Understanding Geographical Patterns of Scientific Collaboration

in the field of Artificial Intelligence

Mohammadmahdi Toobaee

A thesis

in

The Department

of

Concordia Institute for Information Systems Engineering

Presented in Partial Fulfillment of the Requirements

For the Degree of Master of Applied Science (Quality Systems Engineering)

Concordia University

Montréal, Québec, Canada

February 2022

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Mohammadmahdi Toobaee

Entitled: Understanding Geographical Patterns of Scientific Collaboration in the field of Artificial Intelligence.

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

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

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

Signed by the final examining committee:

______________________________ Chair
Dr. C. Wang

______________________________ Examiner
Dr. O. Kuzgunkaya

______________________________Examiner
Dr. C. Wang

______________________________ Supervisor
Dr. A. Schiffauerova

______________________________ Co-supervisor
Dr. A. Ebadi

Approved by ______________________________
Dr. A. Ben Hamza, Director
Concordia Institute for Information Systems Engineering

____
Dr. M. Debbabi, Dean
Faculty of Engineering and Computer Science

Date ______________________________

ABSTRACT

Understanding Geographical Patterns of Scientific Collaboration
in the field of Artificial Intelligence

Mohammadmahdi Toobaee

The role of geographical proximity in facilitating inter-regional or inter-organizational collaborations has been studied thoroughly in recent years. However, the effect of geographical proximity on forming scientific collaborations at the individual level has not been addressed so far. Using co-publication data of AI researchers from 2000 to 2019, first, the effect of geographical proximity on the chance of future scientific collaboration among researchers was studied. The logit regression and machine learning classification results show that geographical distance is an essential impediment to scientific collaboration at the individual level despite the tremendous improvements in transportation and communication technologies during recent decades. Second, the interplay between geographical proximity and network proximity was examined to see whether network proximity can substitute geographical proximity in encouraging long-distance scientific collaborations. The results show that the effect of network proximity on the likelihood of scientific collaboration increases with geographical distance, implying that network proximity acts as a substitute for geographical proximity. Therefore, policies aiming at encouraging long-distance collaborations could positively affect scientific collaboration and future knowledge production.

ACKNOWLEDGMENTS

Foremost, I would like to acknowledge and give my warmest thanks to my supervisor Dr. Andrea Schiffauerova for her guidance, support, and empathy during the past two years. Her enthusiastic encouragement and valuable critiques helped me in all the time of research and writing of this thesis.

I would also like to express my sincere gratitude to Dr. Ashkan Ebadi for his willingness to discuss my research. I could not have imagined having a better mentor, especially in the methodology and data analysis section.

Finally, I would like to dedicate this thesis to my wife Mahnaz and son Mahour. The completion of this study could not have been possible without their unconditional love and support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Collaboration is a key driver of scientific output and performance (Ebadi & Schiffauerova, 2015, 2016). The critical role of collaboration in facilitating the production of new knowledge across different fields of science (J. Adams, 2013; B. F. Jones et al., 2008; Wuchty et al., 2007) and the positive effect of knowledge production on the long-term economic growth (Aghion & Howitt, 1992; C. Jones, 1995) have been studied thoroughly in recent years. Moreover, long-distance scientific collaborations, including international ones, have shown to be even more important as they provide higher quality research productions (J. Adams, 2013; J. D. Adams et al., 2005; Narin et al., 1991). Hence, there is no wonder why worldwide policymakers are encouraging scientific collaboration at regional, national, and even international levels. European Union (EU), for example, launched an ambitious Horizon 2020 program which was the most extensive EU research and innovation program ever with around €80 billion of funding available over seven years (European Commission, 2014). Their main objective was to develop the European research area (ERA) to mitigate the negative effect of geographical distance on EU researchers' collaboration.

One may think that revolutionary developments in transportation and communication technologies could play a crucial role in facilitating collaborations between scientists who are geographically far away from each other, diminishing thereby the effect of geographical distance on collaboration (Castells, 1996; Johnson & Mareva, 2002). Nevertheless, more recent studies show that, despite all technological developments, geography is still among the main determinants of collaboration (see, e.g., Bergé, 2017; Bignami et al., 2020; Morescalchi et al., 2015). Also, despite the acknowledged importance of geographical distance on collaboration, its by-products, such as differences in national systems, make the collaboration even more difficult (Lundvall, 1992). In other words, national borders are found to be another significant barrier for scientific collaborations.

Considering the importance of long-distance collaboration in producing high-quality research on the one hand and the negative effect of geographical distance on forming scientific collaboration, on the other hand, the main question here could be: Can other forms of proximity substitute geographical proximity to encourage scientific collaboration? To answer this question, we need to understand the determinants of scientific collaboration, especially those factors that can help bypass geography.

As different studies discussed it, collaboration requires creating a connection between researchers, and in this sense, it can be considered a social process (Freeman et al., 2014; Katz & Martin, 1997). Then, those connections gradually form a social network, and social networks are the driver of their own evolution over time (Jackson & Rogers, 2007). Therefore, potential network effects that influence the collaboration process should not be neglected. Although some studies worked on the evolution of scientific collaborations' networks (Almendral et al., 2007; Balland, 2012; Barabâsi et al., 2002; Maggioni et al., 2007; Newman, 2001; Wagner & Leydesdorff, 2005), empirical studies on the impact of network proximity on collaboration, and the interplay between geographical proximity and network proximity are scarce. If there is a substitutability pattern between geographical proximity and network proximity, network proximity would partially compensate for the negative effect of geographical distance. In this case, enhancing the network proximity of distant researchers through long-distance funding collaborations would help create new long-distance connections. Otherwise, if geographical proximity and network proximity are proved to be independent, or if there is a complementarity pattern (researchers with greater geographical distance benefit less from network proximity) between them, encouraging long-distance collaborations would not be efficient.

This thesis has two main research objectives. The first objective is to study the effect of geographical distance on scientific collaboration. Besides examining the role of direct geographical distance of partners on the probability of their future collaboration, the impact of institutional proximity (national systems) on future scientific collaboration will be studied. The second objective is to investigate the interaction effect of geographical proximity and network proximity on scientific collaboration. It will provide a clearer picture of the efficiency of funding long-distance collaborations.

We study the geographical patterns of collaboration among researchers in four scenarios to reach those research objectives. In the first scenario, we limit the study to collaborations among Canadian AI[1] researchers. There are at least four advantages for studying geo-patterns of collaboration in Canada: 1) Canada is the second-largest country globally, with several universities and research institutions spread all over the country. Thus, we expect to discern clearer geo-patterns of collaboration among researchers compared to smaller countries, 2) there is a large pool of scientists and researchers in Canada that provides a desirable collaboration network to study, 3) Canada is an industrialized country with advanced communication and transportation infrastructures; therefore, the effect of geographical distance on collaboration can be investigated regardless of the differences in the level of technological advancements, and 4) the provincial government system in Canada would make it possible to study the geo-patterns of collaboration within and between different provinces.

We extend the study to scientific collaborations among AI researchers in Canada and the United States in the second scenario. These two countries have a long history of close scientific collaborations. Thus, extending the study to Canada and the United States, besides all advantages enumerated for the first scenario, provides us with the opportunity to understand the geo-patterns of international collaborations among researchers from bordering countries. Then, to better

[1] Artificial Intelligence

understand how institutional proximity can affect scientific collaborations, we add AI researchers from European countries to the study in the third scenario. The fourth scenario is comprehensive, studying the geo-pattern of collaboration among researchers worldwide.

Since the effect of geography on collaboration has been found to be non-homogenous for different fields of knowledge (Bignami et al., 2020), we avoided mixing several disciplines and focused on collaborations in the field of AI, which is becoming increasingly prominent and more intertwined with existing and new technologies. It is anticipated that by 2030 AI will contribute almost 16 trillion USD to the global economy (PwC, 2017). AI's potential for economic growth and technological innovation, along with its capabilities to improve the efficiency of service delivery through its various tools (Machin learning, Natural language processing, etc.), has encouraged enormous public and private investments in this field. For example, the US government announced an investment of $1 billion in AI and Quantum Information Science research centers (US Government, 2020). In Canada, the significant AI infrastructure investments known as the "superclusters" are expected to contribute to Canada's economy by $66 billion by 2030 (Government of Canada, 2018). Moreover, as of August 2020, the federal government has awarded $1 billion in contributions across the country (Government of Canada, n.d.). Besides, a total of $1.2 billion in public investments have also been committed for the province of Quebec, and over $2 billion has been announced in private investments for Montreal (Brandusescu, 2021).

The organization of this study is as follows. The second chapter covers relevant literature about the notion of proximity and its role in explaining scientific collaboration. The third chapter discusses the data and methodology we used to reach the research objectives. The fourth chapter presents and discusses the results of data analysis. The fifth chapter discusses the main findings of this research work and concludes.

2. LITERATURE REVIEW

2.1. Proximity

Proximity is an influencing factor for knowledge flows in science (Boschma, 2005), and researchers from different disciplines try to analyze and understand complex network-related problems using this notion (André Torre & Gilly, 2000).

Although proximity has been historically associated with location (Gilly & Torre, 2000), different researchers developed the meaning of proximity, as it has gone beyond a spatial connotation (Boschma, 2005). Analyzing the role of technology and R&D[2] on the competitive strategies of multinational pharmaceutical companies, Zeller (2004) specified the categories of spatial, organizational, institutional, cultural, relational, technological, and virtual proximity. In her study, geographical proximity is named spatial proximity. Also, she considered the organization as a corporate unit with its own set of rules and identity. Then, she addressed adherence logic in two elements, i.e., institutional and cultural proximities. She defined institutional proximity as the collection of practices, laws, and rules defined by the geographical setting within a country or region. Then, cultural proximity is defined based on shared cultural background and the consequent norms of behaviour between researchers. She also included personal relationships based on informal structures and facilitated knowledge transfer, as relational proximity. Moreover, Zeller argued that technologically proximate agents could contribute with their respective findings, developments, and know-how; thus, she included technological proximity in her study. The last proximity element that she considered in her framework was virtual proximity which is the similarity between agents in terms of using ICT[3].

[2] Research and development

[3] Information and communication technology

Another proximity classification framework was introduced by Boschma (2005) with five dimensions, i.e., geographical, organizational, social, institutional, and cognitive. In his work, social proximity is identical to relational proximity introduced by Zeller (2004). However, Boschma did not separate cultural and institutional proximities. An institution in this context has two elements: informal institutions referring to cultural factors like ethnics, beliefs, and language, and formal institutions referring to laws and regulations. Besides, cognitive proximity accounts for agents' similarity in terms of their knowledge base.

Boschma & Frenken (2010) argued that the list of proximity dimensions can be extended without changing the meaning of a dimension, as proximity dimensions are analytically orthogonal even though they may be empirically correlated. For example, social proximity is generally higher when organizations are geographically proximate since friendships are more easily established and maintained over short distances. In general, despite differences between studies in terms of the list of proximity dimensions, the factors they use in their studies are shared. However, sometimes they rename them, separate them, or merge diverse aspects in one dimension. Although studying several proximity dimensions can be helpful in understanding networks' evolution, proximity dimensions seem to be substitutes rather than complements (Boschma, 2005). In other words, forming a connection requires at least one dimension of proximity, and if more than one exists, then the contribution of the additional ones is negligible. Thus, even though building networks requires proximity, not every dimension is necessary to form a connection.

2.2. Proximity and collaboration

The concept of proximity is an especially useful framework for analyzing the determinants of collaboration (Cunningham & Werker, 2012; Kirat & Lung, 1999; Andre Torre & Rallet, 2005). Bergé (2017) argued that different dimensions of proximity favor collaboration in two general ways: 1) proximity enhances the chance of potential partners to meet, and 2) it reduces the costs involved in the collaboration. Therefore, it increases the expected net benefits of the collaboration and augments the likelihood of its success.

However, various dimensions of proximity may have different influences on collaboration. The impacts of different types of proximity on collaboration have been studied by Balland (2012). Using data from R&D collaborative projects funded under the European Union 6th Framework Programme from 2004 to 2007, Balland showed geographical, organizational, and institutional proximity favour collaborations, while cognitive and social proximity does not play a significant role.

As is discussed in the introduction section, this thesis has two main objectives. First, to study the effect of geographical proximity on scientific collaboration among AI researchers; second, to examine the substitutability between geographical proximity and network proximity. Thus, we review the literature associated with geographical proximity, network proximity, and their interaction in the following sections. Moreover, following Bergé (2017), we include two more dimensions of proximity, i.e., institutional proximity and cognitive proximity, in the study to understand the evolution of scientific collaboration networks better. Thus, the literature related to institutional proximity and cognitive proximity is also discussed in other sections.

2.3. Geographical proximity

Among different dimensions of proximity, geographical proximity is the most common dimension in the literature (Knoben & Oerlemans, 2006). To understand the relationship between geographical proximity and scientific collaboration, one may deconstruct it as follows.

First, authors have to understand and share complex ideas, concepts and methods in order to collaborate on knowledge production (Gertler, 1995, 2003). Thus, some researchers argued that face-to-face interaction is essential to transfer tacit knowledge and conduct research. Gertler (1995), for example, addressed the importance of closeness between collaborating parties for the successful development and adoption of new technologies via interviews with advanced manufacturing technologies in Southern Ontario. He concluded that co-location is crucial to facilitate face-to-face interactions, consequently enhancing the chance of productive collaborations. Rallet & Torre (1999) also confirmed the importance of physical proximity in forming innovative collaborations and argued that the transfer of tacit knowledge implies frequent face-to-face relations. Howells (2002) discussed the relationship between knowledge and geography in the innovation process as well and highlighted the importance of geographical proximity in transferring tacit knowledge. In another study, Storper & Venables (2004) discussed the main features of face-to-face interactions. They concluded that face-to-face contact is particularly important in creative activities where information is imperfect, rapidly changing, and not easily codified.

Moreover, some studies show face-to-face interaction facilitates coordination, communication, and direct feedback (Beaver, 2001; Freeman et al., 2014), enhancing successful collaboration. Therefore, geographical distance lowers the likelihood of successful collaboration by diminishing knowledge exchange opportunities through face-to-face contact and incurring more significant travel costs (Katz, 1994; Katz & Martin, 1997). Catalini (2018) investigated the relationship between co-location and the rate, direction, and quality of scientific collaboration in a more recent study. His results show that co-location enhances the chance of co-publication by 3.5 times. He argued that co-location has two main advantages; first, it increases the likelihood of serendipitous interaction, and second, it lowers the search cost for new collaborators.

Second, geographical proximity increases the probability of potential partners meeting in the first place via attending social events such as seminars and conferences linked to geographical distance (Bergé, 2017). As an example, van Dijk & Maier (2006) studied the data on participating at the European Regional Science Association congresses and concluded that geographical proximity heightens the chance of attending them. In another study, Breschi & Lissoni (2009) showed that the social embeddedness of researchers decays with geographical distance since they have more knowledge about geographically closer partners. Therefore, we should expect the influence of geographical distance on collaboration to be negative. Several studies in different contexts (co-authorship in scientific collaboration, co-patenting, and cooperation among research institutions) have evidenced this fact.

In the case of co-authorship in scientific publication, Ponds et al. (2007) analyzed the geographical aspects of collaboration in scientific knowledge production, studying the co-publication data in science-based industries (agriculture & food chemistry; biotechnology; fine organic chemistry; analysis, measurement & control technology; optics; information technology; semiconductors and telecommunication) in the Netherlands during 1988 – 2004. Their main finding was that geographical proximity has a significant positive effect on collaboration between academic organizations. Besides, they concluded that geographical proximity could overcome the institutional differences between organizations, necessary for successful collaboration.

Hoekman et al. (2009) also studied the inter-regional research collaborations based on scientific publications and patents among 1,316 regions in 29 European countries. They found a negative and significant association between direct distance and co-publication, implying that the choice for collaboration partners in Europe is not just based on scholarly grounds, and geographical barriers mainly impede it. Also, they found that there is a higher chance of co-publication when regions belong to the same country. Therefore, they concluded that despite the opportunities of international collaborations offered by policymakers in the European Union, most researchers are biased towards domestic partnerships.

In a more recent study, Bergé (2017) used the data on scientific co-publications in the field of chemistry among 132 regions in five large European countries, i.e., Italy, Germany, Spain, France, and the United Kingdom, between 2001 and 2005. He found that increasing the distance between two regions decreases their level of collaboration. In addition, his results showed that national borders have the most impeding effect on collaboration, as when regions are in two different countries, their level of collaboration decreases by 83%.

