Social networks and the geography of innovation and research collaboration : Three essays

Laurent Bergé

▶ **To cite this version:**

Laurent Bergé. Social networks and the geography of innovation and research collaboration : Three essays. Economics and Finance. Université de Bordeaux, 2015. English. NNT : 2015BORD0358 . tel-01278910

HAL Id: tel-01278910
https://theses.hal.science/tel-01278910

Submitted on 25 Feb 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

université de BORDEAUX

THÈSE PRÉSENTÉE
POUR OBTENIR LE GRADE DE

DOCTEUR DE L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE ENTREPRISE, ÉCONOMIE, SOCIÉTÉ (ED N°42)
SPÉCIALITÉ : SCIENCES ÉCONOMIQUES

Par **Laurent BERGÉ**

SOCIAL NETWORKS AND THE GEOGRAPHY OF INNOVATION AND RESEARCH COLLABORATION: THREE ESSAYS

Sous la direction de **Nicolas CARAYOL**, Professeur des Universités

Soutenue le 11 décembre 2015

Membres du jury :

M. **COWAN Robin**,
Professeur, University of Maastricht, *président*

Mme **LE GALLO Julie**,
Professeure, Agrosup Dijon, *rapporteur*

M. **LISSONI Francesco**,
Professeur, Université de Bordeaux, *examinateur*

M. **MAGGIONI Mario**,
Professeur, Università Cattolica del Sacro Cuore, *rapporteur*

The University is not to provide any approval or disapproval regarding the opinions this PhD dissertation includes. These opinions must be considered as being solely those of their authors.

L'Université n'entend ni approuver, ni désapprouver les opinions particulières émises dans cette thèse. Ces opinions sont considérées comme propres à leur auteur.

Remerciements

Mes remerciements s'adressent tout d'abord à mon directeur de thèse, Nicolas Carayol, de m'avoir permis de mener à bien ce travail. Sa patience, son écoute et sa disponibilité à mon égard ont été une aide précieuse.

I am grateful to Robin Cowan, Julie Le Gallo and Mario Maggioni for accepting to be the members of my jury, and for the time they took to evaluate my work.

Durant cette thèse, un grand nombre de personnes ont contribué, de près, de loin, à m'aider à accomplir cette tâche et à devenir un chercheur. En particulier un grand merci à Pascale Roux et Ernest Miguélez pour avoir lu mes travaux et m'avoir apporté leurs conseils tout le long. Une mention spéciale pour Francesco Lissoni qui, en plus de m'avoir aidé sur plusieurs aspects de la thèse, a partagé un dynamisme qui a significativement contribué à me faire changer – pour le mieux. En particulier je lui suis très reconnaissant de m'avoir permis d'aller, l'espace de quelques mois, à la LSE.

Le soutient financier du GREThA, des programmes de recherche et de l'école doctorale Entreprise, Économie et Société, m'a été précieux en rendant ces échanges possibles.

I whish to thank Simona Iammarino and Riccardo Crescenzi for welcoming me in their departement at LSE. I whish also to thank all the fellows PhDs I met once there for the warm welcome they offered me, in particular Davide Luca, Alessandra Scandura and Alex Jaax, whom I am always delighted to meet at conferences.

I am grateful to Thomas Scherngell and Iris Wanzenböck. Collaborating with them was a real pleasure and I did learn a lot in the process.

Un merci à tout mes collègues footballeurs de Grethalia qui m'ont fait vivre un bon – mais rude – moment à affronter la jeunesse étudiante.

Enfin, un grand merci aux collègues de bureau, de couloir, et de laboratoire.

Merci, merci aussi au cercle de l'Happy, pour leur soutien tout particulier mais ô combien agréable.

Je remercie mes parents pour leur patience dans cette longue entreprise et leur soutien constant.

Les derniers mots de ces remerciements vont à Marlène, qui a toujours été là pour m'encourager !

Contents

*This chapter is based on a single authored article.
†This chapter is based on a paper co-authored with Iris Wanzenböck and Thomas Scherngell.
‡This chapter is based on an article co-authored with Nicolas Carayol and Pascale Roux.

List of Figures

List of Tables

Introduction

Understanding innovation is important as it is identified to what makes societies thrive. Indeed, innovation and the generation of new ideas are central elements in the theory of endogenous economic growth (Romer, 1990; Aghion and Howitt, 1992) and are viewed as a key driver of prosperity. As a corollary, innovation is deeply etched in political agendas: for instance the EU programme HORIZON 2020 underlines that "investment in research and innovation is essential for Europe's future" European Commission (2014, p. 5).

A main characteristic of innovation is that the production of innovation is one of the most geographically concentrated of economic activities (e.g., Glaeser, 2011; Carlino et al., 2012). This feature, that innovation tend "naturally" to be geographically clustered, has attracted a significant amount of attention from scholars, to investigate what kind of role co-location did play in the innovation process (see e.g. Audretsch and Feldman, 2004; Feldman and Kogler, 2010; Carlino and Kerr, 2015, for a review).

An important rationale leading to this geographical concentration was identified by *agglomeration economies* (Marshall, 1890). The spatial agglomeration of economic activity is usually associated to three main benefits: economies of scales in the provision of intermediate inputs, job market pooling, and knowledge spillovers (Carlino and Kerr, 2015). The first and second elements are delineated by the market, and benefit in general to every firm by raising their productivity (see for instance Duranton and Puga, 2004, for micro economic rationales). The last element, knowledge spillovers, is more specific to innovation as it concerns the generation and diffusion of ideas. Yet, the meaning of, and mechanisms involved in, knowledge spillovers are more elusive than for the two other agglomeration economies.

The term "knowledge spillover" was coined in reference to what Marshall described as a situation in which "[t]he mysteries of the trade become no mysteries; but are as it were in the air, and children may learn them unconsciously" Marshall (1890, p. 271). Knowledge spillovers can then be summarized by the idea that the knowledge produced by firms is akin to non-rival and non-excludable goods available to other firms in their geographical vicinity. Geographical distance here plays a critical role, as the benefits from these spillovers is assumed to decay with distance. Using the analogy of a firm's production function $F(A, K, L)$, the presence of knowledge spillovers mean that the productivity A is increasing in the firm's geographical proximity to other firms, and for no other reason

than geographical distance. The underlying rationale to knowledge spillovers is that geographical proximity allows the workers to capture the knowledge produced from other workers thanks to non-market related social interactions.

One major difficulty of identifying knowledge spillovers is that they should mostly be the consequences of human interactions and influence, and therefore do not "leave a paper trail" (Krugman, 1991b, p. 53). Challenging this puzzling issue, Jaffe et al. (1993) remarked that one specific kind of interaction did leave a paper trail for legal reasons: citations in patent documents. Indeed, in the patent application process, inventors are required to reference the previous pieces of existing knowledge upon which is built their novel idea. This is therefore equivalent to a disclosure of the inspirations that led to their invention, inspirations that could have been triggered by interactions in the local environment. Hence, Jaffe et al. (1993) did use patent citations as a proxy for knowledge flows and as a means of identifying localized knowledge spillovers. They then employ a case-control methodology to control for the distribution of the patenting activity per sector. Their study evidenced that patent citations were disproportionately localized: inventors tend to cite patents originating from the same area at a higher frequency than what the spatial distribution of the industry would predict. This work is considered as the earliest evidence of localized knowledge spillovers. Despite the methodological issues raised by Thompson and Fox-Kean (2005), the main results from Jaffe et al. (1993) were later confirmed by Murata et al. (2014) who did implement a more general methodology, which includes the two former ones as special cases.

Patent citations tend to be more localized, but what does it really means, is it evidence that knowledge flows are "in the air"? Following the work from Jaffe et al. (1993), later studies criticized the interpretation of this result as a sign of localized knowledge spillovers, and questioned the extent to which such "spillovers" were external to market-based mechanisms (Zucker et al., 1998; Breschi and Lissoni, 2001). The principal suspect leading to this possibly erroneous interpretation is an essential factor in knowledge creation: social networks. Indeed, Breschi and Lissoni (2005, 2009) and Singh (2005) have demonstrated that once the social network of inventors, measured by co-inventions, was controlled for as a channel for diffusion, there were no remaining sign of localized knowledge spillovers as measured à la Jaffe et al. (1993). It is a sign that inventors do not cite randomly the other patents produced in the area, but rather cite the ones produced by inventors in their social network. If localized knowledge spillovers were in fact the consequence of the diffusion of knowledge through connections in the co-patent network, it would no longer be "in the air", i.e. available to all actors in the area. Rather, it would pertain to market interactions, since co-patents are formal relation which can be shaped by the market (for instance via job mobility, or inter-firm collaborations).

Although these studies challenge the concept of knowledge spillovers as being a channel external to the market, they do not rule out the possibility that agglomeration is beneficial for innovation, thanks to the development of the inventors network. The key element

here is that the driver of innovation could be the information the inventors can draw from their social network. Thus these studies rather evidence that social interactions are predominantly localized. Yet, to understand why would these social interactions matter for innovation, one has go beyond the economies of agglomeration and to delve more specifically into the knowledge creation process.

Why would the network, defined by the collaborations between knowledge workers, matter for generating new knowledge? To the importance of collaborations in the innovation process, Jones (2009) provides an appealing thesis. His starting point is the upward increase over time in the age of first invention, in specialization and in the size of teams in science (see, e.g., Royal Society Science Policy Centre, 2011) and innovation (Jones, 2009). He then argues that these observations are evidence that knowledge is getting harder to produce. The central element of his argument is that, since new knowledge is continuously produced, the time devoted to learning in order to reach the frontier of existing knowledge necessarily augments over time. Therefore, to overcome this problem innovators can either: increase the time devoted to education, or narrow their domain of expertise. The consequence is that innovation is getting harder to produce and teamwork is increasingly required to produce new knowledge.

This explanation provides a compelling rationale for the increase of team size over time, and hence underlines the increasing need to collaborate to produce innovation. In addition, beyond Jones' argument, collaborations provide many benefits which increases research output. For instance, teamwork allows researchers to quicken the trial and errors process, allow to better sift which are the good ideas, etc (see e.g., Katz and Martin, 1997; Singh and Fleming, 2010; Freeman et al., 2014).[4] However, focusing on collaboration alone as the main vector of knowledge creation would neglect another perspective on innovation, which is the diffusion of information and ideas. This perspective requires to take a broader look onto collaborations, to step back and consider the whole network of inventors and how they are embedded in their social connections.

To illustrate that team size is not equivalent to the concept of network, Figure 1 represents the co-evolution of both the average number of inventors per patents (i.e. team size) and the average degree per inventor for French patents. Team size is characterized by a slow, but steady, increase over time: it starts from 1.79 in 1985 to reach an average of 2 in 2002. What the figure emphasizes is that the connectedness of inventors over time clearly outpaces the increase in team size. One inventor had an average of 1.8 different collaborators in 1985, this number rose to more than 2.5 in 2002. In this 17 years period, while team size grew only by 12%, inventors connections grew by more than 45%. Thus there is not a direct connection between the two, and inventors have a higher number of different collaborators over time.

[4]Pointing to the importance of collaboration for creation, the motto of the prolific Hungarian mathematician Paul Erdös was "another roof, another proof" (Baker et al., 1990, p. ix), emphasizing for him the importance to move to new places and work with new collaborators.

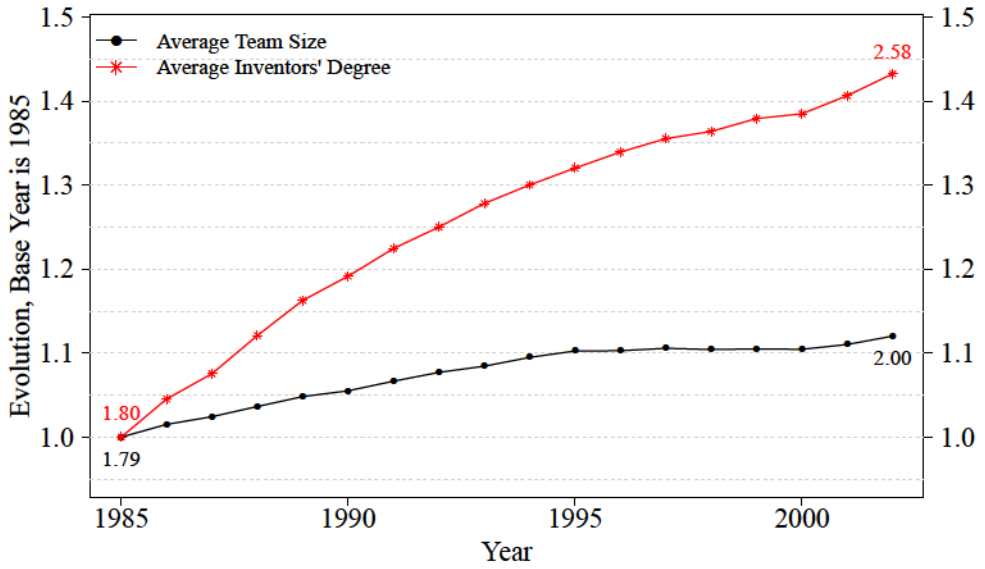

Figure 1 – Evolution of the number of inventors per patents (team size) and of the number of connections per inventor (degree).

Sources and notes: The patent data on French inventors stems from Patstat, and is described in Chapter 3. The inventors' degree is computed using a 5-years-window, i.e., for each time period, the co-patent network is constructed from $t-4$ to t. The degree is then defined as the number of different collaborators an inventor has.

Why would the network, beyond the scope of teamwork, matter for knowledge creation? Burt (1992, 2005) has emphasized that the position of agents within communication networks in organizations can have a great impact on their performance. Specific positions in the network enables to gather a broader set of information and to make better decisions. In a study on micro-finance in Indian villages, Banerjee et al. (2013) show that the network-position of agents who initially participate to a loan program is critical to the diffusion of the adoption of this loan program to other inhabitants of the village.

Does this diffusion pattern is similar for knowledge networks? Theoretically, if knowledge could freely flow through social connections, this diffusion property of social network could enhance increase knowledge creation. In particular, some kind of knowledge is not codified and is specific to a particular firm or worker. Collaboration can provide an access to this specific, tacit, knowledge (e.g., Cowan and Foray, 1997; Cowan et al., 2000). Once an agent dispose of it, he/she can also disseminate it through hi/her contacts. Non only some particular set of knowledge but also ideas can circulate.

In this line of thought, the distribution of network connections could be critical to favour knowledge to diffuse, and therefore be pivotal for the whole performance of the network. For instance, in networks in which agents are at a very short distance, a good idea generated by one agent could quickly reach all other agents. On the contrary, in sparse networks, in which most agents are far apart from each other in the network, good ideas could be very difficult to diffuse through the network connections.

Cowan and Jonard (2004) theoretically investigate this issue by the means of an agent-based model. They show that the diffusion of innovation is facilitated when the network

structure is characterized by a short average network-distance between the agents of the network. There are tantalizing explanations linking the network structure to the production of innovation. However, there is only limited empirical evidence supporting this view. For instance, studying US inventors Fleming et al. (2007) and Lobo and Strumsky (2008) find only inconclusive results regarding the network structure of inventors and the production of innovation.

Thesis purpose and outline. This thesis aims at clarifying the mechanisms involved in knowledge creation, by specifically investigating the role of social networks in this process. This investigation requires to better define the position of agents in networks as well as to shed light on the the link between network position and performance in terms of knowledge creation. Furthermore, since the actors of knowledge creation are highly concentrated in space, this thesis also aims at better understanding the interplay between geography and social networks. The three chapters will cover these issues, each with a specific focus.

The first question that is addressed concerns the determinants of new collaborations. Collaborations plays an significant role in the knowledge creation process and also shape the structure of the network. Therefore, it is important to understand what does determine future collaborations. The first chapter will cover this issue by investigating whether the network of past collaborations can explain the pattern of future collaborations. In particular, the chapter will assess the interplay between the social network and geography as determinants of collaborations.

The second question regards the position of regions in innovation networks. Knowledge creation depends on the combination of different sets of knowledge (Fleming, 2001). To create and combine new sets of knowledge, collaborations with other regions can be critical to renew the local pool of knowledge (Owen-Smith and Powell, 2004). The connections between agents from a given region with agents from other regions form an inter-regional network of collaborations. Furthermore, regions are the locus of policies and there is an increasing need of understanding 'the position of region[s] within the European and global economy' (European Commission, 2012b, p. 18). Therefore, it is important to know whether some regions are well positioned or not in the network. In the second chapter, we will discuss critically how to characterize the position of regions in innovation networks.

Last, we look at whether, and how, the network structure can enhance the innovative outcome. As emphasized earlier, the position of inventors in the network can have an influence on the diffusion of information and ideas. The third chapter then aims at assessing how the position of inventors in the network can influence regional innovation performance.

The remainder of this introduction describes the precise purpose of the three chapters,

along with the methodology they employ and their main findings.

Chapter 1. The first chapter provides an assessment of the influence, and interplay, of both geography and social networks on the formation of scientific collaborations. First, the chapter provides a theoretical discussion of the causes rooting scientific collaborations, with a special emphasis on network-based mechanisms. In particular, the possible consequences of the interaction between social proximity and geography are examined, focusing on the conditions in which they can be substitute or complement. We argue that social proximity should positively influence network formation, but we remain agnostic on whether it substitutes or complements geographic proximity.

Second, an empirical analysis is set up to test the hypotheses laid out by the theoretical discussion. To this end, we use data on scientific collaborations in chemistry, in five EU countries (France, United Kingdom, Germany, Italy, Spain) between 2000 and 2005. A gravity model is used to assess the determinants of inter-regional collaboration flows. Due to the lack of existing measures to assess the inter-regional social proximity, this chapter also contributes methodologically to the literature by introducing such a measure. The measure of inter-regional social proximity we introduce is theoretically grounded and based on micro-level determinants.

The results depict a clear substitutability pattern between geographic proximity and social proximity. Social proximity does positively influence the formation of new collaborations, but its effect is mediated by geography. While being non-significant for the regions most close geographically, the positive effect of social proximity is increasing with geographic distance.

Chapter 2. The second chapter looks at how to cope, in a meaningful way, with the notion of network centrality in the context of inter-regional R&D networks. We provide a theoretical discussion of the notion of network centrality in the context of inter-regional networks. This discussion is complemented with a specific focus on the context of R&D collaboration. We lay out the possible applications of existing network centrality measures to regional networks. We then analyse critically the meaning of these centrality measures in this context. We show that most of the time they cannot be applied to inter-regional R&D networks without suffering from important flaws to their interpretation. From this theoretical context, we then introduce a new measurement whose definition is suited to the context of inter-regional R&D networks. Finally, we use data on the EU co-invention network to illustrate this new centrality measure, and to compare it to other well-known existing measures. Despite some similarities between the different measures, we point to, and comment on, significant differences.

Chapter 3. The third chapter aims at unveiling what kind of network structure favours the most regional innovation. To this end, we first introduce a simple model linking inventors' productivity to their network of collaborators. This model contains three key

parameters. One, referred to as *connectivity*, which scales the network benefits, and two others which rule the network structure: *complementary* (an inventor may benefit more the partner's effort to produce knowledge), and *rivalry* (an inventor may benefit less from his/her partner if the partner has more connections). The model predicts that at equilibrium, the network-related production of the agents are dependent on the square of a measure of their position in the network. This measure of network position is a new form of network centrality which depends on the three elements of *connectivity*, *complementarity* and *rivalry*. Furthermore, this network centrality can be seen as a generalized form of network centrality, since it which encompasses, as special cases, several well known forms of centrality (Bonacich, Page-Rank, degree).

Following this insight, regions disposing of inventors better positioned in the network should perform better. Yet, the network's position which will provide the highest centrality will depend on the parameters of complementarity and rivalry. The empirical analysis will then consist in 1) estimating whether inventors' centrality affect regional innovation performance, and 2) which parameters of the model best fit the data. The estimates of complementarity and rivalry would then provide what kind of network structure favours the most regional innovation. We estimate the model's parameters with Poisson regressions including a various set of fixed-effects. The empirical evidence is based on patent co-authorships between French inventors for the period 1981–2003. We provide two main findings.

The main outcome of this chapter is that, while there is only a slight sign of complementarity, the empirical results strongly evidence that there is no rivalry effect at play: this means that any new network connection between two previously unconnected inventors is always beneficial for regional innovation.

Chapter 1

Network proximity in the geography of research collaboration*

1.1 Introduction

The production of new knowledge is largely viewed as essential in enhancing competitiveness and producing long-term growth (Aghion and Howitt, 1992; Jones, 1995). It is therefore no wonder that it is a central issue for policy makers, at the regional, national and even supra-national scale. This in turn puts at the forefront policies that deal with collaboration in science: indeed, as knowledge becomes more complex and harder to produce (Jones, 2009), scientific activity turns out to be increasingly reliant on collaboration (see, e.g., Wuchty et al. 2007; Jones et al. 2008; Adams 2013 or the Royal Society Science Policy Centre, 2011 for a recent report). In the European Union (EU), the political will towards knowledge creation is being supported by the recent Horizon 2020 programme, which 'should be implemented primarily through transnational collaborative projects' (European Commission, 2013, Article 23). This policy tool aims at developing a European research area (ERA) where collaborations do not suffer from the impediments of distance or national borders, so that EU researchers can act as if they were all working in one and the same country. Such policies are backed by a large EU budget: yet is funding long-distance collaboration efficient? To comprehend this issue, one needs a clear understanding of the determinants of collaboration, and in particular, the factors that help bypass geography.

Despite the trumpeted 'death of distance', due recent developments in the means of communication and in transportation technologies (Castells, 1996), an understanding of geography is still important in explaining collaboration. Co-location facilitates face-to-face contact, eases the sharing of tacit knowledge (e.g., Gertler, 1995; Storper and Venables, 2004) and enhances the likelihood of serendipitous, fruitful collaborations (Catalini, 2012). Furthermore, national borders, a by-product of geography, also play an important role, as differences in national systems render collaboration more difficult (Lundvall, 1992).

*This chapter is based on a single authored article.

A recent stream of literature has shown that geographical distance and national borders are indeed strong impediments to collaboration (e.g., Hoekman et al., 2009; Scherngell and Barber, 2009; Singh and Marx, 2013). Temporal analyses even add that their hindering effects have not decreased over time (e.g., Hoekman et al., 2010; Morescalchi et al., 2015). Returning to the ERA, it seems like the EU's policies have failed to develop an integrated area, in which distant collaborations are eased. However, geography is not the sole determinant of collaboration (Boschma, 2005; Torre and Rallet, 2005; Frenken et al., 2009a; Giuliani et al., 2010). Collaboration is a social process and entails the creation of bonds between researchers (Katz and Martin, 1997; Freeman et al., 2014). Those bonds in turn form a social network, and one salient fact about social networks is that they are a driver of their own evolution (Jackson and Rogers, 2007). Consequently, analyses should not fail to consider potential network effects influencing the collaboration process.

This chapter is a step toward a better understanding of the role of networks in the geography of research collaboration. While the question concerning the determinants of network formation and its link with the notion of proximity has attracted a growing interest over the recent years (e.g., Balland et al., 2013; Boschma et al., 2014a; Balland et al., 2015b), studies focusing on the determinants of research collaboration have mostly been descriptive, a-geographic, or otherwise failed to weld the network together with geography (e.g., Newman, 2001; Barabâsi et al., 2002; Wagner and Leydesdorff, 2005; Almendral et al., 2007; Balland, 2012; Fafchamps et al., 2010; Autant-Bernard et al., 2007a; Maggioni et al., 2007). Thus, the question of substitutability/complementarity between geography and the network has been set aside. There is some evidence on this question provided in other contexts (e.g., Bathelt et al., 2004; Boschma, 2005; Montobbio et al., 2015), but empirical findings on this issue remain scarce. Yet, the answer to this question is important policy-wise. If geographic and network proximity really are substitutable, then heightening the network proximity of distant agents would in turn help them in creating new long-distance links, since network proximity would partly compensate for the loss of geographic proximity. On the contrary, in the case of complementarity, 'forcing' distant collaborations may be inefficient since distant agents would be those who benefit the least from network proximity. Only the former case would support current EU policies, and that is assuming that the network matters at all.

This chapter also contributes to the literature by introducing a new measure of inter-regional network proximity. This measure is defined for each regional dyad and reflects the intensity of indirect linkages between regions. Moreover, this measure can be interpreted from a micro perspective as it can be derived from a simple model of random matching. For a given regional pair, this measure can then be related to the expected number of indirect linkages between the agents of the two regions. This kind of measure is in line with the increasing need to understand 'the position of region[s] within the European and global economy' (European Commission, 2012b, p. 18).

To assess empirically how network proximity affects collaboration, I then make use

of European co-publication data. These data relate to co-publications stemming from five European Union countries (France, Germany, Italy, Spain, the United Kingdom), from the field of chemistry, published between 2001 and 2005. The analysis consists of an estimation of the determinants of flows of collaboration between 17,292 regional dyads from 132 NUTS 2[1] regions, by means of a gravity model (Picci, 2010; Cassi et al., 2015). The results demonstrate an interplay between geography and network proximity: while being negligible or only weakly beneficial to regions located in close proximity, *the importance of network proximity grows with distance*, reaching an elasticity of 0.24 for a distance of 900 km. In other words, network proximity mainly benefits international collaborations. Thus, these results support the claims of EU policy.

The remainder of this chapter is organized as follows: in Section 1.2 the determinants of inter-regional collaborations are discussed, focusing on the role of network-based mechanisms and their possible interplay with non-network forms of proximity; Section 1.3 then presents the estimation methodology and describes the measure of network proximity used in this chapter, along with the model from which it can be retrieved; in Section 1.4, the data set is presented, as well as the econometric strategy; the empirical findings are reported and discussed in Section 1.5; and Section 1.6 concludes the chapter.

1.2 The determinants of inter-regional collaborations

In this section I describe the determinants of scientific collaboration. First, I discuss the static ones, which depend on the characteristics of the researchers, i.e., the nodes of the network, and do not evolve over time. Second, I present the micro-determinants of collaboration stemming from the network. Finally, I discuss the relation between network proximity and geography.

1.2.1 Static determinants of collaboration

When it comes to analysing the determinants of collaboration, the concept of proximity proves to be a very useful framework (Boschma, 2005; Torre and Rallet, 2005; Kirat and Lung, 1999). By distinguishing several types of proximity between agents (such as geographical, institutional, cognitive or organizational), this framework allows one to analyse each of them and to easily assess their interplay. One can distinguish two mechanisms through which proximity, in whatever form, favours collaboration: 1) proximity augments the probability of potential partners to meet and 2) reduces the costs involved in collaboration. In this way, it simultaneously increases the expected net benefits of the collaboration and the likelihood of its success.

The effect of geographical proximity on collaboration can be deconstructed in such a

[1]The Nomenclature of Territorial Units for Statistics (NUTS is the French acronym) refers to EU geographical units whose definition attempts to provide comparable statistical areas across countries. The exhaustive list of regions used in this study is given in Appendix 1.7.4.

way, as follows. First, the context of collaborative production of knowledge may require that the partners share and understand complex ideas, concepts or methods; the collaboration may then involve a certain level of transfer of tacit knowledge. Consequently, face-to-face contact may be important to the effective conducting of the research, as a way of overcoming the problem of sharing tacit knowledge (Gertler, 1995; Collins, 2001; Gertler, 2003). Moreover, face-to-face contact allows direct feedback, eases communication and the mitigation of problems, and facilitates coordination (Beaver, 2001; Freeman et al., 2014). All these elements heighten the probability of the success of a collaboration. Thus, geographical distance, by incurring greater travel costs and fewer opportunities to exchange knowledge by means of face-to-face contact, reduces the likelihood of a successful collaboration (Katz and Martin, 1997; Katz, 1994).

Second, being closer in space enhances the likelihood of potential partners to meet in the first place. Indeed, attendance of social events where researchers meet to share ideas, such as conferences, seminars or even informal meetings, is linked to geographical distance, and thus heightens the chances of finding a research partner at a local scale. For instance, by analysing data on participants at the congresses of the European Regional Science Association, van Dijk and Maier (2006) have shown that a greater distance to the event negatively affects the likelihood of attending it. In addition, the social embeddedness of researchers and inventors has been shown to decay with geographical distance (Breschi and Lissoni, 2009), meaning they will have a better knowledge of potential partners at a closer distance.

Consequently, the effect of geographical distance should be understood as negative. This fact has been evidenced by various recent studies, in different contexts: in the case of co-authorship in scientific publications (Frenken et al., 2009b; Hoekman et al., 2010, 2009), in co-patenting (Hoekman et al., 2009; Maggioni et al., 2007; Morescalchi et al., 2015), and in the case of cooperation among firms and research institutions within the European Framework Programme (Scherngell and Barber, 2009).

Another impediment relating to geography is the effects of national borders. In the context of inter-regional collaboration, national borders are often linked to the notion of institutional proximity (Hoekman et al., 2009). Institutional proximity relates to the fact that 'interactions between players are influenced, shaped and constrained by the institutional environment' (Boschma, 2005, p. 63). Indeed, several features affecting knowledge flows can be perceived at the national level (Banchoff, 2002; Glänzel, 2001). For instance, funding schemes are more likely to exist at a national scale, thus, facilitating collaborations within a single country. In the same vein, workers are more mobile within a country than across countries, and since they may maintain ties with their former partners, their social networks appear to be more developed at the national level (Miguélez and Moreno, 2014). Norms, values and language are also likely to be shared within a country, facilitating collaboration. As a consequence, the literature provides evidence that belonging to the same country significantly eases the collaboration process (e.g.,

Hoekman et al., 2010; Morescalchi et al., 2015).

1.2.2 The role of networks in the process of collaboration

This section discusses a number of network-related mechanisms that help trigger collaboration. The first mechanism playing a role in network evolution is triadic closure, defined as the propensity of two nodes that are indirectly connected to form a link (Carayol et al., 2014). It may be the case that triadic closure occurs because triads, in opposition to dyads, have certain advantages. By reducing individual power, triads can help mitigate conflicts and enhance trust among the individuals (Krackhardt, 1999). The possibility of negative behaviour on the part of one of the agents is more limited, since it can be punished by the third agent, who can sever the relation. These structural benefits offered by a closed triad may in turn lead to triadic closure. This can be an advantage particularly for international collaborations, in which the reliability of different partners may be difficult to assess. In such circumstances,relying on the network and forming a triad – that is to collaborate with a partner of a partner – can be desirable, as it limits opportunistic behaviours, thus reducing the risks associated with the sunk costs of engaging in a collaboration. In a recent study on the German biotechnology industry, ter Wal (2014) has shown that triadic closure among German inventors has become increasingly important over time, as the technological regime has changed and more trust has been needed among partners. In adiition, by examining the behaviour of researchers at Stanford University, Dahlander and McFarland (2013) have shown that having an indirect partner significantly increases the probability of a collaboration.

Another feature of social networks that may influence their evolution is homophily. Homophily can be identified as a compelling feature of social networks; it can be portrayed as 'the positive relationship between the similarity of two nodes in a network and the probability of a tie between them' (McPherson et al., 2001, p. 416). This characteristic has been analysed by sociologists in various contexts – for example, in friendships at school or in working relationships – and it has been shown that similarity among individuals is a force driving the creation of ties. As McPherson et al. (2001, p. 429) put it: 'Homophily characterizes network systems, and homogeneity characterizes personal networks'. Science is no exception: for instance, Blau (1974) has studied the relationships among theoretical high energy physicists, and shown that the similarity of their specialized research interests as well as their personal characteristics, are important factors determining research relationships.

Homophily is not specific to network-related effects. Indeed, the importance of the static determinants of collaboration also rely on homophily. Yet, once the problem is reversed, one can see that the network can influence new connections via homophily. Indeed, take any two agents already connected: they are likely to share at least some similarities that helped them succeed in collaboration. This might, for instance, involve sharing a similar research topic, having the same approach to research questions or simply

being compatible in terms of teamwork (i.e., they are a good match with respect to their own idiosyncratic characteristics). Therefore, if two agents are connected to the same partner, they are likely to be in some way similar to their common collaborator, and consequently to share some similarities themselves. These similarities may in turn favour their future collaboration.

Finally, the network can be seen as a provider of externalities of information, and thus be decisive in determining future collaborations. Indeed, as the need for collaboration becomes more and more acute (Jones, 2009), finding the right partners becomes absolutely critical, but may also be time-consuming. Katz and Martin (1997) point out that time is one of the most important resources for researchers, even before funding. As a consequence, the network can act as a reliable repository of information in which researchers can find their future collaborators (Gulati and Gargiulo, 1999). The role of networks might then best be viewed by analogy to optimization problems: despite not giving the best global match, the network helps to provide the best local match. Researchers are time constrained and are not fully rational, in the sense that they do not dispose of all the required information nor of the ability to gauge all potential matches in order to select the best one. In this situation, 'picking' the best partner in the network vicinity may be a rational and efficient choice. In this vein, Fafchamps et al. (2010) have developed a model describing how researchers obtain information on each other through the network of social connections. They show that the probability to access information on a specific researcher decreases with the network distance. They also find empirically, using data on co-authorship among economists, that being 'closer' in the network positively affects the likelihood of collaboration.

To summarize, the network regroups various mechanisms which favour collaboration, thus affecting its evolution. This yields the following hypothesis:

Hypothesis 1 Network proximity positively affects the creation of new collaborations.

Some precision needs to be applied to the notion of network proximity that will be used throughout this chapter. Although the notion of triadic closure applies specifically to agents who are very 'close' in the network (i.e., they have a common partner), other mechanisms, such as information externalities provided by the network, do not require such proximity and could apply at a greater distance. Thus the notion of network proximity here concerns being connected by indirect social ties. Various distances separate the pairs of agents in the network, and the hypothesis states that the 'closer' the agents are with respect to network distance, the more likely they are to engage in collaboration, as a result of the discussed mechanisms.

However, while network proximity may influence the formation of new collaborations, can its effect be regulated by other factors, like geographical distance or national borders? Or is the effect of network proximity merely independent of these other determinants? This question needs to be investigated in order to unravel the precise mechanism shaping

the landscape of collaboration networks. The next subsection discusses how network proximity and other forms of proximity may be intermingled.

1.2.3 The interplay between the network and other forms of proximity

This section aims to link network proximity to other forms of proximity and to understand their interplay in the collaboration process. For the sake of readability, in this section I will compare network proximity only to geographic proximity. That is to say, geographic proximity is here intended as a shorthand for non-network forms of proximity.

If both network and geographic proximity influence the creation of new collaborations, what might be the net outcome of these two effects? The first case one could consider would be that network proximity benefits homogeneously all prospective partners, meaning an independence between the effects of network and geographical proximity. In other words, the greater the network proximity, the higher the likelihood of a collaboration, at a magnitude independent from geography. However, this independence could only occur if geographical proximity and network proximity functioned at two completely different levels: that is, if the very mechanisms through which they affected collaboration were unrelated. As soon as they are influencing collaboration through the same common mechanisms (like enhancing trust, or facilitating the search for prospective partners), their interplay will not be independent. So if one departs from the hypothesis of independence, one is left with two opposing standpoints in competition.

On the one hand, network proximity can reinforce the benefits of being geographically close. Particularly in cases where agents have a 'taste for similarity', network proximity can foster collaborations in situations in which agents already benefit from geographical proximity. This taste for similarity can be seen as a need to be close in different respects in order to conduct effective research. For instance, in a case where the research is highly subject to opportunistic behaviour, several forms of proximity may function complementarily to mitigate it.

