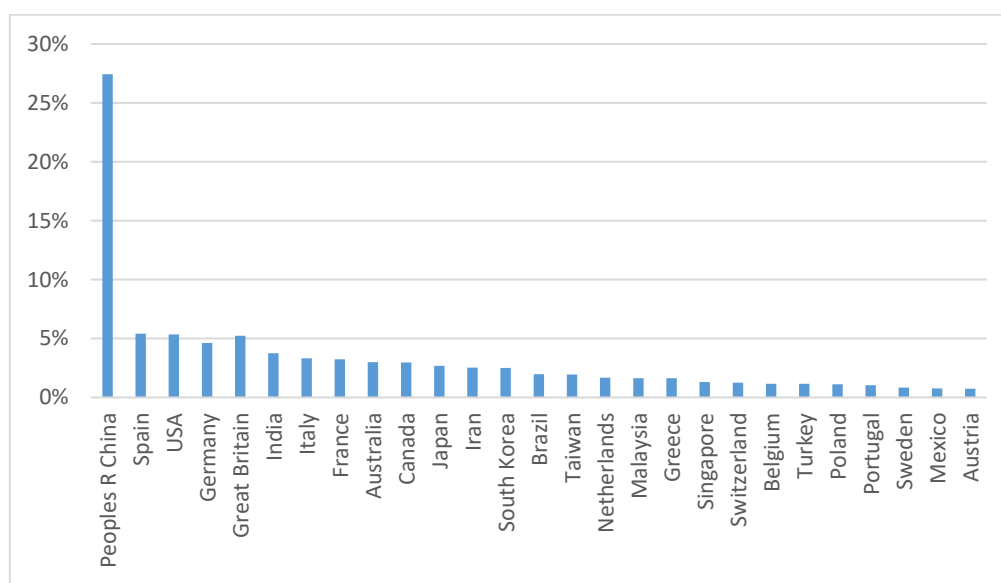
Diagram 4.2: Co-authorship per number of publications.



Source: Own processing

Diagram 4.3 confirms that complex knowledge (which is usually related to emerging technologies or disruptive innovations) is spatially concentrated in the most powerful countries of the world (Balland et al., 2019; Hidalgo & Hausmann, 2009). Another interesting finding, is that in 2015, China contributed to the production of 27% of the scientific papers on AI. This could be a herald of China's aims on AI, as they were stated in 2017 (State Council Gazette, New Generation AI Development Plan, Aug 2017). The plan stipulates that China will be the world leader on AI by the year 2030. A few years earlier, in 2013, President Xi Jinping had stated that "advanced technology is the sharp weapon of the modern state" (Jochheim, , 2021). China's policy on AI could explain the increased volume of publications that originate from this country. It would be interesting to study how the volume amount of publications changes in time, given that both European Union and the United States of America have taken steps to respond to this policy setting up new committees (Ulrich Jochheim, European Parliament Research Service, 2021), and controlling exports regarding AI to China (CSIS, 2022).

Diagram 4.3: Number of Authors per country (percentage calculated on the total number of authors).



Source: Own processing

4.3 Results from the Social Network Analysis

This network is a relatively large network (with a big number of nodes and edges) (Table 4.1). The density of the network is very low, since the Average Degree is low compared to the size of the network (Average Degree: 2.953) (Table 1).

From Table 4.1 we observe that the length of Network Diameter (31) (i.e. the shortest distance between the two farthest points (Scott, 2017) or nodes or authors) (Diameter: 31) is relatively long compared to the relatively low Average Path Length (9.76133795) (Backstrom et al, 2012; Perez & Germon, 2016). So, it seems that the network cohesion is relatively low and that the nature of the network may be fragmented. It should be noted that fragmentation is not a rare phenomenon for co-authorship networks (Rossoni & Hocayen-da-Silva, 2008; Bordons et al., 2013; Pauli et al., 2020).

Table 4.1: The characteristics of the network.

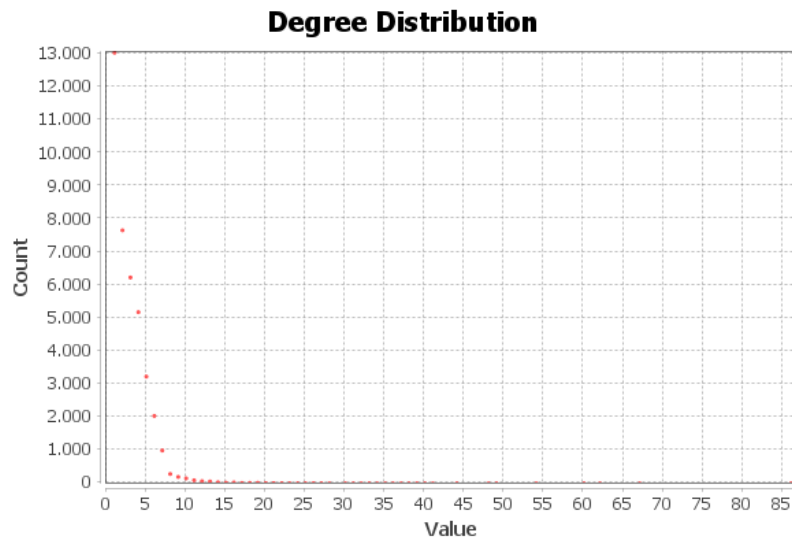
Network Size	
Nodes	39547
Edges	58383
Network Density	
Density	0,0000746621
Average Degree	2,953
Network Diameter	
Diameter	31
Average Path Length	9,761338
Modularity	
Number of Communities	5886