To understand the relationship between geographical proximity and co-patenting, Maggioni et al. (2007) analyzed co-patenting among 109 regions in the five large European countries during 1998 – 2002. Their results clearly showed a negative association between geographical distance and collaboration. Besides, they concluded that contiguity would positively affect the number of co-patenting between regions. In a more recent study, Morescalchi et al. (2015) used data on patents filed with the European Patent Office (EPO) for OECD[4] countries to study the influence of geographical proximity and country borders on inter-regional links in four different networks (co-inventorship, patent citations, inventor mobility, and the location of R&D laboratories) from 1988 to 2009. Their results showed that despite globalization and advancement in communication and transportation technologies, the constraint imposed by geographical distance on R&D inter-regional links has increased.

In the case of cooperation among research institutions, Scherngell & Barber (2009) focused on cross-regional R&D collaborations between 255 regions in 27 European countries to identify the influence of geographical and technological separation on forming R&D collaborative activities. The results showed a negative and significant relationship between direct geographical distance and collaboration. Also, they showed that the intensity of collaboration among regions in different countries is lower than the intensity of collaboration among regions from the same county. Moreover, although they found technological proximity to have a stronger effect on the constitution

[4] Organisation for Economic Cooperation and Development

of R&D collaborations, they concluded that co-localization of organizations in neighbouring regions is an essential factor for the constitution of cross-regional R&D collaborations in Europe.

However, the negative effect of geographical proximity on collaboration could be limited. For instance, Mascia et al. (2017) studied the moderating impact of geographical proximity on the relationship between competitive interdependence and the propensity of organizations to collaborate and exchange resources. They argued that geographical proximity has a differentiated influence on two contrary factors affecting the tendency of organizations to collaborate, i.e., cooperative opportunities and competitive constraints, and concluded that the effect of geographical proximity on collaboration could be limited due to the natural tendency to compete for resources.

In addition, the effect of geography on collaboration could be non-homogeneous. Bignami et al. (2020), as an instance, showed that although geographical distance negatively affects collaboration, its importance is not equal for different types of knowledge that are being transferred in collaboration. They concluded that collaborations in basic science and core knowledge areas are more negatively affected by geographical distance than collaborations within clinical science and exploration knowledge areas.

Also, the results of temporal studies on the relationship between geographical proximity and collaboration over time are antithetical. While Johnson & Mareva (2002) concluded that the importance of geographical distance is diminishing, some studies have reported the effect of geographical distance on collaboration to be unchanged (Hoekman et al., 2010) or even becoming stronger (Morescalchi et al., 2015) over time notwithstanding developments in communication and transportation technologies.

Too much geographical proximity

Whereas geographical distance negatively affects interactive learning and innovation, too much proximity can also be harmful to these purposes (Boschma, 2005). Boschma & Frenken (2010) called this phenomenon "proximity paradox" and explained that it could happen due to the lack of openness and flexibility. They argued that proximity might negatively impact innovation due to the problem of spatial lock-in. When "inward-looking" spreads among agents in a region, the main risk would be their learning ability to be weakened gradually, which may lead to losing their innovative capacity and ability to respond to new developments.

Moreover, Merton (1973) explained scientific parochialism as another risk of geographical proximity. He argued that when researchers limit their interactions to local partners, they will deprive themselves of critical information flows in their field. However, Gittelman (2007) believed that the current scientific publication procedure that involves exchanging papers drafts would prevent scientific parochialism to some extent; it helps scientists keep updated with work from peers, no matter where they are located. In addition, she argued that participating in seminars and conferences may have a significant role in stimulating knowledge exchange and considering collaboration with distant partners.

Other forms of proximity

Some researchers believe that the effect of geographical proximity on collaboration cannot be evaluated in isolation as other forms of proximity are significant in explaining collaboration. Reviewing the literature of scientometric studies that have taken the spatial dimension into account, Frenken et al. (2009) provided an analytic framework for spatial scientometric research. They propose to use the concept of proximity, which distinguishes physical proximity from other forms of proximity (e.g., cognitive, social, organizational, and institutional) as determinants of scientific interaction.

Moreover, Boschma (2005) argued that geographical proximity per se is neither a necessary nor a sufficient condition for collaboration; it facilitates interactive learning by strengthening the other dimensions of proximity. As the other dimensions of proximity may provide alternative solutions to the problem of spatial lock-in in the region, geographical proximity is not a necessary condition. Besides, since the knowledge transfer across large distances requires other forms of proximity to be effective, geographical proximity is not sufficient either. Hence, although economic geographers have emphasized the economic advantages of geographical proximity, they have pointed out that other dimensions of proximity besides geographical proximity are vital in understanding collaboration.

2.4. Network proximity

Network proximity can be defined and measured in different ways. For example, the "shortest path", a fundamental concept in graph theory, is the shortest path of vertices and edges that links two given nodes (Newman, 2001). In a scientific collaboration network where each node represents an author, and each edge represents a co-publication, the shortest path represents how many authors are between two researchers tied up by co-authorship links with their peers. In this case, when the two specific authors have collaborated in a publication, they would be direct neighbors, and the shortest path would be 0. However, Bergé (2017) introduced TENB (Total Expected Number of Bridging Paths) to measure the network proximity between two different geographical districts based on their indirect connections. He argues that using the concept of TENB avoids the problem of reverse causality between co-publication and network proximity at the cost of neglecting possible network proximity originating from direct ties.

Due to the intuitive connection between social bonds and collaboration (Katz & Martin, 1997), the structure of scientific collaboration networks has been investigated in different studies (Almendral et al., 2007; Barabâsi et al., 2002; Fafchamps et al., 2010; Newman, 2001; Wagner & Leydesdorff, 2005). Some network mechanisms which can encourage collaboration have been discussed in the literature. Carayol et al. (2019) examined the notion of triadic closure as the tendency of two indirectly linked nodes to connect. They argue that triads have some advantages over dyads, such as conflict mitigation and trust enhancement, leading to triadic closure (Krackhardt, 1999). Compared with dyads, partners' negative behaviours are less expected in triads, as the third agent who serves the relation can punish it (Bergé, 2017). This structure can become a triadic closure, which is especially useful for international collaborations that assessing the reliability of partners might be difficult in them. In other words, collaborating with a partner of a partner can be favourable since it reduces the risks of collaboration by limiting opportunistic behaviours (Bergé, 2017).

In this vein, Ter Wal (2014) showed that the importance of triadic closure has increased drastically among inventors in Germany since they need more trust between partners due to the change in technological regime. In another study, Dahlander & McFarland (2013), after studying the researchers' behaviour at Stanford University, concluded that the chance of collaboration is positively influenced by having an indirect partner.

Homophily is another characteristic of networks that can affect forming new collaborations. Mcpherson et al. (2001) defined homophily as “the positive relationship between the similarity of two nodes in a network and the probability of a tie between them.” Studying this feature in different contexts, such as forming friendships at school, sociologists have shown that similarity is a force that drives developing new links (Mcpherson et al., 2001). In the context of scientific collaboration, Blau (1974) investigated the relationship among physic scientists and concluded that having similar research interests and personal characteristics positively affect their research relationships.

In a more recent study, Bergé (2017) argued that the collaboration network could trigger new connections through homophily. He believed that in every successful scientific collaboration, partners should have shared some similarities (same research topic, same approach to research questions, etc.). Therefore, if two agents have links with a third agent, there is an excellent chance that they share some similarities, which can lead to their future collaboration.

Finally, researchers can rely on the network to find their potential partners as it is an authentic source of information (Gulati & Gargiulo, 1999). Considering the importance of time for researchers (Katz & Martin, 1997) and increasing the demand for scientific collaboration (B. Jones, 2009), this network function seems crucial for finding the right partners. As time is a primary concern for researchers on the one hand, and assessing all potential matches to find the best collaborator is impossible due to lack of information on the other hand, the most efficient decision could be the best possible option in the network vicinity (Bergé, 2017). In a related study, Fafchamps et al. (2010) showed that when researchers are closer in the network, they have more chances to access information about each other. Moreover, they concluded that being closer in the network enhances the probability of future collaboration.

2.5. Interaction between geography and network proximity

The impacts of geographical and network proximity on collaboration have been discussed in the previous sections. However, one could ask about the net outcome of these two factors. Different hypotheses can be developed here. First, the influence of geographical proximity and network proximity can be independent, meaning that the benefit of network proximity is homogeneous for all potential collaborators regardless of their geographical distance. This independence hypothesis can be the case only if geographical and network proximity influence collaboration via completely unrelated mechanisms. Since the mechanisms through which they affect collaboration (such as improving trust or facilitating searching for a future partner) are the same for both geographical proximity and network proximity, their interaction could not be independent (Bergé, 2017).

Putting the independence hypothesis aside, we need to discuss two contrary patterns: complementarity and substitutability. In the complementarity pattern, a higher level of network proximity will enhance the chance of collaboration when agents are geographically close. Bergé (2017) argued that complementarity can be the case, especially when agents have a "taste for similarity." However, in the substitutability pattern, network proximity is necessary for collaboration among two geographically apart agents since they are not subject to any other forms of proximity. But, if they are geographically close to each other, a high level of network proximity has less importance in encouraging collaboration (Bergé, 2017).

Several studies focused on the effect of geography and network on collaboration, For example, Autant-Bernard et al. (2007) studied the impact of geographical and network proximities on the chance of future collaboration, and Maggioni et al. (2007) investigated the determinants of patenting activity. Both studies concluded that network and geographical proximity would positively affect future collaboration.

In an uncommon study, Bergé (2017) examined the interplay between network proximity and geographical proximity. Using data on scientific co-publications in the field of chemistry among 132 regions in five large European countries, he showed that the effect of network proximity on future collaboration is mediated by geography. Moreover, he found a substitutability pattern between geographical proximity and network proximity showing that network proximity mainly benefits distant collaborations.

The interaction between geography and other forms of proximity

The interaction between geographical proximity and organizational proximity has been investigated in some studies. Analyzing the co-publication data in science-based industries in the Netherlands during 1988 – 2004, Ponds et al. (2007) found a substitutability pattern between geographical and organizational proximities. However, D'Este et al. (2013) did not find such a pattern for university research partnerships in the United Kingdom.

2.6. Institutional proximity

When it comes to collaboration, the effect of national borders is another geographical barrier usually discussed under the context of institutional proximity (Hoekman et al., 2009). Boschma (2005) argued that the institutional environment shapes, affects, and even constrains interactions between agents. Besides, knowledge flow is affected by many factors which can be recognized at the national level (Banchoff, 2002; Glänzel, 2001). Funding schemes, as an example, typically encourage domestic collaborations since they exist at a national scale most of the time.

Moreover, people usually travel within a country rather than across countries, and since they keep ties with their previous partners, their social networks are mainly developed at the national level (Miguélez & Moreno, 2014). In addition, social factors such as language, values, and norms, which generally are shared within a country, facilitate collaboration. Therefore, the fact that being from different countries negatively affects the collaboration (either co-publication or co-patenting) process is demonstrated in the literature (Hoekman et al., 2009, 2010; Morescalchi et al., 2015; Scherngell & Barber, 2009).

2.7. Cognitive proximity

Defined as “the shared knowledge base and expertise of different agents,” different frameworks recognized cognitive proximity as a determinant of collaboration (Boschma, 2005; Zeller, 2004). Boschma (2005) argued that cognitive proximity is crucial to communicate, absorb, comprehend, and process new information between partners. Besides, Boschma & Frenken (2010) argued that agents tend to select close partners in terms of geographical proximity and from the knowledge base point of view. In other words, being capable of absorbing external knowledge is a prerequisite for collaboration among local agents (Boschma, 2005). Moreover, several empirical studies reported the positive effect of cognitive proximity on collaboration (Bergé, 2017; Cunningham & Werker, 2012; Ding, 2011; Jaffe & Hu, 2003; Jaffe & Trajtenberg, 1999). Since AI is not homogeneous and includes many different sub-fields like any other scientific discipline, if agents are from various sub-fields, they might encounter challenges in collaboration. In other words, we should expect cognitive distance to influence scientific collaboration negatively.

2.8. Research gap

As discussed thoroughly in this chapter, several studies have examined the determinants of scientific collaboration among regions or organizations using the notion of proximity. However, research works that address the efficiency of funding long-distance scientific collaborations are scarce. Bergé (2017) worked on the same objective among studies with a relevant topic. Although his work is one of a kind, it is limited in some respects. First, he focused on the research collaboration among regions in five European countries, i.e., Germany, Spain, Italy, France, and the United Kingdom. Second, his study is limited to the co-publications in the field of chemistry between 2001and 2005.

This thesis departs from the previous literature from several aspects. First, we study the geo-pattern of scientific collaborations among researchers, not geographical regions or organizations. Second, we start with studying the geo-pattern of scientific collaboration among Canadian AI researchers and then extend the domain of study to the United States, Europe, and the entire world. This step-by-step strategy will provide a better insight towards understanding the role of different dimensions of proximity on scientific collaboration. Third, this study covers co-publications among AI researchers from 2000 to 2019. As AI has been among the fast-growing fields since 2000 (PwC, 2017), many scientific papers in AI were published after 2000. Therefore, this study is not suffering from a limited time span. Finally, we use regression analysis and machine learning classification to understand the relationship between scientific collaboration and its determinants better.

2.9. Research questions

The following research questions will be answered at the end of this study:

1) How does geographical proximity influence the probability of future collaboration?
2) Can network proximity substitute geographical proximity to encourage long-distance scientific collaborations?

3. METHODOLOGY

3.1.Models

Based on the research objectives defined in the previous chapter, two models are built. The first model is constructed to investigate the impact of geographical proximity and other dimensions of proximity on the probability of future scientific collaboration. Figure 1 demonstrates the conceptual diagram of model 1.

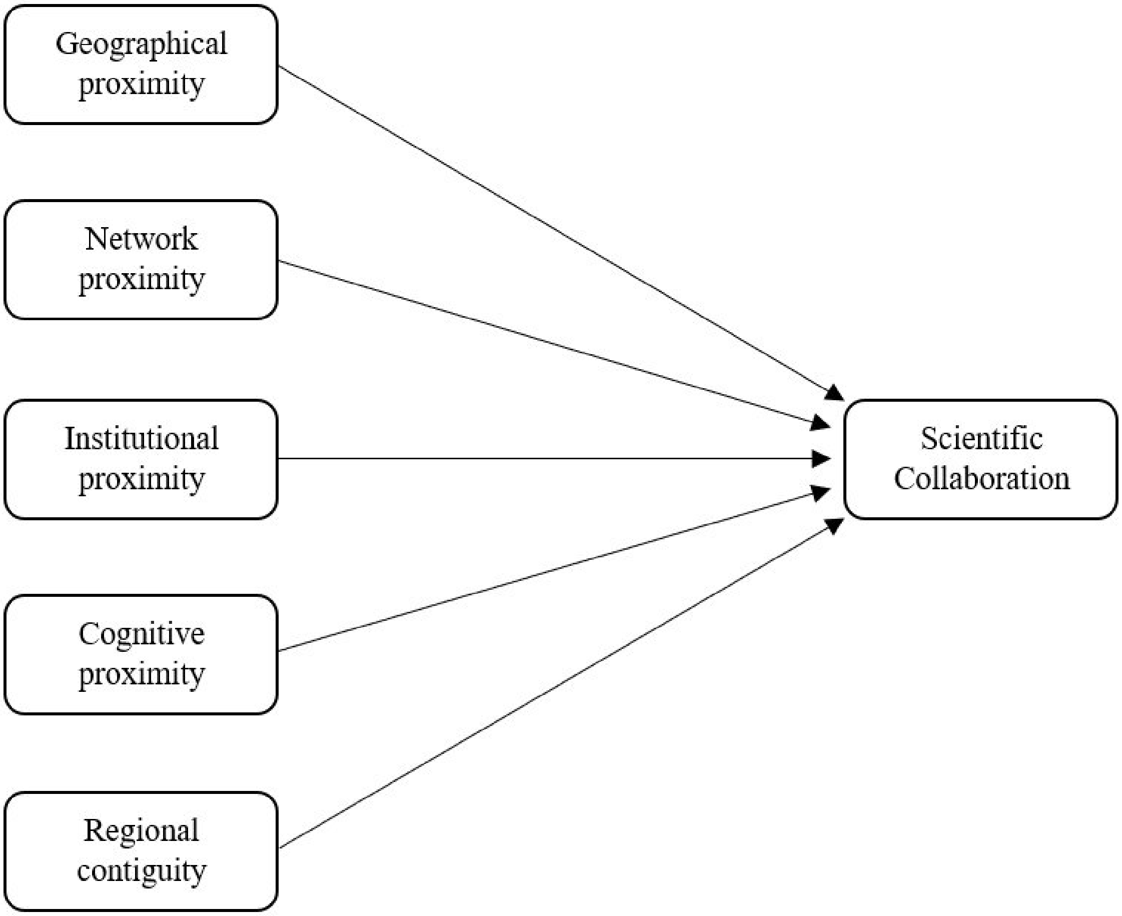

Figure 1. Model 1 conceptual diagram

To investigate the relationship between geographical proximity and scientific collaboration, we developed the following null hypothesis:

$H_0$1: A higher level of geographical proximity does not enhance the likelihood of future scientific collaboration.