On the other hand, the benefits of proximity may suffer from decreasing returns. In this case, network proximity would be a substitute for other forms of proximity. Take the case in which two prospective partners are geographically far apart: for them, network proximity will be crucial to engage in a successful collaboration, as it will be their sole source of proximity. On the contrary, if they are already close to each other then, as a result of the decreasing returns, having close network proximity would matter less, and would therefore not be decisive in triggering effective collaboration. Such effects would depict a pattern of substitutability. Another possible interpretation yielding the same conclusion might be that the net rewards of collaboration may increase with distance (this view is supported by, e.g., Narin et al., 1991; Glänzel, 2001; Frenken et al., 2010; Adams, 2013). In this case, and if the probability of success is still tied to the level of proximity between the agents, this would increase the marginal value of network proximity

for distant collaborations. Thus, this would also depict a substitutability pattern.

The preceding argument then yields these two following competing hypotheses:

Hypothesis 2.a Network proximity is a complement to other forms of proximity.

Hypothesis 2.b Network proximity is a substitute for other forms of proximity.

The interplay between network and non-network proximity has not been completely dealt with in the literature. There have been studies focusing on the role of networks and the role of geography, but few that unravel their interplay. For instance, Maggioni et al. (2007) have compared the effect of network ties (as opposed to purely geographical linkages) as determinants of the regional production of patents. Another example is Autant-Bernard et al. (2007a) who focus on collaborations among firms in the EU's 6^{th} Framework Programme. They asses the effect of network proximity and geographical proximity on the probability of collaboration. Both studies find positive effect for both geographic and network proximity.

In the same vein, other studies have tried to reveal the dependences among different forms of proximity, but not specifically the network form. For instance, Ponds et al. (2007) and d'Este et al. (2013) have studied the relation between organizational proximity and geography. While the former study analyses co-publications in the Netherlands and finds a substitutability pattern, the latter focuses on university-industry research partnerships in the UK and finds no interaction between the two.

This study departs from the previous literature by specifically focusing on network proximity and, more importantly, its relation to geography. In line with previous studies, the focus will be on inter-regional flows of collaborations in Europe (e.g., Scherngell and Barber, 2009; Hoekman et al., 2009; Morescalchi et al., 2015). Yet before outlining the data, I will first present the modelling strategy and spend some time describing the measure used to approximate the notion of network proximity in this paper.

1.3 Empirical strategy and the measure of network proximity

This section introduces the empirical model used in the econometric analysis and then develops the measure that will be used to assess network proximity. As will be shown, the measure can be derived from a model of random matching between agents, thus reflecting the idea of a micro-level measure.

1.3.1 Gravity model

The object of this study is to analyse the determinants of inter-regional collaboration flows. Thus, in line with previous research on this topic, the methodology used will

be the gravity model.[2] The gravity model is a common methodological tool used when assessing spatial interactions in various contexts, such as trade flows or migration flows (Roy and Thill, 2004; Anderson, 2011), and has been recently applied to the context of collaboration (e.g., Maggioni et al., 2007; Autant-Bernard et al., 2007a; Maggioni and Uberti, 2009; Hoekman et al., 2013). In a nutshell, the gravity model reflects the idea that economic interactions between two areas can be explained in terms of the combinations of centripetal and centrifugal forces: while the masses of the regional entities act as attractors, the distance separating them hampers the attraction. This can be written as follows:

$$Interaction_{ij} = Mass_i^{\alpha_1}\ Mass_j^{\alpha_2} F\left(Distances_{ij}\right), \tag{1.1}$$

with $F(.)$ being a decreasing function of the distances. The distance functions are usually of the form $F(x) = 1/x^{\gamma}$ or $F(x) = \exp(-\gamma x)$, depending on the nature of the distance variable x (Roy and Thill, 2004). Traditionally, $Mass_i$ and $Mass_j$ are respectively called 'mass of origin' and 'mass of destination'. In the context of this study, $Interaction_{ij}$ will represent collaboration flows. Within the gravity framework network proximity acts as a centrifugal force.

This study focuses specifically on the role of network proximity and then questions how the position of a particular pair of regions in the network may influence their future linkages. Various studies have applied network analysis tools to assess the position of regions within a network. Some studies cope with the position of regions within the network by making use of centrality measures (see, e.g., Sebestyén and Varga, 2013a,b; Wanzenböck et al., 2015, 2014). Other studies make use of the network, by linking the performance of a given region to the performance of the regions connected to it, in a fashion similar to that of spatial econometrics (see, e.g., Maggioni et al., 2007; Hazir et al., 2014).

To fit into the gravity model framework, and later into the econometric analysis, a measure of inter-regional network proximity should have two properties: first, it should be defined for each pair of region; and second, for the sake of coping with potential endogeneity problems, it should be independent of direct collaborations. Thus, before describing the data and the empirical model, I will first introduce such a measure.

1.3.2 A new measure of inter-regional network proximity

This section introduces a new measure aiming to capture the effect of network proximity in the context of inter-regional collaborations, in line with the gravity model framework. The measure being introduced depends only on inter-regional collaboration flows and functions by asking the following question: How much agents from two different regions are connected to the same agents in other regions? Although defined at the regional level,

[2]For a discussion of the different methodologies used to empirically assess the determinants of knowledge networks at the regional level, see for instance Broekel et al. (2013).

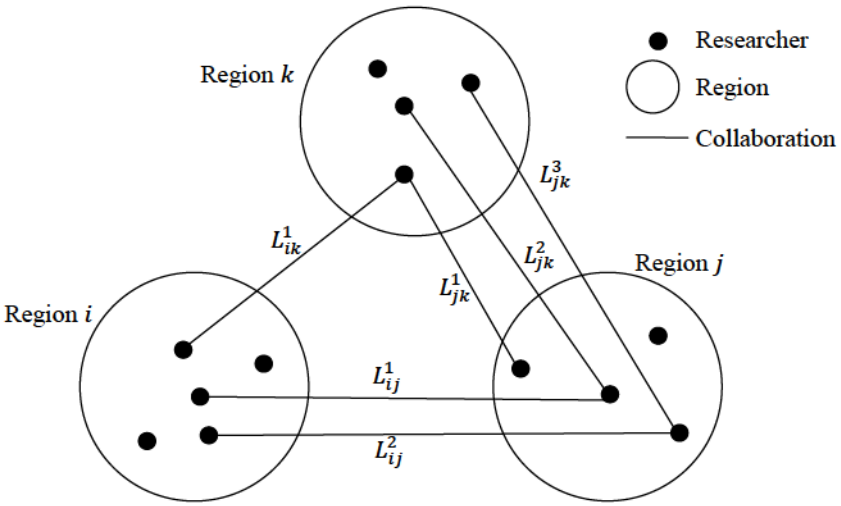

Figure 1.1 – Illustration of a regional network of collaboration and of the notion of bridging paths.

Notes: The figure depicts three bridging paths formed by the following pairs of links: (L^1_{ik}, L^1_{jk}), (L^1_{ij}, L^2_{jk}) and (L^2_{ij}, L^3_{jk}). So the regional dyads (i,j), (i,k) and (j,k) have respectively 1, 2 and 0 bridging paths. For instance, the pair of links (L^1_{ik}, L^1_{jk}) forms a bridging path between regions i and j via the bridging region k because these links are both connected to the same agent in region k, thus creating an indirect connection between agents from i and j. Note that although regions j and k have three direct links, there is no bridging path between them since they have no agent indirectly connected.

the measure actually reflects a micro-level notion, that of 'bridging paths' (i.e., inter-regional indirect connections at the micro level). This measure is referred to as TENB (standing for 'total expected number of bridging paths').

In the remainder of this section, the notion of bridging paths is first introduced, followed by a description of the model from which the measure is derived. Then, I show that the measure is robust to some variation in the model's assumption. Finally, the last subsection discusses the link between the measure as defined at the meso-level and the notion of network proximity.

1.3.2.1 The notion of bridging paths and some notations

First some notations, as they will be useful for defining the concept of bridging paths and will be used in the model in the next subsection. Consider N regions, each populated with n_i researchers. A link between two regions can be defined as a collaboration occurring between two researchers, one from each of those regions. Let g_{ij} be the total number of links between regions i and j. The set of regions to which i is connected (i.e., that have at least one link with i), also called the neighbours of i, is represented by $N_i \equiv \{k | g_{ik} > 0\}$. Finally, let L^a_{ij} represent the a^{th} link, $a \in \{1, \ldots, g_{ij}\}$, between agents from regions i and j, and let L^b_{jk} be the b^{th} link, $b \in \{1, \ldots, g_{jk}\}$, between agents from regions j and k.

Using these notations, a bridging path between region i and j via the bridging region k is defined as a set of two links (L^a_{ik}, L^b_{jk}), such that both links are connected to the same agent in region k. Stated differently, a bridging path exists when one agent from region i and one from j have a common collaborator in region k. The concept is illustrated by

Figure 1.1 which depicts a regional network of collaboration. In this example, the pair of links (L^1_{ik}, L^1_{jk}) forms a bridging path, while others like the pair (L^1_{ik}, L^3_{jk}) do not.

Bridging paths are seen as being a medium for network proximity. The main driver of the idea is that the more two regions have bridging paths, the closer their agents will be with respect to the network, and, *in fine*, they will be more likely to engage in collaboration, thanks to network-based mechanisms.

1.3.2.2 Deriving the measure from a model of random matching

This subsection shows how, by assuming that collaboration between agents stems from a simple random matching process, the expected number of bridging paths between two regions can be derived.

A random matching process. The random matching process used is based on two mild assumptions: 1) a collaboration consists of a match between two agents only; and 2) whenever a collaboration occurs between two regions, the two agents involved are matched at random.

This first assumption is rather functional and is used to make the model simple without being too restrictive. Indeed, the term 'agent' here is intended to be taken as a broad term: it could be either a lone researcher or a team of researchers, since teams can be fairly considered to behave like unique entities (see, e.g., Beaver, 2001; Dahlander and McFarland, 2013). The second assumption is in line with intuition, as it simply implies that for two regions, say i and j, the more observed collaborations there are between i and j, the more likely a randomly picked agent from i will have collaborated with one from j.[3]

Expected number of bridging paths (ENB). Using the information contained in the flows of inter-regional collaborations (i.e., all the g_{ij}) along with the *random matching process* assumptions previously defined, the expected number of bridging paths between two regions via another one (known as the bridging region) can now be derived.

Proposition 1. Under the random matching process, the expected number of bridging paths between regions i and j via the bridging region k is:

$$ENB^k_{ij} = \frac{g_{ik}g_{jk}}{n_k}. \tag{1.2}$$

Proof. See Appendix 1.7.1.

Proposition 1 relates to the expected number of bridging paths stemming from a specific bridging region. However, two regions can have more than just one common

[3]For instance, consider the network in Figure 1.1: if one selects one agent randomly from region i, it is more likely that she/he has collaborated with another agent from j than one from k (because there are two links with the former and only one with the later).

neighbour. The total expected number of bridging paths between two regions i and j is therefore the sum of the bridging paths stemming from all other regions to which i and j are both connected:

$$TENB_{ij} = \sum_{k \in N_i \cap N_j} \frac{g_{ik} g_{jk}}{n_k} \tag{1.3}$$

The measure of network proximity that will be used in this study is the total expected number of bridging paths (TENB). The link between the TENB and the notion of network proximity is discussed in Subsection 1.3.2.4; however before that, the next subsection will elaborate upon the consequences of a variation in the random matching assumption and show that this would imply only a trivial variation.

1.3.2.3 Robustness of the random matching assumption: the case of preferential attachment

Formally deriving the TENB in the previous section required an assumption of random matching: yet what if another kind of mechanism had been considered, like preferential attachment?

Preferential attachment is a feature of social networks that was first evidenced and modelled by Barabási and Albert (1999). It states that, as the network evolves, the new nodes that enter the network tend to link themselves to already well-connected nodes. In actuality, the distribution of the number of links per node in social networks is usually very skewed. The mechanism of preferential attachment, as developed in the model of Barabási and Albert (1999), yields an equilibrium distribution of links similar to real social networks: a power law distribution.[4] As a variation on the previously defined random matching process, I investigate the case of a matching process based on preferential attachment.

A form of preferential attachment. In this case, the matching is not done at random anymore; instead some nodes (researchers) are more likely to create links than others. Formally, the matching mechanism is defined as follows. There are n agents in a given region and they are assumed to be sorted according to their productivity level, so that Agent 1 has the highest productivity level and Agent n the lowest. Let the Greek letter ι, $\iota \in \{1, \ldots, n\}$, be their label. The probability that a new link involves agent ι is defined by $\iota^{-0.5}/\Gamma$ with $\Gamma = \sum_{\iota=1}^{n} \iota^{-0.5}$. For instance, consider a region populated by 10 agents, the probability of being tied to an incoming link is 20% for Agent 1, 14% for Agent 2, etc, and 6% for Agent 10. This can be compared to the random matching process, whereby each agent had the same likelihood of being connected: 10%.

This so-defined mechanism is very similar to the preferential attachment mechanism, except that the probability of creating a new link is exogenous instead of being dependent

[4] The distribution of the number of links per node, i.e., the degree, is assumed to follow a power law of parameter γ if the probability of having a degree k is equal to $f(k) = c \times k^{-\gamma}$ with c being a constant.

on the number of links an agent already has. Notably, as shown in Appendix 1.7.2.1, the expected distribution of the agents' degrees as a result of this process follows a power law of parameter $\gamma = 3$, as in Barabási and Albert (1999).

Now I turn to the derivation of the ENB through such a process, and analyse the difference between this measure and the measure obtained through the random matching process in equation (1.2).

Proposition 2. Under the random matching with preferential attachment, and for large enough values of n_k, the expected number of bridging paths between regions i and j via the bridging region k is as follows:

$$ENB_{ij}^{k,Pref} \simeq ENB_{ij}^{k} \times \frac{\log(n_k)}{4}.$$

Proof. See Appendix 1.7.2.2.

This result implies that, even when a more complex matching mechanism is used, the result is very similar to Proposition 1. Indeed, $ENB_{ij}^{k,Pref}$ is merely an inflation of ENB_{ij}^{k}. Certainly there are some variations as $\log(n_k)$ varies, but the logarithmic form flattens most of these, meaning that the correlation between $ENB_{ij}^{k,Pref}$ and ENB_{ij}^{k} is very high. This goes to show that the measure is robust to such variation in its assumptions.

1.3.2.4 The link between the TENB and the notion of network proximity

This subsection discusses the link between the notion of network proximity and the measure used to approximate it: the TENB. In particular, two points are addressed: an aggregation issue and a truncation issue.

The aggregation issue. In Section 2, the benefits of network proximity were discussed at the individual level. However, the measure created to approximate this notion, the TENB, is actually defined at the meso level. How do our inferences concerning the benefits of network proximity hold up when the concept of the TENB is used, which considers only inter-regional information?

In fact, the inter-regional network is only an aggregated view of micro-economic decisions. Regions do not collaborate with each other, only the agents within them do. Thus, it is conceptually difficult to consider regions simply as individual agents (see, e.g., ter Wal, 2011; Brenner and Broekel, 2011). Yet it would also be inexact to assume that the aggregate flows of collaboration do not convey any information about their micro-structure.

Following this line of thought, the TENB has a particular meaning as it is not simply an aggregate measure but rather can be interpreted as the expected number of indirect ties at the micro level, under mild assumptions. The measure is interpreted as follows:

$TENB_{ij} > TENB_{jk}$ means that the *agents* from the regions i and j are *likely* to be closer, with respect to indirect connections, than the agents from the regions j and k. Thus, the measure actually reflects the likelihood of a pattern at the micro level, in line with the idea of network proximity. Stated differently, a high TENB value *between two regions* is likely to reflect a high level of network proximity *between the agents of these two regions*. Consequently, if the network proximity, as measured by the TENB, has any effect on the inter-regional flows of collaboration, the interpretation should be that this is due to micro-level mechanisms.

The truncation issue. By construction, the measure of the TENB between two regions is based only on the indirect collaborations between them, and is completely independent from any direct collaboration. This implies that the network proximity reflected by the TENB is partial, as it is based on a truncated network. The purpose of this truncation is to avoid a reverse causality issue.

One could argue that the network proximity originating from the direct connections between agents from two regions may also be important in triggering new collaborations. Yet, since the identification of network proximity is based on network connections, direct collaborations between the two regions would directly influence their level of network proximity. As the question is about explaining collaborations, this would create a problem of reverse causality. In consequence, using the TENB means this problem is avoided at the cost of neglecting a possible network proximity originating from direct linkages.

1.4 Data and methodology

This section first explains the construction of the data set and all the variables; Subsection 1.4.3 then goes on to present the full model to be estimated, as well as the estimation procedure. Finally, some descriptive statistics are given.

1.4.1 Data

To measure the intensity of collaboration between two regions, I will make use of co-publication data.[5] Collaboration is here approximated by co-publications as in other studies (e.g., Hoekman et al., 2009; Ponds et al., 2007).

I extracted the information on co-publications from the Thomson-Reuters Web of Science database. This database contains information on the papers published in the majority of international scientific journals, with, for each article, a list of all the participating authors along with their institutions.

[5]Publications can be seen as the result of successful collaborations and therefore by definition they do not reflect all collaborations occurring within a given period. Nonetheless, as Dahlander and McFarland (2013, p. 99) put it, in a study that used extensive data from research collaborations at Stanford University, 'published papers afford a visible trail of research collaboration'.

The data were extracted for a time period ranging from 2001 to 2005, and the geographical scale was restricted to five European Union countries (henceforth EU5): Italy, France, Germany, Spain and the United Kingdom, as in Maggioni et al. (2007). In addition, to avoid the problems that can arise when mixing several disciplines, due to researcher behaviour and publishing schemes that may differ between fields, the analysis has been restricted to one specific field, chemistry, for which some characteristics are presented at the end of this subsection.

For each paper, this database reports the authors' institutions in their by-lines. As there is an address assigned to each institution, it is possible to geographically pinpoint each of them. This localization was mainly done using the postcodes available in the addresses, which should be a very reliable determinant of location. More than 85% of the sample could be assigned a location using the postcodes; the remaining 15% were located using an online map service, based on the name of the city and the country.[6] In the end, 99.6% of the sample was located.[7] Once located, each institution was assigned to a NUTS 2 region with respect to their latitude/longitude coordinates. Across all the EU5 countries, there were 132 NUTS 2 regions in which at least one publication in the field of chemistry has been published.

While this study concentrates on inter-regional collaborations, about half of the articles (64,044) were produced within a single NUTS 2 region. Focusing on the distribution of inter-regional collaborations, there were 23,356 articles produced by institutions located exclusively within the EU5. The articles involving at least one non-EU5 institution amounted to 30% of the sample (37,770 articles), with the country contributing most to these non-EU5 collaborations being the United States (with 7,602 papers). To complete the picture, 6,859 articles involved at least two EU5 institutions as well as at least one non-EU5 institution. In the remaining of this study, while all articles are included, only the links within the EU5 regions are retained, meaning that the links to non-EU5 regions are ignored.[8]

To sum up, the database consisted of all articles from chemistry journals of which at least one author was affiliated to an institution based in the EU5, giving a total of 125,170 publications distributed among 132 NUTS 2 regions and over five years. The analysis will consist of determining the level of collaboration between each pair of these 132 NUTS 2 regions.

[6] The online map service used was Google Maps ©.

[7] Despite its simplicity, the accuracy of the localization based only on the name of the city and the country was quite high. Indeed, I located all the addresses using just these two methods: city/country and postcodes. When comparing the two methods, one can see that less than 1.5% of the NUTS 3 codes differed between the two methodologies. This number falls to less than 0.4% when considering the NUTS 2 codes.

[8] This treatment – the deletion of non-EU5 links – affects the network proximity variable (the TENB). The consequences of this treatment are examined later (in Section 1.4.2), where the empirical construction of the TENB is described.

Some characteristics of the field of chemistry. In this study, I focus on the field of chemistry for several reasons. Firstly, I want to model collaborations through the use of publication data. For such an approximation to be robust, the link between the outcome of chemistry research and publications should be high. As Defazio et al. (2009) mention, 'international refereed journals [in chemistry] play an important role in communicating results' meaning most of the scientific activities in chemistry that provide any kind of result, including collaborations, should leave a paper trail. Thus, scientific articles in this field should allow the bulk of collaborations to be tracked down.

Another particularity I was interested in concerns the productivity of the researchers. Indeed, a researcher's production should be high enough so that new publications can be explained by the behaviour of existing researchers, rather than by the actions of newly active ones. Put differently, as the focus here is on modelling new flows of collaboration with respect to past states of the network, these newly created links should emanate from existing researchers. In the sample I use, the median number of publications per researcher is five in the period 2001–2005, which seems high enough to fit this purpose.[9]

Authors affiliated to multiple institutions could constitute a bias in this study, as some papers could be perceived as inter-regional collaborations while actually involving only one author active in several regions. To appraise the extent to which this could be an issue, I randomly selected 100 articles from the sample and looked, by hand, at the multi-affiliation status of each author. It appeared that multi-affiliations are somewhat rare, as only 12% of the papers had a multi-affiliated author. In addition, the cases of multi-affiliation that would alter the specification of this study would be multi-affiliations within the EU5 (where the inter-regional collaborations are to be measured), and this pattern is even more unusual, as only 1% of the papers were affected.

Lastly, most of the inter-regional papers involved researchers from only two regions from the EU5 countries. As Figure 1.2 shows, two-regions papers account for 82% of the sample while three-regions papers represent a share of 15%. This propensity for two-regions collaborations in chemistry is in line with our random matching process hypothesis that considered matches between agents from two regions only.

1.4.2 Variables

Year range of the variables. As the analysis is cross-sectional, I have constructed the explanatory and dependent variables separately, to avoid any simultaneity bias. The period used to construct the explanatory variables is 2001–2003. This three-year span is used in order to collect enough information on collaboration patterns. The period

[9] In order to infer some statistics relating to the number of publications per researcher, I considered only the researchers who had published an article in 2001, and then counted their publications in the range 2001–2005. To ensure the researchers were working in EU5 institutions, I only selected the ones who had at least one article whose institutions were exclusively within the EU5. Finally, the researchers were identified using their surnames and the initials of their first names. Despite the rough identification of the researchers, this methodology provides an insight into the question of researchers' productivity in chemistry.

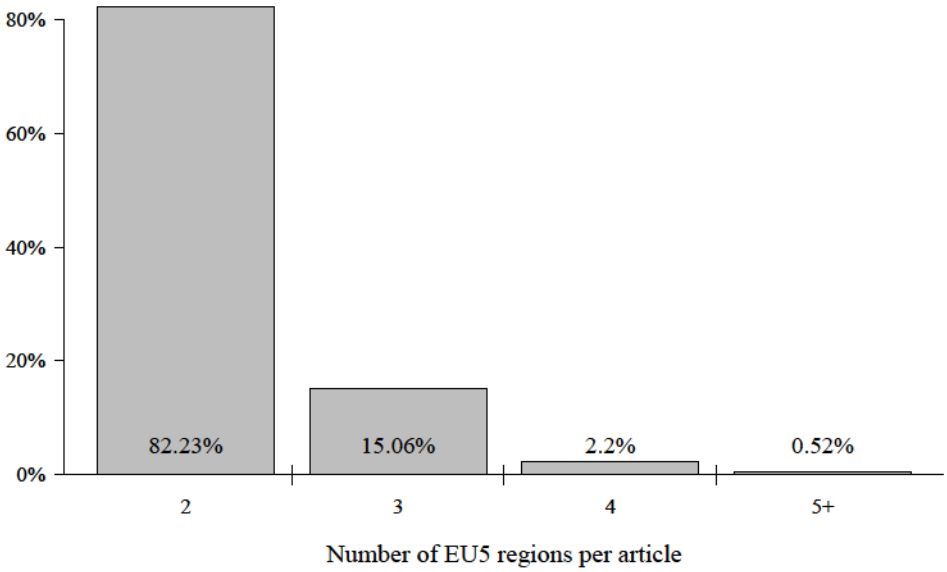

Figure 1.2 – Distribution of the number of regions (from the EU5 countries) per inter-regional article in chemistry.

2004–2005 is then used to build the dependent variable.

Dependent variable. $Copub_{ij}$ is defined as the number of co-publications involving authors from both regions i and j, from the time period 2004–2005. Several methods could have been used to build this variable: most significantly the 'full count' and the 'fractional count' methodologies. The former gives a unitary value for any dyad participating in a publication, while the latter weights each publication by the number of participants, such that the higher the number of participants, the lower the value each dyad receives (for instance if there are n participants, each dyad receives $1/n$). As in other studies (Frenken et al., 2009b; Hoekman et al., 2010), I make use of the full count methodology, since it relates to the idea of participation in knowledge production, rather than net contribution to knowledge production (OST, 2010, p. 541).[10]

Network proximity. The main explanatory variable captures the idea of inter-regional network proximity. Network proximity between two regions is here approximated by the TENB developed in Section 1.3.2 which relates to the number of indirect connections between researchers of different regions. This variable is expected to positively influence future collaborations.

Let $TENB_{ij}$ be the empirical counterpart of the TENB as defined in equation (1.3). As the measure is transposed from the theoretical model to real data, two comments must be made. First, the theoretical model assumes that each collaboration involves only two agents. However, in the data, some articles involve more than two regions from the EU5 countries. To stick to the philosophy of the model, I therefore use only bilateral

[10]Using the fractional count instead of the full count methodology does not alter the results. The results with fractional count are reported on Table 1.5 in appendix.

co-publications, i.e., two-regions articles, to construct $TENB_{ij}$.[11] This in turn implies that $TENB_{ij}$ will be independent from any direct collaborations between regions i and j as it then depends only on the structure of their indirect *bilateral* collaborations. Second, the model uses the number of agents in each region, yet this information is not directly available in the data.[12] As an alternative, I chose the total number of publications of a given region as a way of approximating its number of researchers. Indeed, according to the law of large numbers and for large enough regions, these two values should be proportional. Thus, in the case where the number of researchers is proportional to the number of publications, we would have $Researchers_k = a \times total\_publications_k$ for each region k, with a being the coefficient of proportionality. This approximation circumvents the problem of researchers' identification and will still yield a reliable measure for the TENB, as it should only be proportional to the theoretical value.

Finally, the empirical variable can be defined. Let $bilateral\_copub_{ij}^{2001-2003}$ be the number of bilateral publications (i.e., articles involving agents from only two EU5 regions) between regions i and j, published between 2001 and 2003. In addition, let $total\_publication_k^{2001-2003}$ be the total number of articles produced by researchers in region k. More precisely, it can be defined as the number of articles published between 2001 and 2003 that have at least one author who is affiliated to an institution in region k. The empirical TENB can then be defined as follows:

$$TENB_{ij} = \sum_{k \in N_i \cap N_j} \frac{bilateral\_copub_{ik}^{2001-2003} \times bilateral\_copub_{jk}^{2001-2003}}{total\_publications_k^{2001-2003}}. \tag{1.4}$$

Because of the empirical specification, this variable is best interpreted as a measure of the intensity of network proximity, rather than an exact measure of the number of bridging paths. It is worth noting that the approximation of the number of researchers with the regional mass has no effect on the interpretation of the variable. This is because the interpretation of the coefficients associated with the variable $TENB_{ij}$ in the econometric analysis is done in terms of elasticity, meaning that it is unaffected by a (the coefficient of proportionality).

Furthermore, another advantage of the TENB is that its variation can easily be interpreted. Taking the case of two regions, i and j, from equation (1.4), we can see that an increase of 1% in the number of collaborations between two regions *and their common neighbours* leads to an increase of 2% in the TENB measure.[13] Conversely, an increase

[11] Remember that the links with regions that are not within the EU5 are ignored. Here, bilateral collaboration means articles involving institutions from two – and only two – NUTS 2 regions of the EU5 (regardless of whether there were also institutions from non-EU5 countries involved in the article).

[12] The information provided by Web of Science does not allow the retrieval of the affiliation of researchers. Although each article record does contain all the institutions and researchers who participated in it, researchers and institutions are not matched. Therefore, whenever two or more institutions appear in an article, it is not possible to identify to which institution each researcher belongs.

[13] This footnote shows how to derive this result. Using the notations from Section 1.3.2.1, let g_{ij} represent the collaborations between regions i and j (as the variable $bilateral\_copub_{ij}$), and let n_k be the number of researchers in region k (as the variable $total\_publications_k$). The TENB between regions i

of 1% in the TENB can then be seen as the outcome of a 0.5% increase in collaborations with common neighbours.

Finally, a word on the scope of the measure: the TENB is constructed using information restricted to intra-EU5 collaborations and not accounting for non-EU5 collaborations implies a downward bias on the measure. The TENB reflects the extent to which researchers from two regions share common collaborators in another region. Thus, by deleting non-EU5 links, potential common collaborators from non-EU5 regions are not taken into account: this in turn may lead to a possible underestimation of the TENB for some pairs of regions. So it should be remembered that the inter-regional network proximity measured in this study is somewhat partial, as it stems only from inter-regional collaborations occurring within the EU5 countries.[14]

Other covariates. The variable $GeoDist_{ij}$ was created to capture the impeding effect of geographical distance. It is equal to the 'as the crow flies' distance between the geographic centres (centroids) of the regions, in kilometres. The variable $CountryBorder_{ij}$ is a dummy variable of value '1' when the regions i and j are from different countries and '0' otherwise. To further take into account the notion of geographical proximity, a variable of regional contiguity was created. This variable is aimed at capturing the effects of geography that are not seized by geographical distance alone. The variable $notContig_{ij}$ is then of value '1' when two regions are not contiguous and of value '0' otherwise.

As with any scientific discipline, the field of chemistry is not homogeneous and contains many sub-fields. Thus, two researchers may face difficulty in collaborating if they are from regions specializing in two different sub-fields that differ in various aspects, such as in methodology or research question. (For instance, some regions may specialize in analytical chemistry and others in physical chemistry.) Such differences in sub-field specialization can imply significant differences in terms of collaborative patterns between regional pairs. Consequently, the model includes a cognitive distance variable, which refers to the distance in terms of 'knowledge base and expertise' (Boschma, 2005, p. 63) between the two regions. This variable is intended to account for the distance between the research portfolios of each pair of regions. The sub-fields are identified by the 75 keywords appearing in the chemistry articles of the sample.[15] Let s_{ik} be the share of articles produced by region i containing the keyword k, so that the vector $s_i = (s_{i,1}, \ldots, s_{i,75})$ characterizes region i's research portfolio.[16] The cognitive distance variable is defined

and j is then defined as $TENB_{ij} = \sum_{k \in N_i \cap N_j} (g_{ik} \times g_{jk}/n_k)$. Now assume that there is, *ceteris paribus*, a 1% increase in all collaborations between regions i and j and all their common neighbours, so that $g_{ik}^{new} = 1.01 \times g_{ik}$ and $g_{jk}^{new} = 1.01 \times g_{jk}$ for any common neighbour k. Thus, the new TENB between regions i and j is $TENB_{ij}^{new} = \sum_{k \in N_i \cap N_j} \left(g_{ik}^{new} \times g_{jk}^{new}/n_k\right) = \sum_{k \in N_i \cap N_j} ((1.01 \times g_{ik}) \times (1.01 \times g_{ik})/n_k) \simeq 1.02 \times \sum_{k \in N_i \cap N_j} (g_{ik} \times g_{jk}/n_k) = 1.02 \times TENB_{ij}$.

[14] The extent of this effect is limited by the fact that the bulk of inter-regional collaborations occur internally to the EU5.

[15] A list of all keywords, along with their frequency, is given in Appendix 1.7.5.

[16] Note that, as several keywords can appear in one single article, the sum of the shares, $\sum_k s_{ik}$, may be greater than 1.

as: $CogDist_{ij} = 1 - \text{cor}(s_i, s_j)$, where $\text{cor}(s_i, s_j)$ is the correlation between the research portfolios of regions i and j. This cognitive distance measure is built similarly to the technological distance employed in Jaffe (1986).

Finally, collaborations between researchers from top regions may display different collaborative patterns from the rest of the sample. Presumably, they may display a higher likelihood of collaboration (Hoekman et al., 2009). To control for this, the indicator variable $TopRegions_{ij}$ is included and takes value '1' when both regions i and j are from the top 20 regions in terms of publication (i.e., with respect to the variable $total\_publications_i^{2001-2003}$).

The importance of regional dummies. In the gravity model, regional masses are one essential factor determining the flows of inter-regional interaction. However, the types of regional mass affecting the level of inter-regional collaboration can be numerous. The most obvious one is regional size, as in trade models, here measured in terms of the number of publications. At the same time, relevant masses could also include the number of academic 'stars' in the region, the number of graduate students in chemistry, the quality of research facilities, etc. It is difficult to control for all the relevant regional masses because of their great variety and the limited availability of some of the data. Not properly controlling for them could lead to the model being misspecified as suffering from an omitted variable problem. One convenient way to cope with this problem is to include regional dummies: these dummies would control for any characteristic specific to the region affecting the dependant variable. Consequently, the model includes regional dummies which are able to encompass any kind of regional mass.

1.4.3 Model and estimation procedure

As the dependent variable $Copub_{ij}$ is a count variable, a natural way to estimate equation (1.5) would be via a Poisson regression as in other recent studies (e.g., Agrawal et al., 2014; Belderbos et al., 2014). In the Poisson regression, the dependent variable is assumed to follow a Poisson law whose mean is determined by the explanatory variables. An interesting feature of this estimation is that the conditional variance is equal to the conditional mean. Hence, greater dispersion is allowed as the conditional mean increases, thus hampering potential problems of heteroskedasticity. Furthermore, Santos Silva and Tenreyro (2006) have shown that Poisson regression performs better than other estimation techniques, such as the log-log OLS regression. In particular, they show using simulations, that the estimates obtained in Poisson regressions suffer from less bias than those obtained using other methods.

The structure of the data set, like that of trade models, is dyadic. This means that the statistical unit, i.e., the regions, are both on the left side and on the right side, i.e., can be either the origin or the destination of the flow. When it comes to properly estimating the standard errors of the estimators, this dyadic structure is problematic. Indeed, in most

econometric models, not controlling for the structure of correlation can lead to erroneous standard errors that overstate the precision of the estimators (Cameron and Miller, 2015). As Cameron et al. (2011) demonstrate, by means of a Monte Carlo study, using White's heteroskedasticity-robust covariance matrix may be unreliable, as it can lead to standard errors several times lower than the properly clustered ones. Therefore, in this econometric analysis, the standard errors will be two-way clustered, with respect to the natural clusters of this dataset: the regions of origin and the regions of destination.

Based on the gravity model and on the previously defined variables, the model I will estimate has the following form:

$$\begin{aligned} \mathrm{E}(Copub_{ij}|X_{ij}) &= d_i \times d_j \times (TENB_{ij}+0.01)^{\alpha_1} \times GeoDist_{ij}^{\alpha_2} \\ &\quad \times \exp\left(\alpha_3 notContig_{ij} + \alpha_4 CountryBorder_{ij} + \alpha_5 CogDist_{ij} + \alpha_6 TopRegions_{ij}\right), \end{aligned} \tag{1.5}$$

where X_{ij} represents the set of all explanatory variables, while d_i and d_j are the regional dummies of regions i and j. Note that 0.01 is added to the variable $TENB_{ij}$ as its value may be equal to 0.[17] Furthermore, unlike most gravity models, the masses do not appear as they are specific to each region and therefore absorbed by the regional dummies.

1.4.4 Descriptive statistics

The data set is composed of all the bilateral relations between the 132 NUTS 2 regions, which amounts to 17,292 ($= 132 \times 131$) observations or regional pairs. Table 1.1 shows some descriptive statistics on the data set and the main constructs. Looking at the number of collaborations, one can see that the distribution is uneven, with a coefficient of variation of 3.2. Figure 1.3 depicts the distribution of the collaborations and confirms the skewness of this variable. The maximum of 229 is between the regions Île de France and Rhône-Alpes. The TENB, defined by equation (1.4), is also unevenly distributed, but less so than the number of co-publications, with a coefficient of variation of 2.3. Its maximum value, 28.4, is also obtained between the French regions of Île-de-France and Rhône-Alpes. When considering international dyads only, the maximum is for Cataluña and Île-de-France, with an expected number of bridging paths of 11.5. Table 1.2 shows the correlations among the explanatory variables. The highest correlation is between the geographical distance and national border variables.