Source: Own processing (in Gephi)

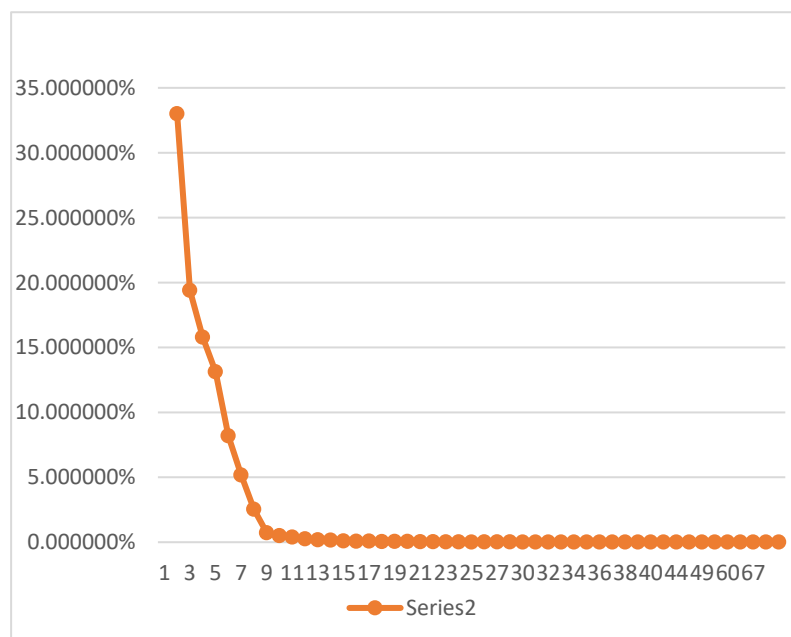
To analyze the coherence of the network more thoroughly we must also examine the degree of each point (of each node or author) of the graph (Diagram 4.4). The degree of a point (or node) is the number of other points connected to it (Scott, 2017; Aktamov & Zhao, 2014). We observe that the sum of the nodes that have a degree from 1 to 4 is 81.3% of all nodes (32153 nodes out of a total of 39547 nodes) (Diagram 4.4a) (Diagram 4.4b). This means that most nodes have a very low degree, which indicates that the network has low cohesion (which is also evident from the low density and low average degree) (Scott, 2017; De Nooy et al., 2005).

Diagram 4.4: Degree Distribution This chart shows the Degree Distribution. y-axis is the percentage of nodes (out of total nodes), and the x-axis is the degree.

a) Degree Distribution: y-axis is the number of nodes, and the x-axis is the degree.



b) Degree Distribution: y-axis is the percentage of nodes (out of total nodes) and the x-axis is the degree.