The second hypothesis can be developed to study the relationship between network proximity and scientific collaboration as:

$H_0$2: A higher level of network proximity does not enhance the likelihood of future scientific collaboration.

However, the effect of institutional proximity on scientific collaboration can be studied at two levels: province level and national level. So, the third null hypothesis can be developed as two separate hypotheses as below:

$H_0$3a: Being from the same province does not enhance the likelihood of future scientific collaboration.

$H_0$3b: Being from the same country does not enhance the likelihood of future scientific collaboration.

The fourth hypothesis is dedicated to the relationship between cognitive proximity and scientific collaboration:

$H_0$4: A higher level of cognitive proximity does not enhance the likelihood of future scientific collaboration.

Like institutional proximity, regional contiguity (sharing a boundary in common) can have two forms: province contiguity and country contiguity. So, the fifth null hypothesis can be developed as two separate hypotheses as below:

$H_0$5a: Province contiguity does not enhance the likelihood of future scientific collaboration.

$H_0$5b: Country contiguity does not enhance the likelihood of future scientific collaboration.

The second model is based on the second objective of the study, i.e., investigating the interaction influence of geographical proximity and network proximity on scientific collaboration. Figure 2 presents the conceptual diagram of the second model.

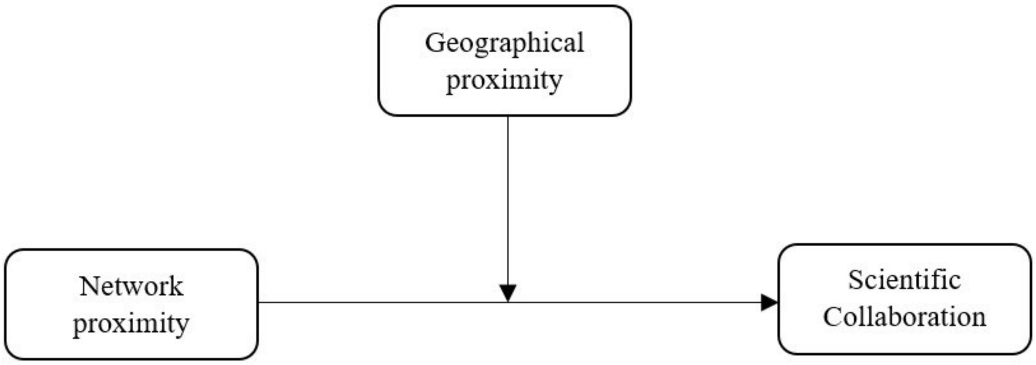

Figure 2. Model 2 conceptual diagram

The hypothesis that can be developed based on the second model is as follows:

$H_0$6: Geographical proximity does not moderate the effect of network proximity on the likelihood of future scientific collaboration.

3.2.Variables

3.2.1. Co-publication

In this study, the dependent variable is co-publication which is employed as a proxy of scientific collaboration in several studies (Dahlander & McFarland, 2013; Hoekman et al., 2009; Ponds et al., 2007). This binary variable is equal to 1 if authors had at least one co-publication during the 2-year time window and 0 otherwise.

3.2.2. Geographical proximity

Following different studies (see, e.g., Bergé, 2017; Cunningham & Werker, 2012), we used direct spatial or "as the crow flies" distance between researchers' affiliations to measure their geographical proximity. To calculate the geographical distance between researchers, we geocoded their geographical location by turning their affiliation addresses into geographical coordinates using Google Geocoding API (Google, n.d.).

Finally, $distance_{i,j}$ as the measure of geographical proximity between author i and author j was calculated using the Haversine formula (Inman, 1835) as follow:

$$a = \sin^2(\frac{\varphi_j - \varphi_i}{2}) + \cos\varphi_i \times \cos\varphi_j \times \sin^2(\frac{\lambda_j - \lambda_i}{2}) \qquad (1)$$

$$c_{i,j} = 2 \times \text{atan2}(\sqrt{a}, \sqrt{1-a}) \qquad (2)$$

$$distance_{i,j} = R \times c_{i,j} \qquad (3)$$

where $distance_{i,j}$ is the geographical distance between author i and author j in kilometres, φ_i and λ_i are, respectively, the latitude and the longitude of the first author's affiliation in radian, φ_j and λ_j are, respectively, the latitude and the longitude of the second author's affiliation in radian, atan2 is the 2-argument arctangent, and R is the average radius of the earth (~ 6,373 kilometres).

3.2.3. Institutional proximity

As institutional proximity usually refers to the collection of practices, laws, and rules defined by the geographical setting (Boschma, 2005), we defined this variable at two levels, i.e., province and country, in this thesis. In the first scenario, we defined a binary variable as a proxy for institutional proximity, capturing the effect of same province collaborations, which takes 0 for two given researchers if they are in the same province, and 1 otherwise. In the second scenario where we extend the study to the United States, we defined two binary variables as proxies for institutional proximity, one to capture the effect of the same province and the other to capture the impact of the same country. Similarly, they take 0 when two researchers are in the same province/country and 1 otherwise. In the third and fourth scenarios where we extend the study to Europe and the entire world, we defined one binary variable to capture the effect of the same country. Again, it takes 0 when two researchers are in the same country and 1 otherwise.

3.2.4. Network proximity

To estimate the network proximity of researchers, we used the “Total Expected Number of Bridging Paths (TENB)” measure, introduced by Bergé (2017). As he developed this metric to measure the network proximity between two different geographical districts based on their indirect connections, we customized it to capture the network proximity between researchers via the following equation:

$$TENB_{i,j} = \sum_{k} \frac{g_{i,k} \times g_{j,k}}{n_k} \qquad (4)$$

where $TENB_{i,j}$ is the total expected number of bridging paths between author i and author j, $g_{i,k}$ is the total number of co-publications between author i and author k, $g_{j,k}$ is the total number of co-publications between author j and author k, n_k is the total number of publications by author k. If

$TENB_{a,b} > TENB_{a,c}$, it means that author a and author b are closer with respect to their indirect connections than author a and author c.

One could argue that the network proximity originating from the direct connections between researchers may also be necessary for triggering new collaborations. Yet, since the identification of network proximity is based on network connections, direct collaborations between the two researchers would directly influence their level of network proximity. As this thesis is trying to explain collaborations, this would create a problem of reverse causality (Bergé, 2017). In consequence, using the concept of TENB means this problem is avoided at the cost of neglecting possible network proximity originating from direct ties.

3.2.5. Cognitive proximity

We implemented the LDA[5] topic modelling technique (Blei et al., 2003) to estimate the cognitive proximity among authors. Topic modelling, in general, is a type of statistical modelling for discovering distinct topics that occur in a collection of documents. LDA is a topic model which helps classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modelled as Dirichlet distributions. In this study, first, we created a vector of topics for each publication. Some data preprocessing steps are required to maximize the likelihood of capturing the main topics discussed in each publication:

Step 1: Concatenating title and abstract

Step 2: Removing punctuation marks

Step 3: Lowercasing words

Step 4: Lemmatizing

Step 5: Removing stop words

[5] Latent Dirichlet Allocation

Step 6: Building Bigrams and Trigrams

Step 7: TF-IFD[6] removal

Step 8: Creating the dictionary

Then, to find the optimum number of the topics, we calculated the coherence measure for a range of topics from 5 to 35. A set of statements or facts is said to be coherent, if they support each other. Thus, a coherent fact set can be interpreted in a context that covers all or most of the facts. In this study, we calculated the coherence value based on Röder et al. (2015). First, the word set t is segmentated into a set of pairs of word subsets S. Second, word probabilities P are computed based on a given reference corpus. Both, the set of word subsets S as well as the computed probabilities P are consumed by the confirmation measure to calculate the agreements ϕ of pairs of S. Last, those values are aggregated to a single coherence value c. Figure 3 shows the coherence value for a given number of topics. As seen, the highest coherence value (0.498) is associated with the number of topics equal to nine.

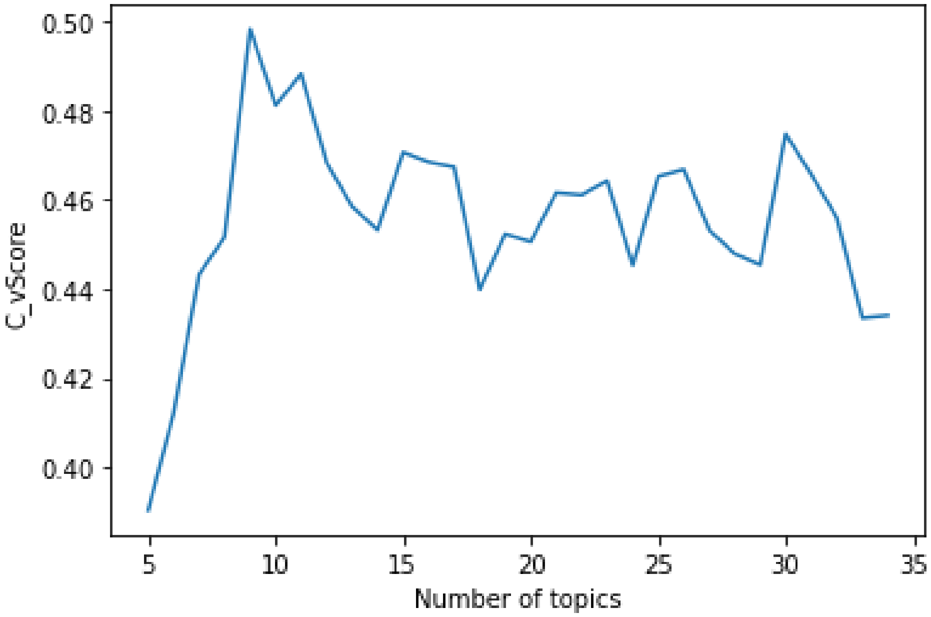

Figure 3. Topic modelling coherence values

[6] Term Frequency - Inverse Document Frequency

Therefore, we chose nine as the number of topics for LDA topic modelling. As a result, a vector with nine elements was generated for each publication. The value of each element shows the weight of the respective topic in that specific publication. To see if topics are distinct enough, we generated the intertopic distance map, which visualizes the topics in a two-dimensional space (Figure 4). The area of these topic circles is proportional to the number of words that belong to each topic across the dictionary. Besides, top-20 most relevant terms for the first topic, as an instance, are presented in Figure 4.

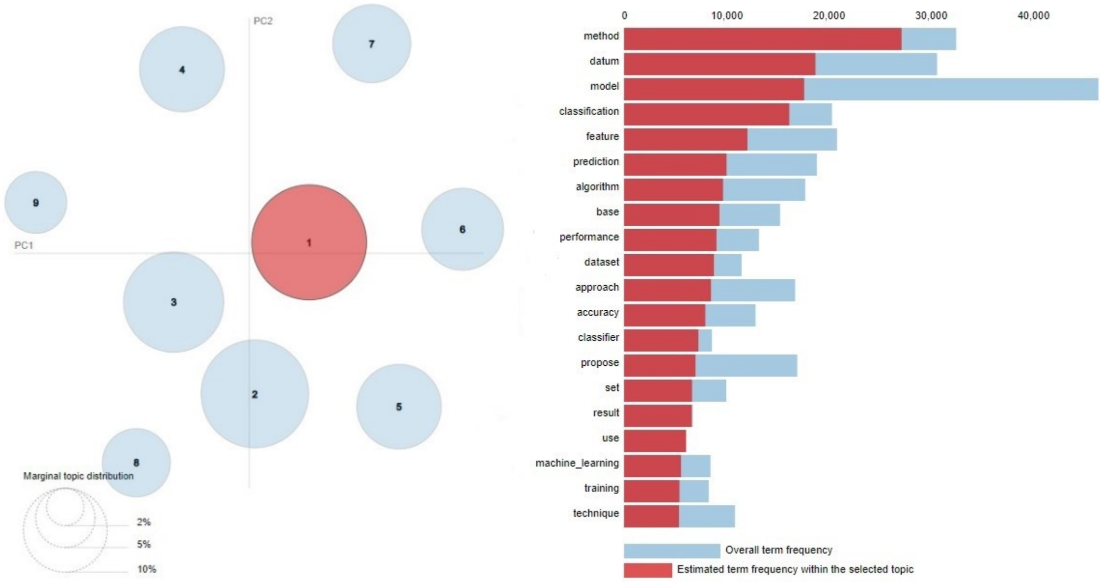

Figure 4. Intertopic distance map for topic one

After attaining the topic vectors for all publications, the knowledge base vector of each author was calculated as the average of topic vectors of all their works published during the past 3-year window. Therefore, for any specific author to have the knowledge base vector, they must have at least one publication during that period.

Finally, the $Cognitive\ distance_{i,j}$ as the measure of cognitive proximity between author i and author j was defined as:

$$Cognitive\ distance_{i,j}\ = 1 - corr(s_i, s_j) \qquad (5)$$

where s_i and s_j are the knowledge base vectors of author i and author j, respectively; and $corr(s_i, s_j)$ is the correlation between the knowledge base vectors of author i and author j. As inferred from equation (5), when two authors have the same knowledge base vectors, the correlation between their knowledge base vectors would be equal to 1; consequently, their cognitive distance is 0, implying the highest level of cognitive proximity. At the other extreme, when the correlation between knowledge base vectors of two authors is equal to -1, their cognitive distance would be equal to 2, implying the lowest level of cognitive proximity.

3.2.6. Regional contiguity

Following Bergé (2017), we added two variables of regional contiguity to capture the effects of geography that are not seized by geographical proximity. These binary variables that capture the impact of collaboration between researchers from contiguous provinces/countries take the value of 0 for two given researchers if they are in two contiguous provinces/countries and 1 otherwise.

3.3.Data

The information on publications was extracted from Elsevier's Scopus. Data collection and preparation involved several steps. First, the bibliographic data, including but not limited to title, abstract, keywords, date of publication, author list, etc., were retrieved from Scopus, filtering in research articles, conference papers, book chapters, and books published from 2000 to 2019. Only publications for which both title and abstract were available were included.

We used the ("artificial intelligence" OR "machine learning" OR "deep learning") search query to extract AI-related publications where at least one of the mentioned phrases appeared in the title, abstract or the keywords section of the publication. The result of running this query was the main database of the study with 45,738 publications by 153,720 authors from 162 different countries during 2000 - 2019.

We filtered the publications based on each scenario before forming the datasets to answer the research questions for different scenarios explained in the previous chapter. In the first scenario, we just included the publications with Canadian authors. The result was a subset of the main database with 670 publications by 1,923 authors. Then, we expanded the data to Canada and the United States in the second scenario. There are 7,180 publications by 22,727 authors from these two countries in this scenario. Adding European publications, there are 20,508 publications by 64,852 researchers from Canada, the United States, and European countries in the third scenario. Finally, the last scenario covers entire publications in the main database.

3.4.Dataset formation

We followed Bergé (2017)'s approach in forming datasets and constructed the dependent and independent variables in different time windows to prevent simultaneity biases. Thus, we considered sliding 2-year and 3-year time windows to calculate the dependent variable and independent variables, respectively.

To form the dataset for each scenario, we first developed the network of potential co-authorship that contains all possible scientific collaborations among authors who had at least one publication during the 2-year time window and at least one publication during the 3-year time window. Figures 5 to 8 compare the number of authors and their collaborations during each time window for each scenario. Consequently, we formed a network for each 2-year time window in which each node represents an author, and each edge represents a possible collaboration between two authors and a dataset record. Then, we calculated the dependent variable and independent variables for each record. For every given pair of authors, if they had at least one co-publication during the 2-year time window, the co-publication variable takes 1, otherwise 0. Independent variables were calculated based on authors' publications during the respective 3-year time window. These steps were repeated for all 2-year time windows as specified in Table 1. With concatenating datasets resulting from each 2-year network, we built the final dataset for each scenario.