1.5 Results

The results are reported in Table 1.3. First, I will focus on model (1), the gravity model which includes all variables but that of network proximity. Consistent with the previous literature (e.g., Hoekman et al., 2009, 2010; Scherngell and Barber, 2009), geography

[17] A low value, 0.01, is added to allow the interpretation in terms of elasticity to hold (as in Fleming et al., 2007). Adding other values imply no qualitative change in the results.

Table 1.1 – Descriptive statistics of the main variables.

	Min	Median	75^{th} percentile	Max	Mean	SD
Co-publications	0	0	1	229	2.21	7.01
Total Publications	1	521.0	909.2	5560	713.35	751.4
TENB	0	0.13	0.45	28.42	0.49	1.13
Geographical Distance	1.09	868.1	1213.3	2595.5	894.4	476.9
Non-Contiguity	0	1	1	1	0.97	0.18
Different Country	0	1	1	1	0.78	0.41
Cognitive Distance	0.01	0.16	0.29	1.06	0.23	0.20
Top 20 Regions	0	0	0	1	0.02	0.15

Notes: Co-publications are based on the period 2004–2005 while all other variables are computed using the period 2001–2003.

Table 1.2 – Correlation matrix of the covariates.

		1	2	3	4	5	6
1	TENB (ln)	1.00					
2	Geographical Distance (ln)	-0.33*	1.00				
3	Non-Contiguity	-0.15*	0.49*	1.00			
4	Different Country	-0.39*	0.73*	0.32*	1.00		
5	Cognitive Distance	-0.59*	0.07*	0.04*	0.05*	1.00	
6	Top 20 Regions	0.26*	-0.00	0.00	0.02	-0.11*	1.00

*: statistically significant at the 1% level (Pearson correlation).

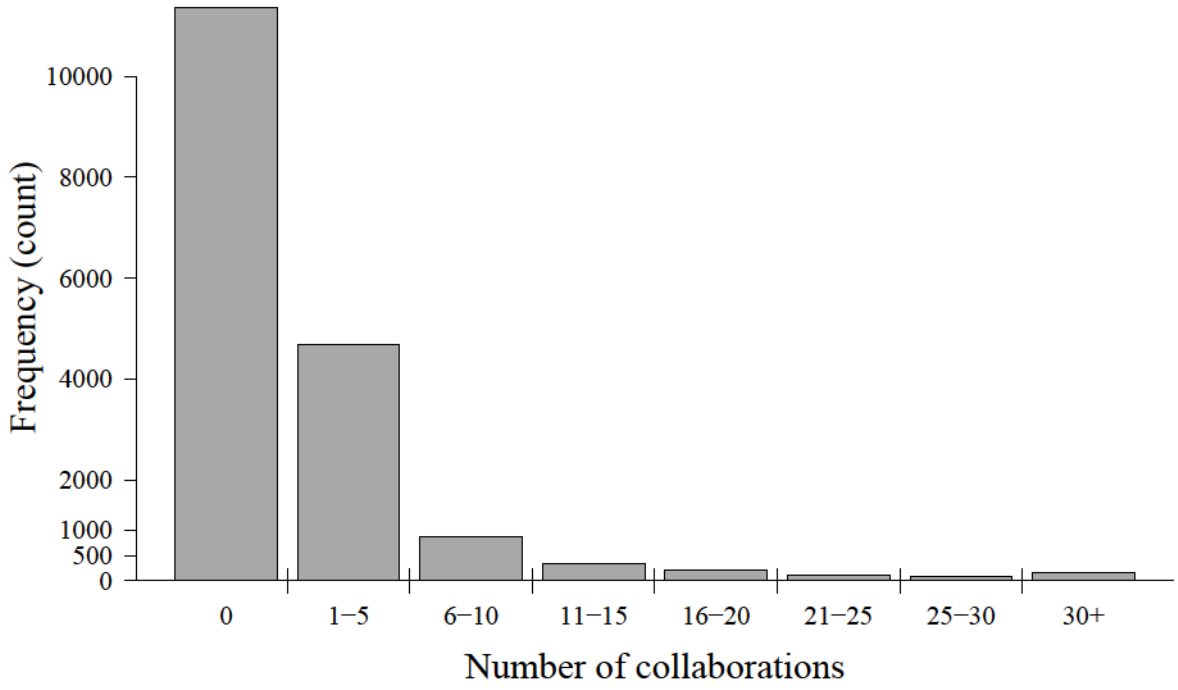

Figure 1.3 – Distribution of EU5 inter-regional collaborations in chemistry for the period 2004-2005.

greatly affects collaboration. The most impeding effect is the national border effect. All else being equal, if two regions are from different countries, their collaboration flows will suffer a decrease of 83% $(1-\exp(-1.801))$. Although the effect of national borders is very strong, the order of magnitude is in line with other estimates in the literature (e.g., Maggioni et al., 2007; Hoekman et al., 2009). Geographical distance is also a hindrance to collaboration: with an elasticity of -0.35, the estimates show that increasing the distance between two regions by 1% decreases their level of collaboration by 0.35%. Seen with a larger variation, when the geographic distance doubles, collaboration decreases by 22% $(1-2^{-0.35})$. Turning to the contiguity effect, as with other distances, it has a non-negligible effect on collaborations: being non-contiguous rather than contiguous reduces the expected number of collaborations by 17%. The cognitive distance exerts a significant negative effect with an estimated elasticity of -1%, meaning regions with different research portfolios will be less likely to collaborate. Finally, contrary to the results on co-publishing in the study of Hoekman et al. (2009), researchers belonging to the top 20 regions do not engage in more collaborations. This may be due to the fact that this model takes better account of the regional masses, thanks to the use of regional dummies.

Now I will turn to the analysis of the results provided by models (2) to (4), where the variable TENB (approximating network proximity) is introduced, along with its interaction with geographical distance. In model (2), only the TENB is introduced in the regression. Its estimated coefficient is 0.244, positive and significant, meaning a 10% increase in the TENB would imply a 2.4% increase in collaboration.[18] This result shows that network proximity does seem to influence network formation in general. However, this positive effect may not be homogeneous and could be mediated by geography.

To test whether network proximity interacts with geography, the interaction with the geographical distance is introduced in models (3) and (4), respectively in a simple and a quadratic form. In these models, the elasticity of the TENB depends on the distance separating the regions. The results of model (3) depict significant estimates for both network proximity and its interaction with geographical distance, with a positive sign for the interaction. Model (4) shows that the coefficient of the interaction with the squared logarithm of the distance is negative. These estimates would seem to imply that the effect of network proximity increases with distance, and possibly decreases after a certain threshold. However, those coefficients cannot be straightforwardly interpreted because they do not represent the total effect of the interaction (see Brambor et al., 2006). The interpretation is helped by Figure 1.4, which represents the estimated elasticity of network proximity with respect to the distance, along with its 95% confidence interval. While network proximity can have a negative impact on co-publications for regions located close to each other, its benefits grow with distance, favouring the most distant regions. As the figure shows, despite a negative coefficient for the quadratic term, the elasticity of the

[18] From Section 1.4.2, a 10% increase in the TENB between two regions can be implied by a 5% increase in collaboration flows between these two regions and their common neighbours.

Table 1.3 – Results of the Poisson regression.

Model:	(1)	(2)	(3)	(4)
Dependent variable:	Co-publications	Co-publications	Co-publications	Co-publications
TENB (ln) [proxy for network proximity]		0.244***	-0.4878***	-1.1098***
		(0.049)	(0.1152)	(0.2992)
TENB (ln) * Geo. Distance (ln)			0.1073***	0.3215***
			(0.015)	(0.0926)
TENB (ln) * Squared Geo. Distance (ln)				-0.0182**
				(0.0076)
Geographical Distance (ln)	-0.3486***	-0.3325***	-0.3778***	-0.4084***
	(0.0376)	(0.0299)	(0.0294)	(0.0301)
$\mathbb{1}_{\{Not\ Contiguous\}}$	-0.1854***	-0.1981***	-0.2474***	-0.2357***
	(0.0532)	(0.0483)	(0.0438)	(0.0413)
$\mathbb{1}_{\{DifferentCountries\}}$	-1.8007***	-1.415***	-1.4949***	-1.4664***
	(0.0584)	(0.1064)	(0.1024)	(0.1027)
Cognitive Distance	-1.0331***	-1.0612***	-1.0127***	-1.0278***
	(0.2536)	(0.2364)	(0.2368)	(0.2388)
Top 20 Regions	0.0714	0.0812*	0.0671	0.0624
	(0.0446)	(0.0465)	(0.045)	(0.046)
Regional dummies (Origin & Destination)	yes	yes	yes	yes
Number of Observations	17292	17292	17292	17292
Adj-Pseudo R^2	0.7115	0.7127	0.7148	0.7149
BIC	45 855.164	45 684.125	45 372.785	45 368.483

Notes: The model estimated is depicted by equation (1.5). The dependent variable is the number of co-publications between pairs of NUTS 2 regions for the period 2004-2005. The explanatory variables are built on 2001-2003. The function $\mathbb{1}_{\{.\}}$ is the indicator function and is used to represent the variables *notContig* and *CountryBorder* defined in Section 1.4.2. The variable TENB approximates network proximity and is defined as a measure of the strength of indirect connections between regions (see Section 1.3.2). Two-way clustered standard errors in parenthesis (see Cameron et al., 2011). Level of statistical significance: * 10%, ** 5%, *** 1%.

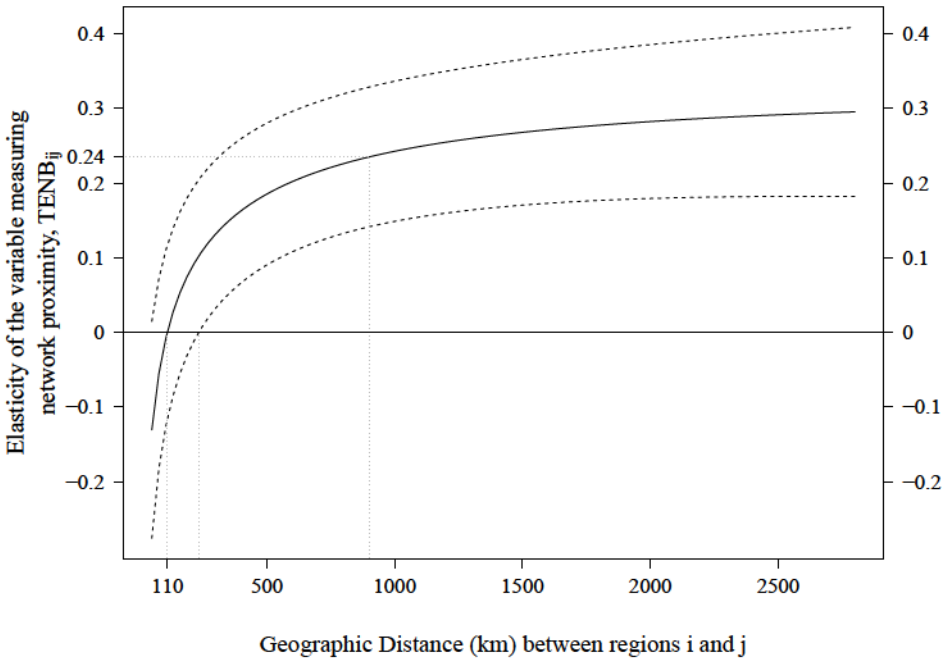

Figure 1.4 – Graph of the interaction between network proximity and geographical distance.
Notes: The graph represents the estimated elasticity of the TENB on co-publications with respect to geographical distance (solid line) along with its 95% confidence interval (dashed lines). This graph is based on the estimates from model (4) of Table 1.3.

TENB strictly increases for distances in the range of those in the sample. The estimates indicate that the effect is even negative for regions located at distances of below 110 km, while the elasticity of the TENB is positive for regions further apart. For instance, the effect starts to be significantly positive at the 5% level for regions at a distance of 233 km. For regions separated by the median distance, 900 km, the elasticity is 0.24, meaning that a 10% increase in the TENB would lead to an increase in co-publications of 2.4%. This result is in line with the hypothesis of substitutability between network proximity and geographical proximity.

As geographical distance per se does not seize all characteristics induced by geography, I will now decompose the effects of the TENB with respect to the national border dummy and the contiguity dummy. The first dummy captures whether regions located in different countries benefit more from network proximity, along with the substitution hypothesis. In addition, in the case of substitution, the effect of network proximity should be greater for non-contiguous regions. The results of these regressions are reported in Table 1.4.

Model (5) considers the sole decomposition with respect to national borders: it shows that network proximity influences international collaborations with an elasticity of 0.23 (significant at the 0.001 level), but does not seem to influence national ones as the coefficient is not statistically significant. Adding the interaction with contiguity yields a more complete picture of the interactions, particularly at the intra-national level. Model (6) reveals that the effect of network proximity on collaborations strictly increases with the loss of other forms of proximity: all else being equal, the elasticity of the TENB is higher when two regions are from different countries instead of from the same country, and when

Table 1.4 – Results of the Poisson regression in which the TENB is interacted with national borders and contiguity.

Model:	(5)	(6)
Dependent variable:	Co-publications	Co-publications
TENB (ln) $* \mathbb{1}_{\{SameCountry\}}$	0.0731	
	(0.0488)	
TENB (ln) $* \mathbb{1}_{\{DifferentCountries\}}$	0.239***	
	(0.0413)	
TENB (ln) $* \mathbb{1}_{\{SameCountry\}} * \mathbb{1}_{\{Contiguous\}}$		-0.0749
		(0.0532)
TENB (ln) $* \mathbb{1}_{\{SameCountry\}} * \mathbb{1}_{\{Not\ Contiguous\}}$		0.1315***
		(0.0443)
TENB (ln) $* \mathbb{1}_{\{DifferentCountries\}} * \mathbb{1}_{\{Contiguous\}}$		0.0456
		(0.1412)
TENB (ln) $* \mathbb{1}_{\{DifferentCountries\}} * \mathbb{1}_{\{Not\ Contiguous\}}$		0.2504***
		(0.0417)
Geographical Distance (ln)	-0.3191***	-0.3035***
	(0.029)	(0.0285)
$\mathbb{1}_{\{Not\ Contiguous\}}$	-0.2115***	-0.4193***
	(0.0461)	(0.049)
$\mathbb{1}_{\{DifferentCountries\}}$	-1.6324***	-1.5885***
	(0.0905)	(0.0889)
Cognitive Distance	-1.089***	-1.0124***
	(0.2389)	(0.2394)
Top 20 Regions	0.0465	0.0369
	(0.0448)	(0.0347)
Regional dummies (Origin & Destination)	yes	yes
Number of Observations	17292	17292
Adj-Pseudo-R^2	0.71438	0.71581
BIC	45447.917	45246.692

Notes: The dependent variable is the number of co-publications between pairs of NUTS 2 regions for the period 2004-2005. The explanatory variables are built on 2001-2003. The function $\mathbb{1}_{\{.\}}$ is the indicator function and is used to represent the variables *notContig* and *CountryBorder* defined in Section 1.4.2. The variable TENB approximates network proximity and is defined as a measure of the strength of indirect connections between regions (see Section 1.3.2). Two-way clustered standard errors in parenthesis (see e.g. Cameron et al., 2011). Level of statistical significance: * 10%, ** 5%, *** 1%.

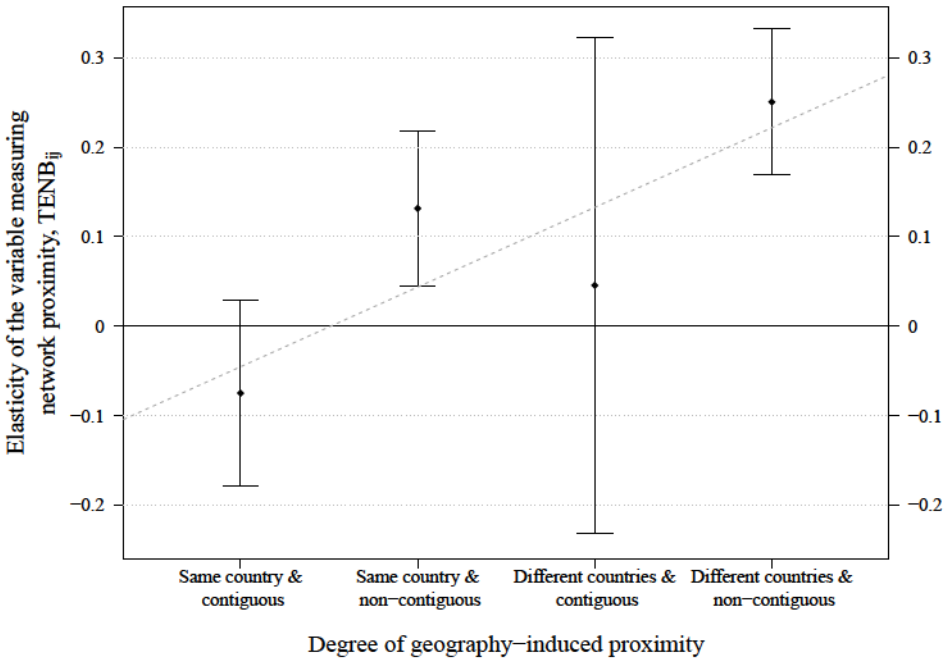

Figure 1.5 – Graph of the link between network proximity and geography-induced proximity.
Notes: The graph reports the elasticity of the TENB on co-publications with respect to different degrees of geography-induced proximity. Both the estimates of the elasticities, as well as their 95% confidence intervals, are represented. This graph is based on the estimates from model (6) of Table 1.4. The linear fit of the estimates represented by the dashed line, depicting an increase of the elasticity with the loss of proximity, is only for visual purpose.

they are non-contiguous instead of contiguous.[19] Figure 1.5 represents these estimates with their 95% confidence intervals. For the most favourable case – that is, when two regions are from the same country and are contiguous – the estimated elasticity is negative (-0.07) but not statistically different from 0. When the two regions lose the benefits of contiguity, the elasticity of the TENB becomes positive, rising to 0.13, while becoming significant at the 1% level. For contiguous regions from different countries the estimated coefficient is low, 0.04, with a large standard error. However, the poor precision of this estimator is possibly due to the very small number of regional pairs in this category (only 30). Finally, in the case of least geographically-induced proximity, namely when two regions are from different countries and are not contiguous either, the benefits induced by network proximity are the highest, with an estimated elasticity of 0.25. These results confirm Hypothesis 2.b, predicting substitutability.

Hence, the main conclusions that can be drawn from the results are twofold. First, the estimates show that network proximity does not have an overall homogeneous effect, but rather acts as a substitute to geographic proximity: the effect of network proximity becomes stronger with distance, whether this be pure geographic distance or another form of geographical distance (namely national borders and non-contiguity). This fact validates Hypothesis 2.b, predicting substitutability. Second, for the regional pairs that benefit most from the forms of proximity induced by geography, the effect is non-significant:

[19] All coefficients of model (6) are significantly different from each other with respect to the t-test.

network proximity is not always beneficial, so Hypothesis 1 is only partially validated. Finally, the TENB used here is a measure of network proximity that is rather conservative, as it neglects direct linkages and is based only on intra-EU5 collaborations: consequently, the effects found in this chapter regarding network proximity are likely to be a lower bound.

1.6 Conclusion

This chapter has investigated the role of networks in the formation of inter-regional research collaborations, as well as its interplay with geography. To this end, a new measure of network proximity was introduced and an empirical study was carried out using a gravity framework.

The first step was to create a measure of network proximity at the inter-regional level. Such a measure, referred to as the TENB, was proposed in Section 1.3.2. This measure fits the gravity framework well as it is independent from direct linkage (preventing any endogeneity issue), and is defined for each dyad of regions. Furthermore, the strength of this measure is that it can be interpreted, under mild conditions, as the expected number of bridging paths between two regions (a bridging path being an inter-regional indirect connection at the micro level).

Next, I empirically assessed the influence of network proximity on network formation using data on co-publications from 132 NUTS 2 regions in the field of chemistry. To that purpose, the TENB variable was embedded within a gravity model estimated using Poisson regressions. Consistent with the existing literature, I found a significant, negative effect of separation variables, such as geographical distance and national borders. The cognitive distance was also found to have a significant hampering effect on collaboration.

Notably, a clear substitutability pattern with geography was revealed: the strength of network proximity rises when the benefits induced by geographic proximity wane. This suggests that network proximity alleviates the impeding effects of distance. In particular, this result underscores the importance of network-related effects in international collaborations. This fact bears great significance in the context of policy making. Indeed, an important characteristic of long-distance collaborations, such as international ones, is that they provide a higher quality of research production (see, e.g., Narin et al., 1991; Adams et al., 2005; Adams, 2013). From this viewpoint, the EU policies aiming at fostering international collaborations could have a sustained positive effect on knowledge production and ease future knowledge flows. As new international connections arise, the network proximity of regions located in different countries increases.[20] This in turn may trigger new international collaborations as a result of network effects, implying that more

[20] Consider two regions in different countries: i and j. If these two were to have a new collaboration, new indirect connections (measured with the TENB) would consequently arise between i and all regions connected to j from j's country, and vice versa. Thus, new international collaborations do indeed increase the network proximity between regions in the two countries.

distant/more yielding collaborations are more likely to be established.

This study has focused on the scientific field of chemistry and has been geographically circumscribed to the EU5, two elements that limit its scope. Thus, natural extensions include the application to other fields of science, to assess whether they display the same pattern of substitutability between geography and the network. Extensions to other geographical areas could also be valuable. In particular, a comparison with US data may be worthwhile to better understand the interplay between network proximity and geographical distance: as there should be no country-border effect for intra-US collaborations, do distance and network proximity still interact there? It might also be interesting to test whether the network-creation force of indirect connections has evolved over time. This dynamic analysis could shed some light on the question of whether the improvement of communication techniques has enforced the 'network proximity' channel for the creation of new links.

1.7 Appendix

1.7.1 Proof of proposition 1

Let L_{ik}^{a} represent the a^{th} link, $a \in \{1,\dots,g_{ik}\}$, between agents from regions i and k, and L_{jk}^{b} to be the b^{th} link, $b \in \left\{1,\dots,g_{jk}\right\}$, between agents from regions j and k. By definition, the pair of links (L_{ik}^{a}, L_{jk}^{b}) forms a bridging path if and only if they are both connected to the same agent in region k (as depicted by figure 1.1). Let the Greek letter ι, $\iota \in \{1,\dots,n_k\}$, designate agent ι from region k. Hence, from the random matching process, we know that the probability that agent ι is connected to any incoming link is $p_\iota = 1/n_k$. Thus, the probability that agent ι is connected to both links L_{ik}^{a} and L_{jk}^{b} is $p_\iota^2 = 1/n_k^2$. Therefore, the pair (L_{ik}^{a}, L_{jk}^{b}) is a bridging path with probability $p = \sum_{\iota=1}^{n_k} p_\iota^2 = 1/n_k$ (summing over all the agents of region k, because each agent can be connected to both links). Let X_{ab} be the binary random variable relating the event that the pair of links (L_{ik}^{a}, L_{jk}^{b}) is a bridging path. This random variable has value 1 with probability p and 0 otherwise, so that its mean is $E(X_{ab}) = p$. The random variable giving the number of bridging paths between regions i and j via region k is then the sum of all variables X_{ab}, a and b ranging over $\{1,\dots,g_{ik}\}$ and $\left\{1,\dots,g_{jk}\right\}$, that is ranging over all possible bridging paths. It follows that the expected number of bridging paths is $ENB_{ij}^{k} = E(\sum_{a=1}^{g_{ik}} \sum_{b=1}^{g_{jk}} X_{ab})$. From the property of the mean operator, it can be rewritten as: $ENB_{ij}^{k} = \sum_{a=1}^{g_{ik}} \sum_{b=1}^{g_{jk}} E(X_{ab}) = \sum_{a=1}^{g_{ik}} \sum_{b=1}^{g_{jk}} p = \sum_{a=1}^{g_{ik}} \sum_{b=1}^{g_{jk}} (1/n_k) = (g_{ik} g_{jk})/n_k$. □

1.7.2 Preferential attachment

In this section I consider the matching mechanism described in section 1.3.2.3. This is a simple matching mechanism where the probability that agents get a new link is based on their productivity level that is exogenous. Consider a region with n agents, all sorted

with respect to their productivity level, then the probability that agent ι connects to an incoming link is $p_\iota = \iota^{-0.5}/\Gamma$ with $\Gamma = \sum_{\iota=1}^{n} \iota^{-0.5}$. In this appendix, I investigate: 1) the distribution of the expected degree of each agent and 2) the derivation of the expected number of bridging paths based on this matching mechanism.

Of course the following analysis can be extended to the case where the probability of connection is more generally defined as: $\iota^{-\alpha}/\Gamma(\alpha)$ with $\Gamma(\alpha) = \sum_{\iota=1}^{n} \iota^{-\alpha}$. I focus on the case $\alpha = 0.5$ as the expected degree distribution corresponds to a power law of parameter $\gamma = 3$ as in Barabási and Albert (1999), which is proven in next section.

1.7.2.1 The expected distribution of the matching mechanism follows a power law

In order to understand what law follows the expected distribution of links along this matching mechanism, I will derive the cumulative distribution function. Say that there are L incoming links, then the expected degree of any agent is simply its probability to get a link times the number of links L. The expected degree of agent ι is then $\left(\iota^{-0.5}/\Gamma\right) \times L$. To get the cumulative distribution function of the expected degree, $F(\mathtt{k}) = P(x < \mathtt{k})$, one has to count the number of agents whose degree is inferior to $\mathtt{k}$, i.e. $\#\left\{\iota |\ \left(\iota^{-0.5}/\Gamma\right) \times L < \mathtt{k}\right\}$. As agents are sorted with respect to their productivity level, one has simply to find out the label ι such that $\left(\iota^{-0.5}/\Gamma\right) \times L = \mathtt{k}$. Indeed, agents having a degree inferior to $\mathtt{k}$ should respect the following condition:

$$\begin{aligned} \left(\iota^{-0.5}/\Gamma\right) \times L &< \mathtt{k} \\ \iota^{-0.5} &< \frac{\mathtt{k}\Gamma}{L} \\ \iota &> \left(\frac{L}{\mathtt{k}\Gamma}\right)^2. \end{aligned} \tag{1.6}$$

Let $\iota(\mathtt{k}) = (L/\Gamma)^2 \mathtt{k}^{-2}$, then the number of agents having a degree inferior to $\mathtt{k}$ is equal to $n - \iota(\mathtt{k})$ as agents such that $\iota \leq \iota(\mathtt{k})$ do not respect the inequality defined by equation (1.6). The share of agents having a degree lesser than $\mathtt{k}$ is then:[21]

$$\begin{aligned} F(\mathtt{k}) &= \frac{1}{n}\left(n - \iota(\mathtt{k})\right) \\ &= 1 - \frac{1}{n}\left(\frac{L}{\Gamma}\right)^2 \mathtt{k}^{-2}. \end{aligned} \tag{1.7}$$

[21] More precisely, the value of the swinging agent is $\iota(\mathtt{k}) = \left\lfloor (L/\Gamma)^2 \mathtt{k}^{-2} \right\rfloor$ where $\lfloor x \rfloor$ is the largest integer not greater than x. The number of agents with a degree inferior to $\mathtt{k}$ is not exactly $n - \iota(\mathtt{k})$, rather, as this number cannot be negative, its value is $\max(n - \iota(\mathtt{k}), 0)$. Now let $\mathtt{k}^*$ be such that $\iota(\mathtt{k}^*) = n$, then it follows that for each $\mathtt{k} < \mathtt{k}^*$ the cumulative is $P(x < \mathtt{k} |\ k < k^*) = 0$. The cumulative distribution function defined by equation (1.7) is defined only for $\mathtt{k} \geq \mathtt{k}^*$ and is 0 otherwise. All these details were skipped for readability.

From the cumulative distribution, one can then derive the distribution by differentiating with respect to $\mathbf{k}$, which yields:

$$f(\mathbf{k}) = \frac{2}{n}\left(\frac{L}{\Gamma}\right)^2 \mathbf{k}^{-3}.$$

This result shows that from a simple connection mechanism based on exogenous probabilities, the expected distribution of links follows a power law of parameter $\gamma = 3$.

A bit of generalization. In the same vein as previously, if one considers that the probability of connection is defined by $\iota^{-\alpha}/\Gamma(\alpha)$ with $\Gamma(\alpha) = \sum_{\iota=1}^{n} \iota^{-\alpha}$ and $\alpha > 0$, the distribution of the expected degree of the nodes is then:

$$f(\mathbf{k}) = \frac{1}{\alpha n}\left(\frac{L}{\Gamma(\alpha)}\right)^{\frac{1}{\alpha}} \mathbf{k}^{-\frac{1+\alpha}{\alpha}}.$$

Expressing the probabilities of connection with respect to the power law parameter, $\gamma = \frac{1+\alpha}{\alpha}$, yields: $\iota^{-\frac{1}{\gamma-1}}/\Gamma_\gamma(\gamma)$ with $\Gamma_\gamma(\gamma) = \sum_{\iota=1}^{n} \iota^{-\frac{1}{\gamma-1}}$; and the distribution function is then:

$$f(\mathbf{k}) = \frac{\gamma-1}{n}\left(\frac{L}{\Gamma_\gamma(\gamma)}\right)^{\gamma-1} \mathbf{k}^{-\gamma}.$$

The distribution of the degrees follows a power law of parameter γ.

1.7.2.2 The derivation of the expected number of bridging paths with preferential attachment

This section strives to derive the expected number of bridging paths between regions from the matching mechanism with preferential attachment. The derivation of the result is based upon a variation of the proof of proposition 1 of section 1.3.2.2. Consider a region k with n_k agents. The number of links between k and regions i and j are g_{ik} and g_{jk} respectively.

Let L_{ik}^a be the a^{th} link, $a \in \{1,\ldots,g_{ik}\}$, between agents from regions i and k, and L_{jk}^b to be the b^{th} link, $b \in \left\{1,\ldots,g_{jk}\right\}$, between agents from regions j and k. By definition, the pair of links (L_{ik}^a, L_{jk}^b) forms a bridging path if and only if they are both connected to the same agent in region k. Let the Greek letter ι designate the agent ι from region k. Hence, the probability that L_{ik}^a and L_{jk}^b are both connected to agent ι is $p_\iota^2 = \left(\iota^{-0.5}/\Gamma\right)^2$. Then the pair (L_{ik}^a, L_{jk}^b) is a bridging path with probability $p = \sum_{\iota=1}^{n_k} p_\iota^2$. Let X_{ab} be the binary random variable relating whether the pair (L_{ik}^a, L_{jk}^b) is a bridging path. It takes value 1 with probability p and value 0 otherwise, so that its mean is $E(X_{ab}) = p$. The random variable giving the number of bridging paths is the sum of all variables X_{ab}, a and b ranging over $\{1,\ldots,g_{ik}\}$ and $\left\{1,\ldots,g_{jk}\right\}$, that is ranging over all possible bridging paths. Then, the expected number of bridging paths is $ENB_{ij}^{k,Pref} = E(\sum_{a=1}^{g_{ik}} \sum_{b=1}^{g_{jk}} X_{ab})$.

From the property of the mean, it can be rewritten as:

$$\begin{aligned} ENB_{ij}^{k,Pref} &= \sum_{a=1}^{g_{ik}} \sum_{b=1}^{g_{jk}} E(X_{ab}) \\ &= g_{ik} g_{jk} \times p. \end{aligned}$$

Now, let us rewrite p, the probability for a pair of links to be a bridging path:

$$\begin{aligned} p &= \sum_{\mathtt{i}=1}^{n_k} p_\iota^2 \\ &= \frac{1}{\Gamma^2} \sum_{\iota=1}^{n_k} \frac{1}{\iota}. \end{aligned}$$

Further, notice that $\Gamma = \sum_{\iota=1}^{n_k} \iota^{-0.5} \simeq \int_1^{n_k} x^{-0.5} \mathrm{d}x = 2 \times (\sqrt{n_k} - 1)$, and that $\sum_{\iota=1}^{n_k} \iota^{-1} \simeq \int_1^{n_k} x^{-1} \mathrm{d}x = \log(n_k)$. Therefore p can be rewritten as:

$$\begin{aligned} p &\simeq \frac{1}{4} \frac{\log(n_k)}{\left(\sqrt{n_k} - 1\right)^2} \\ &\simeq \frac{1}{4} \frac{\log(n_k)}{n_k}, \end{aligned}$$

providing n_k is sufficiently high. From this it follows that the expected number of bridging paths with preferential attachment is approximately equal to:

$$\begin{aligned} ENB_{ij}^{k,Pref} &\simeq \frac{g_{ik} g_{jk}}{n_k} \times \frac{\log(n_k)}{4} \\ &\simeq ENB_{ij}^{k} \times \frac{\log(n_k)}{4}. \end{aligned}$$

which ends the proof of proposition 2. □

1.7.3 Estimation with fractional counting

See Table 1.5.

1.7.4 List of the 132 NUTS 2 regions used in the statistical analysis

CODE	NAME	CODE	NAME	CODE	NAME
DE11	Stuttgart	ES24	Aragón	ITH2	Provincia Autonoma di Trento
DE12	Karlsruhe	ES30	Comunidad de Madrid	ITH3	Veneto
DE13	Freiburg	ES41	Castilla y León	ITH4	Friuli-Venezia Giulia
DE14	Tübingen	ES42	Castilla-La Mancha	ITH5	Emilia-Romagna
DE21	Oberbayern	ES43	Extremadura	ITI1	Toscana
DE22	Niederbayern	ES51	Cataluña	ITI2	Umbria
DE23	Oberpfalz	ES52	Comunidad Valenciana	ITI3	Marche

CODE	NAME	CODE	NAME	CODE	NAME
DE24	Oberfranken	ES53	Illes Balears	ITI4	Lazio
DE25	Mittelfranken	ES61	Andalucía	UKC1	Tees Valley and Durham
DE26	Unterfranken	ES62	Región de Murcia	UKC2	Northumberland and Tyne and Wear
DE27	Schwaben	FR10	Île de France	UKD1	Cumbria
DE30	Berlin	FR21	Champagne-Ardenne	UKD3	Greater Manchester
DE40	Brandenburg	FR22	Picardie	UKD4	Lancashire
DE50	Bremen	FR23	Haute-Normandie	UKD6	Cheshire
DE60	Hamburg	FR24	Centre	UKD7	Merseyside
DE71	Darmstadt	FR25	Basse-Normandie	UKE1	East Yorkshire and Northern Lincolnshire
DE72	Gießen	FR26	Bourgogne	UKE2	North Yorkshire
DE73	Kassel	FR30	Nord - Pas-de-Calais	UKE3	South Yorkshire
DE80	Mecklenburg-Vorpommern	FR41	Lorraine	UKE4	West Yorkshire
DE91	Braunschweig	FR42	Alsace	UKF1	Derbyshire and Nottinghamshire
DE92	Hannover	FR43	Franche-Comté	UKF2	Leicestershire, Rutland and Northamptonshire
DE93	Lüneburg	FR51	Pays de la Loire	UKF3	Lincolnshire
DE94	Weser-Ems	FR52	Bretagne	UKG1	Herefordshire, Worcestershire and Warwickshire
DEA1	Düsseldorf	FR53	Poitou-Charentes	UKG2	Shropshire and Staffordshire
DEA2	Köln	FR61	Aquitaine	UKG3	West Midlands
DEA3	Münster	FR62	Midi-Pyrénées	UKH1	East Anglia
DEA4	Detmold	FR63	Limousin	UKH2	Bedfordshire and Hertfordshire
DEA5	Arnsberg	FR71	Rhône-Alpes	UKH3	Essex
DEB1	Koblenz	FR72	Auvergne	UKI1	Inner London
DEB2	Trier	FR81	Languedoc-Roussillon	UKI2	Outer London
DEB3	Rheinhessen-Pfalz	FR82	Provence-Alpes-Côte d'Azur	UKJ1	Berkshire, Buckinghamshire and Oxfordshire
DEC0	Saarland	FR83	Corse	UKJ2	Surrey, East and West Sussex
DED2	Dresden	ITC1	Piemonte	UKJ3	Hampshire and Isle of Wight
DED4	Chemnitz	ITC3	Liguria	UKJ4	Kent
DED5	Leipzig	ITC4	Lombardia	UKK1	Gloucestershire, Wiltshire and Bristol/Bath area
DEE0	Sachsen-Anhalt	ITF1	Abruzzo	UKK2	Dorset and Somerset
DEF0	Schleswig-Holstein	ITF2	Molise	UKK3	Cornwall and Isles of Scilly
DEG0	Thüringen	ITF3	Campania	UKK4	Devon
ES11	Galicia	ITF4	Puglia	UKL1	West Wales and The Valleys
ES12	Principado de Asturias	ITF5	Basilicata	UKL2	East Wales
ES13	Cantabria	ITF6	Calabria	UKM2	Eastern Scotland
ES21	País Vasco	ITG1	Sicilia	UKM3	South Western Scotland
ES22	Comunidad Foral de Navarra	ITG2	Sardegna	UKM5	North Eastern Scotland
ES23	La Rioja	ITH1	Provincia Autonoma di Bolzano/Bozen	UKM6	Highlands and Islands

1.7.5 List of the keywords used to assess cognitive proximity

The table lists the keywords appearing in the chemistry papers published between 2001 and 2003 as well as their frequency (example of reading: there has been 11,114 papers categorized as 'chemistry, inorganic & nuclear').