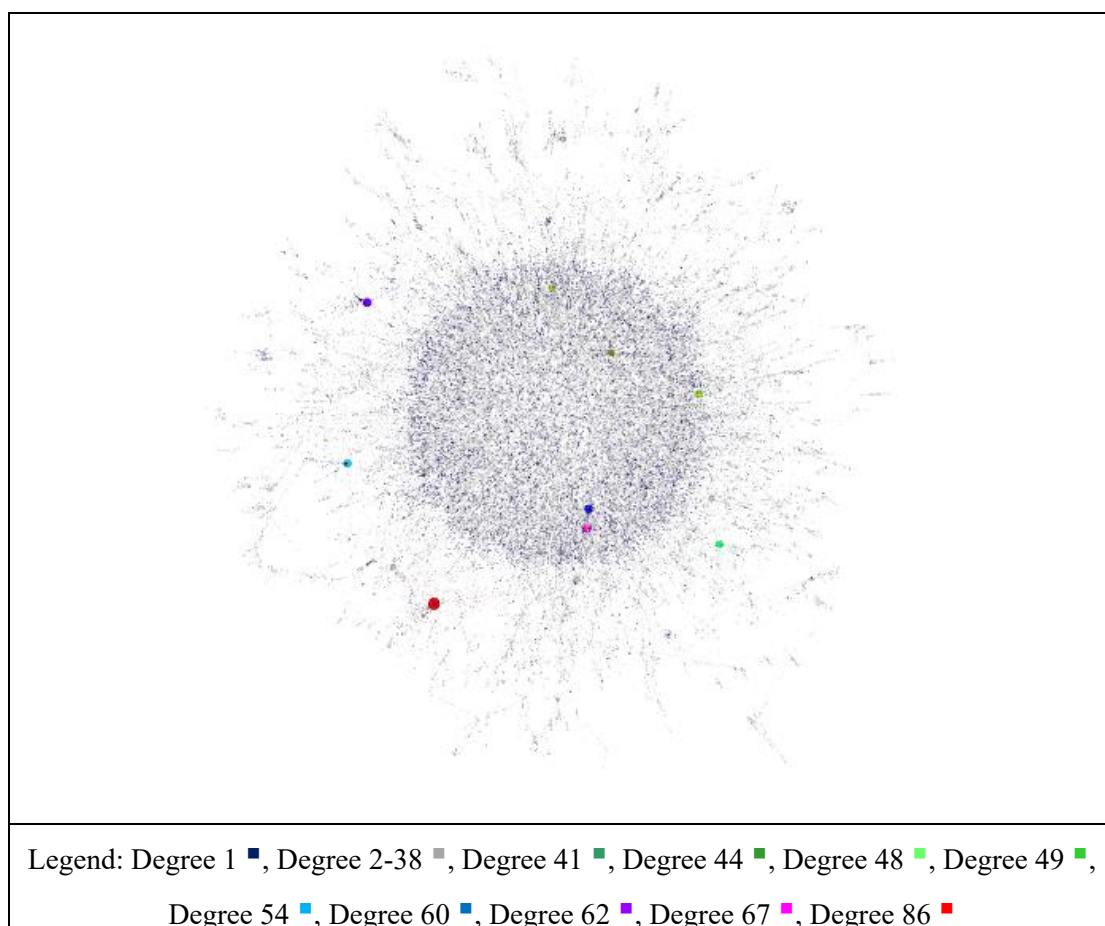
Source: Own processing (in Gephi)

Checking the Modularity of the network, we found that the number of communities is 5886, which is rather high. Modularity is a measure of the structure of networks or graphs which measures the strength of division of a network into modules

(also called groups, clusters, or communities). Therefore, it appears that this network is indeed quite fragmented.

As shown in the Graph 1, a "solid sphere" is formed in the center of the graph. This "solid sphere" is a very large and dense group of nodes that comprise most of the graph (some- times called the core of the network) (Golbeck, 2013). Most of the nodes in the core of the network have a low degree (the nodes whose degree ranges from 1 to 38 represent the 99.96% of all nodes). We also notice in the graph that far from the core through a chain of low-degree nodes there are few nodes that are connected to many nodes with no other connections. These nodes are located at the periphery of the network (Golbeck, 2013). Those nodes in the periphery of the network, which is distant from the core of the network and most nodes, have a high degree (Golbeck, 2013).

Graph 1: Network Visualisation, with nodes colored according to their degree.



Source: Own processing

So, the nodes that have a high degree do not seem to affect the cohesion of the network significantly. This happens because these nodes are few and the degree of these

nodes is low compared to the size of the network. Thus, we don't have a very coherent network in which the nodes with the highest degree would gather around them many nodes with lower degree. From the above and considering that the network has a large number of communities and low density we conclude that it is a fragmented network.

At this point, I must note that the importance of this high fragmentation can be positive. As Rossoni and Hocayen-da-Silva (2008), Pauli et al (2020) and Bordons et al (2015) state, high fragmentation can indicate low cohesion and therefore a greater opening for network expansion. This means that the network actors with the most central position do not affect the network structure and that new actors can participate in the network. In essence, this expansion may mean more possibilities for co-authoring (Pauli et al.,2020). Accordingly, Wanzenboeck et al (2014) found that the fragmentation of the co-authorship network (and in their case they were investigate an international network, as in our case) does not hinder the diffusion of knowledge.

4.4 Results from Spearman Correlation

Table 4.2: Spearman Correlation.

	normal_citations	AuthNum	normal_AvGeoDist	AvOrgDist	AvSocDist	AvCultDist
normal_citations	1.00					
AuthNum	-0.31	1.00				
normal_AvGeoDist	-0.66	0.49	1.00			
AvOrgDist	-0.77*	0.60	0.77*	1.00		
AvSocDist	0.77*	-0.60	-0.77*	-1.00***	1.00	
AvCultDist	-0.60	0.43	0.94***	0.83**	-0.83**	1.00
n=	6					
P						
	normal_citations	AuthNum	normal_AvGeoDist	AvOrgDist	AvSocDist	AvCultDist
normal_citations		0.5441	0.1562	0.0724	0.0724	0.2080
AuthNum	0.5441		0.3287	0.2080	0.2080	0.3965
normal_AvGeoDist	0.1562	0.3287		0.0724	0.0724	0.0048
AvOrgDist	0.0724	0.2080	0.0724		0.0000	0.0416
AvSocDist	0.0724	0.2080	0.0724	0.0000		0.0416
AvCultDist	0.2080	0.3965	0.0048	0.0416	0.0416	

Source: Own Processing

Spearman correlation demonstrates rather interesting results regarding the variables. First of all, average social distance is perfectly negatively correlated to average organizational distance. The p-value equals to 0, so the correlation is significant at 0 level. Absolute negative correlation means that possible changes in the values of one variable can perfectly explain the changes in the other variable. The negative sign indicates that when the values of the first variable increase, the values of the second variable decrease.