Table 1. Time windows to form datasets

Network formation periods	Independent variables estimation periods
2004 - 2005	2001 - 2003
2006 - 2007	2003 - 2005
2008 - 2009	2005 - 2007
2010 - 2011	2007 - 2009
2012 - 2013	2009 - 2011
2014 - 2015	2011 - 2013
2016 - 2017	2013 - 2015
2018 - 2019	2015 - 2017

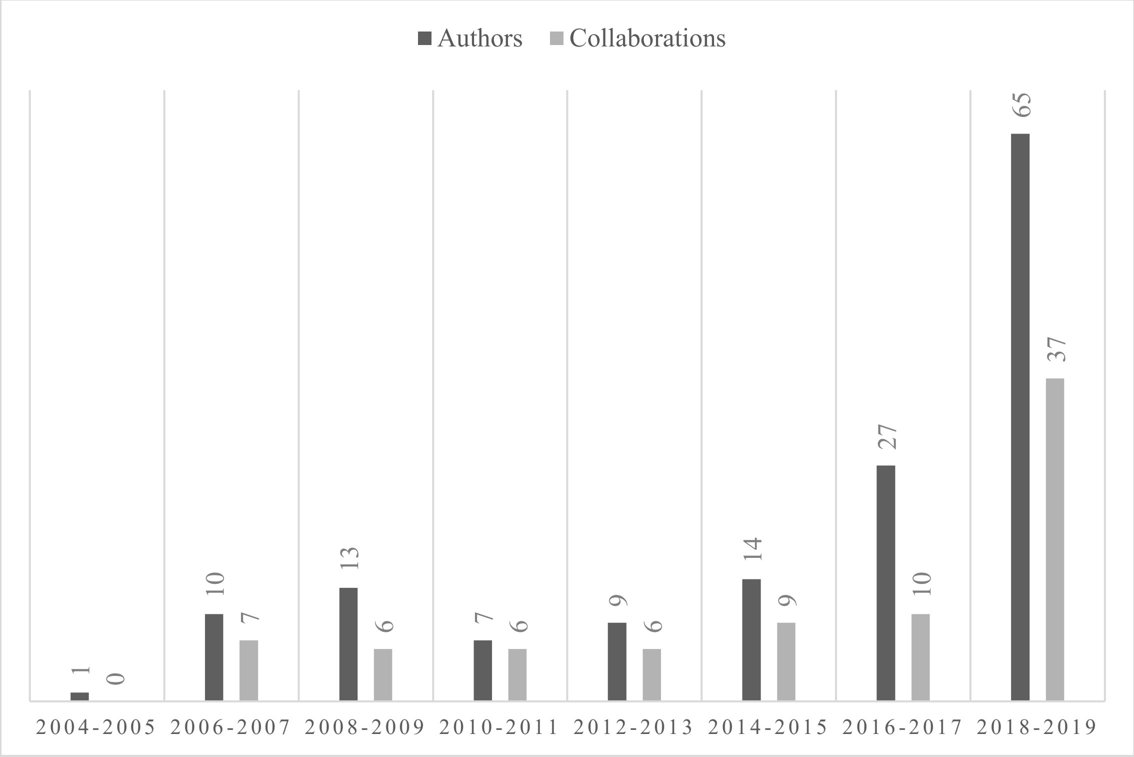

Figure 5. Number of authors and collaborations in the first scenario

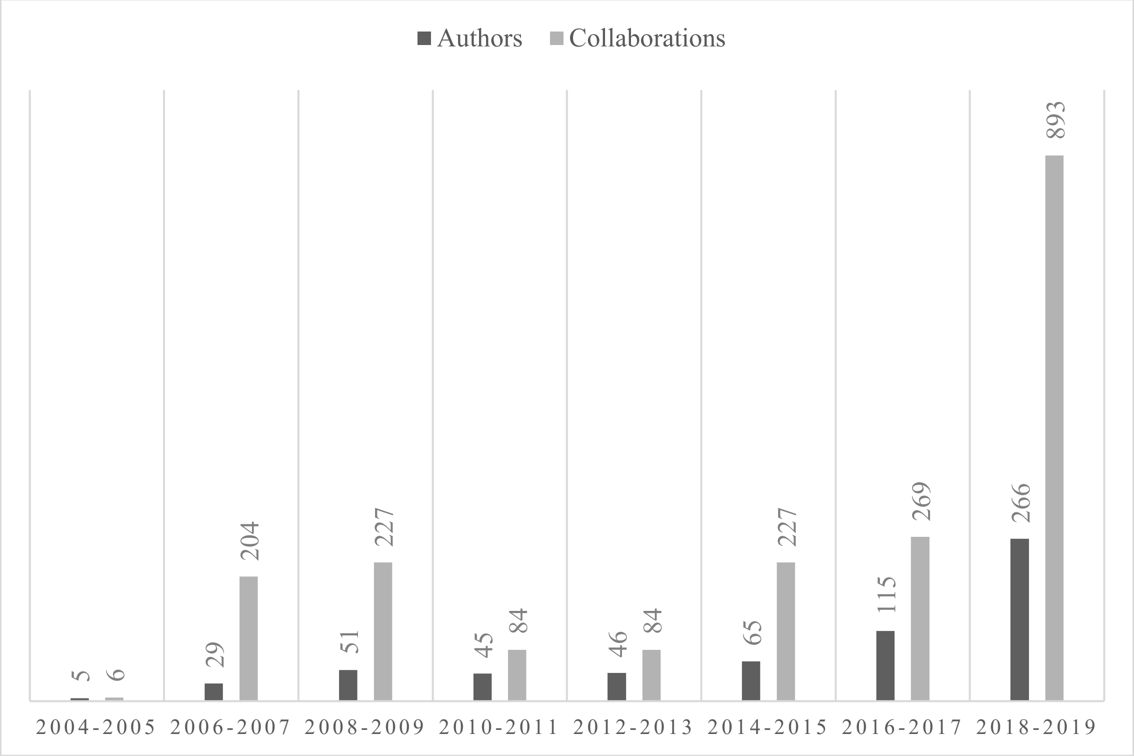

Figure 6. Number of authors and collaborations in the second scenario

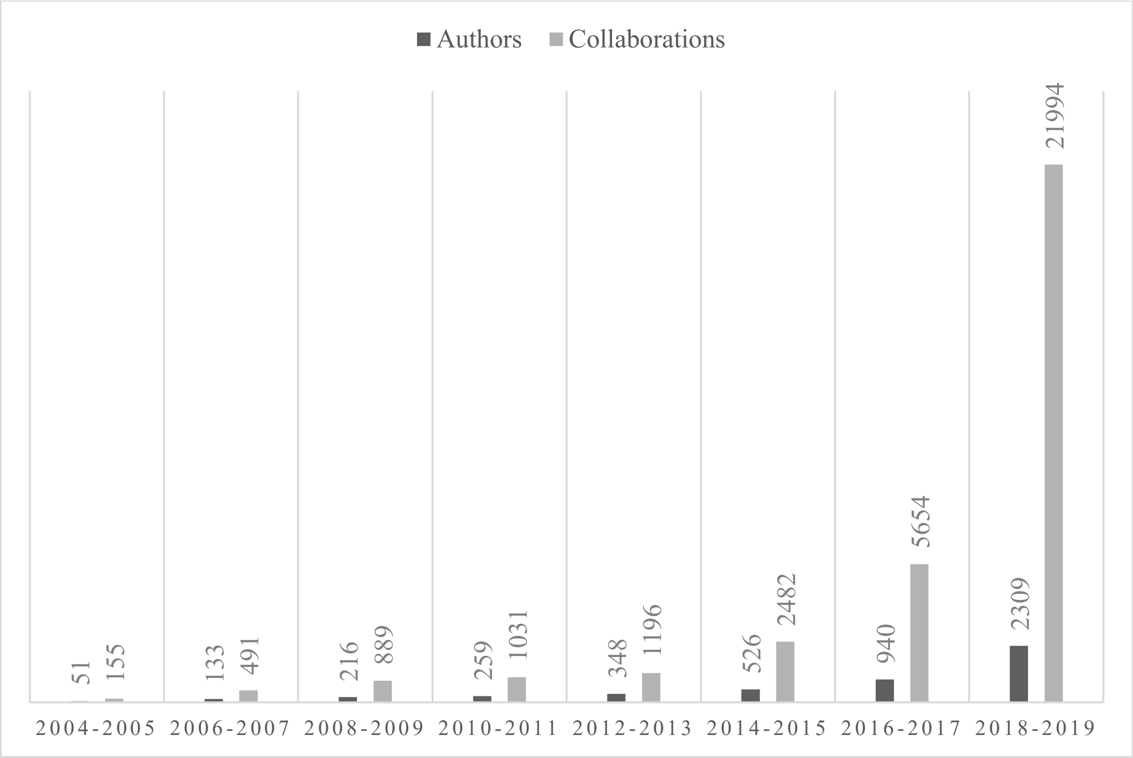

Figure 7. Number of authors and collaborations in the third scenario

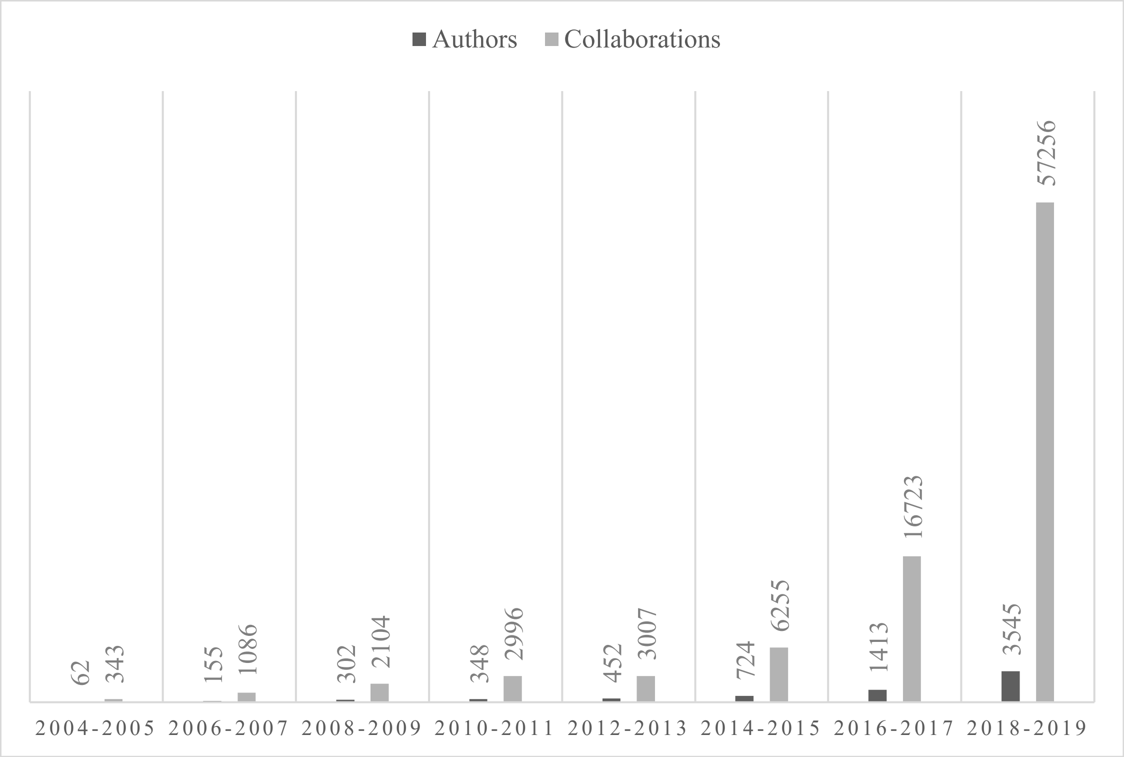

Figure 8. Number of authors and collaborations in the fourth scenario

3.5.Data analysis methods

To address the hypotheses developed in the previous section, we used two data analysis methods: logistic regression and machine learning classification. As these methods use different approaches to answer the same research questions, comparing the outcomes would provide a better insight towards understanding the relationship between dependent variables/features and the independent variable/target.

3.5.1. Logit regression

Logistic regression models the probabilities for problems with a binary dependent variable. It is an extension of the linear regression model for classification problems. The problem of using linear regression for classification problems with two possible outcomes is that it considers classes as numbers (0 and 1) and simply interpolates between the points. Therefore, we cannot interpret the results as probabilities. Also, the linear model extrapolates points meaning that the predictions could be higher than 1 and lower than 0, which does not make sense for a classification problem.

Moreover, there is no meaningful threshold to distinguish one class from another because the predicted outcome is just a linear interpolation between points and not a probability. One solution to tackle those limitations of using linear regression for classification problems is logistic regression. It uses the logistic function to squeeze the output of a linear equation between 0 and 1. The function is defined as follows:

$$logistic(\eta) = \frac{1}{1 + \exp{(-\eta)}} \qquad (6)$$

We can wrap the linear equation into the logistic function to have probabilities between 0 and 1 as the output for classification problems.

$$P(y = 1) = \frac{1}{1 + \exp{(-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n))}} \qquad (7)$$

Since the outcome in logistic regression is a probability between 0 and 1, the interpretation of the coefficients differs from that of the coefficients in linear regression. If we reformulate the equation for the interpretation so that only the linear term is on the right side of the formula, then we have:

$$\ln\left(\frac{P(y=1)}{1-P(y=1)}\right) = \ln\left(\frac{P(y=1)}{P(y=0)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \qquad (8)$$

The term in the ln () function, i.e., probability of class 1 divided by the probability of class 0, is called “odds.” To figure out how the prediction changes when the feature x_j is changed by 1 unit, we need a little shuffling of the terms by applying the exp () function to both sides of the equation (7):

$$\frac{P(y=1)}{1-P(y=1)} = odds = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n) \qquad (9)$$

Looking at the ratio of two predictions:

$$\frac{odds_{x_j+1}}{odds_{x_j}} = \frac{\exp\big(\beta_0 + \beta_1 x_1 + \cdots + \beta_j (x_j + 1) + \cdots + \beta_n x_n\big)}{\exp\big(\beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j + \cdots + \beta_n x_n\big)} \qquad (10)$$

and applying the following rule:

$$\frac{\exp\,(a)}{\exp\,(b)} = \exp(a-b) \qquad (11)$$

we have:

$$\frac{odds_{x_j+1}}{odds_{x_j}} = \exp\big(\beta_j\big(x_j+1\big) - \beta_i x_j\big) = \exp\big(\beta_j\big) \qquad (12)$$

suggesting that for numerical features, if the value of the feature x_j increases by one unit, the estimated odds change by a factor of exp (β_j), and for binary categorical features, changing the feature x_k from the reference category to the other category changes the estimated odds by a factor of exp (β_k).

Considering the Boolean nature of the dependent variable in this study (1 for at least one co-publication for a given pair of researchers, 0 otherwise), we use logit regression to study the relationship between the independent variables and the dependent variable. The model is specified as follows:

$$P(Co\text{-}publication = 1) = F(\beta' X) \qquad (12)$$

where Co-publication is a binary variable that takes 1 if two researchers have at least one co-publication, otherwise 0; the F function is the cumulative probability density function of the logit distribution presented in the equation (6); β' is the vector of the coefficients, and X is the vector of independent variables. To examine the hypotheses developed in the first mode, the vector X includes the natural logarithm of geographical distance, the natural logarithm of TENB, the binary variable representing the institutional proximity, the cognitive distance, and the binary variable representing regional contiguity. Besides, the interaction between geographical distance and TENB was added to the vector X to examine the hypothesis **H$_0$6** from the second model demonstrated in Figure 2.

3.5.2. Machine learning classification

Since the study aims to understand the influence of different dimensions of proximity on forming scientific collaboration, we can formulate the problem as a supervised machine learning classification. In this sense, the target would be the binary variable of co-publication that takes 1 if any given pair of researchers had any co-publication and 0 otherwise. Also, the independent variables explained in the logistic regression analysis section would be the features that predict future co-publication among researchers.

While there are many algorithms for supervised learning out there, we implemented those used to analyze factor contributions in academic data. We find that many authors in the literature make use of a logistic regression classifier (Bethard & Jurafsky, 2010; Dong et al., 2015; Getoor, 2005). In this case, the statistical analysis performed by logistic regression is used iteratively for training as a supervised classifier. Logistic regression classifier is often implemented for classification purposes due to their output consisting of linear combination of the variables with weights, which would provide insights on variable importance (Breiman, 2003). In addition, other prominent models such as support vector machines (SVM) would also have the potential to produce accurate predictions. Yet, SVM is arguably better when applied to regression problems (i.e. when the response variable is of continuous nature) than for binary classification (Bethard & Jurafsky, 2010; Breiman, 2003). We also implement models based on decision trees (DT) classifiers, which were introduced in the 1960's (Dong et al., 2015). Decision trees are one of the most popular methods for data mining, due to their ease of use and interpretation, robustness even with missing values, and their flexibility to use both discrete (categorical) and continuous variables (Song & Lu, 2015). Originating from decision trees, random forests (RF) algorithms (Breiman, 2003) grow an ensemble of trees and lets them vote for the most prominent class. Notably, RF are considered to be accurate classifiers, showing comparable or even better prediction performance than other learning methods (Breiman, 2001; Xu, 2013). They have been extensively used in classification problems (e.g. Dong et al., 2015; Lichtenwalter et al., 2010; Sarigöl et al., 2014) having led to significant improvements in classification accuracy (Breiman, 2003) thanks to their particular advantages, such as robustness against overfitting (a potential problem with decision trees).