Table 1.5 – Poisson regression. The dependent variable is built using the 'fractional count' methodology.

Variables	Fractional co-publications					
TENB (ln)		0.226***	-0.5457***	-1.3187***		
		(0.0519)	(0.1235)	(0.3334)		
TENB (ln) * Geo. Distance (ln)			0.1134***	0.3798***		
			(0.016)	(0.1013)		
TENB (ln) * Squared Geo. Distance (ln)				-0.0226***		
				(0.0083)		
TENB (ln) * $\mathbb{1}_{\{Same\ Country\}}$					0.047	
					(0.0524)	
TENB (ln) * $\mathbb{1}_{\{Different\ Countries\}}$					0.2218***	
					(0.0436)	
TENB (ln) * $\mathbb{1}_{\{Same\ Country\}} * \mathbb{1}_{\{Contiguous\}}$						-0.1132**
						(0.0568)
TENB (ln) * $\mathbb{1}_{\{Same\ Country\}} * \mathbb{1}_{\{Not\ Contiguous\}}$						0.1129**
						(0.0478)
TENB (ln) * $\mathbb{1}_{\{Different\ Countries\}} * \mathbb{1}_{\{Contiguous\}}$						0.0036
						(0.1448)
TENB (ln) * $\mathbb{1}_{\{Different\ Countries\}} * \mathbb{1}_{\{Not\ Contiguous\}}$						0.2353***
						(0.0444)
Geographical Distance (ln)	-0.3751***	-0.3605***	-0.4099***	-0.4479***	-0.3468***	-0.3296***
	(0.0379)	(0.0304)	(0.0298)	(0.0299)	(0.0297)	(0.0288)
$\mathbb{1}_{\{Not\ Contiguous\}}$	-0.1882***	-0.1994***	-0.2491***	-0.2343***	-0.2125***	-0.4397***
	(0.0559)	(0.0512)	(0.0478)	(0.0447)	(0.0493)	(0.0511)
$\mathbb{1}_{\{Different\ Countries\}}$	-1.8434***	-1.4875***	-1.5695***	-1.534***	-1.7161***	-1.6655***
	(0.0602)	(0.1112)	(0.1072)	(0.1066)	(0.0953)	(0.094)
Cognitive Distance	-0.9263***	-0.9565***	-0.899***	-0.9199***	-0.9804***	-0.8932***
	(0.2538)	(0.2355)	(0.2371)	(0.2397)	(0.2391)	(0.239)
Top 20 Regions	0.0929**	0.1027**	0.0874*	0.0814*	0.0663	0.0538
	(0.0468)	(0.0483)	(0.0462)	(0.0474)	(0.0465)	(0.0352)
Regional dummies (Origin & Destination)	YES	YES	YES	YES	YES	YES
Number of Observations	17292	17292	17292	17292	17292	17292
Adj-pseudo R^2	0.64503	0.64591	0.64795	0.64805	0.64744	0.64898
BIC	26834.882	26781.692	26647.021	26647.832	26682.41	26590.875

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Keyword	Count	Keyword	Count	Keyword	Count
Chemistry, Physical	24721	Crystallography	743	Computer Science, Artificial Intelligence	109
Chemistry, Organic	15243	Biophysics	721	Statistics & Probability	109
Chemistry, Multidisciplinary	15089	Plant Sciences	711	Education, Scientific Disciplines	90
Chemistry, Inorganic & Nuclear	11114	Nuclear Science & Technology	606	Agronomy	84
Chemistry, Analytical	10892	Thermodynamics	502	Acoustics	68
Materials Science, Multidisciplinary	5889	Toxicology	498	Oceanography	66
Chemistry, Applied	5250	Biotechnology &, Applied Microbiology	495	Materials Science, Ceramics	65
Physics, Atomic, Molecular & Chemical	5191	Mathematics, Interdisciplinary Applications	404	Biology	56
Chemistry, Medicinal	4089	Geosciences, Multidisciplinary	398	Mathematical & Computational	56
Physics, Condensed Matter	3957	Computer Science, Interdisciplinary Applications	384	Physics, Nuclear	41
Biochemical Research Methods	3626	Nutrition & Dietetics	289	Dermatology	30
Food Science & Technology	2833	Archaeology	268	Materials Science, Characterization & Testing	28
Pharmacology & Pharmacy	2650	Engineering, Environmental	236	Immunology	27
Biochemistry & Molecular Biology	2632	Mineralogy	234	Optics	22
Engineering, Chemical	2591	Materials Science, Textiles	222	Oncology	14
Physics, Applied	2007	Soil Science	191	Engineering, Manufacturing	5
Spectroscopy	1530	Computer Science, Information Systems	179	Geochemistry & Geophysics	5
Agriculture, Multidisciplinary	1493	Integrative & Complementary Medicine	176	Medicine, Legal	4
Electrochemistry	1389	Art	169	Medicine, Research Experimental	4
Nanoscience & Nanotechnology	1107	Radiology, Nuclear Medicine & Medical Imaging	155	Engineering, Electrical & Electronic	2
Environmental Sciences	1055	Energy & Fuels	143	Engineering, Petroleum	2
Metallurgy & Metallurgical Engineering	1017	Physics, Multidisciplinary	143	Genetics & Heredity	2
Materials Science, Coatings & Films	961	Materials Science, Biomaterials	117	Materials Science, Paper & Wood	2
Polymer Science	926	Mechanics	113	Fisheries	1
Instruments & Instrumentation	770	Automation & Control Systems	109	Marine & Freshwater Biology	1

Chapter 2

Centrality of regions in R&D networks: Conceptual clarifications and a new measure[†]

2.1 Introduction

Today it is widely recognized that external knowledge sources have become an essential component for innovating organisations. Both theoretical and empirical literature over the past decade provide evidence for the increasing importance of R&D networks for successful innovation (see, e.g., Powell and Grodal, 2005; Wuchty et al., 2007). Up to now, most studies have emphasized the crucial role of the ability to adopt external knowledge in the form of learning capabilities, such as technical or methodological skills, enabling innovating organisations to apply the externally tapped knowledge in the organisational innovation process. However, recently the importance of a particular relative network positioning to access external knowledge has been highlighted and attracted increasing attention (see, e.g., Ahuja, 2000; Owen-Smith and Powell, 2004). It is assumed that not only the ability to learn, but also a favourable position for a more efficient access to external knowledge is crucial.

From a network theoretical perspective, such a favourable positioning is referred to as centrality of network vertices (Borgatti, 2005), where – in terms of R&D – these vertices represent knowledge producing actors interlinked via edges representing knowledge flows. Actors showing a more central network position will more likely benefit from network advantages. This argument has been taken up at the regional level in recent regional science literature, where regions – constituting the aggregate of its knowledge producing organisations – are treated as relevant units of observation. In this context, the notion of inter-regional R&D collaboration networks has come into use (see, e.g., Autant-Bernard et al., 2007b) where regions are the network nodes representing distinct pools of knowledge, which are assumed to get into motion via the R&D relations between these regions,

[†]This chapter is based on a paper co-authored with Iris Wanzenböck and Thomas Scherngell.

constituting the edges in the network. Such a network representation has developed to an analytical vehicle that has been applied to investigate the geography of R&D networks (Scherngell, 2013), in particular how knowledge diffuses in a multi-regional system (see, e.g., Maggioni et al., 2007; Ponds et al., 2010).

Given this recent focus on regional R&D networks, network analytic measures have been increasingly applied at the regional level in order to characterize the inter-regional connectedness and centrality of a region, by capturing also the structural properties of the network (see, e.g., Sebestyén and Varga, 2013b; Wanzenböck et al., 2015). For observing a region's centrality, up to now the most common analytical approaches from Social Network Analysis (SNA) have been utilized, such as degree centrality or betweenness centrality (Wanzenböck et al., 2014). However, these studies somehow neglect conceptual problems that arise for networks defined at the aggregate level of regions. In particular, such problems are related with the loss of information regarding the structure of network relations and with that, information on the real channels through which knowledge flows. In this context, the question of how to adequately reflect regions in weighted network structures such as R&D networks become even more important.

As we argue in this chapter, the specific characteristics of regions – regarded as aggregate units – have to be taken into account and reflected in some way when designing analytical measurement approaches for regional centrality. Relevant questions in this context are (i) how can we conceive the centrality of regions in a network that is composed of several research actors in its underlying structure, and (ii) what are then the main building blocks that might characterize the centrality of regions, in particular when we talk about R&D networks?

This study is one of the first that deals explicitly with the drawbacks and insufficiencies related with conventional approaches to represent networks and measure centrality at the level of regions. Against this background, the objective is to propose a new measurement approach of regional centrality that is explicitly designed for aggregated networks at the regional level, based on the concept of inter-regional *bridging paths*. Here a bridging path is defined as an indirect connection between two regions via a third 'bridging region'. From a simple random matching process that models the collaborations among the micro-level actors based on the information provided at the aggregated level, we derive a closed form of the expected number of bridges between two regions stemming from a specific bridging region. On this basis we are able to define a new measure of regional centrality that not only depends on the number of links one region has, but also on the structure and intensity of its cross-regional collaborations.

In its fundamentals, our measure of *regional bridging centrality* builds upon several network-and knowledge-related arguments, referring to the role of bridges and the relevance of bridging path between network actors, or the general importance of diversified knowledge sourcing and technological recombinations (see, e.g., Kogut and Zander, 1992; Fleming, 2001; Singh, 2005). Moreover, we show how such a measure defined for aggreg-

ated networks can be meaningfully related to the regional dimension. We demonstrate how our measure of bridging centrality of a region can be easily interpreted as a function of (i) the participation intensity of a region in inter-regional R&D collaborations, (ii) the relative outward orientation in terms of all established network links, and (iii) the diversification of the network-partner-regions and knowledge relations to them. Hence, it views network centrality as a multidimensional problem, and integrates different region-specific aspects of the regional linking structure that might only together determine the visibility and importance of regions in R&D networks.

To illustrate our regional centrality measure we use a large-scale dataset on the European co-patent network in the year 2006 at the NUTS 2 level. The comparative analysis with three common SNA-based measures (degree, betweenness and eigenvector centrality) is based on basic statistics on distribution and correlations between the four centrality measures observed for the regional network. Despite striking similarities in correlations and distributional aspects on a more general level, the in-depth analysis of regional ranks reveals interesting differences which emphasize the advantages of the regional bridging centrality measure, in particular in terms of its interpretative power for region-level analyses.

The remainder of this study is structured as follows: Section 2 discusses in some detail the conventional approach to measure the centrality of regions in R&D networks. Section 3 introduces the concept of bridging paths, constituting the main essence of the measurement approach proposed in this study, before Section 4 formally derives the bridging centrality measure for regions. Section 5 shifts attention to the illustrative example, applying our measure to the European co-patent network and comparing results with conventional measures, before Section 6 concludes with a summary of the main results and some ideas for future research.

2.2 The conventional measurement approach

The notion of the centrality of regions in regional R&D networks has come into use just recently. It is argued that the knowledge creation ability within a region depends to a large extent on the ability of the region-specifc actors to efficiently access region-external knowledge (see, e.g., Bathelt et al., 2004; Graf, 2011). In this regard, inter-regional R&D networks are regarded as effective means, since network links can represent direct channels to a specific (region-external) source of knowledge that actors otherwise would not have access to. Against this background, need has been expressed to derive analytical approaches to measure a region's centrality in such networks, enabling the empirical researcher to characterize whether a region has a favourable position in the network, whether it takes a specific – for instance 'brokering' – role from a global network perspective, or how a region's network positioning changes over time.

However, the concept of network centrality was originally defined at the individual

level in human communication networks and the implications of using this concept at the regional level remain unclear. Therefore, this section intends to clarify the concept of network centrality as applied to inter-regional knowledge networks. We start with examining the origin, meaning, and purpose of network centrality (Subsection 2.2.1), and then lay out the major hurdles facing its transposition to regional R&D networks (Subsection 2.2.2). Finally, the two last subsections provide different ways to adapt well known centrality measures to the regional case, while at the same time keeping focus on their interpretation in the R&D context and pointing out their conceptual limitations.

2.2.1 A short introduction to the notion and context of centrality measures in social networks

The inception of the use of the concept of centrality in social network analysis (SNA) lies on the impetus of Bavelas' early researches (Bavelas, 1948). He was interested in linking the relational position of individuals within working-groups – namely their network-centrality – to their performance and influence over the group. Many empirical studies have followed to investigate if such a link existed in these types of networks, i.e., human communication networks (e.g., Bavelas, 1950; Leavitt, 1951; Faucheux and Moscovici, 1960; Burgess, 1969). The consequence of this line of work was to unveil the potential of the concept of network centrality in SNA.

As the representation of interactions in a network-form is not limited to human communication networks, the notion of centrality was soon extended and applied to various other types of networks. Indeed, this idea of investigating the influence of structural position within networks was promising and has triggered many studies in which the unit of analysis took different forms. Such studies include the application of the notion of centrality on: inter-personal networks within organizations (Beauchamp, 1965), cities in transportation networks (Pitts, 1965), the diffusion of innovation in inter-firm informal communication networks (Czepiel, 1974), the spread of diseases in infection networks (Bell et al., 1999), crime networks (Calvó-Armengol and Zenou, 2004), etc.

Along with these studies, a set of centrality measures has also emerged. Indeed, numerous measures have spawned either to refine existing measures or to adapt them to the networks under scrutiny. Those centrality measures include: the degree centrality, the betweenness (Freeman, 1977), the closeness (Freeman, 1979), the eigenvector (Bonacich, 1972), Katz's prestige (Katz, 1953), Bonacich's measure of power (Bonacich, 1987), etc.

Consequently, as a wide variety of centrality measures has been developed, one should expect that they differ in the meaning they purport and in the contexts they can be applied to. These differences are in fact tied to the very definition of network centrality.

The goal of a centrality measure is to assign to each agent of a network a value related to her/his position within the network. The variety of centrality measures then comes from the fact that each favours a particular network-pattern over others and each carries a 'view' of what being central *should be*. Thus, centrality measures are not neutral: they

rank the agents along some – often hidden – normative viewpoint which should support the aim of the study itself. In other words, different notions of centrality imply different 'competing "theories" of how centrality may affect group process' (Freeman, 1979, p. 238).

Then, the choice of a centrality measure should be dictated by the purpose it is aimed to serve (Borgatti, 2005). This purpose is brought about by the researcher and his research study and is of course highly context dependent. For instance, the kind of centrality measure used in the study of infection networks should be different from the one used in inter-firm cooperation networks.[1] This very idea is also, albeit slightly differently, formulated by Bonacich (1987, p. 1181):

> There are different types of centrality, depending on the degrees to which local and global structures should be weighted in a particular study and whether that weight should be positive or negative. [...] There is no point in subsuming all these situations under one measure.

Therefore, there is no unique and 'best' measure of centrality, no 'one size fits all' centrality measure. One then should remember the implicit choices underlying centrality measures and the context to which they can be applied. Therefore, we are now going to discuss the particular context of regional R&D networks and question whether centrality measures can be applied to it.

2.2.2 Can the concept of network-centrality be applied to R&D networks?

We now delineate two key elements impeding the straightforward application of network-centrality measures to regional R&D networks. *First*, regions are not single entities. Indeed, while being at the centre of the analysis, regions are not the 'actors' taking part to the action of the network. Only the agents that compose the regions are involved in R&D networks (and any kind of inter-regional network more generally). Centrality measures are best suited for situations where the unit of analysis is also the actor of the network. In fact, in the case of regional centrality, there is a strong duality between the micro strata, where lie the actors of the network, and the meso strata, where lies the focus of the centrality measure. Indeed, to assimilate regions as 'actors' would imply to assume that all agents within them would act as one and only one entity; it would require to do 'as if' the region was a single agent, like for instance a single researcher. If this 'as if' hypothesis may be reliable when studying small groups in which information is quickly shared and without depreciation, such as research teams or even – under some conditions

[1] In the study of spreading disease in infection networks, the notion of eigenvector centrality catches best the idea that the central agent, if infected, would spread the fastest the disease across the network (Borgatti, 1995). When studying flows of information in inter-firm communication networks, the closeness centrality reports the best the idea that the central agent would be the first to 'know' the novelties and by then have a technological edge over its competitors (Czepiel, 1974; Freeman, 1979).

– organizations, it no longer holds when looking at complex structures such as regions which are often composed of heterogeneous, non necessarily interacting, agents.[2]

Second, the links in R&D networks involve a particular kind of flow. For instance, in collaboration networks, a link may be the medium of various types of exchanges and could then be interpreted in different ways. If we focus specifically on the notion of knowledge production, the links can represent the access to a specific source of knowledge that agents would have otherwise not have access to, like the possibility to share tacit knowledge with a partner (Collins, 2001). If the focus is more on the dynamics of the collaboration network, links can be seen as vehicles of information, on who would be a suitable and a reliable partner to collaborate with, particularly across regional borders (see, e.g., Gulati and Gargiulo, 1999; Cassi and Plunket, 2015). These two simple different perspectives on how to interpret network links have different implications. In the first case, in which we consider flows of knowledge, the benefits from network-distant agents may decay much more steeply than for the case of flows of information which is acquired and shared more easily. These differences in flows' nature and behaviour are not innocuous regarding the interpretation of centrality measures, as Borgatti (2005, p. 69) has pointed out: 'the importance of a node in a network cannot be determined without reference to how traffic flows through the network'. He has also shown that different centrality measures each carry an implicit different assumption about the kind of flow it is suited for, so that they cannot be applied to any network.

With these details in mind, the next subsection considers the case in which regional R&D networks are seen as weighted networks. Some widely used centrality measures are described as well as: 1) their classic interpretation in the context in which they were originally defined and 2) their interpretation when applied to regional R&D networks. Finally, the last subsection investigates the case in which a region's centrality is inferred by its agents centralities.

2.2.3 Regional R&D networks as weighted networks

The first manner to adapt existing centrality measures to regional R&D networks is to consider the regions as the nodes of the network. Accordingly, the inter-regional R&D collaboration network can be depicted by the matrix G of typical element g_{ij} which represents the number of links between the agents from regions i and j. As collaborations are bilateral and their flow can be higher than one, it yields an undirected weighted matrix G of typical element $g_{ij} \in \mathbb{R}^+$.

We discuss three conventional measurement in this case: the degree-, the eigenvector- and the betweenness-centrality. The properties of these centrality measures are discussed in light of the context of R&D networks.

The first centrality measure, is the degree-centrality. The notion of degree-centrality in

[2]It is to note that Everett and Borgatti (1999) propose an extension of centrality measures to groups but where within-group homogeneity is required to provide a proper interpretation.

SNA was primarily defined as the number of connections an agent had in communications networks and is reviewed in Freeman (1979). As Freeman mentions, early researchers in SNA even considered it as the sole centrality measure, able to summarize the importance of a node in a network. In the case of regional networks, the links between two nodes are typically weighted, the degree of a node can then be defined as the sum of all the links stemming from it.[3] Let d_i be the degree of node i, it is formally defined as follows: $d_i = \sum_j g_{ij}$.[4] Depending on the kind of network under study, the degree can be interpreted as the probability to be reached in a network by a random walk or the ability to infect other agents in a one time period (Borgatti, 2005). However, these two interpretations hardly make sense in the case of R&D networks. Another simple and unambiguous interpretation of the degree is just the dominance of a given region over other regions in terms of R&D collaborations. Depending on the purpose of the study, this interpretation may be relevant. However, in any case, this measure suffers from a major flaw: it does not convey any information on the structure of the network.

Another centrality measure widely used in SNA is the eigenvector centrality. This measure was introduced by Bonacich (1972) and states that the importance of a node is related to the importance of the nodes it is connected to. Contrary to the degree-centrality, the eigenvector-centrality of a given node depends on the information on all the links of the network, meaning the position of the nodes within the global network has an influence on their centrality. Therefore, two nodes with the same degree can have different eigenvector centralities. Formally, the eigenvector-centrality of a node, e_i, is defined by the relation: $\lambda e_i = \sum_j g_{ij} e_j$, with $\lambda > 0$ a proportionality factor. This centrality is self-referential and can be solved by writing it in a matrix-form:

$$\lambda \mathbf{e} = G\mathbf{e}, \tag{2.1}$$

where $\mathbf{e}$ is the vector of all centralities. The vector $\mathbf{e}$ that solves equation (2.1) is the eigenvector of the matrix G associated to the eigenvalue λ.[5] The very idea reflected by this measure is related to node influence. The main driver is that a node will be more influential if it has influence on very influential nodes (the influence being measured by the links between the nodes). While being an appealing feature for studies on individual's influence, this interpretation is strongly impeded by the problem of the micro/meso duality of the regional network. Indeed, assume a region is central thanks to connections to important regions, do its agents – who are the actors of the network – really benefit from their region's centrality? It would imply that every agent within a region would

[3] In the case where the network is directed, like for instance in a patent-citations network, the number of links emanating from a node (e.g., references made to other patents) is called the out-degree while the number of links received (e.g., the number of citations received from other patents) is called in-degree. For undirected networks, such as collaboration networks, the in-degree is equal to the out-degree.

[4] There is a generalization of the degree centrality for weighted networks given by Opsahl et al. (2010) but whose interpretation in this context remains unclear.

[5] By convention, it is standard to use the eigenvector associated to the largest eigenvalue (Bonacich, 1987; Jackson, 2010).

homogeneously benefit from the influence of all other agents of the region, which seems hardly the case. It then happens that this measure is hardly transposable to the regional level.

A third measure commonly applied in SNA is the betweenness-centrality. To define it, we first introduce the notion of network path and shortest path. A path between two nodes i and j is a sequence of K distinct nodes $\{n_1,\dots,n_K\}$ starting from i (i.e., $n_1 = i$), ending with j (i.e., $n_K = j$) and such that each consecutive pair of nodes is connected in the network.[6] The length of a path is the number of nodes composing the path. Then, a shortest path between i and j is a path that has minimal length. Now, let $SP(jk)$ to be the number of shortest paths between nodes j and k, and $SP_i(jk)$ the number of shortest paths between j and k where node i appears. Then the betweenness-centrality of i is defined by the following equation:

$$B_i = \sum_{j \neq i} \sum_{k \neq \{i,j\}} \frac{SP_i(jk)}{SP(jk)},$$

the term in the double sum depicting the share of shortest paths between j and k where i lies on.

This form of centrality was originally defined in the context of communication networks, where links between agents represent information flows. When Freeman introduced this measure, he defined central agents as 'structurally central to the degree that they stand between others and can therefore facilitate, impede or bias the transmission of messages' (Freeman, 1977, p. 36). Alternatively, betweenness-centrality can be seen as how much a node is necessary for flows to connect all other nodes in the network. Despite being computationally easy to apply at the regional level, this measure suffers from major flaws when applied to regional R&D networks. Indeed, for the importance of being in the 'shortest path' to hold, two assumptions are necessary. The first is that the flows necessarily follow the shortest path (which makes the 'central agent' able to retain information and exert some influence). If information (or the adequate flow) does not pass only through shortest paths, this measure becomes much less relevant. In R&D networks, this may not be the case: for instance, when considering information over potential partners obtained via collaboration, that information may not be limited to flow only through shortest paths, just because of the nature of information. The second assumption is that flows do not suffer from any decay. Indeed, at the moment where the relevance of network-flows are reduced with the network-distance, then what is the use of being in the middle of network-paths between agents? In this case, the betweenness of a region, beyond its first or second circle of connections, may be of little use. For instance, if connections materialize access to knowledge sources, it is quite unlikely that agents far apart with respect to the network-distance influence each other. Last, beyond these two limiting

[6] Mathematically, $\{n_1,\dots,n_K\}$ is a path between i and j if $g_{n_k n_{k+1}} > 0$ for all $k \in \{1,\dots,K-1\}$, with $n_1 = i$ and $n_K = j$.

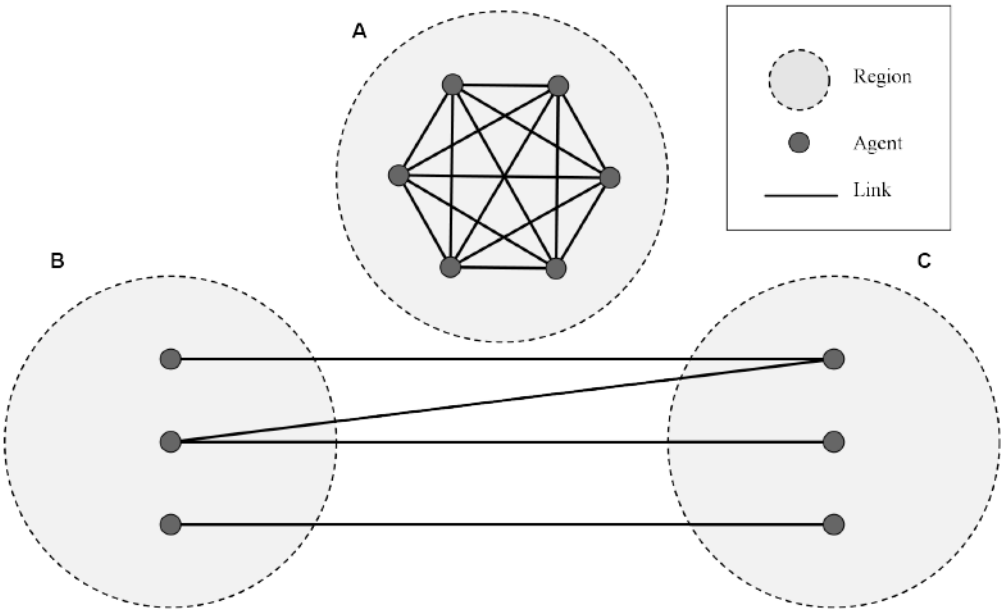

Figure 2.1 – Illustration of a regional network in which a region has a strong internal structure yet no link with the outside.

assumptions, the betweenness measure happens to be much better suited for networks composed of individuals (be it firms or inventors). Similarly to the eigenvector centrality, its interpretation hardly fits the regional scale, since a region with a high betweenness does not necessarily translate into its agents being on shortest paths.

As we have shown, existing measures can be applied to regional R&D networks, when taking regions as the nodes of a weighted network. But the interpretation of the measures and their conceptual meaning is far from being straightforward, if applicable at all. In the next section we show and discuss another way of accounting for regional centrality.

2.2.4 Regions as the aggregate centrality of their actors

A different way to measure regional centrality is to assume that a region's centrality actually refers to the centrality of its agents. Indeed, since regional networks can be seen as the aggregate interactions of the agents from these regions, a natural way to assess a region's centrality could be to link it to the centrality of its agents.

In doing so, the first step is to find the relevant actor of the network. In co-patenting networks, it can either be firms or inventors. The choice depends on whether we believe that the information and knowledge pool of firms is shared among all its inventors. If so, then firms can be considered as the real actors of the network. We will call 'agent' the entity resulting of this choice. Thus, the regional network can then be depicted by a micro-level network formed of the links between the agents, each of them belonging to a region. To build the regional centrality, one has to choose the relevant centrality measure and compute it at the agent's level. Let c_i be the centrality of agent i and let S_r be the

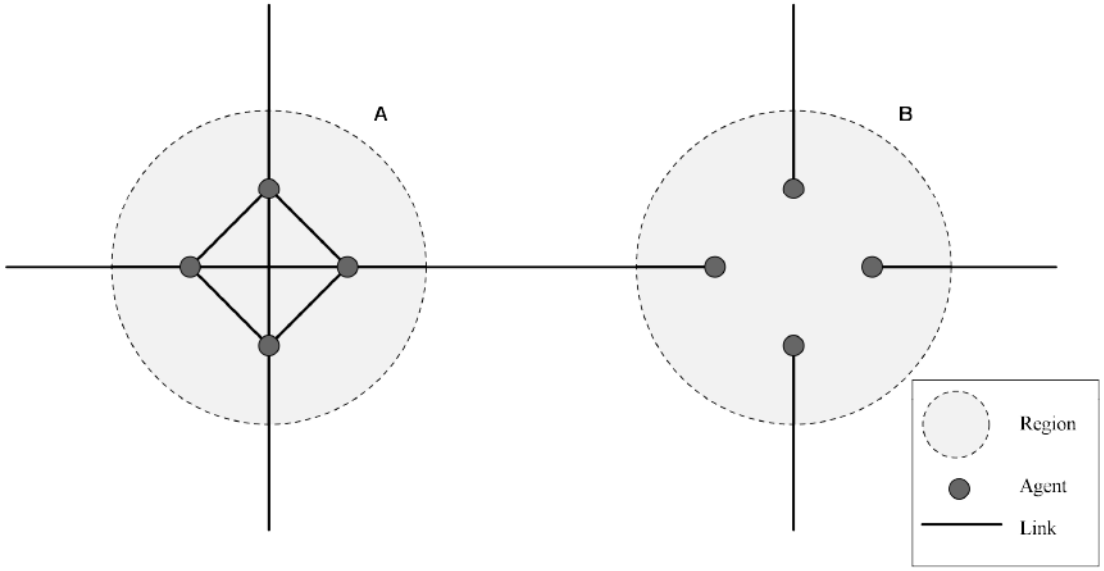

Figure 2.2 – Sample of an inter-regional network. Illustration of two regions with external links, differentiated with respect to their internal links.

set of agents belonging to region r. Then, the centrality of region r, C_r, can be defined merely as the sum of the centrality of its agents, as follows:

$$C_r = \sum_{i \in S_r} c_i.$$

Beyond the problems inherent to the centrality measures discussed earlier (e.g., linked to the nature of the flows), this methodology – aggregation of micro-level network centralities – also involves drawbacks. The main problem stems from the links occurring internally to regions. Indeed, should intra-regional links be counted in the micro-level network? An example of a problematic case is illustrated by Figure 2.1. In this figure, a network of three regions is represented: region A has many agents that are all connected to each other but have not any link with other regions; conversely, the agents from regions B and C have no intra-regional link but do have cross-regional collaborations. Actually, if the network is computed at the micro-level, whatever the measure, region A will have the higher centrality, despite having no inter-regional link whatsoever. This is fundamentally problematic: a measure of regional centrality should not be able to give a high ranking to regions having no external links, simply because it should somewhat relate to the position within the interregional network which is not the case here.

A straightforward solution to this problem would be to 'cut' all intra-regional links: the centrality would then be computed using a network where all internal links would be severed. Yet, this adjustment would also lead to conceptual problems. Take for instance the example illustrated by Figure 2.2. This figure depicts a network of two regions, A and B, that are very similar. They both are composed of four agents and each has a link with

another region. Region A has a strong internal structure: all its agents are connected. On the contrary, no agent from region B has a link within the region. Despite that, the agents of the two regions have different positions in the global network, if all internal links are cut to compute the centralities, then the two regions would be equivalent. Cutting internal links would involve a distortion in the network structure.

Consequently, the major problem raised by the aggregation of micro-level centralities is that intra-regional links cannot either be kept or removed without posing conceptual problems.

As developed in this section, the centrality measures discussed so far all suffer from conceptual drawbacks when applied at the regional level. Given these considerations, there is a need for developing alternative centrality measures applicable for regional R&D networks and resting on more robust conceptual grounds. In what follows, we provide a first attempt for the development of novel measurement approaches that explicitly address the conceptual problems discussed above by taking into account the underlying micro structure of regional R&D networks.

2.3 The concept of bridging paths

There is a strong need for overcoming the duality in analyses of R&D networks of regions concerning the micro level which encompasses the actors participating in R&D collaborations, and the aggregate, i.e. regional, level where the analysis focuses on. As has been discussed in the previous section, major problems arise in applying and interpreting conventional SNA-based centrality measures. The purpose of this section is to provide a new concept that is *meaningful* in the context of inter-regional R&D networks. We introduce the notion of a bridging path denoting a form of indirect connection between regions, i.e. regions are indirectly connected in the network thanks to their micro-level actors. We first define this concept before providing an approach to derive the expected number of bridging paths from aggregate flows of R&D interactions. The expected number of bridging paths between regions will be the major building block of the regional centrality measure we introduce in the next section.

To introduce the concept of bridging paths, consider a network where the nodes are the regions and the connections between the regions represent the R&D interactions between their agents. This represents a weighted network where we define g_{ij} as the number of R&D interactions (i.e. micro-level links) between regions i and j. Further, each micro-level link between two regions is denoted by y_{ij}^{a}, where y_{ij}^{a} represents the a^{th} link between regions i and j with $a \in \{1, \ldots, g_{ij}\}$. A bridging path is then regarded as a set of two links at the micro level connecting three agents from three different regions. Speaking in social network analytical terms, the micro-level agent in one region act as a 'broker' (Burt, 1992) for two other not directly connected actors; he/she has a bridging role in the network of regions linking indirectly the micro-level agents of two other regions. This

triangulation between actors located in three different regions leads to the notion of an inter-regional bridging path. Formally, a bridging path is defined as a set of two links from two different regions, say i and j, with a third one, say k, so that the agents from i and j are both connected to the same agent in k. This means that a pair of links (y^a_{ik}, y^b_{jk}) forms a bridging path if, and only if, y^a_{ik} and y^b_{jk} are connected to the same agent in region k. In other words, agents i and j are indirectly connected thanks to one agent of region k.

This notion is depicted by figure 1 which represents a regional network of three regions. In this figure, the pair of links (y^2_{ik}, y^1_{jk}) is a bridging path between regions i and j stemming from k because the agent from k maintains both links y^2_{ik} and y^1_{jk}. Although both regions j and k do have links with region i, there is no bridging path between them because the agents from i of the links y^1_{ik} and y^2_{ik} are neither connected to y^1_{ij}, y^2_{ij} nor y^3_{ij}. Hence, region i provide not any bridging path between regions j and k in this set-up. We see that the notion of bridging path is about indirect connections. Accordingly, the region with most bridging paths is region j, as it provides two bridging paths between regions i and k.