Furthermore, we can see that there is a strong positive correlation between average cultural distance, and the normalized average geographical distance. P-value

reveals that the correlation is significant at 0.1 level. The correlation indicates that an increase in one of the variables would indicate an increase in the other variable as well.

Average cultural distance is positively correlated to average organizational distance and negatively correlated to average social distance. This makes sense given that average organizational distance and average social distance achieve absolute negative correlation. P-values indicate that each of the correlations is significant at 0.05 level.

Normalized geographical distance is positively correlated to average organizational distance and negatively correlated to average social distance. Each correlation is significant at 0.1 level. Finally, citations number is positively correlated to average social distance and negatively correlated to average organizational distance, with both correlations being significant 0.1 level. These are the only independent variables correlated to the dependent variable. The correlation is an indicator that transmutations in one of these variables could cause transmutations in the dependent variable as well.

We notice that all statistically significant correlations are strong, since all have values greater than 0.7. All dependent variables which measure distances are strongly and statistically significantly correlated to each other. This means that they need to be checked for multicollinearity before they are used in the same multivariable regression model.

The absolute negative correlation between average organizational distance and average social distance indicates that if organizational proximity increases social proximity drops, and vice versa. This could indicate that when scholars choose to collaborate with people who they know, they work for different kinds of organizations. The correlation to the dependent variable could be an indicator that this kind of collaboration is effective.

The fact that AI is a multidisciplinary field, can possibly explain the above assumption. Multidisciplinary fields are characterized by low levels of organizational proximity. In such cases, actors need to increase their interactions which leads to higher levels of social proximity (Rekers & Hansen, 2015).

4.5 Results from Regression and Multivariable Regression Models

Table 4.3 : VIF Test (Number of Authors, Average Geographical Distance, Average Organizational Distance, Average Social Distance, Average Cultural Distance)

Tolerance and Variance Inflation Factor		

Variables	Tolerance	VIF
1 Number of Authors	0.9244546	1.081719
2 Average Geographical Distance	0.5699723	1.754471
3 Average Organizational Distance	0.8421546	1.187430
4 Average Social Distance	0.9921210	1.007942
5 Average Cultural Distance	0.6020107	1.661100

Source: Self Processing

Table 4.4: VIF Test (Average Geographical Distance, Average Organizational Distance, Average Social Distance, Average Cultural Distance)

Tolerance and Variance Inflation Factor		

Variables	Tolerance	VIF
1 Average Geographical Distance	0.5751433	1.738697
2 Average Organizational Distance	0.8839772	1.131251
3 Average Social Distance	0.9953332	1.004689
4 Average Cultural Distance	0.6038927	1.655923

Source: Self Processing

Spearman Correlation showed that there is a high possibility for multicollinearity between the variables. Table 4.3 and Table 4.4 illustrate that VIF Tests give values close to 1, so multicollinearity is not an issue in either of the multivariable regression models.

Both multivariable regression models (table 4.3; table 4.4) are all characterized by very low R-Squared and Adjusted R-Squared, which indicates that neither of them is a good fit. Usually models based on cross-sectional data give low R-Squared, so this could be an explanation. The fact that the network was expanding rapidly in that period, could strengthen the previous assumption. Another possible explanation is that our independent variables are not key variables. Model (3.6) is characterized by slightly higher Adjusted R Squared, so it is the best model to use between the two options, and can explain 0.2% of our sample.

According to both multivariable regression models, Average Organizational Distance is significant at 0 level. Though the value in estimate shows that it slightly affects the dependent variable. Average Organizational Distance is the only significant variable. The results regarding Average Organizational Distance are also confirmed by the regression analysis between citation number and the independent variable (see: Appendix 2). In this case as well, the R Squared is very low, and the impact the variable has on the number of citations is very low.

Table 4.5: Multivariable Regression (Number of Citations~Number of Co-Authors+Average Geographical Distance+Average Organizational Distance+Average Social Distance+Average Cultural Distance)

Call:				
lm(formula = normal_citations ~ AuthNum + normal_AvGeoDist + AvOrgDist + AvSocDist + AvCultDist, data = datacl1)				
Residuals:				
Min	1Q	Median	3Q	Max
-0.0178	-0.0049	-0.0030	0.0003	4.9829
Coefficients:				
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.0033554	0.0014751	2.275	0.0229 *
Number of Co-Authors	-0.0001537	0.0002656	-0.579	0.5628
Average Geographical Distance	0.0003417	0.0007910	0.432	0.6658
Average Organizational Distance	0.0024789	0.0005061	4.899	9.77e-07 ***
Average Social Distance	0.0012287	0.0007925	1.551	0.1210
Average Cultural Distance	0.0002811	0.0010549	0.266	0.7899