There are many measures to evaluate a machine learning classifier. Following several studies using machine learning for link prediction (Aiello et al., 2012; Clauset et al., 2008; Moradabadi & Meybodi, 2017; Schall, 2014), we use AUC[7] to evaluate the classification results. Moreover, AUC

[7] Area Under the Curve

is a classification threshold invariant metric. It measures the quality of the model's predictions irrespective of what classification threshold is chosen.

We implement a multi-step strategy to find the model best fits the dataset and generates the highest possible AUC score. In the first step, we split the dataset into two sets of the train (90%) and test (10%) datasets. The test dataset would be untouched to test the final model performance. As the data is imbalanced in all scenarios, we oversample the data via SMOTE[8] (Chawla et al., 2002), in the second step. Then, we apply various machine learning classifiers (Logistic Regression, Nearest Neighbors, Gaussian Naive Bayes, Support Vector Machines, Random Forest, and Extreme Gradient Boosting) with default hyperparameters to train models through five-fold stratified cross-validation on the train dataset. In each cross-validation fold, the train dataset is further split into a train and a validation dataset containing 80% and 20% of the original train dataset, respectively. Then, the classifier with the highest cross-validation AUC scores will be selected for hyperparameters tuning.

We implement a two-step hyperparameters tuning in this study. First, the random search method is used to find the approximate range for each parameter. Then, the more accurate grid search method is used to fine-tune the parameters found in the random search. After finding the best hyperparameters that maximize the AUC score, the final model will be tested with the test dataset put aside in the first step. To analyze the effect of each feature on predicting future scientific collaboration, we use the concept of SHAP[9] values developed by Lundberg & Lee (2017). They introduced this game-theoretic approach that assigns each feature an importance value for a particular prediction.

[8] Synthetic Minority Over-sampling Technique

[9] SHapley Additive exPlanations

4. RESULTS AND DISCUSSION

In this chapter, the results of the data analysis are provided. First, the descriptive statistics are presented. Then, after discussing the correlation analysis matrices, the logistics regression analysis and machine learning classification results are presented and discussed.

4.1.Descriptive statistics

This section presents the descriptive statistics of each scenario via tables 2 to 5. In the first scenario, which covers collaborations among Canadian authors, there are 2,702 observations. The dataset is imbalanced as observations associated with pairs of authors with at least one co-publication form only 3% of total observations. Analyzing the statistics of geographical distance among authors is very informative; although the maximum geographical distance in the dataset is more than 4,400 kilometres, the maximum geographical distance for authors who had a co-publication is only 817 kilometres. Moreover, 75% of all collaborations happened among authors whose affiliations were close geographically (less than 7 kilometres). Comparing the average of TENB between the entire dataset and the subset of collaborations also shows the role of network proximity in collaboration. The average of TENB for the pair of authors with at least one co-publication is 26 times greater than the average of TENB for the entire dataset. Comparing the average and maximum cognitive distance among the whole dataset and the subset of collaborations also implies the critical role of cognitive proximity in forming scientific collaboration. The average cognitive distance among authors with co-publication is around 17 times smaller than the average cognitive distance in the entire dataset. Besides, the maximum cognitive distance in the whole dataset is more than three times greater than the maximum cognitive distance for authors in the subset of collaboration. Belonging to the same province also seems essential to form collaborations, as only 5% of total collaborations happened among authors from different provinces.

Table 2. Descriptive statistics – First scenario

	Co-publication		Geo. distance (km)		TENB		Cog. distance		Different provinces		Not contiguous	
	Entire dataset	Collaborations	Entire dataset	Collaborations	Entire dataset	Collaborations	Entire dataset	Collaborations	Entire dataset	Collaborations	Entire dataset	Collaborations
mean	0.03	1.00	1,460.63	62.72	0.06	1.56	0.83	0.05	0.72	0.05	0.70	0.95
std	0.17	0.00	1,341.75	165.35	0.36	0.98	0.42	0.12	0.45	0.22	0.46	0.22
min	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	0.00	1.00	335.64	0.00	0.00	1.00	0.51	0.00	0.00	0.00	0.00	1.00
50%	0.00	1.00	793.20	0.16	0.00	2.00	0.91	0.00	1.00	0.00	1.00	1.00
75%	0.00	1.00	2,883.09	6.84	0.00	2.00	1.18	0.01	1.00	0.00	1.00	1.00
max	1.00	1.00	4,440.56	817.86	3.00	3.00	1.74	0.57	1.00	1.00	1.00	1.00
Total number of observations: 2,702 Total number of collaborations: 81												

Analyzing the statistics of variables in the other scenarios shows the same pattern. In all scenarios, the average geographical distance among authors in the collaboration subset is shorter than the average geographical distance in the entire dataset. Moreover, the average TENB of the whole dataset in all scenarios is considerably lower than the average TENB for their respected subset of collaboration, implying the critical role of network proximity in forming scientific collaborations. Also, comparing the average and maximum values for cognitive distance among authors in the collaboration subsets with the same parameters for authors in the entire datasets clearly shows that cognitive proximity influences forming scientific collaboration. Besides, discussing the average of different country binary variables show that authors from the same country are more likely to collaborate. Only 1% of total collaborations happened among authors from different countries in the second scenario. Although this figure is higher for the third (36%) and the fourth (40%) scenarios, it still shows that institutional proximity is an essential factor in scientific collaboration.

Table 3. Descriptive statistics – Second scenario

	Co-publication		Geo. distance (km)		TENB		Cog. distance		Different provinces		Different countries		Not contiguous	
	Entire dataset	Collaborations	Entire dataset	Collaborations	Entire dataset	Collaborations	Entire dataset	Collaborations	Entire dataset	Collaborations	Entire dataset	Collaborations	Entire dataset	Collaborations
mean	0.04	1.00	1,874.81	586.71	0.11	2.42	0.84	0.16	0.92	0.29	0.15	0.01	0.97	0.99
std	0.20	0.00	1,279.83	1131.26	0.78	2.92	0.41	0.30	0.28	0.46	0.36	0.11	0.17	0.10
min	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	0.00	1.00	760.37	0.00	0.00	0.00	0.54	0.00	1.00	0.00	0.00	0.00	1.00	1.00
50%	0.00	1.00	1,600.59	0.73	0.00	1.18	0.94	0.00	1.00	0.00	0.00	0.00	1.00	1.00
75%	0.00	1.00	3,006.78	613.60	0.00	3.5	1.18	0.16	1.00	1.00	0.00	0.00	1.00	1.00
max	1.00	1.00	7,157.89	7157.89	20.00	20.00	1.81	1.71	1.00	1.00	1.00	1.00	1.00	1.00
Total number of observations: 47,599 Total number of collaborations: 1,994														

Table 4. Descriptive statistics – Third scenario

	Co-publication		Geo. distance (km)		TENB		Cog. distance		Different countries		Different continents		Not contiguous	
	Entire dataset	Collaborations	Entire dataset	Collaborations	Entire dataset	Collaborations	Entire dataset	Collaborations	Entire dataset	Collaborations	Entire dataset	Collaborations	Entire dataset	Collaborations
mean	0.01	1.00	4,336.88	1,870.27	0.06	5.39	0.88	0.26	0.78	0.36	0.50	0.19	0.93	0.96
std	0.10	0.00	3,137.02	2,719.77	1.03	8.61	0.39	0.26	0.41	0.48	0.50	0.40	0.26	0.21
min	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	0.00	1.00	1,168.52	0.58	0.00	0.33	0.59	0.05	1.00	0.00	0.00	0.00	1.00	1.00
50%	0.00	1.00	4,327.86	381.23	0.00	5.39	0.97	0.26	1.00	0.00	0.00	0.00	1.00	1.00
75%	0.00	1.00	6,979.76	2,746.68	0.00	5.39	1.20	0.26	1.00	1.00	1.00	0.0	1.00	1.00
max	1.00	1.00	13,713.99	11,594.67	48.83	48.83	1.97	1.75	1.00	1.00	1.00	1.00	1.00	1.00
Total number of observations: 3,372,222 Total number of collaborations: 33,892														

Table 5. Descriptive statistics – Fourth scenario

	Co-publication		**Geo. distance (km)**		**TENB**		**Cog. distance**		**Different countries**		**Different continents**		**Not contiguous**	
	Entire dataset	Collaborations	Entire dataset	Collaborations	Entire dataset	Collaborations	Entire dataset	Collaborations	Entire dataset	Collaborations	Entire dataset	Collaborations	Entire dataset	Collaborations
mean	0.01	1.00	7,050.29	2,399.09	1.97	170.3	0.89	0.14	0.88	0.40	0.70	0.26	0.95	0.93
std	0.11	0.00	4,360.46	3,079.16	27.32	189.5	0.39	0.26	0.33	0.49	0.46	0.44	0.21	0.26
min	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	0.00	1.00	2,943.61	242.80	0.00	1.00	0.62	0.00	1.00	0.00	0.00	0.00	1.00	1.00
50%	0.00	1.00	7,593.68	1,229.41	0.00	94.16	0.98	0.04	1.00	0.00	1.00	0.00	1.00	1.00
75%	0.00	1.00	9,857.26	3,461.81	0.00	261.7	1.20	0.14	1.00	1.00	1.00	1.00	1.00	1.00
max	1.00	1.00	19,966.06	19,438.04	491.7	491.7	1.97	1.75	1.00	1.00	1.00	1.00	1.00	1.00
Total number of observations: 7,762,616 Total number of collaborations: 89,766														

4.2. Correlation analysis

Tables 6 to 9 present the correlation matrix for different scenarios. There is a negative correlation between co-publication and geographical distance in all scenarios, implying the negative effect of geographical distance on scientific collaboration. In all scenarios, the strong and positive correlation between co-publication and TENB reflects the critical role of network proximity on scientific collaboration. Besides, as expected, the correlation between co-publication and cognitive distance in all scenarios is negative.

Table 6. Correlation's matrix – First scenario

	Co-publication	**Geo. distance**	**TENB**	**Cog. distance**	**Different provinces**	**Not contiguous**
Co-publication	1.0^{***}					
Geo. distance	-0.18^{***}	1.0^{***}				
TENB	0.74^{***}	-0.16^{***}	1.0^{***}			
Cog. distance	-0.33^{***}	0.17^{***}	-0.32^{***}	1.0^{***}		
Different provinces	-0.26^{***}	0.59^{***}	-0.24^{***}	0.12^{***}	1.0^{***}	
Not contiguous	0.09^{***}	0.39^{***}	0.09^{***}	0.03	-0.4^{***}	1.0^{***}

Level of statistical significance: ***1%

Table 7. Correlation's matrix – Second scenario

	Co-publication	**Geo. distance**	**TENB**	**Cog. distance**	**Different provinces**	**Different countries**	**Not contiguous**
Co-publication	1.0^{***}						
Geo. distance	-0.21^{***}	1.0^{***}					
TENB	0.62^{***}	-0.15^{***}	1.0^{***}				
Cog. distance	-0.35^{***}	0.11^{***}	-0.27^{***}	1.0^{***}			
Different provinces	-0.47^{***}	0.40^{***}	-0.34^{***}	0.17^{***}	1.0^{***}		
Different countries	-0.08^{***}	0.07^{***}	-0.05^{***}	0.01	0.13^{***}	1.0^{***}	
Not contiguous	0.02^{***}	0.18^{***}	0.02^{***}	0.02^{***}	-0.05^{***}	-0.21^{***}	1.0^{***}

Level of statistical significance: ***1%

In the first two scenarios in which the binary variable of different provinces is included in the model, the negative correlation between co-publication and being from different provinces confirms the effect of institutional proximity on scientific collaboration. However, as expected, there is a positive and strong correlation between geographical distance, different countries, and different continents. For example, the correlation among geographical distance and different continents in the third and fourth scenarios are 92% and 80%, respectively. Thus, to avoid multicollinearity problems, we exclude the different continents variable from those scenarios in further analyses.

Table 8. Correlation's matrix – Third scenario

	Co-publication	**Geo. distance**	**TENB**	**Cog. distance**	**Different countries**	**Different continents**	**Not contiguous**
Co-publication	1.0***						
Geo. distance	-0.08***	1.0***					
TENB	0.52***	-0.04***	1.0***				
Cog. distance	-0.16***	0.03***	-0.11***	1.0***			
Different countries	-0.10***	0.45***	-0.05***	0.04***	1.0***		
Different continents	-0.06***	0.92***	-0.03***	0.02***	0.52***	1.0***	
Not contiguous	0.01***	0.27***	0.01***	0.01***	-0.15***	0.28***	1.0***

Level of statistical significance: ***1%

Table 9. Correlation's matrix – Fourth scenario

	Co-publication	**Geo. distance**	**TENB**	**Cog. distance**	**Different countries**	**Different continents**	**Not contiguous**
Co-publication	1.0***						
Geo. distance	-0.12***	1.0***					
TENB	0.67***	-0.08***	1.0***				
Cog. distance	-0.20***	0.05***	-0.16***	1.0***			
Different countries	-0.20***	0.48***	-0.16***	0.08***	1.0***		
Different continents	-0.13***	0.80***	-0.11***	0.05***	0.56***	1.0***	
Not contiguous	0.01***	0.25***	0.02***	0.01***	-0.08***	0.29***	1.0***

Level of statistical significance: ***1%

4.3.Logistic regression results

In this section, the results of logistic regression analysis are discussed. We provide the results of each scenario for two models explained in the methodology section. Table 10 shows the logistic regression analysis results for the first scenario. In the first model, all the independent variables were included. The negative and significant coefficient of geographical distance confirms the hindering effect of geographical distance on scientific collaboration, as a 10% increase in the geographical distance among authors would decrease the chance of their collaboration by 7%. The positive effect of network proximity on scientific collaboration is reflected in the positive and significant coefficient of TENB; a 1% increase in TENB would increase the probability of collaboration by 4%. Also, the negative and significant coefficient of cognitive distance shows the negative influence of cognitive distance on scientific collaboration. However, the coefficients of different provinces and not contiguous variables are not significantly different from zero.

In the second model, the interaction between geographical distance and TENB is added. The coefficient of the interaction term is positive and significant, implying that the effect of TENB on scientific collaboration is even more powerful when authors have a farther geographical distance.

Table 10. Logistic regression results – First scenario

Model	**(1)**	**(2)**
Dependent variable	Co-publication	Co-publication
Geo. distance (ln)	-0.69*** (0.10)	-0.89*** (0.14)
TENB (ln)	4.06*** (0.41)	2.64*** (0.54)
Geo. distance (ln) × TENB (ln)		0.44*** (0.15)
Cog. distance	-5.60*** (0.82)	-5.36*** (0.82)
Different provinces	1.11 (0.75)	1.50 (0.87)
Not contiguous	0.69 (0.28)	1.00 (0.81)
Number of observations	2,878	2,878
Pseudo R^2	0.85	0.86
BIC	296.41	259.05

Level of statistical significance: ***1%

The result of logistic regression in the second scenario is provided in Table 11. The negative and significant coefficients of geographical distance and cognitive distance confirm the hindering effects of these two variables on scientific collaboration. Also, the positive and significant coefficient of TENB refers to the critical role of network proximity on scientific collaboration. Moreover, the different countries variable added in the second scenario has a negative and significant coefficient. It implies being from different countries (Canada or the United States) would decrease the probability of collaboration by 77%[10]. Like the first scenario, adding the interaction between geographical distance and TENB to the second model, the positive and significant coefficient of the interaction term confirms that when authors are farther geographically, the role of TENB in forming scientific collaboration is more powerful.