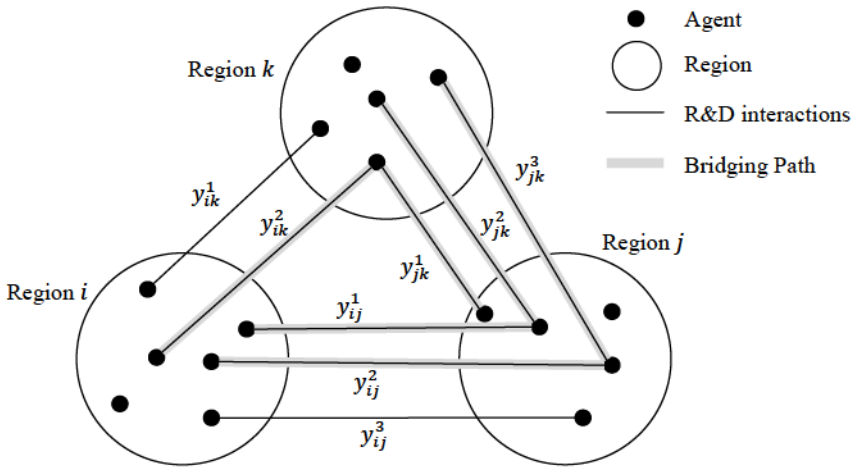

Figure 2.3 – Illustration of the notion of bridging paths

Notes: The figure depicts three bridging paths formed by the following pairs of links: (y^2_{ik}, y^1_{jk}), (y^1_{ij}, y^2_{jk}) and (y^2_{ij}, y^3_{jk}). So the regional dyads (j,k), (i,k) and (i,j) have respectively 0, 2 and 1 bridging paths stemming from regions i, j and k, respectively.

The relevance of the bridging paths concept can be quite directly underlined by means of basic theoretical considerations in innovation research. The creation of new knowledge is often viewed as a recombination of existing knowledge (see, e.g., Kogut and Zander, 1992; Fleming, 2001; Cassiman and Veugelers, 2006). It implies that the source from which the agents draw their knowledge will have an impact on their ability to generate interesting ideas and new knowledge. In the case where a region is isolated, where collaborations occur mainly within the region, the knowledge pool may become redundant and even lead to lock-in situations (see, e.g., David, 1985; Arthur, 1989). Collaborations with agents from other regions allow to benefit from different knowledge bases (see, e.g., Singh, 2005; Berliant and Fujita, 2012), help moderate the problem of redundancy and generate

more radical innovations. From this viewpoint, bridging paths provide better knowledge opportunities to regions. Then regions from which stem many bridging paths could be seen as key players in the network with actors potentially benefiting from a more diversified knowledge pool.

Moreover, bridging paths may also be of significance when we consider network formation processes. Indeed, several recent studies have put at the forefront the consideration that the structure of network links plays an important role in explaining future states of the network (see, e.g., Barabâsi et al., 2002; Jackson and Rogers, 2007). Recent research in the context of R&D networks has shown that two actors are more likely to collaborate together if they share a common collaborator (that is if they are indirectly linked in the network, see, e.g., Fafchamps et al., 2010; ter Wal, 2014). Hence, bridging path create network proximity and opportunity for (triadic) closure so that there are good reasons to assume that bridging paths matter for the evolution of the whole network. Indeed, if bridging paths represent indirect connections between agents from different regions, then we can assume that those regions which provide the bridging paths are in a position to facilitate the connectivity between other regions in the network. Bridging paths can then be seen as important for regions not only in the context of accessing a diversified knowledge pool, but also in a network formation perspective as it helps the formation of inter-regional connections and with that inter-regional diffusion of knowledge.

2.4 A new measure of regional centrality

Proposing the significance of the bridging path concept for measuring regional centrality in regional R&D networks, the question arises at this point how this concept can be incorporated into regional centrality measures. Usually, empirical researchers focusing on regions as units of observations, and by this, on regional R&D networks, face the problem that the underlying micro structure of the network may be either undefined or unobservable. Concerning the latter, one may consider the example of co-patenting networks (see, e.g., Lata et al., 2015), for which the relevant actors are individual persons (inventors) that are hardly identifiable as homogeneous nodes over time. Thus, we introduce a model of random matching. It allows us to approximate the underlying micro-structure by deriving an expected number of bridging paths (ENB) between two regions, using only the aggregate flows of collaborations between regions.[7]

Our random matching process relies on two basic assumptions: (i) collaborations occur between two agents, and (ii) when a collaboration occurs, the two agents are matched at random. By this, it reflects the ex post probability to be matched, i.e. the probability that two agents for two particular regions have been matched conditional to the structure of the inter-regional flows of collaborations. The very intention is to give a baseline for

[7]This model is an adaptation of the one in Bergé (2015). In fact, the methodology is very similar to the one used by Bloom et al. (2013), which provides a measure of technological similarity between firms' patenting activity introducing a model which considers random encounters between pairs of scientists.

a micro-network that was likely to occur, with respect to what is observable at the meso level. Thus, random matching is used to infer the structure of the micro network by using only the information included the links between the regions.

On this basis, it is now possible to derive the expected number of bridging paths stemming from a given region by using directly the aggregate flows of collaborations occurring between regions. First, denote by n_i the number of actors active in R&D collaboration in region i. Then the expected number of bridging paths, ENB^i_{jk}, between the two regions j and k stemming from the bridging region i along the random matching process is:[8]

$$ENB^i_{jk} = \frac{g_{ij} g_{ik}}{n_i}. \tag{2.2}$$

The expression related by equation (2.2) simply states that the more connections two regions, j and k, have with a third common region, i, the more likely they will have indirect connections at the micro level (bridging paths) thanks to the actors located in i.

Based on this, we are able to construct a new measure of the centrality of regions in R&D networks, denoted as *regional bridging centrality (BC)*. The BC is defined as the number of bridging paths stemming from a region between all dyads of the network. Formally, this means that the BC of region i is equal to:

$$BC_i \;=\; \sum_{j \neq i} \sum_{k \neq i,j} ENB^i_{jk}, \tag{2.3}$$

where ENB^i_{jk} is defined by equation (2.2).

The interesting point of our measure is that its definition can be pretty much simplified and interpreted meaningfully in a regional context. Assume that the number of agents (n_i) is proportional to the number of projects (g_i); then equation (2.3) decomposes to a notion of centrality of a region that entails a combination of three different components, reflecting i) a region's *participation intensity,* ii) a region's *relative outward orientation* and iii) a region's *diversification of network links.*[9] It is defined as

$$BC_i \;=\; \bar{g}_i \, s_i \, (1 - h_i), \tag{2.4}$$

where

$\bar{g}_i$ is the number of outer collaborations (i.e. outer degree, that is $\bar{g}_i = g_i - g_{ii}$ which is the total number of collaborations of i, noted g_i, excluding the internal ones, noted g_{ii}). It refers to a region's *participation intensity* in inter-regional collaborations, which affects positively the centrality of the region. It is a general measure of how well a region is embedded in the particular R&D network. Note that a region's

[8] For a formal proof, see Bergé (2015).

[9] The formal proof is given in Appendix 2.7.1.

size will amplify the probability of yielding more bridges between other regions. The participation intensity could therefore be interpreted as a broad measure of the relational capacity of the regional network nodes, which should be taken into account.

s_i is the share of outer collaborations with $s_i = \bar{g}_i / g_i$. It can be related to the *relative outward orientation* of all established network linkages, i.e. the relative degree of external R&D interactions. It refers to the openness of a region with respect to knowledge sourcing strategies. Given the fact that the BC focuses on the capacity of one region to link other regions, a high number of region-internal collaborations would have a negative influence as it potentially reduces the number of actors connecting different regions.

h_i refers to the Herfindahl-Hischman (HH) index of the distribution of i's outer collaborations defined as $h_i = \sum_{j \neq i} (g_{ij} / \bar{g}_i)^2$. The term $1 - h_i$ varies between 0 and 1 according to the degree of *diversification of network links* to other regions, and indicates how a region's collaborations are distributed along its neighboring regions in the network. In this case, the more the collaborations are concentrated, the less the region is central. This is because concentration offsets the benefits of outer connections as it reduces the actors' possibility to build bridges among different regions. Also it relates to the fact that the more the outer collaboration pool is diversified over different regions, the more the region can draw its knowledge from different sources.

One central promising property of the measure is that it takes account of the peculiar characteristics of regional networks. Indeed, regional networks are characterised by the structure of region-internal and region-external links and this feature cannot be dealt with adequately by using a single (a-spatial) SNA centrality measure. A region's ability to benefit from new ties in the R&D network or exploit external knowledge sources via the links may be determined by all three components together. Outward orientation and higher diversification in particular may help a region to develop and renew the regional knowledge base faster, or prevent lock-in situations in certain technologies (see, e.g., Breschi and Lenzi, 2015).

2.5 An illustrative example: an application to the European co-patent network

Given the promising features of the regional bridging centrality (BC) measure as defined in the previous section, an application to empirical regional R&D networks is required in order to illustrate the behaviour of the measure as compared to the conventional ones. To this end, we will employ co-patent data, comparing the regional BC with three other

commonly used centrality measures, that is degree, eigenvector and betweenness centrality.[10] We use the European co-patent network, a network of inter- and intra-regional collaborations in patent production observed at the regional level. A co-patent, that is a collaboration issuing a patent grant, is a visible trail of a successful R&D collaboration and is defined as an invention implying at least two inventors. This data are extracted from the REGPAT database (Maraut et al., 2008) and consist of all patents applied for at the European patent office (EPO) in the year 2006. We make use of the information contained in each patent record to build the co-patent network. Particularly, we use the address contained in each inventor's byline to map every patent to a set of NUTS 2 regions. That is, the NUTS 2 regions represent the place of residence of the inventors when they applied the patent. We consider that the flow of inter-regional collaborations between two regions consists of all patents having at least one inventor from each of these two regions. Collaborations occurring strictly within the regions are counted as intra-regional patent.

The network consists of collaboration flows between 245 NUTS 2 regions. This cross-regional co-patenting network is based on a total of 40,142 patents, of which 16,661 are inter-regional collaborations linking the 245 NUTS 2 regions. As a starting point, the three components of the BC are described by table 2.1a. The participation intensity is on average 237, which means that the regions show on average 237 co-patent links to other regions in the network. This is much higher than the median of 100, confirming the right-skewed distribution of the number of co-patent links the individual regions hold to other regions.

More interestingly is the relative outward orientation. Here, the median is 71%, meaning that for half the regions, more than 71% of their patents are of inter-regional nature, being invented with at least one partner outside the regions. Also diversification is relatively high, with an average at 85%, meaning that the co-patents are rather distributed along several regions. Hence, the regions resort – on average – to a rich portfolio of partner regions leading to a diversified structure of inter-regional knowledge exchanges in patenting. In contrast to the participation intensity, the other two components, the relative outward orientation and the structure, are slightly left skewed, and can be seen as moderators of the scale of a region. Indeed, being a large region with a high network participation intensity does not necessarily lead to a high centrality value, if either the share of intra-regional collaborations is very large or inter-regional links are concentrated among only a few regions.

Table 2.1 reports some statistics on the BC measure as compared to the conventional measures, and the correlations among them. Note that all measures are normalized so

[10]The degree is here calculated as the number of unique projects the agents of a region are involved in. The eigenvector and the betweenness centrality are computed using the package `igraph` available in the statistical software `R`. Both these two measures are based on the weighted regional co-patent network where the nodes are the regions and where the linkages between any two regions are the number of patents co-invented by agents from these two regions. Due to the nature of the network, we used the weighted version of both the betweenness and the eigenvector centrality.

Table 2.1 – Descriptive statistics of the components of the BC and of the centrality measures applied on co-patenting data.

(a) Descriptive statistics of the three components of the bridging centrality measure.

	Min	Q1	Median	Q3	90%	Max	Mean	SD	Skewness	Kurtosis
Participation intensity	1	26	100	280	559	2333	237.16	376.77	3.09	10.78
Relative outward orientation	0.2000	0.600	0.737	0.835	0.907	1	0.714	0.16	-0.45	-0.14
Diversification	0	0.831	0.893	0.925	0.945	0.972	0.850	0.14	-3.85	18.47

(b) Summary statistics.

	Min	Q1	Median	Q3	90%	Max	Mean	SD	Skewness	Kurtosis
Bridging Centrality	0.0000	0.0083	0.0332	0.0965	0.2231	1.0000	0.0881	0.1490	3.3465	13.2502
Degree	0.0000	0.0116	0.0406	0.1384	0.2597	1.0000	0.1081	0.1722	3.0310	10.1794
Eigenvector	0.0000	0.0010	0.0035	0.0208	0.0927	1.0000	0.0434	0.1285	4.9103	26.8212
Betweenness	0.0000	0.0022	0.0118	0.0595	0.1719	1.0000	0.0598	0.1307	4.3528	23.0673

(c) Correlations.

	Bridging Centrality	Degree	Eigenvector	Betweenness
Bridging Centrality	1.0000	0.9080	0.9311	0.6886
Degree	0.9080	1.0000	0.8141	0.8411
Eigenvector	0.9311	0.8141	1.0000	0.5641
Betweenness	0.6886	0.8411	0.5641	1.0000

Notes: The *participation intensity* is the outer degree. The *relative outward orientation* is the share of outside collaborations over all collaborations, it varies between 0 and 1. The *diversification* is $1 - h_i$ with h_i being the Herfindahl index of the distributions of region i's collaborations over all other regions; it varies between 0 and 1, the more the collaborations are concentrated, the lower is the measure. 9

that the highest value is one and the lowest zero.[11] While there is no large difference in the summary statistics provided by table 2.1b, it can still be noted that the eigenvector-centrality is clearly the more skewed of the four measures. Table 2.1c further shows that the correlation between the bridging centrality and the other measures ranges from 70% to 93%. Those high levels are reassuring as they show that the BC does not completely reorder the regional positioning. The difference in the distribution of the four centrality measures compared is also illustrated by figure 2 which reports the cumulative distribution of each measure. We can see that the eigenvector-centrality, except at the very beginning of the distribution, is on the top of all other measures while the BC lies between the degree and the betweenness. The differences in distribution are higher at the beginning of the distribution (below 0.50) than at the end where the distribution of the BC, the degree and the betweenness are much closer. Yet, the differences with existing measurements are real and it is worthwhile to point out to changes occurring to some particular regions. Moreover, it becomes obvious from this basic statistics that the bridging centrality is a combination of three components. It depends not only the scale of a region, like it might be the case for the degree centrality, or the quality of partners, i.e. whether they are located at the very core of the network, as for the eigenvector centrality. Therefore, it might be of particular interest how differently the three components are distributed across the individual regions.

[11] Formally, the transformation applied to each centrality measure is: $(x - x_{min})/(x_{max} - x_{min})$.

Table 2.2 – Centralities of the top 30 regions for the co-patent network, ranked by bridging centrality.

	NUTS 2	Bridging Centrality value (rank)	Degree Centrality value (rank)	Eigenvector Centrality value (rank)	Betweenness Centrality value (rank)
Karlsruhe	DE12	1.00 (1)	0.90 (3)	1.00 (1)	0.46 (7)
Darmstadt	DE71	0.85 (2)	0.86 (5)	0.79 (3)	0.53 (5)
Rheinhessen-Pfalz	DEB3	0.80 (3)	0.66 (8)	0.89 (2)	0.25 (12)
Düsseldorf	DEA1	0.80 (4)	0.82 (6)	0.62 (4)	0.51 (6)
Köln	DEA2	0.78 (5)	0.73 (7)	0.59 (6)	0.55 (4)
Oberbayern	DE21	0.57 (6)	0.93 (2)	0.39 (7)	0.93 (2)
Stuttgart	DE11	0.51 (7)	0.87 (4)	0.59 (5)	0.34 (8)
Northwestern Switzerland	CH03	0.50 (8)	0.44 (13)	0.18 (16)	0.24 (14)
Freiburg	DE13	0.48 (9)	0.55 (9)	0.35 (9)	0.20 (20)
Arnsberg	DEA5	0.42 (10)	0.39 (17)	0.31 (10)	0.06 (62)
Berlin	DE30	0.40 (11)	0.42 (14)	0.22 (13)	0.18 (22)
Tübingen	DE14	0.38 (12)	0.47 (12)	0.35 (8)	0.15 (31)
Île de France	FR10	0.34 (13)	1.00 (1)	0.06 (36)	1.00 (1)
Münster	DEA3	0.33 (14)	0.29 (19)	0.22 (12)	0.16 (28)
Mittelfranken	DE25	0.33 (15)	0.40 (16)	0.16 (18)	0.11 (37)
Alsace	FR42	0.30 (16)	0.26 (26)	0.13 (19)	0.22 (16)
Zurich	CH04	0.30 (17)	0.32 (18)	0.11 (22)	0.18 (23)
Schwaben	DE27	0.29 (18)	0.28 (20)	0.21 (14)	0.06 (58)
Brandenburg	DE40	0.28 (19)	0.23 (29)	0.16 (17)	0.03 (92)
Hannover	DE92	0.25 (20)	0.25 (27)	0.12 (21)	0.08 (50)
Unterfranken	DE26	0.24 (21)	0.26 (24)	0.23 (11)	0.05 (64)
Rhône-Alpes	FR71	0.24 (22)	0.54 (10)	0.06 (37)	0.32 (10)
Hamburg	DE60	0.24 (23)	0.21 (35)	0.10 (25)	0.09 (47)
Prov. Vlaams-Brabant	BE24	0.23 (24)	0.19 (38)	0.04 (44)	0.13 (33)
Espace Mittelland	CH02	0.22 (25)	0.26 (25)	0.09 (26)	0.06 (60)
Koblenz	DEB1	0.22 (26)	0.18 (40)	0.19 (15)	0.07 (57)
Schleswig-Holstein	DEF0	0.22 (27)	0.22 (32)	0.10 (23)	0.04 (72)
Prov. Antwerpen	BE21	0.21 (28)	0.21 (34)	0.04 (42)	0.19 (21)
Lüneburg	DE93	0.20 (29)	0.17 (43)	0.07 (34)	0.10 (44)
Région de Bruxelles, Brussels Hoofdstede	BE10	0.20 (30)	0.14 (61)	0.02 (56)	0.06 (59)

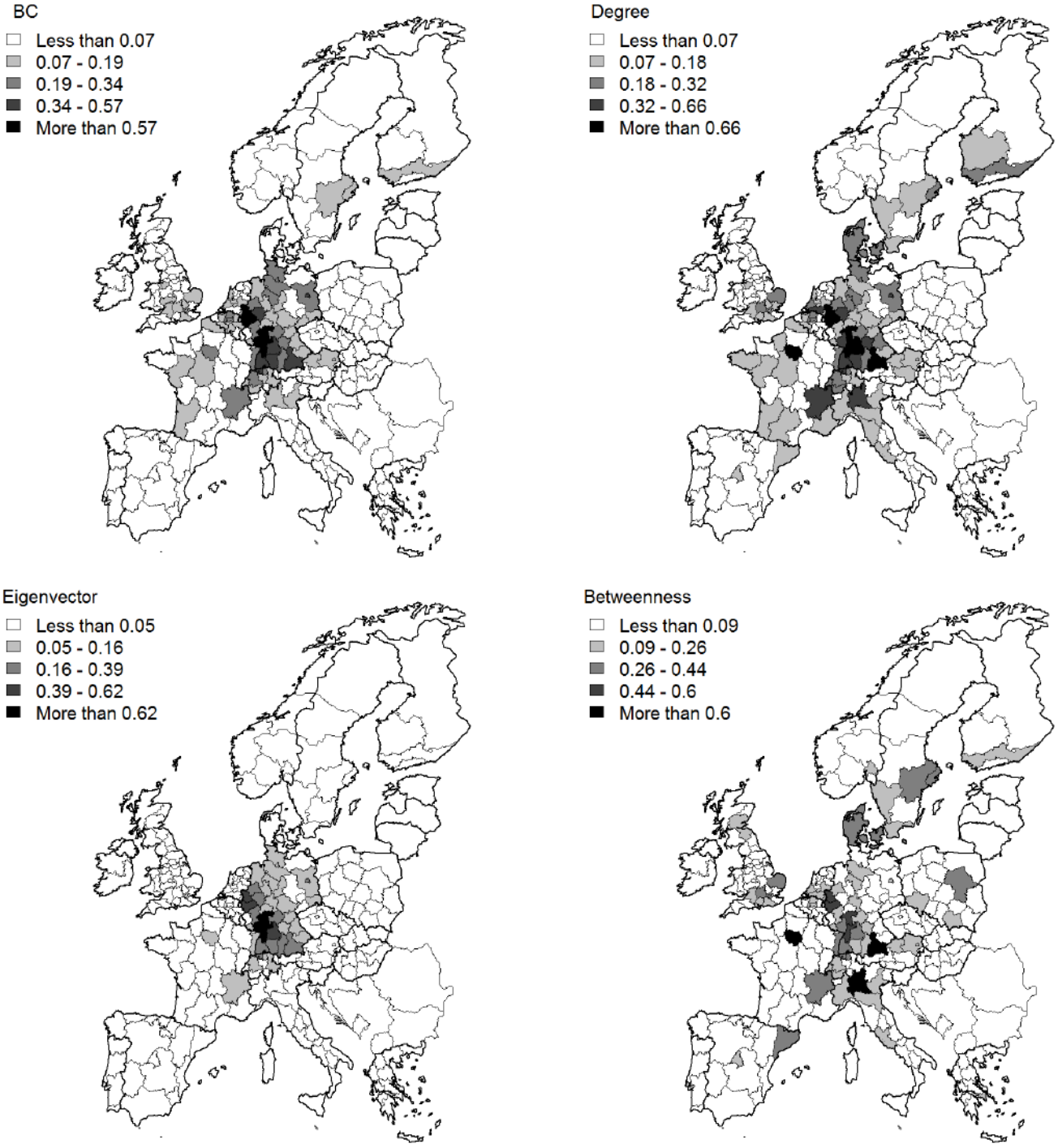

Figure 2.4 – Spatial distribution of the four centrality measures among the 242 NUTS 2 regions.

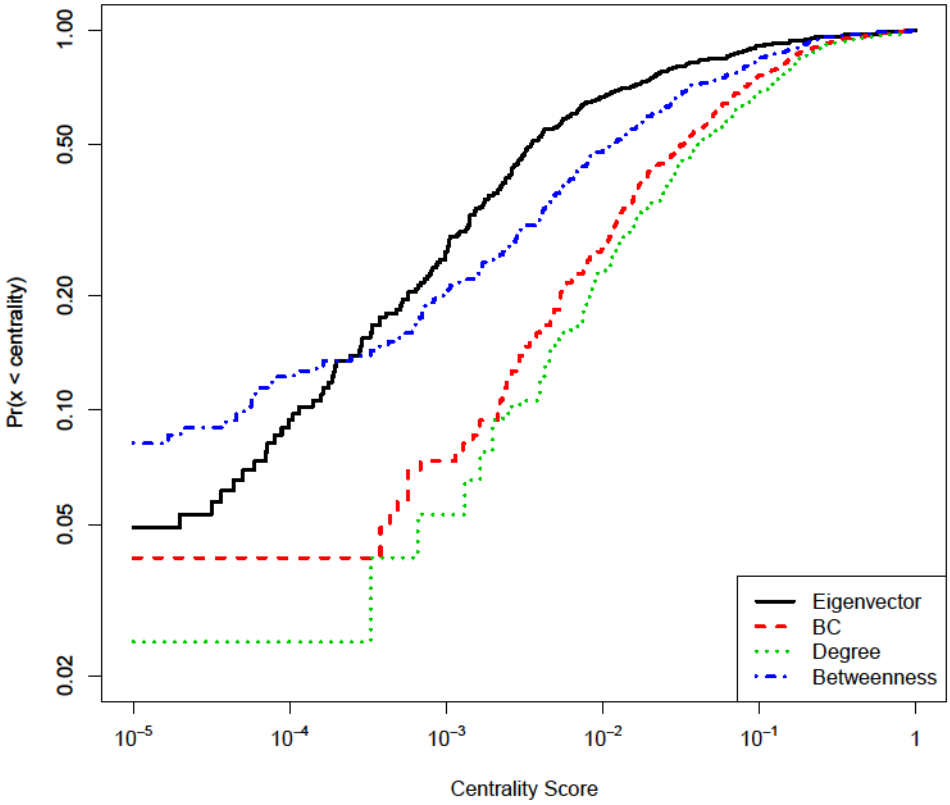

Figure 2.5 – Cumulative distributions of the centrality measures in log-log.

Table 2.2 represents the top 30 centralities ordered by the bridging centrality. We focus on commenting the most salient differences. As highlighted by Figure 2.4, the ranking is clearly dominated by German regions which rank highest for most measures. Interestingly, we find 13 German regions among the 15 best ranked regions for the bridging centrality. This results from the fact that they show both a high participation intensity as well as high openness from an inter-regional perspective; they show a high absolute as well as relative number of inter-regional co-patents. However, the concentration tendency and high clustering of co-patenting activities at the national level of Germany may point to the fact that economic linkages at the national level prevail. Likely explanations are low language / cultural barriers as well as lower transaction costs. These factors seem to promote the high regional bridging centrality in German regions.

Another interesting case is the region of Île de France (FR10) which ranks at the 13^{th} position for the bridging centrality, while being ranked first with respect to its degree centrality. We see that the measure of degree centrality may overstate its position in the inter-regional co-patent network. Despite its highly distributed structure of collaborations (it has a low HH index of 0.04), this region is highly reliant on internal collaborations (the outer share of collaborations is only 45%) that it fails to provide much bridging paths to the inter-regional R&D network. By contrast, the eigenvector centrality may understate the importance of FR10; it ranks only 36 as it is linked to the network core regions at a lower degree. For the same reason as for FR10, some regions that are ranked high in

the degree centrality end up much lower in the BC; i.e. they show high embeddedness in the inter-regional R&D network but are less open and diversified in the structure of their inter-regional collaboration, thus receiving lower values of bridging centrality.

Following the criteria of openness and diversification, interesting is also the case of Brussels (BE10) which ranks after the 56^{th} place for all centrality measures other than the bridging centrality. With the BC, BE10 ranks 30^{th}, gaining at least 26 places compared to other measures. Yet, the SNA-based centrality measures may underestimate its positioning in the inter-regional co-patent network: due to its very high outward orientation (its outer share is 94%) and a highly distributed structure of collaborations (it has a low HH index of 0.07), this region is likely to provide many bridging paths to the network and may therefore be an important bridge for the whole network and for inter-regional knowledge diffusion.

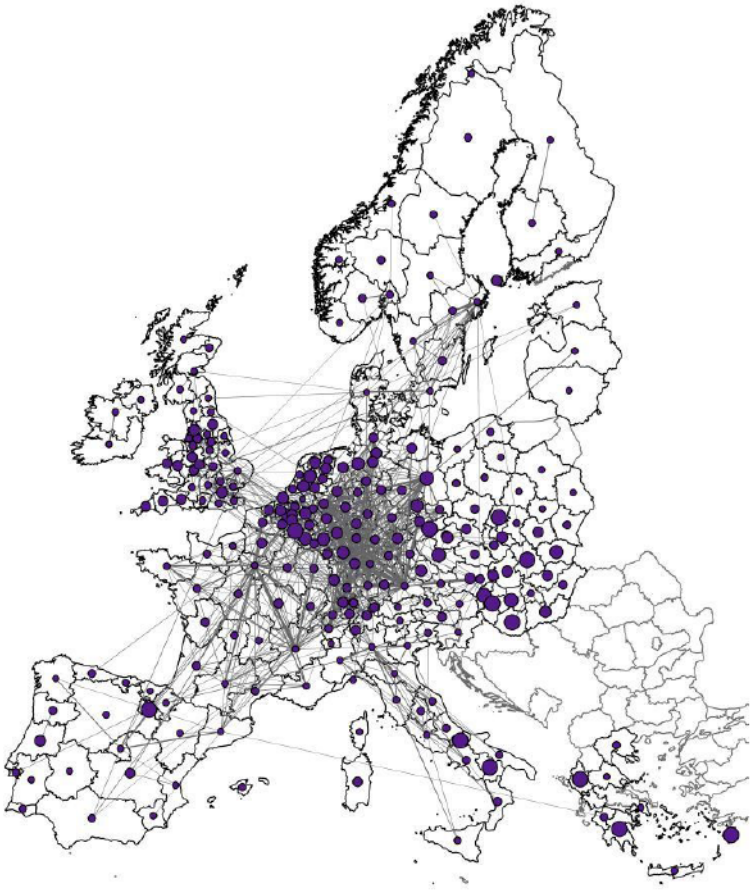

Figure 2.6 – The European co-patent network

Notes: Node size corresponds to the relative outward orientation of a region, line width corresponds to the number of co-patents between two region.

Figure 3 illustrates the European co-patent network for the European NUTS 2 regions, with the node size corresponding to the relative outward orientation of a region. It confirms the very dense network structure between core regions clustered in Germany, which hold intensive connections among each other. From a regional perspective, the bridging centrality is high for these regions, i.e. they yield high values for all three components, despite the fact that most of the links are confined at the national level.

Furthermore, we observe a high relative outward orientation of some South and Eastern European regions. In terms of established co-patent links they seem to be highly open, which could be explained by the lack of internal collaboration structures. Nevertheless, inter-regional linkages are generally weak for these regions.

2.6 Concluding remarks

The notion of centrality is ubiquitous in debates on the role of regions in R&D networks. Quantitative approaches to measure regional centrality, however, are often based on micro-level centrality measures as introduced in social network analysis (SNA). Empirical analysis of regional networks requires accounting for the network structure originally defined at the micro level or by the linkages between different organisations, which often limits the usefulness and conclusive identification of regions in the network. A further unavoidable problem relates to the considerable loss of information regarding network structure and meaning when regions are regarded only as aggregate units. In this study we address this micro / meso-level duality in how we view regional networks and the region's structural network positioning is usually defined, questioning the conventional measurement approaches for region-level analysis.

By introducing the notion of regional bridging centrality we suggest a new approach for assessing the centrality of regions in R&D networks that is able to cope with the regional dimension in measuring the centrality. Based on the concept of bridging paths, i.e. a set of two links connecting three actors in three different regions, we develop a measure of centrality that satisfies the requirements of both R&D networks and region-level applications: A bridging path between regions characterizes a situation where regional actors represent bridges or brokers in the network of regions as they connect indirectly the actors located in two other regions. Such a triangulation in regional networks, as we argue, is a key issue for knowledge recombinations and the extension of a region's knowledge base.

We further show that centrality in terms of bridging centrality can be viewed as a function of (i) the participation intensity in inter-regional collaborations, (ii) its openness to other regions (i.e. the relative outward orientation of network links), and iii) the diversification of links to other regions. With these three components – which are both intuitive and computationally simple – we argue that regional network centrality has to be viewed from a multidimensional perspective. Only with such an integrative perspective we can achieve a better understanding of the role of certain regions in inter-regional R&D networks.

The comparative analysis with three standard SNA-centrality measures confirms the performance and usefulness of our measure of regional bridging centrality. We chose the inter-regional co-patent network for European regions as illustrative example. Despite observing similar patterns in basic statistics like correlations of the centralities or the skewness, we were able to show striking and interesting differences in the structure of the

inter-regional co-patent linkages across regions. The results reveal that thinking only of the degree of participation is not enough. Rather, the most central regions show simultaneously high embeddedness, high relative outward orientation and high diversification of their network links (e.g. Karlsruhe). In contrast, regions that may be strongly embedded (i.e. high participation intensity) may show low openness and diversification of links, thus yielding lower centrality values (e.g. Île de France). Hence, a region's outward orientation and the diversification of its network links moderates the influence of regional scale on network centrality. This is a major strength of the measure proposed in this study, and it paves the way for future studies to examine the role of certain regions in networks of inter-regional knowledge flows. Viewing network positioning of regions in terms of regional bridging centrality might further elevate our understanding of which regions are the most central, show high visibility and at the same time are most important for the network and the inter-regional diffusion of knowledge.

Furthermore, the bridging centrality measure may contribute to the development of a multi-dimensional typology of regions, based on structural network criteria according to their levels of embeddedness, openness and diversification of links in inter-regional networks. Such a typology might enhance our understanding of how different the roles of regions in networks might be, and how they contribute to the arrangement and evolution of the inter-regional structure. This is one of our main points for a future research agenda. Moreover, it seems natural that an application of the bridging centrality measure on other types of knowledge networks according to different technological fields might reveal interesting patterns of the most central network nodes. Hence, the measure of bridging centrality is not limited to the context of R&D collaborations but may prove to be useful also for the application in other types of network structures, such as inter-regional trade flows or inter-regional economic value chains, also regarding their evolution over time.

2.7 Appendix

2.7.1 Obtaining the Bridging Centrality

Assume that the number of agents of region i, n_i, and the number of projects of that region, g_i, are proportional so that $n_i = \alpha g_i$. Then the bridging centrality can be rewritten as follows:

$$
\begin{aligned}
BC_i &= \sum_{j\neq i}\sum_{k\neq i,j} ENB^i_{jk} \\
&= \sum_{j\neq i}\sum_{k\neq i,j} \frac{g_{ij}g_{ik}}{\alpha g_i} \\
&= \frac{1}{\alpha}\frac{1}{g_i}\sum_{j\neq i}\left[g_{ij}\sum_{k\neq i,j} g_{ik}\right] \\
&= \frac{1}{\alpha}\frac{1}{g_i}\sum_{j\neq i} g_{ij}\left(\bar{g}_i - g_{ij}\right) \\
&= \frac{1}{\alpha}\frac{\bar{g}_i^2}{g_i} - \frac{1}{g_i}\sum_{j\neq i} g_{ij}^2 \\
&= \frac{1}{\alpha}\frac{\bar{g}_i^2}{g_i}\left(1 - \sum_{j\neq i}\left(\frac{g_{ij}}{\bar{g}_i}\right)^2\right) \\
&= \frac{1}{\alpha}\bar{g}_i s_i\left(1-h_i\right)
\end{aligned}
$$

Further, as the α is common to all regions, we lose no generality to setting it to $\alpha = 1$. Which yields the result. □

Chapter 3

How does the structure of the inventors network affect regional inventive performance? Evidence from France[‡]

3.1 Introduction

Social and economic networks are involved in numerous mechanisms susceptible to ground agglomeration economies, and their role appears to be reinforced when related to innovation. The literature suggests they are at play in the three types of mechanisms that ground agglomeration economies (sharing, matching and knowledge spillovers, according to the classification developed by Duranton and Puga (2004)). Sharing: Local companies are likely to benefit not only from sharing larger local production factor markets, but also from sharing denser social connections between agents on the factor supply side. Social networks are known to facilitate the circulation of information on the availability of jobs (Granovetter, 1973; Calvó-Armengol and Zenou, 2004), in particular for finding highly qualified jobs (Granovetter, 1995). Concerning the capital market factor, inter-individual networks are also known to be crucial determinants of the syndication behaviour in the venture capital business (Sorenson and Stuart, 2001). Matching: Job search models suggest that interpersonal relations may reduce the costs of acquiring information on potential matches and increase the quality of matches. Hanaki et al. (2010) show that inventors who had professional relations with the employees of his/her new employer are more productive and have longer tenure. Intermediaries in the network may also play the role of reference persons thereby mitigating information asymmetry problems prior to the matches (Montgomery, 1991). Knowledge spillovers: It is probably the mechanism in which social networks have the most prominent role. Marshall (1890) explained how locally generated ideas can be diffused through social and professional interactions

[‡]This chapter is based on an article co-authored with Nicolas Carayol and Pascale Roux.

and thereby gain in being improved and combined with other suggestions. Knowledge spillovers in local environments nearly always involve non-market social relations through which knowledge actually diffuses.[1]

However, Carlino and Kerr (2015), who review studies on agglomeration and innovation, argue that "there is very little insight into how knowledge is transmitted among individuals living in close geographic proximity. Presumably this occurs through both professional and social networks, but this has not been confirmed" (Carlino and Kerr, 2015). Though many case studies have emphasized that social networks ground the performance of clusters (Saxenian, 1991; Porter, 1998) and are likely to sustain knowledge spillovers (cf. Singh, 2005; Agrawal et al., 2006; Breschi and Lissoni, 2009), only a few large empirical studies have investigated the role of social networks on local innovation using explicit social network data. To our knowledge the only exceptions are (Fleming et al., 2007) and (Lobo and Strumsky, 2008), based on nearly identical US patent data from the late seventies to 2002. They regress, at the MSA level, patents counts against network variables built using co-invention patterns and other controls. Both studies findings are somewhat negative concerning the influence of professional networks on regional innovative performance while more traditional agglomeration features prove to be much more related to innovation.