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Residual standard error: 0.04598 on 13103 degrees of freedom				
Multiple R-squared: 0.002461, Adjusted R-squared: 0.00208				
F-statistic: 6.466 on 5 and 13103 DF, p-value: 5.202e-06				

Source: Self Processing

Table 4.6: Multivariable Regression (Number of Citations~Average Geographical Distance+Average Organizational Distance+Average Social Distance+Average Cultural Distance)

Call:				
lm(formula = normal_citations ~ normal_AvGeoDist + AvOrgDist + AvSocDist + AvCultDist, data = datacl1)				
Residuals:				
Min	1Q	Median	3Q	Max
-0.0172	-0.0050	-0.0031	0.0003	4.9833
Coefficients:				
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.0028121	0.0011378	2.471	0.0135 *
Average Geographical Distance	0.0002983	0.0007874	0.379	0.7049
Average Organizational Distance	0.0024152	0.0004939	4.890	1.02e-06 ***
Average Social Distance	0.0012548	0.0007912	1.586	0.1128
Average Cultural Distance	0.0003152	0.0010533	0.299	0.7648

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Residual standard error: 0.04597 on 13104 degrees of freedom				
Multiple R-squared: 0.002436, Adjusted R-squared: 0.002131				
F-statistic: 7.999 on 4 and 13104 DF, p-value: 1.952e-06				

Source: Self Processing

All regression models (see Appendix 1; Appendix 2; Appendix 3; Appendix 4) are characterized by very low R-Squared and Adjusted R-Squared, which means that none of them is a good fit. Average Organizational Distance (see: Appendix2) is significant at 0 level. Average Geographical Distance (see: Appendix1) and Average Cultural Distance (see:Appendix4) are both significant at 0.01 level. All three variables are characterized by low estimate values, which means they slightly affect the dependent variable. Average Social Distance is not significant.

5. Conclusions

In the present research I study a co-authorship network in the field of AI as a knowledge network. I mainly focus on international scientific publications on AI. The study is based on data derived from Web of Science, regarding publications in AI for the years from 1970 to 2020.

I use descriptive statistics to analyze the emergence of the network from 1970 to 2020. Then I focus on year 2015 and I examine countries in which the researchers work. Then, I use SNA to explore how the network's structure affects the relations between the network's actors. Finally, I conduct multivariable regression to probe if and how different dimensions of proximity affect the actor's performance.

From the empirical findings of the research, I conclude that:

- In the technological field of AI the annual number of publications has increased from 83 in 1970 to 74433 in 2020. The annual increase has been slight from 1970 to 2000, moderate from 2000 to 2017, and rapid from 2017 to 2020. The initial slight annual increase can be contributed to limitations of computational power, and the establishment of the new scientific concept. The moderate and sharp increases can be explained by the emergence and spread of deep learning and the progress computational power.
- Publications in AI are mainly a product of cooperation. Since the early 1970s co-authorship has been the major trend in the field, with the proportion of co-authored publications increasing annually. The change has been dramatic since it rose from about 30% in 1970, to 95% in 2020. Co-authorship might be related to increased scientific productivity.
- In 2015 authors from China produced 27% of the total publications. It is evident that the countries considered rich have the greatest contribution to the production of new and more complex knowledge (Balland et al., 2019; Hidalgo & Hausmann, 2009).
- From the SNA results I find that the nature of the network is fragmented.
- The fragmented nature of the network indicates low cohesion, and this usually means (for co-authorship networks) a greater openness to network expansion (Rossoni & Hocayen-da-Silva 2008; Pauli et al., 2020; Bordons et al., 2015).

This means that the network actors with the most central position do not affect the network structure and that new actors can participate in the network. In fact, this expansion may mean more possibilities for co-authoring (Pauli et al., 2020). Also, this fragmentation of the co-authorship network is not likely to hinder the diffusion of knowledge (Wanzenboeck et al., 2014).

- Spearman Correlation shows that all variables which measure the dimensions of proximity are strongly and significantly correlated. Average organizational distance and average social distance are absolutely negatively correlated (-1.00) and the correlation is significant at 0 level. With regard to the concept of proximity, this indicates that when social proximity increases, organizational proximity decreases. These two variables are the only independent variables that are correlated to the dependent variable and can have an impact on the number of citations. This could indicate that high social proximity combined with low organizational proximity could lead to publications with higher citation number.
- Regression analysis, and multivariable regression analysis show that our models are characterized by very low R Squared and Adjusted R Squared values, so none of them is a good fit. This, may be due to the use of cross-sectional data. According to all models, Average Organizational Distance is statistically significant at 0 level and has a slight positive impact on the number of citations. Regression models demonstrate that Average Geographical Distance and Average Cultural distance are both significant at 0.01 level, and each of them has a marginal positive impact on the number of citations.