Table 11. Logistic regression results – Second scenario

Model	**(1)**	**(2)**
Dependent variable	Co-publication	Co-publication
Geo. distance (ln)	-0.53*** (0.02)	-0.58*** (0.07)
TENB (ln)	5.99*** (0.16)	2.14*** (0.22)
Geo. distance (ln) × TENB (ln)		0.66*** (0.04)
Cog. distance	-2.72*** (0.09)	-2.74*** (0.09)
Different provinces	-0.50*** (0.13)	-0.49*** (0.43)
Different countries	-1.48*** (0.40)	-1.49*** (0.42)
Not contiguous	1.84*** (0.22)	2.05*** (0.08)
Number of observations	50,156	50,156
Pseudo R^2	0.75	0.76
BIC	7,694.35	7,534.40

Level of statistical significance: ***1%

The results in the third and fourth scenarios are almost the same as in the first two scenarios. The negative and significant effects of geographical distance and cognitive distance, and the positive and significant effect of TENB, on scientific collaboration can be seen in the regression results.

[10] (1 - exp (-1.48))

Table 12. Logistic regression results – Third scenario

Model	(1)	(2)
Dependent variable	Co-publication	Co-publication
Geo. distance (ln)	-0.36*** (0.00)	-0.38*** (0.00)
TENB (ln)	6.56*** (0.03)	3.97*** (0.05)
Geo. distance (ln) × TENB (ln)		0.41*** (0.01)
Cog. distance	-2.41*** (0.01)	-2.40*** (0.01)
Different countries	-0.34*** (0.01)	-0.34*** (0.01)
Not contiguous	0.83*** (0.01)	0.91*** (0.01)
Number of observations	3,672,155	3,672,155
Pseudo R^2	0.72	0.72
BIC	638,721.85	637,076.32

Level of statistical significance: ***1%

Table 13. Logistic regression results – Fourth scenario

Model	(1)	(2)
Dependent variable	Co-publication	Co-publication
Geo. distance (ln)	-0.39*** (0.00)	-0.40*** (0.00)
TENB (ln)	7.21*** (0.03)	4.24*** (0.05)
Geo. distance (ln) × TENB (ln)		0.43*** (0.01)
Cog. distance	-2.59*** (0.01)	-2.58*** (0.01)
Different countries	-0.36*** (0.01)	-0.34*** (0.01)
Not contiguous	1.22*** (0.01)	1.29*** (0.01)
Number of observations	8,440,135	8,440,135
Pseudo R^2	0.75	0.75
BIC	1,263,501.13	1,260,965.54

Level of statistical significance: ***1%

Moreover, the negative impact of different countries on scientific collaboration is evident in these two scenarios as well. Ceteris paribus, if authors are from two different countries, their chance of collaboration would decrease by 30%[11] to 29%[12]. Besides, the coefficient of the variables representing the interaction between geographical distance and TENB is positive and significantly different from zero in the third and fourth scenarios.

[11] (1 - exp (-0.36))

[12] (1 - exp (-0.34))

4.4.Machine learning results

The machine learning classification results are presented in this section. For the first scenario, Table 14 shows the results of applying different classifiers on the train dataset. The results show that the Extreme Gradient Boosting (so-called XGBoost) provides the best AUC score among other classifiers.

Table 14. Cross-validation results – First scenario

Classifier	Accuracy	AUC	Precision	Recall	F1 score
Logit regression	0.97 (± 0.01)	0.98 (± 0.01)	0.53 (± 0.06)	0.96 (± 0.05)	0.68 (± 0.06)
Gaussian Naive Bayes	0.95 (± 0.01)	0.98 (± 0.02)	0.40 (± 0.08)	0.90 (± 0.05)	0.55 (± 0.08)
Nearest Neighbors	0.98 (± 0.00)	0.95 (± 0.04)	0.67 (± 0.07)	0.89 (± 0.09)	0.76 (± 0.03)
Support Vector Machines	0.98 (± 0.01)	0.97 (± 0.01)	0.58 (± 0.08)	0.95 (± 0.05)	0.71 (± 0.06)
Random Forest	0.99 (± 0.01)	0.98 (± 0.02)	0.86 (± 0.11)	0.93 (± 0.06)	0.89 (± 0.06)
Extreme Gradient Boosting	1.00 (± 0.00)	0.99 (± 0.01)	0.92 (± 0.07)	0.95 (± 0.03)	0.93 (± 0.04)

For both random search and grid search hyperparameters tunings, the range and the best hyperparameters are provided in Table 15. In the random search, we consider 200 fits to find the approximate range of each parameter. Then, we fine-tuned the parameters in a grid search with 7,290 fits. The best AUC score achieved in the hyperparameters tuning is 0.996. However, to assess the model's actual performance, the model was tested using the test dataset we put aside in the first step. The result of the model evaluation using the test dataset and the confusion matrix are presented in Table 16 and Table 17, respectively.

Table 15. Hyperparameters tuning – First scenario

Classifier: Extreme Gradient Boosting (XGBoost)				
Hyperparameter	Random search range	Best parameter	Grid search range	Best parameter
max_depth	3, 4, 5	5	4, 5	4
learning_rate	Uniform (0.1, 0.6)	0.55	0.53, 0.55, 0.57	0.55
subsample	Uniform (0.1, 0.9)	0.87	0.85, 0.87, 0.89	0.87
colsample_bytree	Uniform (0.1, 0.9)	0.78	0.75, 0.77, 0.79	0.75
colsample_bylevel	Uniform (0.1, 0.9)	0.93	0.90, 0.93, 0.95	0.90
n_estimators	100, 200, … 1000	500	450, 500, 550	500
gamma	0, 0.1, 0.2, …, 0.5	0	0	0
scale_pos_weight	1, 2, 3, … 30	9	7, 8, 9	8
Total number of fits	200		7,290	
Best AUC score	0.996		0.996	

Table 16. Final model evaluation with the test dataset – First scenario

Accuracy	0.99
AUC	0.94
Precision	1.00
Recall	0.88
F1 score	0.93

Table 17. Confusion matrix - First scenario

		Predict	
		0	1
Actual	0	263	0
	1	1	7

After reviewing the model performance, the importance of each feature in predicting the co-publication should be discussed. Figure 9 demonstrates the beeswarm plot of SHAP values for the first scenario. This plot is designed to display an information-dense summary of how the features in a dataset impact the model's output. The features are sorted from the most important (on the top) to the least important (at the bottom).

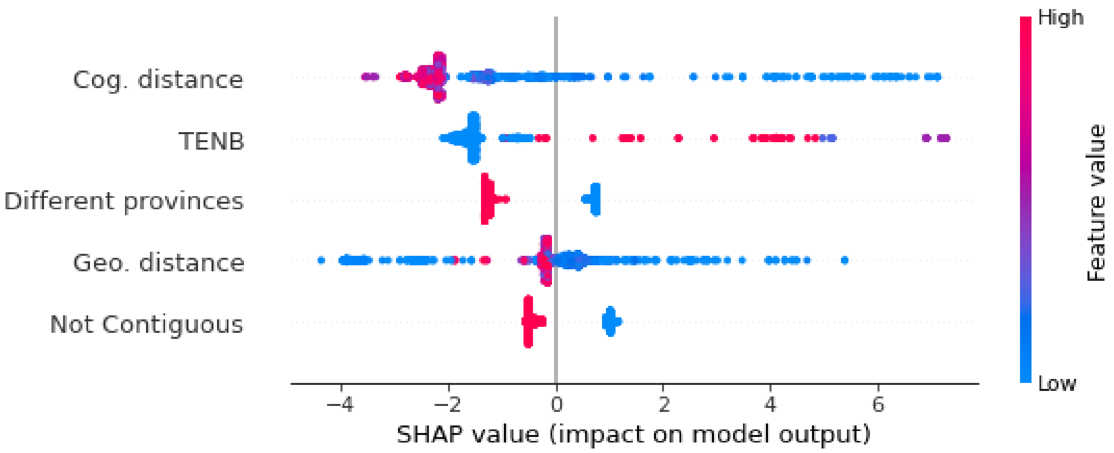

Figure 9. Beeswarm plot – First scenario

As shown in Figure 9, cognitive distance has been the most important feature in predicting scientific collaborations. Moreover, a high density of observations with a high level of cognitive distance on the left side of the diagram, which associates with negative SHAP values, confirms the negative effect of cognitive distance on the likelihood of scientific collaboration. On the contrary, observations with a higher level of TENB are mainly associated with higher SHAP values, implying the positive role of network proximity in forming scientific collaboration. For binary variables, i.e., different provinces and not contiguous, the pattern is even more evident; observations in blue represent the same province/contiguous provinces collaborations, while the observations in red represent different provinces/not contiguous collaborations. As expected, for those two binary variables, the blue dots are concentrated on the right side of the diagram. In contrast, the red dots are concentrated on the left side, implying the negative influence of being from different provinces/not contiguous provinces on forming scientific collaborations. Regarding the geographical distance, although observations with close geographical distance are associated with both high and low SHAP values, the high density of observations with high geographical distance on the left side of the diagram is discernible.

The results of applying different classifiers on the dataset of the second scenario are presented in Table 18. XGBoost provides the best AUC score among all classifiers like the first scenario.

Table 18. Cross-validation results – Second scenario

Classifier	Accuracy	AUC	Precision	Recall	F1 score
Logit regression	0.96 (± 0.01)	0.97 (± 0.01)	0.51 (± 0.02)	0.87 (± 0.02)	0.65 (± 0.01)
Gaussian Naive Bayes	0.93 (± 0.01)	0.96 (± 0.01)	0.36 (± 0.04)	0.90 (± 0.02)	0.51 (± 0.04)
Nearest Neighbors	0.97 (± 0.00)	0.94 (± 0.01)	0.58 (± 0.02)	0.86 (± 0.03)	0.69 (± 0.02)
Support Vector Machines	0.93 (± 0.00)	0.86 (± 0.02)	0.35 (± 0.01)	0.71 (± 0.03)	0.47 (± 0.02)
Random Forest	0.98 (± 0.00)	0.97 (± 0.01)	0.74 (± 0.02)	0.90 (± 0.03)	0.81 (± 0.02)
Extreme Gradient Boosting	0.99 (± 0.00)	0.98 (± 0.01)	0.93 (± 0.01)	0.83 (± 0.04)	0.88 (± 0.02)

The results of hyperparameters tuning are provided in Table 19. Although an AUC score of 0.984 is achieved in the random search hyperparameters tuning, the model performance is improved by 0.005 after fine-tuning the hyperparameters in the grid search.

Table 19. Hyperparameters tuning – Second scenario

Classifier: Extreme Gradient Boosting (XGBoost)				
Hyperparameter	Random search range	Best parameter	Grid search range	Best parameter
max_depth	4, 5	5	5	5
learning_rate	Uniform (0.1, 0.6)	0.22	0.20, 0.22, 0.24	0.22
subsample	Uniform (0.1, 0.9)	0.60	0.58, 0.60, 0.62	0.58
colsample_bytree	Uniform (0.1, 0.9)	0.90	0.88, 0.90, 0.92	0.88
colsample_bylevel	Uniform (0.1, 0.9)	0.32	0.30, 0.32, 0.34	0.3
n_estimators	100, 200, … 1000	200	150, 200, 250	250
gamma	0, 0.1, 0.2, …, 0.5	0.2	0.1, 0.2, 0.3	0.3
scale_pos_weight	1, 2, 3, … 30	1	1, 2	1
Total number of fits	200		7,290	
Best AUC score	0.984		0.989	

Table 20. Final model evaluation with the test dataset – Second scenario

Accuracy	0.98
AUC	0.95
Precision	0.70
Recall	0.92
F1 score	0.80

Table 21. Confusion matrix - Second scenario

		Predict	
		0	1
Actual	0	4,485	76
	1	15	184

The importance of each feature on the model’s output in the second scenario is depicted in Figure 10. Cognitive distance has the most important role in predicting scientific collaborations in the second scenario, like the first scenario. The density of red points on the left side of the diagram implies more cognitive distance decrease the probability of scientific collaboration. Geographical distance is in second place in terms of importance in the second scenario. As expected, a closer geographical distance is associated with a higher chance of scientific collaboration. TENB has third place in this scenario. The high density of observations with low values of TENB on the left side of the diagram refers to the critical role of network proximity in forming scientific collaborations. The pattern is apparent regarding the different provinces. Observations associated with authors from different provinces are mainly concentrated on the left side of the diagram, implying that when authors are from different provinces, the chance of collaboration is lower. The same pattern can be seen for the different counties. Observations associated with authors from different countries are spread on the left side of the diagram, confirming the negative effect of being from different countries on the likelihood of scientific collaboration.

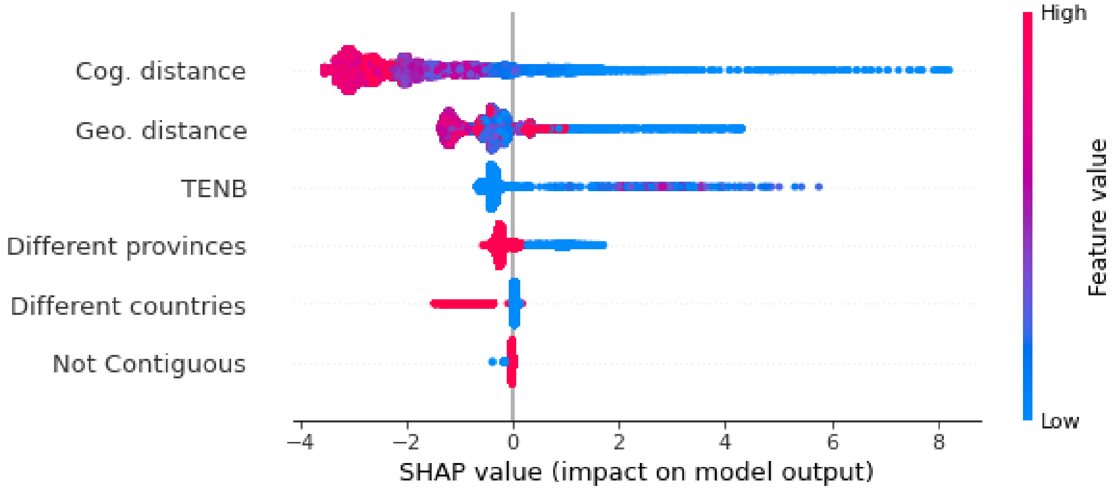

Figure 10. Beeswarm plot – Second scenario

The results of applying different classifiers on the dataset of the third scenario are presented in Table 22. XGBoost provides the best AUC score among all other classifiers like the first and second scenarios. The results of hyperparameters tuning for this scenario are also provided in Table 23. Although in the random search hyperparameters tuning, an AUC score of 0.979 is achieved, the model performance is improved by 0.003 after fine-tuning the parameters in the grid search.

Table 22. Cross-validation results – Third scenario

Classifier	Accuracy	AUC	Precision	Recall	F1 score
Logit regression	1.00 (± 0.00)	0.95 (± 0.00)	0.91 (± 0.00)	0.73 (± 0.02)	0.81 (± 0.00)
Gaussian Naive Bayes	1.00 (± 0.00)	0.95 (± 0.00)	0.90 (± 0.00)	0.73 (± 0.02)	0.81 (± 0.00)
Nearest Neighbors	0.99 (± 0.00)	0.93 (± 0.00)	0.61 (± 0.00)	0.81 (± 0.00)	0.70 (± 0.00)
Support Vector Machines	0.98 (± 0.01)	0.91 (± 0.02)	0.31 (± 0.08)	0.71 (± 0.06)	0.42 (± 0.07)
Random Forest	1.00 (± 0.00)	0.96 (± 0.00)	0.79 (± 0.00)	0.84 (± 0.00)	0.81 (± 0.00)
Extreme Gradient Boosting	1.00 (± 0.00)	0.97 (± 0.00)	0.96 (± 0.00)	0.75 (± 0.00)	0.84 (± 0.00)

Table 23. Hyperparameters tuning – Third scenario

Classifier: Extreme Gradient Boosting (XGBoost)				
Hyperparameter	Random search range	Best parameter	Grid search range	Best parameter
max_depth	4, 5, 6	4	4	4
learning_rate	Uniform (0.1, 0.6)	0.35	0.33, 0.35, 0.37	0.33
subsample	Uniform (0.1, 0.9)	0.93	0.90, 0.93, 0.95	0.93
colsample_bytree	Uniform (0.1, 0.9)	0.77	0.75, 0.77, 0.79	0.75
colsample_bylevel	Uniform (0.1, 0.9)	0.69	0.67, 0.69, 0.71	0.67
n_estimators	100, 200, … 1000	700	650, 700, 750	750
gamma	0, 0.1, 0.2, …, 0.5	0.4	0.3, 0.4, 0.5	0.4
scale_pos_weight	1, 2, 3, … 30	3	3, 4	3
Total number of fits	200		7,290	
Best AUC score	0.979		0.982	

Table 24. Final model evaluation with the test dataset – Third scenario

Accuracy	0.99
AUC	0.90
Precision	0.77
Recall	0.79
F1 score	0.78

Table 25. Confusion matrix - Third scenario

		Predict	
		0	1
Actual	0	333,041	792
	1	666	2,723

The impact of each feature on the model’s output in the third scenario is depicted in Figure 11. Like the first and the second scenarios, cognitive distance is the most important feature in predicting the scientific collaborations in the third scenario. The concentration of the red dots on the left side of the diagram confirms that a higher cognitive distance between authors decreases the probability of scientific collaboration. Like the second scenario, geographical distance is in the second place of importance.