In this study, we aim at reassessing whether inventors networks do favour regional innovation, and if so, how. Building on the literature on network centrality (Katz, 1953; Bonacich, 1972; Brin and Page, 1998; Ballester et al., 2006), we propose simple and flexible micro-foundations, conceptualizing how inventors productivity can be affected by their network of social connections. This model contains three key ingredients that will be tested in the empirical part of the chapter: connectivity, complementarity and rivalry. Connectivity is the most basic assumption. It states that inventors productivity is (hypothetically positively) affected by connections to other inventors. Complementarity posits that an inventor efforts' productivity depends positively on the efforts that his/her partners put into knowledge production. This assumption captures the idea that all partners are not the same: more active neighbours are contributing more to ego's inventive productivity. Rivalry posits that the benefit one can draw from each collaboration is inversely related to the number of connections of that partner. This captures the idea that incoming knowledge or information flows are reduced when partners are more connected. Consequently, rivalry implies that new connections incur some local negative externality. As we will show in this chapter, this very simple set-up implies that, at the equilibrium, inventors produce efforts that are proportional to their centrality in the network (as in Ballester et al., 2006). The form of network centrality implied by the model is new and encompasses existing centrality measures as specific cases, such as the Degree, the Bonacich and the Page-Rank. The main characteristic of this centrality measure is that it is ruled

[1]Otherwise, knowledge may also be transferred through market transactions (e.g. lobar mobility, specialized business services), but then networks are back into the picture through matching mechanisms (Porter, 1990).

by the chosen levels of connectivity, complementarity and rivalry, which will estimated in our empirical exercise, highlighting how networks presumably affect innovation. In particular, complementarity increases the importance of being directly connected to more central agents in the network and therefore captures the positive effect of being connected to stars in the network.

Our empirical evidence is based on panel data of nearly one hundred thousand French inventors and their collaborations for the period 1981-2003 previously cleaned, disambiguated and matched with company mandatory surveys data (Carayol et al., 2015). The analysis is carried out at the regional level, identified by French employment areas (EAs), combined to seven technological fields. Our micro-founded methodology allows us to refrain from regressing regional performance on several network statistics (often highly correlated) in an ad-hoc manner. We estimate a model in which the future patent production of a given employment area×technology is a function of the average network centrality of the inventors of this EA-technology. The structure of the data allows us to include a various set of controls such as EA-technology and time-technology fixed effects.

The results show that, first of all, the inventors productivity are indeed positively affected by their connections in the network (*connectivity*): inventors' centrality has a positive influence on regional innovation. This first result demonstrates that the collaboration network of inventors are not just equilibrium properties that would be shaped by other physical, social or economic conditions. This is for instance implicitly assumed in economic geography models such as the one developed by Helsley and Strange (2004), in which agents meet randomly in cities and these meetings in turn support learning opportunities with some probability or knowledge exchanges. For them, the architecture of the connections is not relevant, neither for meeting probabilities nor for the expected benefits of these meetings. Our results show instead that the architecture of the professional collaboration networks of inventors are not just a by-product of agglomeration but relevant state variables explaining future innovation.

Our second main empirical result is negative, in that we find no evidence of rivalry in the way networks sustain innovation. This finding, which is robust to a long list or robustness checks, is in support of an interpretation of the network benefits in terms of diffusion and contact rather than in terms of shared time and efforts in common projects (which are rival). One implication is that network connections do not imply any negative externality on neighbours. The third main result, although less systematic, supports that there is some complementarity effect at play in the network. This complementarity effect is verified only when either: the most prolific inventors are excluded as recipients of these effects (top-five or top-one percent), only the most innovative EA-fields are considered, only intra-regional networks are accounted for, or larger spatial units of analysis are used (NUTS 3 instead of EA). Lastly, we also show that the efforts-based view of our microeconomic model, which sustains the complementarity of efforts interpretation of previous result, is justified. Indeed, a slight modification of the model, not retaining this

feature, predicts a relation between the network and invention which is not supported by the data.

The next section reviews the literature on collaboration and the benefits provided from the network at the inventor and at the regional level. Section 3.3 presents the model which links inventors' productivity to their network. The data, variables and the econometric strategy are described in Section 3.4. The results are presented in Section 3.5. The last section provides concluding remarks.

3.2 Why would inventors networks matter for regional innovation

In this section, our objective is to explain why and how inventors network should matter for regional innovativeness. Researchers have highlighted the importance of localized knowledge flows as crucial drivers of local inventivity. This is also consistent with a strong assumption made in economic geography according to which knowledge spillovers ground the formation of industrial and innovative clusters (Section 3.2.1). Knowledge flows through (various) interpersonal relations of researchers and/or engineers, which are essentially tied nearby (Section 3.2.2). Finally, there are many reasons to think that, as an "invisible college" of academic researchers exist, communities of inventors would account for inventiveness. Paradoxically, the few studies that have sought to reveal this influence are not conclusive (Section 3.2.3).

3.2.1 Agglomeration, knowledge spillovers and regional innovation

What explains the innovativeness of a region? A first response can be indirectly found in Marshall (1890), followed by Jacobs (1961) and Jaffe (1986), who described three reasons for which productive activities tend to cluster in space: the proximity to dedicated suppliers and services, the presence of a lobar market pooling and the existence of localized knowledge spillovers. Those benefits to agglomeration seem to play fully for innovation activities since it has been proved in many ways and in different contexts (national, sectoral) that innovative (R&D) activities and innovations are even more concentrated than are manufacturing industries (e.g., Audretsch and Feldman, 1996; Carrincazeaux et al., 2001; Buzard and Carlino, 2013).

Recently, Carlino and Kerr (2015) survey theoretical models examining how these mechanisms could operate for innovation. Using Duranton and Puga (2004) taxonomy, they set the reasons why sharing common inputs (such as skilled labour, specialized business services, entrepreneurial finance), benefiting from labour market pooling (which improves the quality of matches and the mobility of workers) and from learning (thanks to

local information diffusion) are each relevant for innovative activities. Unfortunately, the empirical identification of the outcome of each of these mechanisms in terms of knowledge generation is problematic and indeed, there is a severe lack of empirical evidence on these issues (Carlino and Kerr, 2015).

Among those mechanisms, knowledge spillovers are of critical importance in explaining local creativity and innovation. This has been also pointed out in New Growth Theory models, which emphasize the role of knowledge flows as crucial drivers of economic growth (e.g. Lucas, 1988; Romer, 1990). In parallel, abundant empirical evidences, mostly relying on knowledge production functions, highlight that such knowledge spillovers are spatially and technologically bounded (see Jaffe, 1986; Anselin et al., 1997; Orlando, 2000; Autant-Bernard, 2001; Feldman, 1999, for a review). Jaffe et al. (1993) or Almeida and Kogut (1997) have taken a different route to examine how knowledge flows in space, using patents data. They use patent citations to trace the flows of knowledge from one invention to another. They compute and compare the probabilities of patents citing prior patents with inventors from the same metropolitan area against a randomly drawn control sample of cited patents. They show that citations are (according to samples) two, three or even six times more likely to come from the same area than control patents.

3.2.2 Inventors' relationships vs. co-location

If knowledge spillovers mainly occur locally, ideas are not circulating "in the air". Especially when complex, tacit or advanced, their diffusion impose interpersonal relations between researchers and/or engineers which may span firm or institutions boundaries (Fleming and Marx, 2006). According to Saxenian (1996), those relations are often informal. Moreover, the Silicon Valley was characterized by a high mobility of workers across companies. This led, at least partly, to a superior and lasting innovative performance of this region, contrasting with the decline of the Boston-Route 128 cluster. Thus co-location is only part of the story: this has been proven more systematically thanks to the availability of patents and publications data which make (at least in part) visible both these flows of information, through citations, and of collaborations between inventors, through co-patenting or co-publication. Breschi and Lissoni (2003) and Singh (2005), respectively for US and European inventors, show that interpersonal connections are the support of such knowledge spillovers and that knowledge flow decreases sharply with social distance. These authors found that being located in the same area has little or not supplementary impact on the probability of knowledge flow between inventors that already have close network ties. The importance of relations between inventors as support of knowledge transmission was also evidenced by Fleming and Marx (2006) on the basis of interviews of a representative sample of U.S. patent inventors.

Indeed, it turns out that inventors' relationships are mainly tied locally. Using European patent data, Carayol and Roux (2007) analyse the geographic distance separating invent-

ors who co-invent a patent. They found that more than 75% of such connections are achieved between inventors that live at less than 50 km from each other while less than 4% of the connections are formed between agents who live at more than 550 km from each other[2]. One explanation is linked to the costs of maintaining connections which would increase with distance: closely located agents incur lower costs to establish communications and to coordinate, in terms of transporting costs and time. Another explanation is the fact that geographic proximity between inventors and their agglomeration favour the chance of meeting and matching. Thus, geographic proximity or co-location seem to be imperfect proxies of local relationships which generate knowledge spillovers. On the other hand, geographic space is non neutral for the process of relations formation. In sociology and other professional contexts, there is evidence of a positive effects of geographical proximity on inter-organizational relations formation (Powell and Grodal, 2005; Sorenson and Stuart, 2008). Thus relationships and geography are likely to overlap.

There are various channels through which the relations between inventors foster their productivity. Innovation consists fundamentally in a recombination of existing knowledge that eventually leads to a new device or knowledge (e.g. Nelson and Winter, 1982; Basalla, 1988; Henderson and Clark, 1990; Sorenson and Fleming, 2004; Sorenson et al., 2006). Fleming (2002, p. 1072) considers an innovation as any new "combination or rearrangement of components [...], regardless of [...] its success", where the term 'component' is meant to have a broad meaning, encompassing abstract ideas and physical objects. Thus, the role of collaboration can be underlined by this need of recombination. As the set of knowledge a researcher masters is bounded, collaboration makes possible, by combining different knowledge sets, to extend the set of knowledge that is accessible for the team and would have not been attainable alone (see e.g., Arora and Gambardella, 1994; Weitzman, 1998; Fleming, 2002; Jones, 2009; Lee et al., 2015).

In addition, collaboration may have a lasting effect on inventors' productivity. In the collaboration process, inventors can learn new skills or benefit from the specific knowledge possessed by their collaborators (Bercovitz and Feldman, 2011). For instance, in a survey on scientists' incentives to collaborate, Freeman et al. (2014, table 4) report that, for more than 85% of the sample, "learning from each other" was an essential motivation to collaborate. Furthermore, social interactions, mainly in the form of face to face contacts, are the only medium to allow the sharing of tacit knowledge embodied in scientists and not existing in a codified form (Dasgupta and David, 1994; Cowan and Foray, 1997). Thus, only such interactions allow to draw from a specific, not available to all, set of knowledge. This echoes to the idea that some specific sets of tacit knowledge are akin to 'club goods' so that the only possible way to access this knowledge would be from collaboration with scientists from this club (Breschi and Lissoni, 2003).

Furthermore, while there are immense possibilities of knowledge recombinations, only

[2]They also found that most geographically mobile inventors remain in the same area: nearly 86% of mobile inventors have a maximal distance between their different locations which is less than 50 km.

a few may yield any important commercial outcome. From the inventor's or firm's viewpoints, as research is costly, it is then essential to efficiently identify which idea may lead to a valuable innovation. In this situation, interactions plays a critical role since they allow the individuals to enter a process of quick exchange and assessment of ideas implying that low value ideas are more efficiently spotted and sifted out (e.g., Fleming et al., 2007; Singh and Fleming, 2010; Lee et al., 2015). Stated differently, inventors must steer their talent and energy to the right direction, otherwise, their efforts could lead to dead ends or new knowledge that is not valuable (Singh and Fleming, 2010).

Learning effects combined with a better selection of ideas brought about by collaboration imply that inventors benefiting from many social connections can have a better knowledge of which ideas or future research paths can be valuable. In other words, social connections also allow the inventors to be more aware of which direction to sail in the "uncharted sea of technological possibilities" (Schumpeter, 1943, p. 103). These effects may be lasting beyond the collaboration term.

Finally, the social connections may also foster the inventors' future productivity through another channel: they also constitute a repository of information in which individuals can draw information on possible partners (e.g., Gulati and Gargiulo, 1999; ter Wal, 2014). This in turn may help them in finding future partners that are likely to be good matches, and subsequently increase the productivity of their future collaboration.

3.2.3 Inventors networks and regional innovation

Some authors have started to investigate theoretically the impact of the full network of research relations on the innovation within regions or systems. They aim to identify which network characteristics are most relevant to explain firms' or regional innovative performance, looking at both their local and global structural properties. Some of them focused on the "small-world" property of social networks (Watts and Strogatz, 1998), which is the coincidence of high local clustering and short global social distance separating agents. Cowan and Jonard (2004) model knowledge diffusion between agents who barter different types of knowledge. They examine the relationship between the knowledge network architecture and diffusion performance. They found that the efficiency of knowledge transmission is affected by the global architecture of connections among agents: diffusion performance is maximal when the network exhibits both high local clustering while some few relationships are long distance. It is generally argued that this tension between in the one hand, local clustering which fosters communication and enforces cooperation, and on the other hand, distant connections which bring non-redundant connections, is at the core of the creative outcomes of small worlds (e.g., Fleming and Marx, 2006).

Using the analogy of innovation as a recombination process helps us to see this point. Inventors of a given region develop inventions based on their set of existing knowledge. If they do collaborate only internally to the region, the set of knowledge to be recombined

may end up to be redundant, and this in turn may exhaust the set of possible valuable innovations (Berliant and Fujita, 2008, 2012). The network can compensate this lack of renewal of existing knowledge if some agents do collaborate with other regions (Bathelt et al., 2004). These external (to the region) connections can bring in new possibilities of innovation, as the knowledge available in other regions is likely to be differentiated. Berliant and Fujita (2012) theoretically studied the development of knowledge embodied in scientists across regions. They show that inter-regional connections are critical to enhance the productivity of the economy by combining different knowledge sets. Thus, the inventors network matter to regional productivity as linkages to other regions may renew the knowledge base of the region and benefit regional innovation.

Furthermore, the density of the regional network is important since it can help ideas to flow more easily across the connections of the network, increasing regional productivity. Indeed, collaborations allow the inventors to be more efficient to select valuable ideas and to be more aware of the valuable directions of research. These two benefits of collaboration can diffuse along social connections (Sorenson and Fleming, 2004; Sorenson et al., 2006) and make the whole network more productive. All the more, social connections can help the scientists to be more aware of who could be a good partner for them. In this vein, the denser the network, the more each scientist can draw information on their possible partners. Therefore, the allocation of inventors across teams may be more efficient overall in regions with dense networks, consequently increasing future regional productivity.

Though we have good reasons to expect that the whole web of relations matters for regional innovation. Paradoxically, the few empirical studies examining this link have not been conclusive. Fleming et al. (2007) have investigated whether regions whose internal inventors network displayed a "small-world" structure are more inventive than others. They rely on US patent data from the late seventies to 2002 and regress, at the metropolitan statistical area (MSA) level, patents counts against network variables and other controls. They find no evidence of such pattern: social average distance is negatively correlated with innovation while clustering and the interaction between the two variables are not significant. On nearly identical data, (Lobo and Strumsky, 2008) more explicitly study the separate effects of inventors agglomeration and of their collaboration networks on local patenting behaviour. They find that when agglomerative features of the MSA are controlled for, structural characteristics of the network have small effects on metropolitan patenting. Moreover, they find a slightly significant negative effect of density of the inventors network on regional innovation. Different studies in other contexts (scientific or artistic productions for instance) are not more conclusive (e.g., Uzzi and Spiro, 2005; Guimera et al., 2005; Smith, 2006).

Those puzzling results challenge the way social networks and their outcomes for individuals are apprehended. In the following section, we thus propose a model allowing us to go inside the black box of social networks and to examine how they matter for local innovation. These micro-foundations will guide our empirical investigations.

3.3 A model of patent production at the inventor level

This section intends to provide a simple view of how inventors' productivity may be affected by their social connections, and hence how the network structure may influence innovation. First Section 3.3.1 describes the model and Section 3.3.2 illustrates its implications.

3.3.1 The model

The model developed in this section is close to the one in Ballester et al. (2006), in which each agent's productivity is linked to his/her network. It is stylized and integrates only network-related characteristics. We implicitly assume that the effect the network has on inventors' productivity is independent from any other determinant of productivity and are thus discarded from the present analysis.[3] Later on, in the empirical section, other factors that may affect inventor's productivity will be controlled for.

Consider a network of n nodes, where the nodes represent inventors and the connections linking the nodes are interactions between inventors. These interactions can be seen as professional connections based on past or present collaborations. The set of all nodes and links can be represented by the symmetric matrix g, which ith line and jth column entry $g_{ij} = 1$ if inventors i and j are linked, and $g_{ij} = 0$ otherwise. Self-relations are excluded ($g_{ii} = 0$). Further, if $g_{ij} = 1$, then inventors i and j are referred to as *neighbours* and $N_i \equiv \{j \,|g_{ij} = 1\}$ represents the set of all i's neighbours. Further, the number of links of an inventor is noted $d_i \equiv \sum_{j=1}^{n} g_{ij} = \#N_i$ and is referred to as i's *degree*.

Let y_i be the inventive productivity of agent i, also considered as equal to i's gross payoffs. It is modelled in a simple fashion, relying on the one hand upon the efforts he/she exerts to produce inventions, noted e_i, and, on the other hand, upon the productivity of these efforts, noted ψ_i. We consider the most simple and intuitive form to combine these two components into y_i, a multiplicative form, so that:

$$y_i = e_i \cdot \psi_i. \tag{3.1}$$

Turning to the utility function, we assume, as in Ballester et al. (2006), that the amount of effort exerted has a negative and convex effect on utility.[4] This leads to the following utility function:

$$u(e_i, \psi_i) = e_i \psi_i - \frac{e_i^2}{2}. \tag{3.2}$$

[3] This approach is comparable to the one developed by Calvó-Armengol et al. (2009), who, studying peer-effects at school, assumed that pupils could produce two kind of efforts: network-related effort and non-network-related effort. Similarly, we are interested in network-related efforts only.

[4] Note that including a parameter in the utility function which sets the negative effect of effort would imply no change in the results, as shown in Appendix 3.7.2.

The main ingredient of the model is the productivity component, which rely upon the inventor's network. As seen in the previous section, network connections are beneficial to producing innovation through various mechanisms. The way we conceptualize productivity is based on two main ideas. First, productivity can be increasing with the efforts of the connected agents. Whatever the type of spillover involved – be it the learning, diffusion or selection of ideas – the increase in the connected agents' efforts in invention generation enhances the returns from one's own efforts. This implies that efforts would be complementary. Second, the benefits gained from collaboration can be decreasing with the partner's degree. This idea is very similar to the one in the co-author model of Jackson and Wolinsky (1996). In their model, because of time constraint, each author efforts are divided among all the projects he/she is involved in. Thus each new connection involves a negative externality on partners. We associate this idea to the notion of rivalry.

Finally, we integrate those mechanisms in this following simple expression for productivity:

$$\psi_i(g, \mathbf{e}) = 1 + \lambda \sum_{j \in N_i} \frac{e_j^{\alpha}}{d_j^{\beta}}, \tag{3.3}$$

where $\mathbf{e}$ is the vector of all efforts, and λ, α and β are parameters. We suppose that, when not connected to anyone in the network, any inventor has the same productivity level, normalized to the unity.[5] The parameters λ, α and β each has an influence on inventors' productivity and each carry a different meaning. We hereby summarize the three parameters with the perspective of which kind of network structure they underlie:

λ **(Connectivity):** scales the benefits to productivity stemming from the network. If $\lambda = 0$, the inventor's network has no effect on his/her productivity. The higher λ, the higher the benefits drawn from the network as compared to the autonomous part normalized to one.

α **(Complementarity):** scales the benefits to productivity stemming from the complementarity of efforts. The higher it is, the more one will benefit from the partner's effort. It should be noted that the effort of one agent may 'spread' its external benefits further than his/her direct connections. Indeed, an increase in the effort of one inventor will rise the productivity of his direct connections, which will lead them to increase their efforts which will in turn rise the productivity of their neighbours, etc. If $\alpha = 0$, then the partner's position in the network does not matter whatsoever. It would imply that the benefit stemming from any partner would be identical.

β **(Rivalry):** scales how much the access to one's effort is rival. A high value of this parameter means that the benefits stemming from the network are in fact rival, so that highly connected agents will only slightly benefit to their colleagues. On the contrary, in the case where β is low, for instance $\beta = 0$, then there is no rival effect:

[5] Note that setting the default productivity level to any other value has not implication on the results, as shown in Appendix 3.7.2.

the external effect will fully benefit to each of his/her collaborators. In this line of thought, a low value of β could be interpreted as if the flows involved in network connections were more related to the diffusion of information and contact processes than really involving shared time spent together.

We now look at the equilibrium efforts and subsequent inventors' productivity. If each inventor maximizes his/her utility while taking the effort of all other inventors as given, then the Nash equilibrium yields the following equilibrium: $e_i^* = \psi_i(g, \mathbf{e}^*), \forall i$, and network-related production y_i^* thus equals $y_i^* = \psi_i^2(g, \mathbf{e}^*)$. It turns out that[6]

$$y_i^*(g, \lambda, \alpha, \beta) = c_i^2(g, \lambda, \alpha, \beta), \tag{3.4}$$

where $c_i(g, \lambda, \alpha, \beta)$ is a centrality measure that depends only on the position of the inventor within the network and on the three parameters λ, α and β. This network centrality is defined by the following relation:

$$c_i(g, \lambda, \alpha, \beta) = 1 + \lambda \sum_{j \in N_i} \frac{c_j^\alpha(g, \lambda, \alpha, \beta)}{d_j^\beta}. \tag{3.5}$$

This form of centrality links the centrality of i to the centrality of his/her partners in a fashion closely related to the Bonacich centrality. In fact, the centrality measure defined by Equation (3.5) is a generalized form of centrality that includes various forms of classical centrality measures. Depending on α and β, this centrality relates to either the degree centrality (Bavelas, 1948), the Bonacich centrality (Bonacich, 1987) or the Page-Rank centrality (Katz, 1953; Brin and Page, 1998). The relation with the existing forms of centrality is given by Table 3.1. These existing centrality measures can be obtained for different values of parameters α and β: for instance, the degree centrality can be seen as a situation in which there is no complementarity nor rivalry at play, while at the other end the Page-Rank centrality refers to a situation in which both complementarity and rivalry occur.

Finally, an interesting property of the centrality defined by Equation (3.5) is that, contrary to the Bonacich or the Page-Rank which are defined only for a limited set of λ,[7] whenever $\alpha < 1$, this centrality is defined for any $\lambda \geq 0$.[8]

The main insight stemming from the model is that, if there is any effect from the network ($\lambda > 0$), inventors should be influenced by their network centrality. This centrality in turn can favour different kinds of network positioning depending on the parameters α and β. The main endeavour of the remaining of the chapter will be to use these model's

[6] The proof is given in Appendix 3.7.1.

[7] More details on the restriction on λ are given in Section 3.4.3, describing the empirical construction of the variable.

[8] The proof is given in Appendix 3.7.3.

Table 3.1 – Summary of the existing centrality measures to which the centrality depicted by Equation (3.5) is equivalent, depending on the values of the parameters α and β. The last column provides the formula of the centrality when the parameters α and β are set as in the first column.

(α, β)	Centrality name	Related formula
$(0,0)$	Degree centrality	$c_i = 1 + \lambda d_i, \ \forall i$
$(1,0)$	Bonacich centrality	$c_i = 1 + \lambda \sum_{j \in N_i} c_j, \ \forall i$
$(1,1)$	Page-Rank centrality	$c_i = 1 + \lambda \sum_{j \in N_i} \frac{c_j}{d_j}, \ \forall i$

results to test: 1) if the inventors network affect regional innovation and if so 2) which kind of network structure most favours innovation.

Next subsection illustrates the link between α, β, and network position.

3.3.2 Illustrations

In this section we first illustrate which type of individual network-position is ranked highest along different types of centrality. Then, using data on co-inventions, the centrality of regional networks is illustrated.

To illustrate the differences in network structure implied by the parameters, we now compare the centrality values of the agents of the network represented in Figure 3.1. This figure depicts a network consisting of 7 agents, four of which are fully connected with each others (agents 2, 3, 4 and 5) and two of which have only one connection (nodes 6 and 7). Agent 1 is connected to these two groups.

We now discuss the centrality measures with respect to the values of α and β, considering the four types of agents: 1, 2, 3 and 6. The centralities are computed with $\lambda = 0.2$ and are reported in Table 3.2a. Assume $(\alpha, \beta) = (0,0)$, so that there is no complementarity nor rivalry in the network, i.e., the benefits stemming from any connection is the same. The centrality of the agents then rely only on their number of connections. In this case, agents 1 and 2 have four connections and therefore have the highest centrality, followed by agents 3 and 6.

Now consider the case of complementarity with no rivalry: $(\alpha, \beta) = (1,0)$. In this situation, one's productivity depends positively on the partner's effort, and as the rise

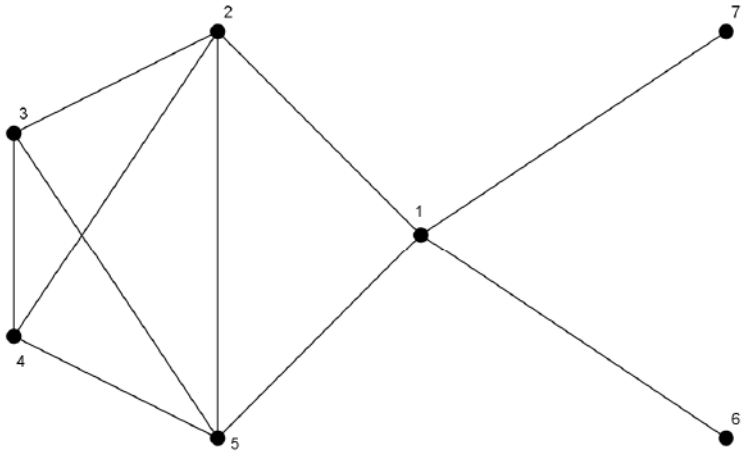

Figure 3.1 – Stylized example of an inventor network.

in productivity spurs one's effort, this new effort will in turn increase the partner's productivity, which will increase the partner's effort, etc. Therefore, complementarity implies that interconnected agents are the ones who benefit the more from this type of centrality. This means that, despite having the same number of connections as Agent 1, Agent 2 will have the highest centrality because he/she lies in a clique (a fully connected subnetwork).

Finally, take the case in which both complementarity and rivalry occurs: $(\alpha, \beta) = (1,1)$. Although there is complementarity, the rival effect means that the benefit from the network decreases with the degree of the partner. Thus, rivalry implies that connections to isolated agents are more beneficial than connections to agents in a clique. Hence, Agent 1 will have the highest centrality in this situation.

So, as the example shows, the network position which provides the highest benefit from the network is ruled by the parameters α and β. The question we will ask in this study therefore is: What is the type of position which favours the most innovation? The idea is to assess whether a region which has more inventors in a given network-position performs better than others.

Let us now concretely illustrate this point with Figure 3.2, which depicts three regional innovation networks. These network are based on co-inventions in the technological field of 'chemicals' for the period 1991–1995.[9] The inventors are the nodes of the network; the ones from the region are coloured in red. All the links to the inventors of the area-technology are represented. Inventors from other technological fields are pictured as triangles.

In the simple model we introduced earlier, the network-related production of each inventor was equal to their squared centrality. Thus, for each geographic area, we compute the average squared centrality of their inventors. This centrality refers to the position of

[9]For more information on the data, see Section 3.4.1.

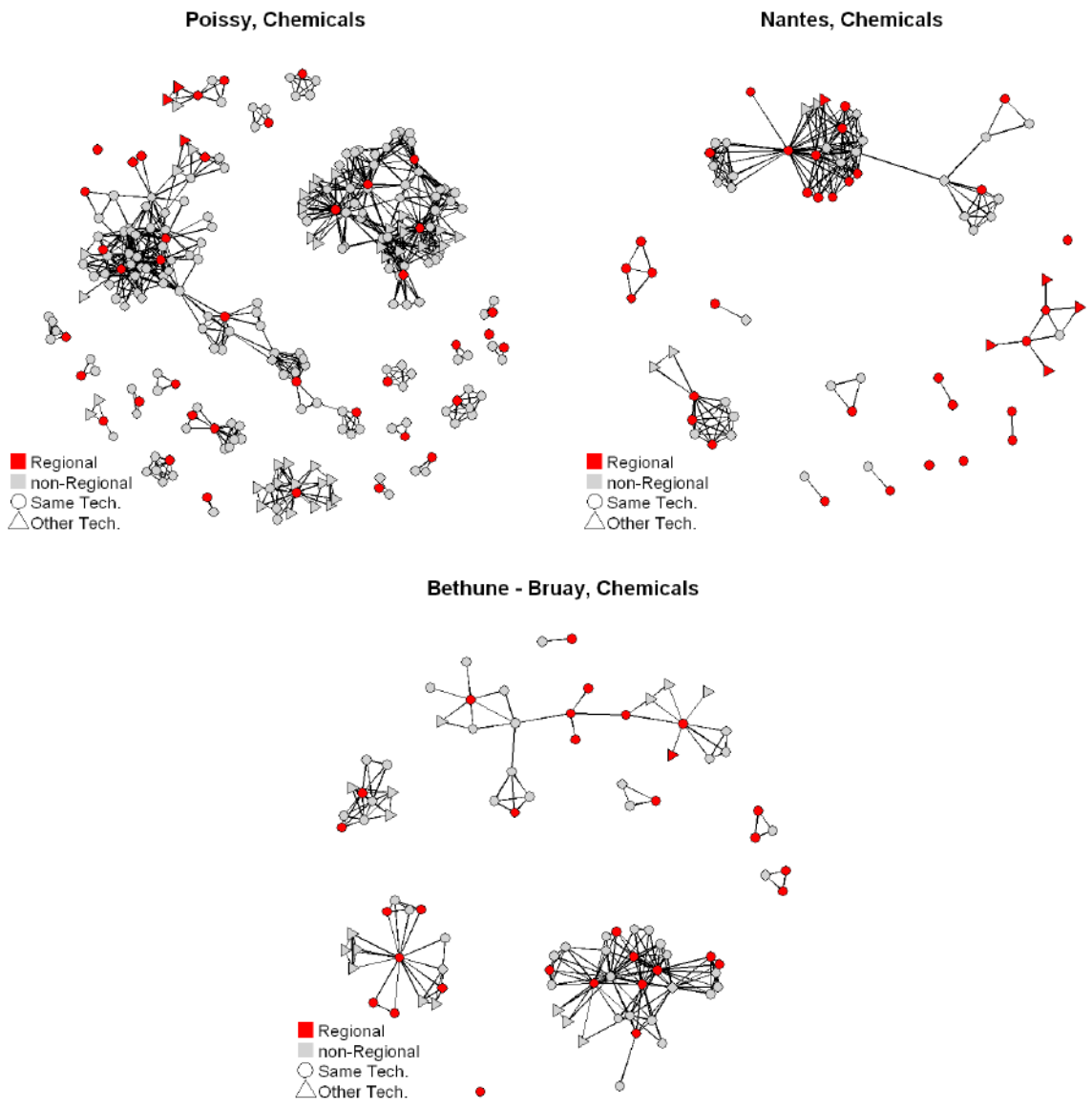

Figure 3.2 – Illustration of three regional networks.

Table 3.2 – Centrality measures.

(a) Centrality measures, as defined by Equation (3.5), for the nodes depicted in Figure 3.1. The value of λ is set to 0.2.

Agent Type	Degree	Bonacich	Page-Rank
1	**1.8**	3.05	**1.56**
2	**1.8**	**3.51**	1.30
3	1.6	3.00	1.21
6	1.2	1.61	1.07

(b) Average squared centrality of the regional networks depicted in Figure 3.2. The value of λ is set to 0.04.

Region	Inventors	Degree	Bonacich	Page-Rank
Poissy	40	**1.69**	2.86	1.10
Nantes	33	1.45	**3.44**	1.07
Bethune Bruay	31	1.51	1.81	**1.11**

the regional inventors in the *global* network, when all the links between all inventors are taken into account. The centrality for the three regions are given in Table 3.2b.

As shown by the table, each of these areas favours a different type of centrality. The area of 'Poissy' has the highest degree, the highest Bonacich is found in 'Nantes' while inventors from 'Bethune Bruay' have the highest Page-Rank centrality.

3.4 Data and econometric strategy

3.4.1 Data

Our empirical evidence is based on the French co-invention network. The data consist of all European patent applications of which at least one inventor declared an address in France, and which were first applied for between January 1981 and December 2003. Using patent data raises a cleaning issue due to homonymy problems on inventors' names or to spelling errors. Indeed, the proper identification of inventors cannot be neglected since small identity errors are likely to imply significant changes in network measures. For instance, homonymy errors leading to consider that two different persons are the same can lead to erroneously link different communities of inventors. Several matching algorithms have been developed to cope with the problem of inventors' identification (for a review, see Miguélez and Gòmez-Miguélez, 2011). The inventors' disambiguation is done using the methodology developed by Carayol and Cassi (2009). They introduced a Bayesian

methodology for estimating the probability that two inventors are the same given a series of observables provided by the data, and then calibrated their model using a benchmark of inventors surveyed one by one.

Network from patents. The model of patent production is based on interactions between knowledge workers. However, these interactions cannot be directly observed. Instead, as routinely done in network studies using patent data (see e.g., Singh, 2005; Agrawal et al., 2006; Carayol and Roux, 2007; Fleming et al., 2007; Lobo and Strumsky, 2008; Breschi and Lissoni, 2009), the network of social interactions is indirectly drawn from the patent records. The underlying assumption is that all inventors appearing in a patent record did interact with each other; or stated differently, co-inventions necessarily imply interactions. Thus, two inventors will be connected in the network if they both appear in a patent record. The two drawbacks of this assumption are: 1) false negatives: interactions occurring with knowledge workers that are not in the patent document are ignored and 2) false positives: two inventors may appear in a patent document without having significantly interacted. In particular, the case of very large patenting teams is a concern for the second problem. While assuming 'interactions from co-inventions' seems a plausible stance for inventions involving a few inventors, this is much more problematic for patents involving large teams. Because the number of connections evolves with the square of the team size (as every pair of inventor of the team is connected), then large teams imply a very large number of connections, and by then is likely to introduce many false positive relations. In consequence, large teams may artificially (and wrongly) inflate the density of the network. To deal with this problem, we decided to withdraw from the sample every patent having strictly more than 8 inventors. All the more, very large teams is a peculiar pattern in our sample: by doing so, we exclude less than 0.2% of all patents.[10] Finally, the patents database consists of 124,825 patents and 97,287 unique inventors.