5.1 Proposals for Future Research

Expanding the research to the rest of the 51-year period may show different results. However, this is limited by the availability of author data, as there are not valid metadata for older publications. Still, there may be sufficient data for interesting probing since 2000, when journal publications became available online. This could lead to interesting results, as we would be able to examine how improvements in computational power and the emergence of deep learning influenced the evolution of the co-authorship network.

References

- Abbasi, A., & Altmann, J. (2011). A social network system for analyzing publication activities of researchers. In *On collective intelligence* (pp. 49-61). Springer Berlin Heidelberg.
- Abbasi, A., Chung, K. S. K., & Hossain, L. (2012). Egocentric analysis of co-authorship network structure, position and performance. *Information Processing and Management*, 48(4), 671-679, 2w0.
- Acedo, F. J., Barroso, C., Casanueva, C., & Galán, J. L. (2006). Co-authorship in management and organizational studies: An empirical and network analysis. *Journal of management studies*, 43(5), 957-983.
- Aggarwal, C. C. (2011). *An introduction to social network data analytics* (pp. 1-15). Springer US.
- Aktamov, S., & Zhao, Y. (2014). Impact of network centrality positions on innovation performance of the firm: evidence from china automobile industry. *Business Management and Strategy*, 5(1), 164.
- Antonelli, C. (Ed.). (2011). *Handbook on the economic complexity of technological change*. Edward Elgar Publishing. 380-382.
- Backstrom, L., Boldi, P., Rosa, M., Ugander, J., & Vigna, S. (2012, June). Four degrees of separation. In *Proceedings of the 4th Annual ACM Web Science Conference* (pp. 33-42).
- Balland, P. A., De Vaan, M., & Boschma, R. (2013). The dynamics of interfirm networks along the industry life cycle: The case of the global video game industry, 1987–2007. *Journal of Economic Geography*, 13(5), 741-765.
- Balland, P. A., Boschma, R., & Ravet, J. (2019). Network dynamics in collaborative research in the EU, 2003–2017. *European planning studies*, 27(9), 1811-1837.
- Bathelt, H., & Glückler, J. (2018). Relational research design in economic geography. *The New Oxford Handbook of Economic Geography*. Oxford, 179, 195.
- Blanc, H., & Sierra, C. (1999). The internationalisation of R&D by multinationals: a trade-off between external and internal proximity. *Cambridge Journal of Economics*, 23(2), 187-206.
- Blundell, R., Dearden, L., Meghir, C., & Sianesi, B. (1999). Human capital investment: the returns from education and training to the individual, the firm and the economy. *Fiscal studies*, 20(1), 1-23.
- Bordons, M., Aparicio, J., González-Albo, B., & Díaz-Faes, A. A. (2015). The relationship between the research performance of scientists and their position in co-authorship networks in three fields. *Journal of Informetrics*, 9(1), 135-144.
- Boschma, R. (2005). Proximity and innovation: a critical assessment. *Regional studies*, 39(1), 61-74.
- Boschma, R., & Frenken, K. (2010). The spatial evolution of innovation networks: a proximity perspective. In *The handbook of evolutionary economic geography*. Edward Elgar Publishing.
- Boschma, R., Frenken, K., Bathelt, H., Feldman, M., & Kogler, D. (2012). Technological relatedness and regional branching. *Beyond territory. Dynamic geographies of knowledge creation, diffusion and innovation*, 29, 64-68.
- Broekel, T., & Boschma, R. (2012). Knowledge networks in the Dutch aviation industry: the proximity paradox. *Journal of economic geography*, 12(2), 409-433.