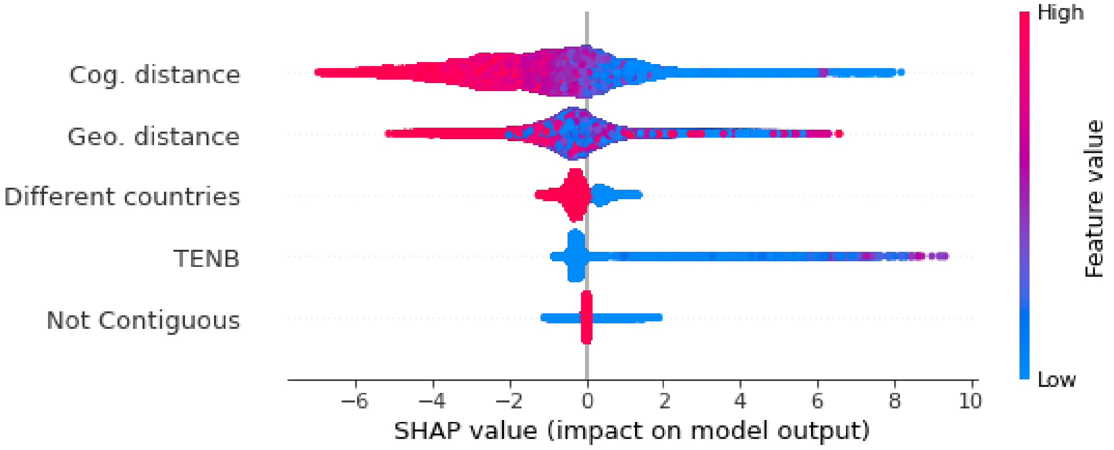

Figure 11. Beeswarm plot – Third scenario

As expected, a farther geographical distance is associated with a lower chance of scientific collaboration. For the different countries, the pattern is evident. Observations related to authors from different countries are mainly concentrated on the left side of the diagram, implying that when authors are from different countries, the chance of collaboration is lower. TENB has the fourth place in this scenario. The association between higher values of TENB and a higher chance of scientific collaboration is discernible.

The results of applying different classifiers on the dataset of the fourth scenario are presented in Table 26. Like the other scenarios, XGBoost provides the best AUC score among all other classifiers.

Table 26. Cross-validation results – Fourth scenario

Classifier	Accuracy	AUC	Precision	Recall	F1 score
Logit regression	1.00 (± 0.00)	0.96 (± 0.00)	0.95 (± 0.00)	0.77 (± 0.00)	0.85 (± 0.00)
Gaussian Naive Bayes	1.00 (± 0.00)	0.96 (± 0.00)	0.95 (± 0.00)	0.77 (± 0.00)	0.08 (± 0.00)
Nearest Neighbors	0.99 (± 0.00)	0.95 (± 0.02)	0.69 (± 0.00)	0.86 (± 0.00)	0.77 (± 0.00)
Support Vector Machines	0.99 (± 0.01)	0.94 (± 0.02)	0.55 (± 0.08)	0.78 (± 0.06)	0.62 (± 0.07)
Random Forest	1.00 (± 0.00)	0.96 (± 0.00)	0.79 (± 0.00)	0.84 (± 0.00)	0.81 (± 0.00)
Extreme Gradient Boosting	0.90 (± 0.01)	0.97 (± 0.00)	0.96 (± 0.00)	0.79 (± 0.00)	0.87 (± 0.00)

The results of hyperparameters tuning for the fourth scenario are provided in Table 27. Although in the random search hyperparameters tuning, an AUC score of 0.968 is achieved, the model performance is improved by 0.008 after fine-tuning the parameters in the grid search.

Table 27. Hyperparameters tuning – Fourth scenario

Classifier: Extreme Gradient Boosting (XGBoost)				
Hyperparameter	Random search range	Best parameter	Grid search range	Best parameter
max_depth	4, 5, 6	4	4	4
learning_rate	Uniform (0.1, 0.6)	0.35	0.33, 0.35, 0.37	0.33
subsample	Uniform (0.1, 0.9)	0.93	0.90, 0.93, 0.95	0.93
colsample_bytree	Uniform (0.1, 0.9)	0.77	0.75, 0.77, 0.79	0.75
colsample_bylevel	Uniform (0.1, 0.9)	0.69	0.67, 0.69, 0.71	0.67
n_estimators	100, 200, … 1000	700	650, 700, 750	750
gamma	0, 0.1, 0.2, …, 0.5	0.4	0.3, 0.4, 0.5	0.4
scale_pos_weight	1, 2, 3, … 30	3	3, 4	3
Total number of fits	200		7,290	
Best AUC score	0.968		0.976	

Table 28. Final model evaluation with the test dataset – Fourth scenario

Accuracy	0.99
AUC	0.91
Precision	0.83
Recall	0.82
F1 score	0.83

Table 29. Confusion matrix - Fourth scenario

		Predict	
		0	1
Actual	0	765,811	1,474
	1	1,577	7,400

The impact of each feature on the model's output in the fourth scenario is depicted in Figure 12. Like all other scenarios, cognitive distance is the most important feature in predicting scientific collaborations. The concentration of the red dots on the left side of the diagram confirms that a higher cognitive distance between authors decreases the probability of scientific collaboration. Geographical distance is in the second place of importance. As expected, a farther geographical distance is associated with a lower chance of scientific collaboration. TENB has the third place in this scenario. The association between higher values of TENB and a higher chance of scientific collaboration is discernible. For the different countries, the pattern is clear. Observations associated with authors from different countries are mainly concentrated on the left side of the diagram, implying that when authors are from different countries, the chance of collaboration is lower.

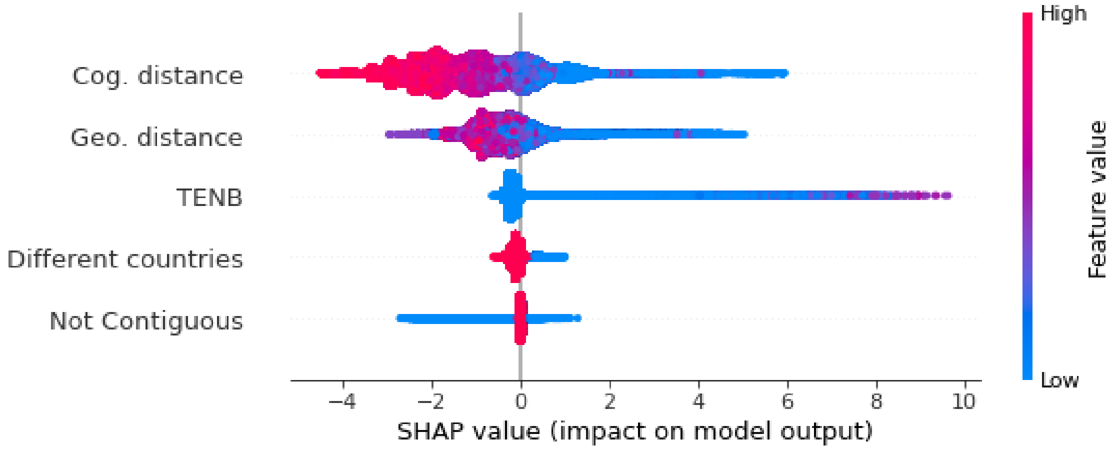

Figure 12. Beeswarm plot – Fourth scenario

4.5.Discussion

The results of logistic regression analysis and machine learning classification were provided in the previous chapter. Now, we want to discuss the results to see if the null hypotheses developed in chapter three can be rejected.

The logistic regression results show a negative and significant association between geographical distance and the probability of scientific collaboration in all scenarios. All else being equal, a 10% increase in the geographical distance would decrease the chance of collaboration between authors by 3 to 7%. SHAP feature importance analysis also confirms the logistic regression results. As clearly illustrated in all beeswarm diagrams, observations related to authors with farther geographical distance are mainly associated with lower SHAP values, implying a lower chance of collaboration for those authors. These findings are in line with several previous studies (Bergé, 2017; Frenken et al., 2009; Maggioni et al., 2007; Morescalchi et al., 2015; Ponds et al., 2007; Scherngell & Barber, 2009) that found a negative association between geographical distance and scientific collaboration. Thus, we can reject the first null hypotheses:

$H_0$1: A higher level of geographical proximity does not enhance the likelihood of future scientific collaboration.

The second null hypothesis developed in this study was about the association between network proximity and scientific collaboration. The regression results show a positive and significant relationship between network proximity and the chance of scientific collaboration in all scenarios. Ceteris paribus, 1% increase in the level of network proximity would enhance the opportunity of scientific collaboration by up to 7.2%. The feature importance analysis confirms the logistic regression results. Higher values of TENB are generally associated with higher SHAP values, implying that when authors are closer in the network, they are more likely to have at least one co-publication. This result aligns with Bergé (2017) that used the same measure (TENB) to gauge network proximity. Therefore, we can reject the second null hypothesis:

$H_0$2: A higher level of network proximity does not enhance the likelihood of future scientific collaboration.

In this study, two variables represented institutional proximity, i.e., different provinces included in the first and the second scenarios, and different countries included in all scenarios except the first one. Although the coefficients of different provinces in both scenarios are not significantly different from zero, the feature importance analysis clearly shows an association between being from different provinces and a lower chance of scientific collaboration. However, the negative and significant association between different countries and the likelihood of scientific collaboration is much more discernible in the third and the fourth scenarios where several countries are included. All else being equal, when authors are from different countries, their chance of collaboration would decrease by around 30%. The feature importance analysis confirms this result, as observations related to authors from different countries are mainly associated with lower SHAP values. This finding confirms many previous studies (Bergé, 2017; Hoekman et al., 2009, 2010; Morescalchi et al., 2015; Scherngell & Barber, 2009) that found being from different countries would negatively affect collaboration. So, both third null hypotheses can be rejected:

$H_0$3a: Being from the same province does not enhance the likelihood of future scientific collaboration.

$H_0$3b: Being from the same country does not enhance the likelihood of future scientific collaboration.

The fourth null hypothesis was about the effect of cognitive proximity on the chance of scientific collaboration among authors. The negative and significant coefficient of cognitive distance in all scenarios clearly shows that authors with higher cognitive distance are less likely to collaborate.

However, the magnitude of this effect becomes smaller in scenarios that include more countries. The feature importance analysis confirms the logistic regression results regarding the cognitive distance. According to beeswarm diagrams, cognitive distance is the most important feature in predicting future scientific collaborations in all scenarios. Besides, observations related to authors with higher cognitive distance are generally associated with lower SHAP values, implying a lower chance of scientific collaboration. This result is in line with several empirical studies that report the negative effect of cognitive distance on collaboration (Bergé, 2017; Cunningham & Werker, 2012; Ding, 2011; Jaffe & Hu, 2003; Jaffe & Trajtenberg, 1999). Therefore, the fourth null hypothesis can be rejected:

H$_0$4: A higher level of cognitive proximity does not enhance the likelihood of future scientific collaboration.

The fifth null hypotheses were about regional contiguity and scientific collaboration. As can be seen in the regression results, in the first scenario where not contiguous variable identifies if authors are from contiguous provinces, the coefficient is not significantly different from zero, implying that contiguity of authors' provinces does not affect the chance of scientific collaboration. Besides, the coefficient is positive and significant in the second scenario, where not contiguous variable again identifies if authors are from contiguous provinces. Therefore, we cannot reject the following null hypothesis:

H$_0$5a: Province contiguity does not enhance the likelihood of future scientific collaboration.

In the third and the fourth scenarios that not contiguous variable measures contiguity of authors' countries, not contiguous variable has a positive and significant coefficient. Thus, the following hypothesis cannot be rejected:

H$_0$5b: Country contiguity does not enhance the likelihood of future scientific collaboration.

The last null hypothesis developed in this study was about the substitutability of network proximity and geographical proximity. As the logistic regression results show, the coefficient of interaction between geographical distance and network proximity is positive and significant in all scenarios, implying that geographical proximity moderates the relationship between network proximity and the probability of scientific collaboration. In other words, as the effect of network proximity increases with geographical distance, network proximity does not seem to have a homogeneous overall impact. Instead, it acts as a substitute for geographical proximity. Thus, we can reject the sixth null hypothesis:

$H_0$6: Geographical proximity does not moderate the effect of network proximity on the likelihood of future scientific collaboration.

5. CONCLUSION

In this study, we examined the relationship between the geographical proximity of researchers and the likelihood of scientific collaboration among them. Using the co-publication data in Artificial Intelligence during 2000 – 2019, we studied the geographical patterns of scientific collaboration among AI researchers in four scenarios. In the first scenario, we only included Canadian AI researchers. Then, in the second scenario, AI researchers from the United States were added to the study. To have a more comprehensive understanding of geographical patterns of scientific collaboration, the scope of the study was extended to European countries and all countries around the world, in the third and the fourth scenarios, respectively. The logistic regression and machine learning classification results clearly show that geographical distance is among the main barriers to scientific collaborations at the individual level.

However, examining the relationship between the interaction of geographical proximity and network proximity, and the probability of scientific collaboration revealed a substitutability pattern. The logistic regression results show that the influence of network proximity rises when researchers are geographically farther. This result aligns with Bergé (2017) that found a substitutability pattern between network proximity and geographical proximity.

The substitutability pattern between geographical proximity and network proximity has an important implication from a policy-making point of view. Supporting long-distance scientific collaborations not only could result in higher quality research productions (see, e.g., J. Adams, 2013; J. D. Adams et al., 2005) but also may increase indirect connections among researchers, which in turn will trigger new scientific collaborations. Forming new long-distance collaboration increases the network proximity (measured by TENB) of researchers who had a scientific collaboration with the researchers in the new collaboration. This, in turn, may trigger new collaborations because of network effects, implying that more distant/more yielding collaborations are more likely to be established. In this sense, policies aiming at encouraging long-distance collaborations could positively affect knowledge production and ease future knowledge flows.

This study contributes to understanding the influence of various forms of proximity on the probability of scientific collaboration at the individual level. Although the effect of different types of proximity on scientific collaborations among geographical regions or organizations has been studied before, to the best of my knowledge, this is the first comprehensive study that examines the role of geographical proximity, network proximity, and their interaction on the chance of scientific collaborations.

Future works

This thesis has focused on the field of Artificial Intelligence. Thus, extending the study to other areas of science will be helpful to see whether they display the same pattern of substitutability between geographical proximity and network proximity. Moreover, when it comes to studying the determinants of scientific collaboration at the individual level, adding more individual explanatory variables (gender, age, seniority, language, religion, etc.) to the model can be helpful to predict future collaborations more precisely.

REFERENCES

Adams, J. (2013). The fourth age of research. *Nature*, *497*(7451), 557–560. https://doi.org/10.1038/497557a

Adams, J. D., Black, G. C., Clemmons, J. R., & Stephan, P. E. (2005). Scientific teams and institutional collaborations: Evidence from U.S. universities, 1981-1999. *Research Policy*, *34*(3), 259–285. https://doi.org/10.1016/j.respol.2005.01.014

Aghion, P., & Howitt, P. (1992). A Model of Growth Through Creative Destruction. *The Econometric Society*, *60*(2), 323–351.

Aiello, L. M., Barrat, A., Schifanella, R., Cattuto, C., Markines, B., & Menczer, F. (2012). Friendship prediction and homophily in social media. *ACM Transactions on the Web*, *6*(2). https://doi.org/10.1145/2180861.2180866

Almendral, J. A., Oliveira, J. G., López, L., Mendes, J. F. F., & Sanjuán, M. A. F. (2007). The network of scientific collaborations within the European framework programme. *Physica A: Statistical Mechanics and Its Applications*, *384*(2), 675–683. https://doi.org/10.1016/j.physa.2007.05.049

Autant-bernard, C., Billand, P., Frachisse, D., & Massard, N. (2007). Social distance versus spatial distance in R&D cooperation: Empirical evidence from European collaboration choices in micro and nanotechnologies. *Papers in Regional Science*, *86*(3), 495–519. https://doi.org/10.1111/j.1435-5957.2007.00132.x

Balland, P. A. (2012). Proximity and the Evolution of Collaboration Networks: Evidence from Research and Development Projects within the Global Navigation Satellite System (GNSS) Industry. *Regional Studies*, *46*(6), 741–756. https://doi.org/10.1080/00343404.2010.529121

Banchoff, T. (2002). Institutions, inertia and European Union research policy. *Journal of*

Common Market Studies, *40*(1), 1–21. https://doi.org/10.1111/1468-5965.00341

Barabâsi, A.-L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A*, *311*(3), 590–614.