Spatial unit. As we are interested in investigating local inventive performance, the spatial unit requires to be defined. The most accurate regional aggregation unit is the French 'zones d'emploi' which corresponds to employment areas (EA). These areas are statistical constructs and are not limited by administrative boundaries. They are constructed such as to maximize the share of inhabitants who live and work in companies located in the same EA.[11] They usually take the form of an area surrounding large towns and can be related to U.S. metropolitan statistical areas, although being much smaller in terms of geographical area and population.[12] As inventors personal addresses are given in patent data, we were able to associate to each inventor in each patent a metropolitan French

[10] In total, only 227 patents had more than 8 inventors and were thus excluded.

[11] Continental France is split in 297 'zones d'emploi'. In 2012, the average share of inhabitants working in the same area as defined by the 'zone d'emploi' was 77.5%. More information on those statistical areas can be found in Jayet (1985). In this chapter we use the 2010 definition of the 'zones d'emploi'.

[12] The average area of the EAs is 1,818 km^2, and the average population is 82,05.(INSEE, 2010)

town, and then to an employment area. Though some inventors are geographically mobile (about twelve thousand), they however mostly remain in the same areas: nearly 79% of mobile inventors have a maximal distance between their different locations of less than 20 km. When inventors are mobile across EAs, they are assumed to be fully associated to each area.

Technological fields. In order to account for specific patenting schemes depending on the technological field, the empirical analysis will be carried out controlling for technological fields. When filing a patent to the EPO, the patent holder has to assign it to one or several technological classes which correspond to an international patent classification (IPC) code. The IPC codes are then transformed into 7 aggregate technological fields according to the Observatoire des Sciences et Techniques (OST).[13]

The statistical unit of analysis is the area-field, i.e. the EA combined to the technological field.

3.4.2 Empirical model and estimation procedure

3.4.2.1 An aggregated area-field production function approach

The production of innovations of each area-field is assumed to follow similar patterns. As in standard regional knowledge production function approaches, the innovative outcomes are obtained from a common production function with similar elasticities across units of observation (Fleming et al., 2007; Lobo and Strumsky, 2008). The main input considered here is the contribution of inventors which is assumed to constitute the backbone of innovation production. The basic relation describing the area-field innovation production is thus given by the following equation:

$$Y_{a,f} = A_{a,f} \cdot l_{a,f}^{\gamma}, \tag{3.6}$$

in which $Y_{a,f}$ is the area-field innovation output, $A_{a,f}$ contains all specific factors affecting regional production and $l_{a,f}$ is the sum of the contributions of all inventors associated to the area-field. γ is a positive parameter.

If we let $invSet_{a,f}$ (resp. $Inv_{a,f}$) denote the set (resp. the number) of inventors associated to area a and technological field f, we are able to introduce the individual inventive contributions of local inventors as defined in the previous section, as follows:

$$l_{a,f} = \sum_{i \in invSet_{a,f}} e_i \psi_i(g, \mathrm{e}). \tag{3.7}$$

The efforts agents exert and the direct influence of their neighbours are however typically

[13] The 7 categories, referred to as OST7, are: 'Electronics', 'Instruments', 'Chemicals', 'Drugs, Medicine', 'Industrial process', 'Machinery, Transport', 'Consumer goods, Construction'. More information on the transition between IPC to OST7 can be found in OST (2010).

not observable in the data. Our micro-foundations suggest that they are affected by their network so that, in the equilibrium, they are equal to the square of agents network centrality (cf. Equation 3.4), which depends on the inventors network and on the three parameters of connectivity, complementarity and rivalry. In the equilibrium, we thus have:

$$l_{a,f} = \sum_{i \in invSet_{a,f}} c_i^2(g, \lambda, \alpha, \beta) \tag{3.8}$$

In this equation, it is clear that variable $l_{a,f}$ is in fact the combination of two elements: 1) a size effect, since it is the sum over all inventors associated to the are-field, and 2) a network effect, since the production of each inventor is assumed to depend on their network position. However, as Bettencourt et al. (2007) have shown, the number of inventors is a major determinant of patent production and is tightly linked to regional size. If one wants to identify the network effect, it should be separated from the size effect. In consequence, the variable $l_{a,f}$ is broken down as the product of the number of inventors in the area-field $Inv_{a,f}$ and their average equilibrium contribution, noted $\overline{c_{a,f}^2(g, \lambda, \alpha, \beta)}$, and both are assumed to have distinct exponents in the area-field innovation production function, as follows:

$$Y_{a,f} = A_{a,f} \cdot Inv_{a,f}^{\sigma} \cdot \left(\overline{c_{a,f}^2(g, \lambda, \alpha, \beta)}\right)^{\tau}. \tag{3.9}$$

3.4.2.2 Estimation procedure

As described in the next section, regional production will be measured in terms of patent-citations, which is a count variable. A natural way to estimate the model is to use a Poisson model. Indeed, contrary to linear models which lead to biased coefficient estimates when dealing with count data, a Poisson model copes suitably with this issue and also allows dealing with over-dispersion issues (see e.g., Santos Silva and Tenreyro, 2006). Further, to limit the problem of omitted variables and so as to fully exploit the panel data structure of our evidence, we employ a fixed effect Poisson estimation where the unit of observation is the area-field, so that every unobservable effect specific to the area and the technology will be controlled for. Further, time-field dummies are also accounted for to cope with systematic time changes in the production patterns of each technological field over time. To avoid endogeneity issues, the dependent variable is in $t+1$ so that the explanatory variables explain the production of the subsequent year (as in Fleming et al., 2007). Finally, the equation that will be estimated is:

$$E\left(Y_{a,f,t+1}\right) = d_{a,f} \cdot d_{f,t} \cdot Controls_{a,f,t} \cdot Inv_{a,f,t}^{\sigma} \cdot \left(\overline{c_{a,f}^2(g_t, \lambda, \alpha, \beta)}\right)^{\tau}, \tag{3.10}$$

where $Y_{a,f,t}$ is the dependent variable, and $d_{a,f}$ and $d_{f,t}$ are area-field and time-field dummies for the EA a, the technological field f and the year t. The dummies allow us to control for any effect specific to the area-field and for any yearly variation related to the technological fields (see Agrawal et al., 2014; Menon, 2015). The variables in $Controls_{a,f,t}$

are all other determinants of regional patent production, and include agglomeration economies variables as well as other network-related controls. As the explanatory variables of the model enter the Poisson regression in a logarithmic form, we add 0.01 to the variables whose value is 0 (similarly to Fleming et al., 2007). We have introduced the time reference in regressors $Inv_{a,f,t}$ and $\overline{c^2_{a,f}(g_t, \lambda, \alpha, \beta)}$ which effects are estimated separately. The precise empirical construction of these variables is detailed further in the next subsection.

3.4.3 Variables

First, a foreword on some general features of the dataset. The dependent variable is created such as to not occur simultaneously to the explanatory variables. Indeed, in this study, the aim will be to predict the future regional innovation production based on past characteristics. Turning to the explanatory variables, the network-based variables will be constructed using a five-year-window, from $t-4$ to t. This period of time is used to gather enough information on the network patterns of the area-fields, as patenting can be considered as a rare event (Lobo and Strumsky, 2008). Last, the dataset will be composed of all area-fields and all years. For some area-fields, it is possible that there is no patent produced in the five-year-window, such that some network-based variables cannot be created. When this occurs, we set these variables to 0 (see Fleming et al., 2007).

Dependent variable. The measure of a region's innovation output will be drawn from patent data. However, patent count alone may not be sufficient to account for innovation as the value of patents is of great variability (see e.g., Trajtenberg, 1990; Lanjouw et al., 1998; Hall et al., 2005). A way to account for patent quality is to measure how much the knowledge embodied in a patent has been used in later innovations. When a patent is applied for, it has to reference the sources of knowledge that were necessary to produce this innovation (see e.g., Criscuolo and Verspagen, 2008, for a review). In consequence, those references to earlier work (citations) reveal which patent has been useful in generating new innovations. Further, the positive relationship between patent value and citations received has been demonstrated in various studies (see e.g., Trajtenberg, 1990; Harhoff et al., 1999; Hall et al., 2005). Thus, in order to have a finer grained measure of innovation, each patent will be weighted by 1 plus the number of citations it receives, in line with various studies dealing with regional innovation (e.g., Agrawal et al., 2014; Kaiser et al., 2015b).

A five years window is used to construct the number of citations a patent receives, allowing this number to be comparable across patents from different years. As the most recent patents from our sample are from 2003, we need information on citing-patents until 2008. Further, as the aim is to depict a patent's quality, the citing-patents should not be restricted to French patents only. The citations-related data is drawn from another

data set: the CRIOS-Patstat database which compiles data on all EPO-filed patents (see Coffano and Tarasconi, 2014, for a description of the database).

The number of citations a patent filed in year t receives (the cited-patent) is defined as the number of EPO-patents whose application-date lies between t and $t+5$ and that cites the application number of the cited-patent.[14] Further, in order to avoid citations that do not stem from the 'usefulness' of the patent but rather from other factors unrelated to quality, we withdraw every citation coming from patents either from the same inventor or from the same company.[15] The location of the patents is based on the inventors' addresses, so that the dependent variable for area a and technological field f will be the number of citation-weighted patents filed in year $t+1$ in technological field f that have at least one inventor located in area a.

Centrality. In the model, the equilibrium production of the inventors depend on their squared centrality. This centrality in turn depends on three parameters which provide information on how the network influence inventors' productivity. Regional centrality measures will be computed for different values of those parameters to unveil which network pattern most favours innovation in the area-field.

The variables of centrality at the regional level are built in two steps. In the first step, the centrality of all inventors is computed by using the whole co-invention network on a 5 years window. This defines g_t the network between all the inventors having patented between year $t-4$ and t, no matter the technological class they patented in, built by assigning a link between two of these inventors if the patented together during that period. Then, from this network and after setting the parameters α, β and λ, we compute the centrality of each inventor $c_i(g_t, \alpha, \beta, \lambda)$ according to Equation (3.5).

Depending on the value of α, we use two different methods to compute the centrality. When $\alpha = 1$, the centrality has a closed-form and can be obtained as:

$$\mathbf{c}(g, \lambda, \alpha = 1, \beta) = \left(I - \lambda \tilde{G}(\beta)\right)^{-1} \mathbf{1},$$

where $\mathbf{1}$ is a n vector of ones, n being the number of inventors; $\mathbf{c}(g, \lambda, \alpha = 1, \beta)$ is the vector of all centralities; and $\tilde{g}(\beta)$ is the matrix of typical element $\tilde{g}_{ij}(\beta) = g_{ij}/d_j^{\beta}$, where g_{ij} equals 1 if i and j are connected and 0 otherwise. For instance, $\tilde{g}(\beta)$ is equal to the adjacency matrix for the Bonacich centrality and the column standardized adjacency matrix for the Page-Rank centrality. Further, when $\alpha = 1$, there is a restriction on λ, as the centrality is not defined when λ is greater than the inverse of the largest eigenvalue

[14] The 5 years window is accurate to the day. As the day, month and year of application are available for each patent, we are able to keep only the citing-patents which were filed no later than 1,825 (365×5) days after the cited-patents.

[15] Thanks to the algorithm from Pezzoni et al. (2014), each patent of the CRIOS-Patstat database has an identification number for the inventors who filed it and the companies which own it. The 'self-citations' were cleaned thanks to those identification numbers.

of the matrix $\tilde{g}(\beta)$.[16]

When $\alpha \in [0,1[$, the centralities are computed using an iterative algorithm. Let c_i^k to denote the centrality of inventor i at iteration k (the parameters α, β and λ are omitted for readability). Each centrality is first initialized to 1, such that $c_i^0 = 1,\ \forall i$. Then the following calculus is performed until numerical convergence:[17]

$$c_i^{k+1} = 1 + \lambda \sum_{j \in N_i} \frac{\left(c_j^k\right)^\alpha}{d_j^\beta}.$$

At the end of the process, each inventor's centrality respects the definition of Equation (3.5) up to a negligible numerical error. Contrary to the previous case, the only restriction on λ is that λ is only required to be positive.

The second step is to aggregate these inventors' centralities at the regional level. Each inventor is assumed to contribute to the centrality of each area-field he/she has patented in. So the equilibrium 'network-related production' produced by the inventors of an area-field will be equal to the average squared centrality of every inventor having patented at least once over the 5-years-window period. Formally, the regional average 'network-related production' is given by:

$$\overline{c_{a,f}^2(g_t,\alpha,\beta,\lambda)} = \frac{1}{Inv_{a,f,t}} \times \sum_{i \in invSet_{a,f,t}} c_i^2(g_t,\alpha,\beta,\lambda),$$

where $invSet_{a,f,t}$ (resp. $Inv_{a,f,t}$) be the set (resp. number) of inventors having patented in area a, technological field f and year t. Note that for area-fields with no inventor, the centrality is not defined. We thus assign the centrality to the value of 1 for these area-fields, as 1 is the minimal value possibly attained by the centrality.

Covariates. The main input of patent generation is creative individuals, and in this subset of the population, the inventors themselves. The variable *inventors* is the number of unique inventors having patented at least one patent in the area-field over the 5 years period ($Inv_{a,f,t}$). This variable aims also at capturing the pure agglomeration of innovative activity effects.

Patents are not locally bounded, they can be the outcome of inter-regional collaborations. If so, the number of inventors of a region as a control may be not sufficient to capture the inputs to knowledge creation as it would neglect the inventors outside the region who have also contributed to produce the patents of the area. To control for this,

[16] The centrality measure makes sense only if every centrality is positive. Further, the matrix $\left(I - \lambda \tilde{G}(\beta)\right)^{-1}$ is non-negative (implying that its product with 1 is positive) only if $\lambda < 1/s$ where s is the largest eigenvalue of the non-negative matrix $\tilde{G}(\beta)$ (remind that this matrix contains only positive elements as it is a variation of the adjacency matrix). See theorem III* of Debreu and Herstein (1953) for a formal proof.

[17] The algorithm stops when the maximum absolute difference between two successive centralities ($\max_i\{c_i^{k+1} - c_i^k\}$) is smaller than 0.0001.

we include the *share of outside collaborators* which is the number of external (to the area) inventors divided by the total number of inventors having participated to the patents of the area.

The distribution of the patent resources along different technologies may influence the efficiency of knowledge production. If agglomeration economies are at work, the concentration of patents in some particular technological fields may enhance the productivity of research in those fields. Thus we include the variable *technology Herfindahl*, defined as the Herfindahl index of the patents produced in the area distributed among 30 technological classes.[18] This variable is defined at the area-year level and its formal definition is $\sum_{c=1}^{30} s_c^2$, where s_c is the share of patents in the 30-categories-technological-class c.

The econometric analysis will control for the specificity of the technological fields with area-field dummies. Yet, even when controlling for a technology, some regions may still be specialized in domains within this technology which are more recent and more fertile in new ideas, and thus in new patents. Those technologies are possibly less likely to cite old knowledge (Lobo and Strumsky, 2008). To control for this effect, we include the *technology age* variable, which is the average number of references cited by the patents produced in the area-field.

One important issue that needs to be controlled for stems from the nature of the collaboration network. The collaboration network is based on a bipartite graph: only the connections between inventors and patents are observed while connections between inventors are not directly available. Inter-inventor connections are then reconstructed from this bipartite graph, where two inventors will be connected whenever they co-author a patent. In this context, team size has a large influence on the network structure because all inventors within a team are connected. Even though the inventors network structure and team sizes are two different concepts, the increase in team size may still rise inventors' average centrality by increasing their number of connections. In consequence, if larger teams produce more patents, and if larger teams also implies more centrality, then the effect captured by the centrality variable may be spurious. What aims to be unveiled in this study is the effect of individual networks on regional innovation and not the mere pooling of inputs as characterized by large team sizes. Therefore, the model includes the *average team size*, defined as the average number of inventors per patent produced in the area-field-year.

Finally ,we also integrate economic variables. To do so, we use plant-level data stemming from French annual business surveys over the period 1985–2003.[19] These mandatory surveys provide information regarding employment for all manufacturing firms of more than 20 employees. All the more, it reports the precise location along with the level of

[18]The classification leading to 30 technological classes, referred to as OST30, is based on the IPC code of the patents and is a finer grained version of the OST7 classification. See OST (2010) for more information.

[19]The sources are the data from the 'enquetes annuelles d'entreprises', which are collected by the French ministry of industry jointly with the French institute for national statistics (INSEE).

employment of each French plant of these firms. We create, for each EA, the variables of the *number of workers* and of the *number of plants of more than 200 employees*. At last, we create an index of *employment diversity* in the EA. This index of employment diversity is based on an Herfindahl index at the 3-digits sectoral level (noted s). It is defined by the following equation:

$$Employment\_diversity_t^a = \ln\left(1/\sum_s\left(\frac{\#employees_t^{s,a}}{\#employees_t^a}\right)^2\right),$$

with $\#employees_t^{s,a}$ the number of employees in sector s, area a and year t, and $\#employees_t^a = \sum_s \#employees_t^{s,a}$.

3.4.4 Descriptive statistics

The data is composed of 297 EAs, 7 technological fields and 18 years (1985–2002 for the explanatory variables and 1986–2003 for the dependent variables). Some area-fields do not have any patent whatsoever during the whole time period and are therefore discarded from the sample.[20] The sample ends up consisting of 1,940 area-fields. Table 3.3a presents the summary statistics for the main variables. The average area-field is rather small, as it produces 4 patents a year and contains an average of 16 unique inventors in a 5-years-window. We can notice that the distribution is skewed as the median area-field produces only 1 patent per year and includes only 3 inventors in a 5-years-window. Looking at the spatial distribution of collaborations, we see that inter-regional collaborations is a common feature since the share of non-regional inventors participating to the regional patents is on average 28%. The teams of inventors producing innovations are rather small as the average team size is of 1.75 on average at the area-field level.

The correlation between the variables are reported in Table 3.3b. The highest correlation, of 98%, is between the number of workers and the number of large plants. As these variables are simply used as controls, we keep them both in the sample.

3.5 Results

Does the network influence regional performance, and if so, how? The simple model developed in Section 3.3 provides a rationale for such an effect motivated by a micro level perspective. As inventors collaborate to produce patents, they exert an influence on their collaborators, influence that will shift their partners' productivity. The model then predicts that the productivity shifts will depend on a form of network centrality. This centrality is ruled by three parameters each of which carries a specific meaning on how the network may influence innovation. We provide, in a first subsection, an investigation assessing the influence of social networks on inventors' productivity assuming different

[20] Note that due to the presence of area-field fixed-effects, they would have been dropped in the estimation.

Table 3.3 – Descriptive statistics and correlations.

(a) Descriptive statistics.

	Min	Median	Q3	Max	Mean	S.D.
Citations-Weighted Number Of Patents	0.000	1.000	4.000	1560.000	7.963	44.104
Number of Patents	0.000	1.000	3.000	847.000	4.186	22.125
Number of Inventors	0.000	3.000	9.000	3814.000	16.354	89.280
Average Team Size	0.000	1.833	2.476	8.000	1.754	1.238
Share of Outside Collaborators	0.000	0.273	0.500	0.875	0.281	0.255
Technology Age	0.000	10.429	14.000	77.333	9.837	6.982
Technology Herfindahl (OST30)	0.000	0.145	0.247	1.000	0.202	0.171
Number of Plants of 200+ Employees	0.000	6.000	12.000	398.000	11.042	26.279
Number of Workers	31.000	6003.000	12709.000	569152.000	11750.365	27787.162
Employment Diversity (3-digits)	0.168	2.222	2.583	3.828	2.155	0.608
Average Squared Degree Centrality	1.000	1.107	1.209	5.018	1.138	0.156
Average Squared Bonacich Centrality	1.000	1.121	1.257	9.300	1.186	0.277
Average Squared Page-Rank Centrality	1.000	1.057	1.083	1.599	1.051	0.044

(b) Correlations.

		1	2	3	4	5	6	7	8	9	10	11
1	Number of Inventors	1										
2	Average Team Size	0.081	1									
3	Share of Outside Collaborators	0.033	0.842	1								
4	Technology Age	0.070	0.486	0.387	1							
5	Technology Herfindahl (OST30)	-0.121	-0.328	-0.297	-0.278	1						
6	Number of Plants of 200+ Employees	0.775	0.106	0.051	0.101	-0.208	1					
7	Number of Workers	0.776	0.106	0.051	0.101	-0.203	0.988	1				
8	Employment Diversity (3-digits)	0.175	0.199	0.165	0.172	-0.368	0.306	0.258	1			
9	Average Squared Degree Centrality	0.121	0.764	0.640	0.305	-0.251	0.133	0.133	0.166	1		
10	Average Squared Bonacich Centrality	0.108	0.652	0.534	0.216	-0.203	0.118	0.119	0.127	0.895	1	
11	Average Squared Page-Rank Centrality	0.106	0.698	0.629	0.425	-0.288	0.130	0.128	0.185	0.800	0.614	1

Notes: The explanatory variables based on the network are constructed using a 5-years-window (it concerns the following variables: number of inventors , average team size, share of outside collaborators, technology age , technology Herfindahl, and the network-centrality variables). The centrality measures are computed with $\lambda = 0.04$.

typical forms of influence (different types of centralities). Secondly, we introduce a non-linear estimation procedure allowing us to investigate the issue of how inventors networks are affecting invention.

3.5.1 Assessing the influence of network centrality on regional patenting

The first step of this study is to test whether network centrality has any effect on local innovation. As this centrality varies with its parameters, the analysis will first be conducted by setting these parameters to ad-hoc values. Three well know forms of centrality will be used: the Degree, the Bonacich and the Page-Rank. These forms of centrality are encompassed by the general form stemming from the model and can be considered as limiting cases: no complementarity nor rivalry ($\alpha = 0,\ \beta = 0$); complementarity but no rivalry ($\alpha = 1,\ \beta = 0$); and complementarity and rivalry ($\alpha = 1,\ \beta = 1$). For the Bonacich and the Page-Rank centralities, there is a restriction on λ, because when $\alpha = 1$ the complementarity effect can be explosive when λ is too high. The most restrictive case is the Bonacich centrality as it is not defined when λ is greater than the inverse of the highest eigenvalue of the adjacency matrix. In the collaboration network defined by our sample the maximum value for λ is $\lambda^{max} = 0.05$. For the sake of comparability, we then set λ to 0.04 for the three measures. After creating these variables, whose summary statistics are reported in Table 3.3a, we estimate Poisson regressions with fixed-effects.

The results of the Poisson estimation are reported in Table 3.4. We first focus on Model (1) which is a benchmark model excluding network centrality variables, before going on to comment the main results in Models (2) to (4) on the network centrality variables. As usual in studies on regional patenting, the number of inventors has a strong positive effect. We find that a 10% increase of the number of inventors leads to a 2.7% increase in the number patents (plus the number of citations they have received). More intriguing is the negative effect of the average patenting team size. The estimates suggest that a 10% increase of the average team size in the EA-field would imply a decrease of 1.8% in patenting. This result is nonetheless in line with other estimates in the literature (e.g., Lobo and Strumsky, 2008 find a negative effect, Breschi and Lenzi, 2012 find a non significant effect). Having access to knowledge from outside the region should be valuable since it creates new possibilities of knowledge combination. Accordingly, the share of non-regional inventors taking part to regional patents has a positive and significant coefficient. Regional specialization, as measured by the *technology Herfindahl*, has a positive and significant effect on regional innovation. The age of the technology developed in the EA-field, as measured by average number of references to older patents, has a negative and significant effect. This means that EA-fields that become specialized into 'younger' technologies are likely to produce more innovations. Economic variables such as the number of large plants or the diversity index of the workforce seem not to influence

regional patenting. On the other hand, the number of workers within the EA does foster regional patenting (as the number workers is highly correlated to the number of large plants, the interpretation is limited).

Turning to the centrality measures, in Models (2) to (4), the network centrality of regional inventors seems to positively influence regional patent production as the three have a positive and significant coefficient. The only difference between the three measures is that they are based on different values of α and β. For instance, the Page-Rank centrality is the only one of the three measures which accounts for rivalry among inventors' connections. Based on the models' fit statistics, such as the BIC, this form of centrality (in Model (4)) is the one which is the least well-fitted to the data, as compared to Models (2) and (3). This hints that there is possibly no rivalry effect occurring, but needs to be further investigated, as it will be in the next subsection.

To conclude the first part of the empirical study, those results tell us that the positioning of regional inventors in the global network matters for regional innovation. However, these different forms of network centrality tell different stories about how the network may benefit to inventors.

3.5.2 How does network structure affect innovation?

In this section, we first introduce a methodology to estimate the structural parameters of the model, i.e. λ, α and β. The basic results obtained are then provided in Section 3.5.2.2, complemented with extensions and robustness checks in Section 3.5.2.3.

3.5.2.1 Estimation procedure

As seen in the previous section, local inventors' network centrality positively influences innovation. Now, our goal is to estimate the parameters which tell the story behind the network (α, β, λ). Unfortunately, these parameters cannot be obtained by traditional linear techniques. Indeed, changing the value of any of these parameters imply non trivial changes on every inventor's network centrality. Stated differently, the network centrality variable cannot be expressed as a linear combination of its parameters with some other exogenous variables. To cope with this problem, we then apply non-linear estimation techniques.

Similarly to models with linear right-hand sides, the estimated coefficients are simply the set of parameters that maximizes the likelihood as follows:

$$\arg\max_{\boldsymbol{\theta},\lambda,\alpha,\beta} \mathcal{L}(\mathbf{Y}|\mathbf{X},\boldsymbol{\theta},\lambda,\alpha,\beta),$$

where $\mathcal{L}$ is the likelihood of the Poisson distribution, $\mathbf{X}$ is the set of all controls (including the dummies) and $\boldsymbol{\theta}$ their coefficients. The main difference with linear estimations lies on the relation the three parameters λ, α and β have together. This can be seen by writing

Table 3.4 – Poisson estimations at the EA level.

Dependent Variable	Citations-Weighted Number Of Patents (t+1)			
Model:	(1)	(2)	(3)	(4)
Area-Field-Specific Variables				
# Inventors (ln)	0.2684***	0.2692***	0.2709***	0.2671***
	(0.0205)	(0.0204)	(0.0204)	(0.0205)
Average Team Size (ln)	-0.1774***	-0.2446***	-0.2189***	-0.2092***
	(0.0341)	(0.0359)	(0.0352)	(0.0351)
Share of Outside Collaborators	0.1393*	0.0931	0.1139	0.1181
	(0.0756)	(0.0769)	(0.0763)	(0.0763)
Technology Age (ln)	-0.0666***	-0.0314	-0.0454**	-0.0532**
	(0.0204)	(0.0215)	(0.0211)	(0.0208)
Area-Specific Variables				
Technology Herfindahl (OST30)	0.2975**	0.2448*	0.258*	0.2854**
	(0.1352)	(0.136)	(0.1356)	(0.1366)
# Plants of 200+ Employees (ln)	-0.011	-0.0128	-0.0126	-0.0114
	(0.0259)	(0.0257)	(0.0258)	(0.0258)
# Workers (ln)	0.3551***	0.3537***	0.3566***	0.3558***
	(0.0561)	(0.056)	(0.056)	(0.056)
Employment Diversity (3-digits)	-0.0259	-0.0277	-0.0351	-0.023
	(0.036)	(0.0359)	(0.0359)	(0.0359)
Network Centrality Variables				
Squared Degree Centrality (ln)		0.6767***		
		(0.1246)		
Squared Bonacich Centrality (ln)			0.3033***	
			(0.0715)	
Squared Page-Rank Centrality (ln)				1.333***
				(0.4166)
Dummies				
EA × Tech	YES	YES	YES	YES
Time ×Tech	YES	YES	YES	YES
Fit statistics				
Observations	34920	34920	34920	34920
Adj-pseudo R^2	0.87339	0.87353	0.87348	0.87343
BIC	177240.109	177072.242	177136.811	177192.555

Notes: Fixed-effects Poisson estimation. The dependent variable is in $t+1$ while the variables based on patent data are built using a 5 years window from $t-4$ to t. The three squared centrality variables: Degree, Bonacich and Page-Rank refer, respectively, to the following general-form centrality variables $\overline{c^2_{a,f,t}}(\lambda, \alpha = 0, \beta = 0)$, $\overline{c^2_{a,f,t}}(\lambda, \alpha = 1, \beta = 0)$ and $\overline{c^2_{a,f,t}}(\lambda, \alpha = 1, \beta = 1)$. They are the average squared centrality of the inventors of the area-field and were computed along the methodology described in Section 3.4.3 and using $\lambda = 0.04$. White heteroskedasticity-robust standard-errors in parenthesis. Level of statistical significance: ***, **, * means significance at the 1%, 5% and 10% level.

the relation between the dependent and the explanatory variables, as follows:

$$E(Y_{a,f,t}) = \exp\left(\sum_k \theta_k \ln\left(X^k_{a,f,t}\right) + \ln\left(\overline{c^2_{a,f}\left(g_t, \lambda, \alpha, \beta\right)}\right)\right),$$

where, contrary to any θ_k, the centrality parameters are not linearly associated to a variable. As usual for maximum likelihood models, the estimates are obtained by using a maximization algorithm. In this context, different from any exogenous variable $X^k_{a,f,t}$, the variable $\overline{c^2_{a,f}\left(g_t, \lambda, \alpha, \beta\right)}$ needs to be computed anew at each iteration of the maximization process. Note that to avoid estimation issues, the elasticity of $\overline{c^2_{a,f}\left(g_t, \lambda, \alpha, \beta\right)}$ is set to 1.[21] Indeed, we are not interested in estimating the precise elasticity of the network effect, but rather in the value of the structural coefficients. The identification of the network effect does not suffer from the constraint on the elasticity, since the key element λ, which represents connectivity, can be properly estimated. The statistical significance of λ is crucial as if λ is equal to 0, then there is no effect whatsoever of the network and the other two parameters, α and β, cannot be interpreted.

The interpretation of the parameters along the model defined in Section 3.3 is valid only for positive values of these parameters. Further, the value of α cannot be greater than 1, since when $\alpha \geq 1$ the centrality is not defined for any positive λ. Consequently, the estimation will be run with the following constraints: $\lambda \geq 0$, $\alpha \in [0; 0.99]$, $\beta \geq 0$.

Finally, even though these parameters enter the model in a non-linear form, they end up to be asymptotically normally distributed (see Wooldridge, 2010, theorem 12.3, p. 407) and their heteroskedasticity-robust covariance is defined as the Huber-White sandwich variance estimator (Wooldridge, 2010, p. 416). To make this estimation, we used the statistical software *R* in combination with the package 'FENmlm' which estimates maximum likelihood models with fixed-effects and allows for non-linear right hand sides.[22]

[21] This is rather a purely technical point and lies on the approximation $\ln(1+\epsilon) \approx \epsilon$ when ϵ is low. In the case in which λ has a low value, the approximation applies and yields an identification issue if there is also a coefficient of elasticity in the model. For instance, assume for simplicity that we are only interested in estimating λ, so that α and β are given. Further, take the case of a region with only one inventor, then the squared centrality of this region can be written as $1 + 2\lambda\tilde{c}(\lambda) + \lambda^2\tilde{c}(\lambda)^2$, with $\tilde{c}(\lambda) = \sum_{j \in N} c_j(\lambda)^\alpha / d_j^\beta$ an N the set of collaborators of this inventor. For λ low enough, we have $\ln[1 + 2\lambda\tilde{c}(\lambda) + \lambda^2\tilde{c}(\lambda)^2] \approx \lambda[2\tilde{c}(\lambda) + \lambda\tilde{c}(\lambda)^2] \approx 2\lambda\tilde{c}(\lambda)$. Now assume we include γ, an elasticity coefficient, to the network centrality variable. The relation to be estimated becomes $E(y_{a,f,t}) = \exp\{\sum_k \theta_k \ln(X^k_{a,f,t}) + \gamma \ln[\tilde{c}_{a,f,t}(\lambda)^2]\}$ which can be simplified, when λ is low enough, to $E(y_{a,f,t}) = \exp\{\sum_k \theta_k \ln(X^k_{a,f,t}) + 2\gamma\lambda\tilde{c}_{a,f,t}(\lambda)]\}$. Thus the two parameters cannot be properly identified, as for any estimated set of parameter (γ, λ), there exists another set (γ', λ') yielding – at least numerically – the same fit. Indeed, take $\lambda' = \lambda + \tau$, for τ not too large. Noting that $\tilde{c}_{a,f,t}(\lambda + \tau) - \tilde{c}_{a,f,t}(\lambda)$ is increasing with λ (thus is lower for lower values of λ), then setting $\gamma' = \gamma\lambda/\lambda'$ such that $\gamma\lambda = \gamma'\lambda'$ would yield the same likelihood up to a negligible numerical error. To conclude, if the 'real' λ (the one to be found by the estimation) is close to 0, then the numerical estimation faces an identification issue when combined with an elasticity coefficient γ.

[22] This package is available in the comprehensive *R* archive network (CRAN), at the following link: https://cran.r-project.org/web/packages/FENmlm/.

3.5.2.2 Results

The basic results of this estimation on our data are reported in Table 3.5. We expect to find a positive coefficient of the connectivity, λ, as the first results have shown that both the Degree and the Bonacich centrality have a positive and significant effect. Indeed, the general effect of the network is positive and significant, with the estimated value of λ being close to 2%. This first result means that the level of interaction of the inventors in the global network does increase regional innovation. Now, what about the structural parameters, ruling the network shape?

First, the results tend to suggest that there is no rivalry effect at play. The estimated coefficient of rivalry β, is at its lower bound: 0, even if not significant. This suggests that connections in the network inflict no negative externality. Stated differently, the interactions are not more beneficial when more exclusive. Turning to the estimated value of complementarity α, it is positive, yet the precision of the parameter is weak as the standard-error is high and fails to be significant. This tells us that complementarity is not significantly affecting invention.

All in all, it appears that it matters for invention that agents are connected to other agents (λ is positive and significant), but it does not matter whether they have to share these connections with others (β is null), nor that are connected to more central agents or not (α is positive but not significant). This result tends to suggest that the network-position of the inventors in the global network does matter for innovation. What is most important is just how many connections they have. Up to this point, to whom in particular inventors are connected, including how many and which connection their partners have seems to be much less important.

3.5.2.3 Robustness checks and extensions

We now propose several extensions of these first investigations aiming at:

- checking the robustness of the first two results obtained so far (network connectivity effect and no rivalry),
- testing whether the absence of complementarity effect is always verified,
- checking whether the efforts-based model we introduced is justified.

Other dependent variables To ensure that the results do not rest upon the choice of the citations-weighted number of patents as the dependent variable, we run the econometric analysis on other measures of regional innovation. In fact, we break down the measure of the citations-weighted number of patents in two: 1) the number of patents and 2) the *citations-only-weighted number of patents*, so that patents with no citations have a null weight. While the former variable reflects the idea of quantity of regional innovation production, the latter seizes more the idea of quality while the original dependent variable

Table 3.5 – Non-linear Poisson estimations to determine the value of the centrality parameters. Three different dependent variables.