- Chaminade, C., De Fuentes, C., Harirchi, G., & Plechero, M. (2016). The geography and structure of global innovation networks: global scope and regional embeddedness. In *Handbook on the Geographies of Innovation* (pp. 370-381). Edward Elgar Publishing.
- Christensen, C. M. (2013). *The innovator's dilemma: when new technologies cause great firms to fail*. Harvard Business Review Press.
- Christensen, C. M., Anthony, S. D., & Roth, E. A. (2004). *Seeing what's next: Using the theories of innovation to predict industry change*. Harvard Business Press.
- Coulon, F. (2005). *The use of Social Network Analysis in Innovation Research: A literature review*. PhD Candidate. Division of Innovation – LTH Lund University. Sweden.
- De Nooy, W., Mrvar, A., & Batagelj, V. (2005). *Exploratory network analysis with Pajek*. Cambridge University Press. 62-64.
- Drucker, P. F. (1969). The knowledge society. *New Society*, 13(343), 629-631.
- Forti, E., Franzoni, C., & Sobrero, M. (2013). Bridges or isolates? Investigating the social networks of academic inventors. *Research Policy*, 42(8), 1378-1388.
- Furman, J. L., Porter, M. E., & Stern, S. (2002). The determinants of national innovative capacity. *Research policy*, 31(6), 899-933.
- Golbeck, J. (2013). Analyzing the social web. *Newnes*. p.27, 37
- Grillitsch, M., Asheim, B., & Trippel, M. (2018). Unrelated knowledge combinations: The unexplored potential for regional industrial path development. *Cambridge Journal of Regions, Economy and Society*, 11(2), 257-274.
- Heringa, P. W., Hessels, L. K., & van der Zouwen, M. (2016). The influence of proximity dimensions on international research collaboration: an analysis of European water projects. *Industry and Innovation*, 23(8), 753-772.
- Hidalgo, C. A., & Hausmann, R. (2009). The building blocks of economic complexity. *Proceedings of the national academy of sciences*, 106(26), 10570-10575.
- Hoffmann, C. P., Lutz, C., & Meckel, M. (2014). Impact factor 2.0: Applying social network analysis to scientific impact assessment. In 2014 47th Hawaii International Conference on System Science (HICSS), New York, 2014, *Anais do 47th Hawaii International Conference on System Science*, Hawaii, EUA. Recuperado de <https://www.computer.org/csdl/proceedings/hicss/2014/2504/00/2504b576.pdf>
- Iammarino, S., & McCann, P. (2006). The structure and evolution of industrial clusters: Transactions, technology and knowledge spillovers. *Research policy*, 35(7), 1018-1036.
- Inglehart, R & C. Welzel. (2005). *Modernization, Cultural Change and Democracy: The Human Development Sequence*. New York: Cambridge University Press.
- Knoben, J., & Oerlemans, L. A. (2006). Proximity and inter-organizational collaboration: A literature review. *international Journal of management reviews*, 8(2), 71-89.
- Kudic, M. (2015). *Innovation networks in the German laser industry: Evolutionary change, strategic positioning, and firm innovativeness*. Springer. Springer International Publishing 4th Ed.
- Kumar, S. (2015). Co-authorship networks: a review of the literature. *Aslib Journal of Information Management*.
- Laato, S., Vilppu, H., Heimonen, J., Hakkala, A., Björne, J., Farooq, A., ... & Airola, A. (2020, October). Propagating ai knowledge across university disciplines-the design of a multidisciplinary ai study module. In *2020 IEEE Frontiers in Education Conference (FIE)* (pp. 1-9). IEEE.

- Laperche, B., & Sommers, P. (2010). Innovation networks and clusters: The knowledge backbone. Peter Lang. 182-184.
- Li, E. Y., Liao, C. H., & Yen, H. R. (2013). Co-authorship networks and research impact: A social capital perspective. *Research Policy*, 42,(9), 1515-1530, 2013.
- Li, M., Porter, A. L., & Suominen, A. (2018). Insights into relationships between disruptive technology/innovation and emerging technology: A bibliometric perspective. *Technological Forecasting and Social Change*, 129, 285-296.
- Lundvall, B. A. (1992). National systems of innovation: towards a theory of innovation and interactive learning.
- Malerba, F., & Vonortas, N. S. (Eds.). (2009). *Innovation networks in industries*. Edward Elgar Publishing.
- Malerba, F. (1992). Learning by firms and incremental technical change. *The economic journal*, 102(413), 845-859.
- Nilsson, M. (2019). Proximity and the trust formation process. *European Planning Studies*, 27(5), 841-861.
- Pauli, J., Basso, K., Gobi, R. L., & Bilhar, A. (2020). The effect of co-authorship network density on the performance of postgraduate programs. *BBR. Brazilian Business Review*, 16, 576-588.
- Perez, C., & Germon, R. (2016). Graph creation and analysis for linking actors: Application to social data. In *Automating open source intelligence* (pp. 103-129). Syngress.
- Phelps, C., Heidl, R., & Wadhwa, A. (2012). Knowledge, networks, and knowledge networks: A review and research agenda. *Journal of management*, 38(4), 1115-1166.
- Rekers, J. V., & Hansen, T. (2015). Interdisciplinary research and geography: Overcoming barriers through proximity. *Science and Public Policy*, 42(2), 242-254.
- Roberts, P. W. (1999). Product innovation, product–market competition and persistent profitability in the US pharmaceutical industry. *Strategic management journal*, 20(7), 655-670.
- Rossoni, L., & Hocayen-da-Silva, A. J. (2008). Cooperação entre pesquisadores da área de administração da informação: evidências estruturais de fragmentação das relações no campo científico. *Revista de Administração*, 43(2), 138-151.
- Rosenkopf, L. (2000). *Managing Dynamic Knowledge Networks*. Wharton on managing emerging technologies (p. 337-357). New York: Wiley.
- Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology?. *Research policy*, 44(10), 1827-1843.
- Schmidt, A. L., & Junker, C. (2016). The role of science in disruptive innovation. Conference: Academic Proceedings 2016 University-Industry Interaction Conference: Challenges and Solutions for Fostering Entrepreneurial Universities and Collaborative Innovation, At Amsterdam.
- Scott, J. (2017). *Social Network Analysis*. Sage 4th Ed.
- Sonnenwald, D. H. (2007). Scientific collaboration. *Annu. Rev. Inf. Sci. Technol.*, 41(1), 643-681.
- Torre, A., & Gilly J.P. (1999). On the analytical dimension of Proximity Dynamics. *Regional Studies*, vol. 34, n°2, 169-180.
- Tsouri, M., Hansen, T., Hanson, J., & Steen, M. (2022). Knowledge recombination for emerging technological innovations: The case of green shipping. *Technovation*, 114, 102454.