Bergé, L. R. (2017). Network proximity in the geography of research collaboration. *Papers in Regional Science*, *96*(4), 785–815. https://doi.org/10.1111/pirs.12218

Bethard, S., & Jurafsky, D. (2010). Who should I cite: learning literature search models from citation behavior. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 609–618. https://doi.org/http://doi.acm.org/10.1145/1871437.1871517

Bignami, F., Mattsson, P., & Hoekman, J. (2020). The importance of geographical distance to different types of R&D collaboration in the pharmaceutical industry. *Industry and Innovation*, *27*(5), 513–537. https://doi.org/10.1080/13662716.2018.1561361

Blau, J. R. (1974). Patterns of Communication Among Theoretical High Energy Physicists Author (s): Judith R . Blau Published by : American Sociological Association Stable URL : https://www.jstor.org/stable/2786390 American Sociological Association is collaborating with JST. *Sociometry*, *37*(3), 391–406.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(4–5), 993–1022. https://doi.org/10.1016/b978-0-12-411519-4.00006-9

Boschma, R. (2005). Proximity and innovation: A critical assessment. *Regional Studies*, *39*(1), 61–74. https://doi.org/10.1080/0034340052000320887

Boschma, R., & Frenken, K. (2010). The spatial evolution of innovation networks: A proximity perspective. *The Handbook of Evolutionary Economic Geography*, *June 2014*, 120–138. https://doi.org/10.4337/9781849806497.00012

Brandusescu, A. (2021). *Artificial intelligence policy and funding in Canada: Public investments,* *private interests*. 1–62. https://www.mcgill.ca/centre-montreal/files/centre-montreal/aipolicyandfunding_report_updated_mar5.pdf

Breiman, L. (2001). Statistical modeling: The two cultures. *Quality Engineering*, *48*, 81–82.

Breschi, S., & Lissoni, F. (2009). Mobility of skilled workers and co-invention networks: An anatomy of localized knowledge flows. *Journal of Economic Geography*, *9*(4), 439–468. https://doi.org/10.1093/jeg/lbp008

Carayol, N., Bergé, L., Cassi, L., & Roux, P. (2019). Unintended triadic closure in social networks: The strategic formation of research collaborations between French inventors. *Journal of Economic Behavior and Organization*, *163*, 218–238. https://doi.org/10.1016/j.jebo.2018.10.009

Castells, M. 1942-T. A.-T. T.-. (1996). *The rise of the network society* (2nd ed., w). Wiley-Blackwell. http://site.ebrary.com/id/10355273

Catalini, C. (2018). Microgeography and the direction of inventive activity. *Management Science*, *64*(9), 4348–4364. https://doi.org/10.1287/mnsc.2017.2798

Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res. (JAIR)*, *16*, 321–357. https://doi.org/10.1613/jair.953

Clauset, A., Moore, C., & Newman, M. E. J. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature*, *453*(7191), 98–101. https://doi.org/10.1038/nature06830

Cunningham, S. W., & Werker, C. (2012). Proximity and collaboration in European nanotechnology. *Papers in Regional Science*, *91*(4), 723–742. https://doi.org/10.1111/j.1435-5957.2012.00416.x

D’Este, P., Guy, F., & Iammarino, S. (2013). Shaping the formation of university-industry research collaborations: What type of proximity does really matter? *Journal of Economic*

Geography, *13*(4), 537–558. https://doi.org/10.1093/jeg/lbs010

Dahlander, L., & McFarland, D. A. (2013). Ties that last: Tie formation and persistence in research collaborations over time. *Administrative Science Quarterly*, *58*(1), 69–110. https://doi.org/10.1177/0001839212474272

Ding, Y. (2011). Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of Informetrics*, *5*(1), 187–203. https://doi.org/10.1016/j.joi.2010.10.008

Dong, Y., Johnson, R. A., & Chawla, N. V. (2015). Will This Paper Increase Your h -index? *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. https://doi.org/10.1145/2684822.2685314

Ebadi, A., & Schiffauerova, A. (2015). How to become an important player in scientific collaboration networks? *Journal of Informetrics*, *9*(4), 809–825. https://doi.org/10.1016/j.joi.2015.08.002

Ebadi, A., & Schiffauerova, A. (2016). How to boost scientific production? A statistical analysis of research funding and other influencing factors. *Scientometrics*, *106*(3), 1093–1116. https://doi.org/10.1007/s11192-015-1825-x

European Commission. (2014). *Webseite: What is Horizon 2020?* http://ec.europa.eu/programmes/horizon2020/en/what-horizon-2020

Fafchamps, M., van der Leij, M. J., & Goyal, S. (2010). Matching and Network Effects. *Journal of the European Economic Association*, *8*(1), 203–231. https://doi.org/10.1111/j.1542-4774.2010.tb00500.x

Freeman, R. B., Ganguli, I., & Murciano-Goroff, R. (2014). Why and Wherefore of Increased Scientific Collaboration. In *National Bureau of Economic Research Working Paper*. https://doi.org/10.7208/chicago/9780226286860.003.0001

Frenken, K., Hardeman, S., & Hoekman, J. (2009). Spatial scientometrics: Towards a cumulative research program. *Journal of Informetrics*, *3*(3), 222–232. https://doi.org/10.1016/j.joi.2009.03.005

Gertler, M. S. (1995). “Being There”: Proximity , Organization , and Culture in the Development and Adoption of Advanced Manufacturing Technologies. *Economic Geography*, *71*(1), 1–26.

Gertler, M. S. (2003). Tacit knowledge and the economic geography of context, or the undefinable tacitness of being (there). *Journal of Economic Geography*, *3*(1), 75–99. https://doi.org/10.1093/jeg/3.1.75

Getoor, L. (2005). *Link-based Classification BT - Advanced Methods for Knowledge Discovery from Complex Data* (S. Bandyopadhyay, U. Maulik, L. B. Holder, & D. J. Cook (eds.); pp. 189–207). Springer London. https://doi.org/10.1007/1-84628-284-5_7

Gilly, J., & Torre, A. (2000). Proximity Relations : Elements for an Analytical Framework. *Industrial Networks and Proximity*, 1–17.

Gittelman, M. (2007). Does geography matter for scienee-based firms? Epistemic communities and the geography of research and patenting in biotechnology. *Organization Science*, *18*(4), 724–741. https://doi.org/10.1287/orsc.1070.0249

Glänzel, W. (2001). National characteristics in international scientific co-authorship relations. *Scientometrics*, *51*(1), 69–115. https://doi.org/10.1023/A:1010512628145

Google. (n.d.). *Geocoding API.* https://developers.google.com/maps/documentation/geocoding/overview

Government of Canada. (n.d.). *Search Grants and Contributions*. https://search.open.canada.ca/en/gc/?_ga=2.258231562.1488605449.1607055646-2066370278.1596646179

Government of Canada, P. M. O. (2018). *Prime Minister announces investment in artificial intelligence to create over 16,000 jobs for Canadians*. https://www.newswire.ca/news-releases/prime-minister-announces-investment-in-artificial-intelligence-to-create-over-16000-jobs-for-canadians-702095332.html

Government, U. (2020). *The Trump Administration Is Investing $1 Billion in Research Institutes to Advance Industries of the Future*. https://www.quantum.gov/the-trump-administration-is-investing-1-billion-in-research-institutes-to-advance-industries-of-the-future/

Gulati, R., & Gargiulo, M. (1999). Where do interorganizational networks come from? *American Journal of Sociology*, *104*(5), 1439–1493. https://doi.org/10.1086/210179

Hoekman, J., Frenken, K., & Tijssen, R. J. W. (2010). Research collaboration at a distance: Changing spatial patterns of scientific collaboration within Europe. *Research Policy*, *39*(5), 662–673. https://doi.org/10.1016/j.respol.2010.01.012

Hoekman, J., Frenken, K., & van Oort, F. (2009). The geography of collaborative knowledge production in Europe. *Annals of Regional Science*, *43*(3 SPEC. ISS.), 721–738. https://doi.org/10.1007/s00168-008-0252-9

Howells, J. R. L. (2002). Tacit knowledge, innovation and economic geography. *Urban Studies*, *39*(5–6), 871–884. https://doi.org/10.1080/00420980220128354

Inman, J. (1835). *Haversine Formula – Calculate geographic distance on earth*. https://www.igismap.com/haversine-formula-calculate-geographic-distance-earth/

Jackson, B. M. O., & Rogers, B. W. (2007). *Meeting Strangers and Friends of Friends : How Random Are Social Networks ? Author (s): Matthew O . Jackson and Brian W . Rogers Published by : American Economic Association Stable URL : http://www.jstor.com/stable/30035025 REFERENCES Linked references* . *97*(3), 890–915.

Jaffe, A. B., & Hu, A. G. Z. (2003). Patent citations and international knowledge flow: the cases of Korea and Taiwan. *International Journal of Industrial Organization*, *21*(6), 849–880. http://biblio.iztapalapa.uam.mx:2060/science?_ob=ArticleURL&_udi=B6V8P-48BC1SB-1&_user=1695873&_coverDate=06%2F30%2F2003&_rdoc=6&_fmt=full&_orig=browse&_srch=doc-info(%23toc%235876%232003%23999789993%23431275%23FLA%23display%23Volume)&_cdi=5876&_sort=d&_d

Jaffe, A. B., & Trajtenberg, M. (1999). International knowledge flows: evidence from patent citations. *Economics of Innovation and New Technology*, *8*(1–2), 105–136. https://doi.org/10.1080/10438599900000006

Johnson, D. K. N., & Mareva, M. (2002). *It's a Small(er) World: The Role of Geography and Networks in Biotechnology Innovation* (2002-01). http://hdl.handle.net/10419/23232

Jones, B. F., Wuchty, S., & Uzzi, B. (2008). Multi-University Research Teams: Shifting Impact, Geography, and Stratification in Science. *Science*, *322*(5905), 1259–1262. https://doi.org/10.1177/107554709401500403

Jones, C. (1995). R & D-Based Models of Economic Growth. *Journal of Political Economy*, *103*(4), 759–784.

Katz, J. (1994). Geographical proximity and scientific collaboration. *Scientometrics*, *31*(1), 31–43. https://doi.org/10.1007/BF02018100

Katz, J., & Martin, B. (1997). What is research collaboration? *Research Policy*, *26*(1), 1–18. https://doi.org/10.1016/S0048-7333(96)00917-1

Kirat, T., & Lung, Y. (1999). Innovation and Proximity. *European Urban and Regional Studies*, *6*(1), 27–38. https://doi.org/10.1177/096977649900600103

Knoben, J., & Oerlemans, L. A. G. (2006). Proximity and inter-organizational collaboration: A literature review. *International Journal of Management Reviews*, *8*(2), 71–89.

https://doi.org/10.1111/j.1468-2370.2006.00121.x

Krackhardt, D. (1999). The Ties That Torture: Simmelian Tie Analysis in Organizations. *Research in the Sociology of Organizations*, *16*.

Lichtenwalter, R. N., Lussier, J. T., & Chawla, N. V. (2010). New Perspectives and Methods in Link Prediction. *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 243–252. https://doi.org/10.1145/1835804.1835837

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Lundvall, B.-A. 1941-. (1992). National systems of innovation : towards a theory of innovation and interactive learning. In *TA - TT* -. Pinter Publishers ; http://www.gbv.de/dms/hbz/toc/ht004483061.PDF

Maggioni, M. A., Nosvelli, M., & Uberti, T. E. (2007). Space versus networks in the geography of innovation: A European analysis. *Papers in Regional Science*, *86*(3), 471–493. https://doi.org/10.1111/j.1435-5957.2007.00130.x

Mascia, D., Pallotti, F., & Angeli, F. (2017). Don’t stand so close to me: competitive pressures, proximity and inter-organizational collaboration. *Regional Studies*, *51*(9), 1348–1361. https://doi.org/10.1080/00343404.2016.1185517

Mcpherson, M., Smith-lovin, L., & Cook, J. M. (2001). *B IRDS OF A F EATHER : Homophily in Social Networks*.

Merton , R. K. (1973). *The sociology of science : theoretical and empirical investigations*. University of Chicago Press.

Miguélez, E., & Moreno, R. (2014). What attracts knowledge workers? The role of space and social networks. *Journal of Regional Science*, *54*(1), 33–60. https://doi.org/10.1111/jors.12069

Moradabadi, B., & Meybodi, M. R. (2017). Link prediction in fuzzy social networks using distributed learning automata. *Applied Intelligence*, *47*(3), 837–849. https://doi.org/10.1007/s10489-017-0933-0

Morescalchi, A., Pammolli, F., Penner, O., Petersen, A. M., & Riccaboni, M. (2015). The evolution of networks of innovators within and across borders: Evidence from patent data. *Research Policy*, *44*(3), 651–668. https://doi.org/10.1016/j.respol.2014.10.015

Narin, F., Stevens, K., & Whitlow, E. (1991). *Scientific Co-operation in Europe and the Citation of Multinationally Authored Papers*. *21*(3), 313–323.

Newman, M. E. J. (2001). The structure of scientific col laboration networks. *Proceedings of the National Academy of Sciences*, *98*(2), 404–409. https://doi.org/10.1073/pnas.021544898

Ponds, R., van Oort, F., & Frenken, K. (2007). The geographical and institutional proximity of research collaboration. *Papers in Regional Science*, *86*(3), 423–443. https://doi.org/10.1111/j.1435-5957.2007.00126.x

PwC. (2017). *Sizing the prize - PwC's Global Artificial Intelligence Study: Exploiting the AI Revolution. What's the real value of AI for your business and how can you capitalise?* PwC. https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html

Rallet, A., & Torre, A. (1999). Is geographical proximity necessary in the innovation networks in the era of global economy? *GeoJournal*, *49*(4), 373–380. https://doi.org/10.1023/A:1007140329027

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures.

WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining, 399–408. https://doi.org/10.1145/2684822.2685324

Sarigöl, E., Pfitzner, R., Scholtes, I., Garas, A., & Schweitzer, F. (2014). Predicting scientific success based on coauthorship networks. *EPJ Data Science*, *3*(1), 9. https://doi.org/10.1140/epjds/s13688-014-0009-x

Schall, D. (2014). Link prediction in directed social networks. *Social Network Analysis and Mining*, *4*(1), 1–14. https://doi.org/10.1007/s13278-014-0157-9

Scherngell, T., & Barber, M. J. (2009). Spatial interaction modelling of cross-region R&D collaborations: Empirical evidence from the 5th EU framework programme. *Papers in Regional Science*, *88*(3), 531–546. https://doi.org/10.1111/j.1435-5957.2008.00215.x

Song, Y.-Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2), 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044

Storper, M., & Venables, A. J. (2004). Buzz: Face-to-face contact and the urban economy. *Journal of Economic Geography*, *4*(4), 351–370. https://doi.org/10.1093/jnlecg/lbh027

Ter Wal, A. L. J. (2014). The dynamics of the inventor network in german biotechnology: Geographic proximity versus triadic closure. *Journal of Economic Geography*, *14*(3), 589–620. https://doi.org/10.1093/jeg/lbs063

Torre, André, & Gilly, J. P. (2000). On the analytical dimension of proximity dynamics. *Regional Studies*, *34*(2), 169–180. https://doi.org/10.1080/00343400050006087

Torre, Andre, & Rallet, A. (2005). Proximity and localization. *Regional Studies*, *39*(1), 47–59. https://doi.org/10.1080/0034340052000320842

Wagner, C. S., & Leydesdorff, L. (2005). Network structure, self-organization, and the growth of international collaboration in science. *Research Policy*, *34*(10), 1608–1618.

https://doi.org/10.1016/j.respol.2005.08.002

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, *316*(5827), 1036–1039.

Xu, R. (2013). Improvements to random forest methodology. *Dissertation (Doctor of Philosophy) Iowa State University*, 1–88. http://lib.dr.iastate.edu/etd/13052/

Zeller, C. (2004). North Atlantic innovative relations of Swiss pharmaceuticals and the proximities with regional biotech arenas. *Economic Geography*, *80*(1), 83–111. https://doi.org/10.1111/j.1944-8287.2004.tb00230.x