- Van der Valk, T., Chappin, M. M., & Gijsbers, G. W. (2011). Evaluating innovation networks in emerging technologies. *Technological Forecasting and Social Change*, 78(1), 25-39
- Verspagen, B. (2007). Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research. *Advances in complex systems*, 10(01), 93-115.
- Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications* (Vol. 8). Cambridge university press. 101-103.
- Wanzenboeck, I., Scherngell, T., & Brenner, T. (2014). Embeddedness of regions in European knowledge networks: a comparative analysis of inter-regional R&D collaborations, co-patents and co-publications. *The Annals of Regional Science*, 53, 337-368.
- Welzel, C., Inglehart, R., & Kligemann, H. D. (2003). The theory of human development: A cross-cultural analysis. *European Journal of Political Research*, 42(3), 341-379.
- Jochheim, U. (2021). China's ambitions in artificial intelligence. European Parliament Research Service. September 2021 ([https://www.europarl.europa.eu/RegData/etudes/ATAG/2021/696206/EPRS_ATAG\(2021\)696206_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/ATAG/2021/696206/EPRS_ATAG(2021)696206_EN.pdf), accessed 19/3/2023).
- Allen, G. (2022). Choking off China's Access to the Future of AI. CSIS. 11/10/2022 (<https://www.csis.org/analysis/choking-chinas-access-future-ai>, accessed 19/3/2023).
- State Council Gazette. Plan for Development of the New Generation of Artificial Intelligence. China's State Council. Issue No. 22 Serial No. 1597 (Aug 10 2017) (http://english.www.gov.cn/archive/state_council_gazette/2017/08/10/content_281475782151912.htm, accessed 19/3/2023).

Appendices

Appendix1 : Number of Citations~Average Geographical Distance

Call:				
lm(formula = normal_citations ~ normal_AvGeoDist, data = datacl1)				
Residuals:				
Min	1Q	Median	3Q	Max
-0.0115	-0.0051	-0.0034	-0.0001	4.9928
Coefficients:				
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.0054835	0.0005662	9.685	<2e-16 ***
Average Geographical Distance	0.0014720	0.0005977	2.463	0.0138 * D

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Residual standard error: 0.04601 on 13107 degrees of freedom				
Multiple R-squared: 0.0004626, Adjusted R-squared: 0.0003863				
F-statistic: 6.066 on 1 and 13107 DF, p-value: 0.01379				

Source: Self Processing

Appendix 2: Number of Citations~Average Organizational Distance

Call:				
lm(formula = normal_citations ~ AvOrgDist, data = datacl1)				
Residuals:				
Min	1Q	Median	3Q	Max
-0.0170	-0.0050	-0.0031	0.0002	4.9830
Coefficients:				
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.0044965	0.0005438	8.268	< 2e-16 ***
Average Organizational Distance	0.0024933	0.0004644	5.369	8.05e-08 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Residual standard error: 0.04597 on 13107 degrees of freedom				
Multiple R-squared: 0.002194, Adjusted R-squared: 0.002118				
F-statistic: 28.83 on 1 and 13107 DF, p-value: 8.054e-08				

Source: Self Processing

Appendix 3: Number of Citations~Average Social Distance

Call:				
lm(formula = normal_citations ~ AvSocDist, data = datacl1)				
Residuals:				
Min	1Q	Median	3Q	Max
-0.0117	-0.0053	-0.0035	-0.0002	4.9937
Coefficients:				
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.0052635	0.0010145	5.188	2.15e-07 ***
Average Social Distance	0.0010198	0.0007901	1.291	0.197

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Residual standard error: 0.04602 on 13107 degrees of freedom				
Multiple R-squared: 0.0001271, Adjusted R-squared: 5.08e-05				
F-statistic: 1.666 on 1 and 13107 DF, p-value: 0.1968				

Source: Self Processing

Appendix 4: Number of Citations~Average Cultural Distance

Call:					
lm(formula = normal_citations ~ AvCultDist, data = datacl1)					
Residuals:					
	Min	1Q	Median	3Q	Max
	-0.0127	-0.0051	-0.0035	-0.0002	4.9929
Coefficients:					
		Estimate	Std. Error	t value	Pr(> t)
	(Intercept)	0.0061060	0.0004390	13.909	<2e-16 ***
	Average Cultural Distance	0.0016693	0.0008193	2.038	0.0416 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
Residual standard error: 0.04602 on 13107 degrees of freedom					
Multiple R-squared: 0.0003167, Adjusted R-squared: 0.0002404					
F-statistic: 4.152 on 1 and 13107 DF, p-value: 0.04161					

Source: Self Processing