Dependent Variables:	Citations-Weighted Number Of Patents (t+1)	Number Of Patents (t+1)	Citations-Only-Weighted Number Of Patents (t+1)
Model:	(5)	(6)	(7)
Network Centrality Parameters			
λ (Connectivity)	0.024***	0.0215***	0.0271**
	(0.0089)	(0.007)	(0.0134)
α (Complementarity)	0.4166	0.1061	0.5738
	(0.4637)	(0.3142)	(0.4715)
β (Rivalry)	0	0	0
	(0.2294)	(0.1916)	(0.2958)
Area-Field-Specific Variables			
# Inventors (ln)	0.2701***	0.2771***	0.2617***
	(0.0205)	(0.0173)	(0.0299)
Average Team Size (ln)	-0.2436***	-0.2686***	-0.2194***
	(0.0355)	(0.0285)	(0.0561)
Share of Outside Collaborators	0.097	0.1844***	-0.0132
	(0.0767)	(0.0595)	(0.1228)
Technology Age (ln)	-0.0322	-0.0182	-0.0479
	(0.0213)	(0.0173)	(0.0332)
Area-Specific Variables			
Technology Herfindahl (OST30)	0.2452*	0.3604***	0.1589
	(0.1357)	(0.1142)	(0.1954)
# Plants of 200+ Employees (ln)	-0.0128	-0.0163	-0.0108
	(0.0257)	(0.0201)	(0.0407)
# Workers (ln)	0.3546***	0.3998***	0.2856***
	(0.0559)	(0.0456)	(0.0827)
Employment Diversity (3-digits)	-0.0276	-0.0325	-0.0385
	(0.0358)	(0.0298)	(0.0517)
Dummies			
EA × Tech	YES	YES	YES
Time ×Tech	YES	YES	YES
Fit statistics			
Observations	34920	34920	31230
Adj-pseudo R^2	0.87353	0.84427	0.82428
BIC	177086.341	118655.115	129716.266

Notes: Fixed-effects Poisson estimation. There is only 31,230 observations in Model (7) since the fixed-effects strategy implies that area-fields for which the dependent variable is 0 for the whole period 1985–2002 are dropped. The dependent variable is in $t+1$ while the variables based on patent data are built using a 5 years window from $t-4$ to t. White heteroskedasticity-robust standard-errors in parenthesis. Level of statistical significance: ***, **, * means significance at the 1%, 5% and 10% level.

was a mix of the two. The results of these estimations are reported in the Models (6) and (7) of Table 3.5.

For both dependent variables the coefficient of connectivity, λ, is positive and the coefficient of rivalry, β, remains equal to 0. The main difference is that in Model (6), for patent counts, the complementarity coefficient is equal to 0 while the coefficient of complementarity in Model (7) is equal to 0.5. Although the coefficient is not significant in Model (7), it suggests that complementarity may be more relevant when looking at more qualitative measures of innovations such as the pure number of citations.

Dynamic model The level of regional innovation can be lasting over time so that past levels of the dependent variable may influence future outcomes. To cope with this possible issue, we introduce a dynamic component in the model (see e.g., Windmeijer, 2008). Let $Y_{a,f,t}$ to represent the lag of the dependent variable for area a and technological field f (remind that the dependent variable in the estimation is $Y_{a,f,t+1}$). Following the methodology in Crépon and Duguet (1997), we assume two separate effects of $Y_{a,f,t}$ on later outcomes. Namely, when $Y_{a,f,t}$ is equal to 0, the effect on $Y_{a,f,t+1}$ is supposed to be different than when $Y_{a,f,t}$ is strictly positive. Accordingly, we include $\ln(Y_{a,f,t})$ as a regressor when $Y_{a,f,t}$ is strictly positive and include a dummy taking value 1 when $Y_{a,f,t}$ is zero. The results are provided in Model (8) of Table 3.6 and present no important difference with the main results. The auto-regressive terms both have a positive effect and the coefficient of connectivity is lower even though still statistically significant.

Spatial dependence Now we consider the possibility that the production of patents may be spatially autocorrelated across EAs. Spatial dependence would imply that the patenting in one EA is correlated to the patenting in the neighbouring EAs. Thus we performed a Moran I's test for each year-technology in the sample. The coefficient of spatial correlation is 0.2, although not high, it is yet statistically different from 0. A closer look to the data shows that the small sign of spatial correlation pattern is driven by the French region of Île de France (IDF). When withdrawing the EAs contained in the region of IDF, the coefficient of spatial correlation drops to .06 and is no longer significant.

Ideally, the analysis should take care of this spatial auto-correlation pattern in the estimation framework. However, because of the specific functional form to estimate, this is not possible. Therefore, we redo the econometric analysis when omitting the EAs contained in the region of IDF, leaving a sample free of any spatial correlation.[23] The results of this estimation are reported in Model (17) of Table 3.9. The results present no difference with the baseline model. This is a sign that the initial results are not driven by this small spatial auto-correlation pattern.

[23]The region of Île de France contains the following EAs: Cergy, Coulommiers, Créteil, Etampes, Evry, Meaux, Melun, Nemours, Orly, Paris, Poissy, Provins, Rambouillet, Saclay, Versailles.

Table 3.6 – Robustness checks: introducing a dynamic component in the model and changing the sample.

Dependent Variable ($Y_{a,f,t+1}$):	Citations-Weighted Number Of Patents (t+1)		
Sample:	Full	Mean (#Inventors per year) ≥ 5	Mean (#Inventors per year) ≥ 10
Model:	(8)	(9)	(10)
Network Centrality Parameters			
λ (Connectivity)	0.0168**	0.019***	0.0252***
	(0.008)	(0.005)	(0.0064)
α (Complementarity)	0.4621	0.9481***	0.9423***
	(0.5681)	(0.2621)	(0.1493)
β (Rivalry)	0	0	0
	(0.2782)	(0.1362)	(0.1125)
Area-Field-Specific Variables			
# Inventors (ln)	0.153***	0.4901***	0.5952***
	(0.0211)	(0.034)	(0.046)
Average Team Size (ln)	-0.1069***	-0.3276***	-0.3585**
	(0.0356)	(0.105)	(0.1472)
Share of Outside Collaborators	-0.0567	0.404**	0.7213***
	(0.0763)	(0.1597)	(0.2409)
Technology Age (ln)	-0.041**	0.0356	-0.0183
	(0.0206)	(0.0569)	(0.0794)
Area-Specific Variables			
Technology Herfindahl (OST30)	0.0859	1.0892***	1.1898**
	(0.1327)	(0.3935)	(0.5024)
# Plants of 200+ Employees (ln)	-0.0107	0.0346	0.1117
	(0.0256)	(0.0608)	(0.0743)
# Workers (ln)	0.3348***	0.2351***	0.1784*
	(0.0545)	(0.081)	(0.0942)
Employment Diversity (3-digits)	-0.0296	0.0076	-0.0209
	(0.0347)	(0.045)	(0.054)
Dynamic Component			
$\log(Y_{a,f,t})$	0.1496***		
	(0.0108)		
$\mathbb{1}_{Y_{a,f,t}=0}$	0.0611**		
	(0.0272)		
Dummies			
EA $\times$ Tech	YES	YES	YES
Time $\times$Tech	YES	YES	YES
Fit statistics			
Observations	34920	5346	2682
Adj-pseudo R^2	0.87446	0.89573	0.91419
BIC	175935.457	54812.618	31357.63

Notes: Fixed-effects Poisson estimation. The dependent variable is in $t+1$ while the variables based on patent data are built using a 5 years window from $t-4$ to t. White heteroskedasticity-robust standard-errors in parenthesis. Level of statistical significance: ***, **, * means significance at the 1%, 5% and 10% level.

Restricting to most active area-fields In the econometric model, we estimate the network effect by looking at the average squared regional centrality. Small regions, in terms of number of inventors, may suffer from a higher variability of the network centrality variable as each inventor has a higher influence on the regional variable. To stymie this possible problem, we run the analysis on the sample of area-fields which are the most innovative and thus suffer much less from this 'limited number of inventors' problem.

In the restricted samples, we want to keep more active area-fields, the ones having a sufficient number of inventors per year. The selection of these area-fields is based on their yearly average number of inventors. Table 3.6 reports the results when the sample is restricted to area-fields with a yearly average of, in Model (9), more than 5 inventors (192 area-fields), and, in Model (10), more than 10 inventors (149 area-fields). Consistently with the previous results, the network effect is positive and the rivalry coefficient is equal to 0. However, the coefficient of complementarity, in both estimations, is now close to 1 (.95 and .94 respectively) and is statistically significant. Thus, in the most innovative area-fields, it seems that there is a complementarity effect at play so that the inventors benefit of being connected to more central partners.

Further, these two restricted samples are much less subject to the high variability in terms of average squared centrality brought about by small area-fields, in which there is only a handful of inventors over the whole period of analysis. Therefore, the results may be more robust with these restricted samples than with the full sample.

Controlling for stars With the type of centrality indexes we are using, the distribution of centrality in the population might be very skewed. We are worried that the results could be driven by a specific set of inventors: star inventors that would also be outliers in terms of centrality. While this is not a problem per se, it would raise an issue in terms of interpretation of the results. Indeed, we here assess the shape of the regional network and the interpretation is done in terms of the overall productivity of regional inventors. Thus, if only the centrality of stars led the results, this would partly flaw this interpretation. To control for this, we run a new analysis in which the average network centrality is computed excluding star inventors.[24]

More precisely, the new centrality variable is obtained in two steps. In the first step, the centrality of all inventors is computed using the whole network, stars included. The difference with the original variable lies on the second step where we average the squared centrality of only non-stars inventors at the area-field level. Formally, let $InvSet_{a,f,t}^{no\_Star}$ be the set of inventors of the area-field that are not defined as stars. Then the non-stars

[24]The 'average squared centrality' for area-fields in which only star-inventors reside is set to 1 (which is the minimum value for this variable), as for area-fields in which there is no inventor at all.

regional average network centrality is defined as:

$$\overline{\left(c_{a,f,t}^{no\_Star}\right)^2} = \frac{1}{Inv_{a,f,t}^{no\_Star}} \times \sum_{i \in InvSet_{a,f,t}^{no\_Star}} c_{i,t}^2.$$

Star inventors are defined anew for each year t, based on their production between $t-4$ and t. An inventor is designated as a 'star' if the number of patents he/she produced in a given 5-years-window is strictly greater than the top 5% percentile, Model (11), or top 1% percentile, Model (12).[25] The results of these estimations are reported in Table 3.7. We find that the estimated connectivity λ, is still positive and significant and the estimated rivalry is still equal to 0. Interestingly, the estimated complementarity becomes positive and significant in the two estimations. This coefficient is now, in both estimations, close to 1 (.89 and .94 respectively). These values are very close to what we have obtained when excluding less active area-fields. It implies that being connected to central agents is beneficial to the non-star inventors.

A different spatial aggregation unit A common issue arising when dealing with discrete geographical units is the moving areal unit problem (MAUP). Because space is continuous and geographical units are discrete per nature, the results can be reliant on the choice of these geographical units. To limit this issue and assure the robustness of our results, the econometric analysis of the baseline model is replicated using NUTS 3 geographical units. In France, the NUTS 3 regions correspond to the 'départements' which were defined by the French administration. They are larger aggregates than the EA: there are 94 continental France NUTS 3 regions as opposed to 297 EA.[26]

The estimates for this geographical unit are reported in Model (16) of Table 3.9. The results are similar to the main results at the EA level. The coefficient of connectivity, λ, is positive and significant with an order of magnitude almost identical to the baseline model. Similarly, the coefficient of rivalry is found to be equal to 0. The main difference with the baseline results concerns again the complementarity coefficient, α, which has a value of 0.92 and is statistically significant.

Restricting to regional networks The previous investigations have focused on the inventors positions in the global network and shown that inventors' network connections are indeed beneficial to local innovation. Now we tackle a slightly different question: Does the *internal* regional network structure affects innovation performance?

To address this question, we employ the same methodology as previously. The only difference is that the inventors' centrality will be computed using the intra-regional links

[25] The cut-off for being in the top 5% inventors is 4 patents for the period 1985-1988, 5 patents for 1989-1999 and 6 patents for 2000-2002. The cut-off for being in the top 1% inventors increases gradually from 8 patents in 1985 to 12 patents in 2002.

[26] The average areas in squared kilometres are: 5,745 for the NUTS 3 regions and 1,818 for the EA.

Table 3.7 – Robustness check: average squared network centrality of non-stars only.

Dependent Variable:	Citations-Weighted Number Of Patents (t+1)	
Star Definition:	Top 5% Inventors	Top 1% Inventors
Model:	(11)	(12)
Network Centrality Parameters (Regional Squared Centrality Of Non-Stars Only)		
λ (Connectivity)	0.0251***	0.0201***
	(0.0079)	(0.0062)
α (Complementarity)	0.8908***	0.944***
	(0.2694)	(0.3415)
β (Rivalry)	0	0
	(0.1678)	(0.1738)
Area-Field-Specific Variables		
# Inventors (ln)	0.2659***	0.2677***
	(0.0204)	(0.0204)
Average Team Size (ln)	-0.2409***	-0.2331***
	(0.0366)	(0.0361)
Share of Outside Collaborators	0.0861	0.1001
	(0.0774)	(0.0769)
Technology Age (ln)	-0.0297	-0.0355
	(0.022)	(0.0216)
Area-Specific Variables		
Technology Herfindahl (OST30)	0.2733**	0.2716**
	(0.1352)	(0.1355)
# Plants of 200+ Employees (ln)	-0.0135	-0.0128
	(0.0257)	(0.0257)
# Workers (ln)	0.3606***	0.3625***
	(0.056)	(0.0559)
Employment Diversity	-0.0274	-0.0269
	(0.0357)	(0.0358)
Dummies		
EA × Tech	YES	YES
Time ×Tech	YES	YES
Fit statistics		
Observations	34920	34920
Adj-pseudo R^2	0.87348	0.87347
BIC	177149.191	177165.616

Notes: Fixed-effects Poisson estimation. The dependent variable is in $t+1$ while the variables based on patent data are built using a 5 years window from $t-4$ to t. White heteroskedasticity-robust standard-errors in parenthesis. Level of statistical significance: ***, **, * means significance at the 1%, 5% and 10% level.

only. This means that, to be connected in the intra-regional network, two inventors must appear at least in one patent record in which they both have their addresses in the same EA. For instance, if an inventor has only collaborated with inventors outside of his/her region, he/she will be considered as isolated in the intra-regional network.

Table 3.8 presents the results with the intra-regional network. Model (13) reports the main regression, Model (14) restricts the sample to the most innovative area-fields, and Model (15) excludes the top 5% inventors from the calculation of the average regional centrality. Consistently across all three estimations, connectivity is found to be significantly positive and rivalry coefficient is null, as in the main model. It should be noted that, in all three estimations, the complementarity coefficient is now equal to its maximal value (.99) and it is significant.

3.5.2.4 An alternative model without efforts

We would now like to test the dependency of our results to one important assumption of the model which is not directly tested empirically. In the simple model introduced in Section 3.3, agents are assumed to exert efforts, the returns of which are affected by their network connections. However, these efforts are not observable in the data. We thus here propose a simple and alternative model without such efforts. Let us rather assume that the inventive contribution of agent i is directly equal to his/her productivity: $y_i = \psi_i$, and let us assume that productivity has an autonomous and a network based component (in a very similar fashion as in the main model), given by the following equation:

$$\psi_i(g) = 1 + \lambda \sum_{j \in N_i} \frac{\psi_j(g)^{\alpha}}{d_j^{\beta}}, \tag{3.11}$$

Now α receives a slightly different interpretation: it scales the extent to which an agent's productivity increases with the one of her/his partners. Without any form of maximization, agents' productivities are again obtained as fixed-point, solution of the system of equations induced by Equation 3.11, for all i. It turns out that, as in the main model, we obtain:

$$\psi_i(g) = c_i(g, \lambda, \alpha, \beta), \tag{3.12}$$

where c_i is a centrality measure that depends only on the position of the inventor within the network and on the three parameters λ, α and β. In this simple model without efforts, agents' productivities are equal to their outcomes y_i^*. Therefore, we obtain that the only difference between the main model, based on efforts, and this one, is that here inventors' productivities are equal to their centrality whereas in the main model, they are equal to the square of their centrality. We have thus tested, in Model (18) (Table 3.9), the same model as Model (5) but using the average centrality in the area-field instead of the average squared centralities. We find that none of the parameters are significant, thus justifying the use of the full model based on agents' efforts.

Table 3.8 – Poisson regression. The centrality is based on the intra-regional network only.

Dependent Variable:	Citations-Weighted Number Of Patents (t+1)		
Sample:	Full	Mean(#Inventors per year)≥ 10	Full
Other information:	–	–	Top 5% Inventors Are Excluded
Model:	(13)	(14)	(15)
Network Centrality Parameters			
λ (Connectivity)	0.0387***	0.041***	0.0401***
	(0.0103)	(0.0138)	(0.0106)
α (Complementarity)	0.99***	0.99***	0.99***
	(0.0674)	(0.0987)	(0.0661)
β (Rivalry)	0	0	0
	(0.0939)	(0.0661)	(0.0594)
Area-Field-Specific Variables			
# Inventors (ln)	0.2502***	0.5733***	0.2516***
	(0.0203)	(0.0462)	(0.0206)
Average Team Size (ln)	-0.2466***	-0.4347**	-0.2411***
	(0.0373)	(0.1754)	(0.0391)
Share of Outside Collaborators	0.3204***	1.0637***	0.3042***
	(0.0824)	(0.3036)	(0.087)
Technology Age (ln)	-0.0198	-0.002	-0.0228
	(0.0225)	(0.0789)	(0.0239)
Area-Specific Variables			
Technology Herfindahl (OST30)	0.2436*	1.2165**	0.2765**
	(0.1355)	(0.4932)	(0.1359)
# Plants of 200+ Employees (ln)	-0.0148	0.0938	-0.0128
	(0.0256)	(0.0742)	(0.0256)
# Workers (ln)	0.3739***	0.2108**	0.3691***
	(0.0555)	(0.0928)	(0.056)
Employment Diversity (3-digits)	-0.0262	-0.0001	-0.0264
	(0.0356)	(0.0537)	(0.0358)
Dummies			
EA × Tech	YES	YES	YES
Time ×Tech	YES	YES	YES
Fit statistics			
Observations	34920	2682	34920
Adj-pseudo R^2	0.87354	0.91419	0.87348
BIC	177069.361	31357.643	177150.686

Notes: Fixed-effects Poisson estimation. The dependent variable is in $t+1$ while the variables based on patent data are built using a 5 years window from $t-4$ to t. White heteroskedasticity-robust standard-errors in parenthesis. Level of statistical significance: ***, **, * means significance at the 1%, 5% and 10% level.

Table 3.9 – Other robustness checks.

Dependent Variable:	Citations-Weighted Number Of Patents (t+1)		
Information:	NUTS 3	EAs From IDF Are Excluded	Centrality Is Not Squared
Model:	(16)	(17)	(18)
Network Centrality Parameters			
λ (Connectivity)	0.025***	0.0234**	0.0572
	(0.0054)	(0.0102)	(0.0501)
α (Complementarity)	0.9291***	0.5501	0.4893
	(0.1377)	(0.5219)	(0.6833)
β (Rivalry)	0	0	0
	(0.1008)	(0.2618)	(0.5836)
Area-Field-Specific Variables			
# Inventors (ln)	0.3699***	0.2428***	0.2693***
	(0.0288)	(0.0215)	(0.0205)
Average Team Size (ln)	-0.3612***	-0.2153***	-0.2553***
	(0.0553)	(0.0369)	(0.0366)
Share of Outside Collaborators	0.0501	0.0456	0.0875
	(0.1058)	(0.0791)	(0.0776)
Technology Age (ln)	-0.0077	-0.0335	-0.0253
	(0.0336)	(0.0219)	(0.0217)
Area-Specific Variables			
Technology Herfindahl (OST30)	1.2874***	0.3132**	0.2374*
	(0.3392)	(0.1354)	(0.1355)
# Plants of 200+ Employees (ln)	-0.0009	-0.0337	-0.0134
	(0.0707)	(0.0256)	(0.0257)
# Workers (ln)	0.271***	0.1478*	0.358***
	(0.0909)	(0.078)	(0.0558)
Employment Diversity (3-digits)	-0.0597	-0.1184***	-0.0272
	(0.0367)	(0.0457)	(0.0357)
Dummies			
EA × Tech	YES	YES	YES
Time ×Tech	YES	YES	YES
Fit statistics			
Observations	11772	33030	34920
Adj-pseudo R^2	0.88406	0.75243	0.87355
BIC	92206.806	160704.023	177062.564

Notes: Fixed-effects Poisson estimation. The dependent variable is in $t+1$ while the variables based on patent data are built using a 5 years window from $t-4$ to t. White heteroskedasticity-robust standard-errors in parenthesis. Level of statistical significance: ***, **, * means significance at the 1%, 5% and 10% level.

3.6 Conclusion

This study aims to unveil whether the inventor's network has an influence on local innovation and, most of all, how does network structure affect inventivity. Departing form the existing literature, we introduced a stylized model linking an inventor's productivity to his/her network connections which associates inventors' productivity to the square of their centrality in the network. Centrality is formulated generically so that the network can affect inventivity in different ways. These forms have been empirically estimated on longitudinal French inventors and company data over twenty years. Thanks to fixed effect panel Poisson regressions at the employment area – technological domain level and to non-linear regression techniques, we show that network connections do indeed matter for regional innovation. We also show that direct connections benefit agents in a non rival manner. Therefore, it seems that the benefits provided by network connections are likely to be similar in nature to information. That is to say, they are relatively easily transferable without impairing the use of other neighbours. For instance, such information-related network-benefit could consist in the transfer of valuable ideas and of promising research paths. Moreover, we find some form of complementarity so that agents are more innovative when connected to more central partners. This complementarity effect is empirically verified when the most prolific inventors are excluded as recipients of these effects (top-five or top-one percent), when only the most innovative EA-fields are considered, when only intra-regional networks are accounted for, or when larger spatial units of analysis are used (NUTS 3 instead of EA). Non-rivalry of connections and complementarity in networks provide a new, network-based, justification for knowledge spillovers. It also tends to highlight the role of star inventors since their numerous partners benefit, in a non rival manner, from their central position.

This study has limitations. First of all, the empirical analysis is carried out using data on French inventors only. It is possible that the patterns of collaborations and the nature of the exchanges involved in collaborations vary across countries. Natural extensions of this work include the application to other countries and geographical scales. In addition, the model introduced in this chapter is very stylized and considers only the benefits an inventor can gain from the network. Therefore it neglects other relevant determinants to regional innovation that are later controlled for in the econometric study. Further theoretical work could integrate in the model other, non network-related, factors of inventors' productivity.

3.7 Appendix

3.7.1 Network-related production at the Nash equilibrium

This section shows how to derive the main result of Section 3.3.1, being that the equilibrium network-related production of inventors are equal to their squared centrality. At the

Nash equilibrium, each inventor maximizes his/her utility taking the behaviour of other inventors as given. They set their effort such as to maximize their utility, as follows:

$$\arg\max_{e_i} \; e_i \psi_i \left(g, \mathbf{e}_{-i}\right) - \frac{e_i^2}{2},$$

with $\mathbf{e}_{-i}$ the vector of all efforts except that of i. At equilibrium, the first order condition must hold, yielding the following equations:

$$\begin{aligned} &\psi_i \left(g, \mathbf{e}_{-i}\right) - e_i = 0, \; \forall i, \\ \Leftrightarrow\; &1 + \lambda \sum_{j \in N_i} \frac{e_j^\alpha}{d_j^\beta} - e_i = 0, \; \forall i, \end{aligned}$$

which provides the formula specifying the effort at equilibrium:

$$e_i^* = 1 + \lambda \sum_{j \in N_i} \frac{\left(e_j^*\right)^\alpha}{d_j^\beta}, \; \forall i.$$

This equilibrium effort is exactly the definition of the centrality defined by Equation (3.5): $e_i^* = c_i \left(g, \lambda, \alpha, \beta\right)$.

Now that we have the equilibrium effort, we look at the equilibrium network-related production. The first order condition holds at equilibrium, implying the following equality:

$$e_i^* = \psi_i^* \left(g, \mathbf{e}_{-i}^*\right), \; \forall i.$$

That is to say, at equilibrium, an inventor's effort equals his/her productivity. Thus, the network-related production is equal to the squared centrality, from the following equivalences (and dropping the indices):

$$y_i^* = e_i^* \psi_i^* \left(g, \mathbf{e}_{-i}^*\right) = \left(e_i^*\right)^2 = c_i^2, \; \forall i.$$

3.7.2 A more general form of the model

This section follows the model developed in Section 3.3.1 and shows that the introduction of new parameters in this model imply no significant change to the results.

Consider the following new productivity and utility functions:

$$\psi_i = \gamma_1 + \lambda \sum_{j \in N_i} \frac{e_j^\alpha}{d_j^\beta},$$

$$u_i = e_i \psi_i - \frac{\gamma_2}{2} e_i^2,$$

where γ_1 is the inventor's own productivity without any collaborator and γ_2 is a parameter scaling the disutility of the effort. Those modifications imply no significant change to the result.

Indeed, with those new parameters, the equilibrium effort of each inventor, $e_i^*(g,\lambda,\alpha,\beta,\gamma_1,\gamma_2)$, must respect the following system of equations:

$$e_i^* = \frac{\gamma_1}{\gamma_2} + \frac{\lambda}{\gamma_2}\sum_{j\in N_i}\left(e_j^*\right)^\alpha d_j^{-\beta},\ \forall i.$$

Denoting $\gamma = \gamma_1/\gamma_2$ and dividing by γ yields:

$$\frac{e_i^*}{\gamma} = 1 + \frac{1}{\gamma}\times\frac{\lambda}{\gamma_2}\sum_{j\in N_i}\left(e_j^*\right)^\alpha d_j^{-\beta},$$

$$\Leftrightarrow \frac{e_i^*}{\gamma} = 1 + \gamma^{-(1-\alpha)}\frac{\lambda}{\gamma_2}\sum_{j\in N_i}\left(\frac{e_j^*}{\gamma}\right)^\alpha d_j^{-\beta}.$$

Note that by writing $\tilde{e}_i^* = e_i^*/\gamma$, $\tilde{e}_i^*$ respects the centrality defined by Equation (3.5) and thus can be written as $\tilde{e}_i^* = c_i\left(g,\gamma^{-(1-\alpha)}\lambda/\gamma_2,\alpha,\beta\right)$. This shows that we have the following equivalence:

$$e_i^*(g,\lambda,\alpha,\beta,\gamma_1,\gamma_2) = \gamma c_i\left(g,\gamma^{-(1-\alpha)}\frac{\lambda}{\gamma_2},\alpha,\beta\right).$$

Thus including the two parameters, γ_1 and γ_2, to the productivity and the utility functions would merely lead to the introduction of a proportionality coefficient to the centrality measure at equilibrium without providing any distributional change.

3.7.3 Existence of the centrality when $\alpha \in [0;1[$

In this section we demonstrate the following proposition:

Proposition 1. When $\alpha \in [0;1[$, the centrality defined by Equation (3.5) has a positive solution for any $\lambda \geq 0$. That is to say, there exists a strictly positive vector $c \in \left(\mathbb{R}^{+*}\right)^n$, such that Equation (3.5) holds.

As when $\lambda = 0$, and when $\alpha = 0$, the proof is trivial, in what follows we consider only $\lambda > 0$ and $\alpha \in]0;1[$. The demonstration of the proposition is based on three lemmas.

Definition 1. Let the function $f:\ \mathbb{R}^n \to \mathbb{R}^n$ be defined by

$$f(x) = \mathbf{1} + \lambda G x^\alpha,$$

where $x \in \mathbb{R}^n$, $\mathbf{1}$ is a n-vector of ones, $\lambda > 0$ and $\alpha \in]0;1[$ are fixed scalars, and G is a $n \times n$ matrix of typical element g_{ij} such that $g_{ij} \geq 0$, $\forall i,j$. Finally, x^α is defined as the vector whose elements are $(x^\alpha)_i = x_i^\alpha$.

Definition 2. Let $c_t \in \mathbb{R}^n$ be the sequence such that $c_0 = \mathbf{1}$ and $c_{t+1} = f(c_t)$.

To prove Proposition 1, we just need to show that the function f has a positive fixed point, i.e., there exists $x^* > \mathbf{0}$ (with $\mathbf{0}$ the n-vector of zeros) such that $f(x^*) = x^*$.[27] And showing that f has a fixed point is similar to showing that the sequence c_t is increasing and convergent.

Lemma 1. The sequence c_t is increasing.

Proof of Lemma 1. We denote x_i the i^{th} element of x and $f_i(x)$ the i^{th} element of $f(x)$. Let us look at the fist derivative of f:

$$\frac{\partial f_i(x)}{\partial x_j} = \alpha\lambda g_{ij} x_i^{-(1-\alpha)}, \ \forall i,j. \tag{3.13}$$

As α, λ and g_{ij} are positive, then $\alpha\lambda g_{ij} \geq 0$. Consequently, for any $x > \mathbf{0}$, the function f is increasing (as any partial derivative defined in Equation (3.13) is positive). Further, the first element of the sequence, c_0, is strictly positive by definition. Thus, $c_1 = f(c_0) \geq c_0$. By mathematical induction, we have $c_{t+1} = f(c_t) \geq c_t, \ \forall t$. Hence, the sequence c_t is increasing.□

Now let $J = \mathbf{1}\mathbf{1}'$ be the $n \times n$ matrix of ones, $\hat{g} = \max\{g_{ij}\}$ and $\hat{G} = \hat{g} \cdot J$. That is to say, the matrix $\hat{G}$ is the $n \times n$ matrix composed only of the maximum element of G. Thus, by definition $\hat{G} \geq G$.[28] In addition, by the definition of $\hat{G}$, we have $\hat{G} \cdot x = n\hat{g} \cdot \bar{x} \cdot \mathbf{1}$ where $n\hat{g}$ is a scalar and $\bar{x} = \sum_i x_i/n$ is the mean of the vector x.

Let the function $h: \ \mathbb{R}^n \to \mathbb{R}^n$ be defined by

$$\begin{aligned} h(x) &= \mathbf{1} + \lambda \hat{G} x^\alpha \\ &= \mathbf{1} + \lambda n\hat{g} \cdot \bar{x^\alpha} \cdot \mathbf{1}, \end{aligned}$$

with $\bar{x^\alpha} = \sum_i x_i^\alpha/n$.

The function h is defined such that for any $x > \mathbf{0}$, $h(x) \geq f(x)$. Indeed, note that

$$\begin{aligned} h(x) - f(x) &= \left(\mathbf{1} + \lambda \hat{G} x^\alpha\right) - \left(\mathbf{1} + \lambda G x^\alpha\right) \\ &= \lambda \left(\hat{G} - G\right) x^\alpha \\ &\geq \mathbf{0}. \end{aligned}$$

Lemma 2. For any $\lambda > 0$, the sequence defined by $s_0 = \mathbf{1}$ and $s_{t+1} = h(s_t)$ converges to a fixed point of h. That is to say, there exists $s^* > \mathbf{1}$ such that $s^* = h(s^*)$.

[27] The comparison $x > y$, x and y two vectors , means: $x_i > y_i, \forall i$.
[28] The comparison $A \geq B$, with A and B two matrices, means: $A_{ij} \geq B_{ij}, \forall i,j$.

Proof of Lemma 2. The vector s is a fixed point of h if and only if:

$$\begin{aligned} s &= h(s) \\ &= \mathbf{1} + \lambda \hat{G} \cdot s^{\alpha} \\ &= \mathbf{1} + \lambda n \hat{g} \cdot \bar{s^{\alpha}} \cdot \mathbf{1}. \end{aligned}$$

Assume s is a fixed point of h. Let us look at the i^{th} element of s:

$$\begin{aligned} s_i &= 1 + \lambda n \hat{g} \cdot \bar{s^{\alpha}} \\ &= 1 + \lambda n \hat{g} \cdot \sum_j s_j^{\alpha}/n \\ &= 1 + \lambda \hat{g} \cdot \sum_j s_j^{\alpha}, \ \forall i. \end{aligned}$$

As this equation is true for all i, it implies that all elements of s are equal (as each element of s is identically defined). Assume the fixed point exists and let y be the unique element of s (i.e., $s_i = y, \ \forall i$), the previous equation simplifies to:

$$y = 1 + \lambda n \hat{g} \cdot y^{\alpha}. \tag{3.14}$$

Thus, proving that s is a fixed point of h is equivalent to show that Equation (3.14) has a solution.

Let $g(y) = y - \lambda n \hat{g} \cdot y^{\alpha} - 1$, the first and second derivatives of g are:

$$\frac{\partial g(y)}{\partial y} = 1 - \alpha \lambda n \hat{g} \cdot y^{-(1-\alpha)},$$

$$\frac{\partial^2 g(y)}{\partial^2 y} = \alpha(1-\alpha) \lambda n \hat{g} \cdot y^{-(2-\alpha)}.$$

Remind that $\alpha \in]0;1[$, and λ, n, $\hat{g}$ are positives. As $g(1) < 0$, to show that there is a solution, we just need to show that there exists values of y such that $g(y) > 0$. Note that the first derivative of g is positive whenever $y > (\alpha n \lambda)^{1/(1-\alpha)}$ and is increasing as the second derivative of g is strictly positive. This implies that g crosses the x-axis just once in $\mathbb{R}^+$. (Similarly, note that $\lim_{y \to +\infty} g(y) = +\infty$.) So, necessarily, there exists one unique y^* such that $g(y^*) = 0$ and $g(y) > 0, \ \forall y > y^*$. Hence, the vector s^*, defined by $s_i^* = y^*, \ \forall i$, is a fixed point of h.□

The two first lemmas are used to prove Lemma 3.

Lemma 3. The function f has a positive fixed point. That is to say, there exists $x > \mathbf{0}$ such that $f(x) = x$.

Proof of Lemma 3. By the definition of s_t, we have $s_t \geq c_t, \ \forall t$. Indeed, $s_0 = c_0 = \mathbf{1}$ and $(h(s_0) =) s_1 \geq c_1 (= f(c_0))$ because $h(x) \geq f(x), \ \forall x > \mathbf{0}$, so that $s_t \geq c_t, \ \forall t$ follows by

mathematical induction. From Lemma 2, the sequence s_t converges to the fixed point s^* of h. It implies that s^* is an upper bound of c_t.

The sequence c_t is increasing (Lemma 1) and is upper bounded by s^*, therefore it is convergent. Let c^* be the point to which the sequence c_t converges. The vector c^* is a fixed point of f.□

We can now turn back to the proof of the proposition.

Proof of Proposition 1. The centrality, as defined by Equation (3.5), is equal to:

$$c_i(g,\lambda,\alpha,\beta) = 1 + \lambda \sum_{j \in N_i} \frac{c_j^{\alpha}(g,\lambda,\alpha,\beta)}{d_j^{\beta}}, \ \forall i.$$

Let $\tilde{G}(\beta)$ be the matrix of typical element $\left(\tilde{G}(\beta)\right)_{ij} = d^{-\beta}$ if nodes i and j are connected in the network and $\left(\tilde{G}(\beta)\right)_{ij} = 0$ otherwise. Dropping the parameters, the centrality can be rewritten in matrix form:

$$c = \mathbf{1} + \lambda \tilde{G}(\beta)\, c^{\alpha}, \tag{3.15}$$

with c the vector of all centralities. Define the function f_c such that $f_c(x) = \mathbf{1} + \lambda \tilde{G}(\beta)\, x^{\alpha}$. Since $\lambda > 0$, $\alpha \in]0;1[$, and every element of $\tilde{G}(\beta)$ is greater than or equal to 0, the function f_c respects the definition of the function f given in Definition 1. Therefore, from Lemma 3, the function f_c has a positive fixed point. It implies that there exists $c > \mathbf{0}$ that solves Equation (3.15), c is the centrality.□

Conclusion

The main objective of this thesis is to clarify the mechanisms involved in knowledge creation, by specifically investigating the role of social networks in this process. To this end, the literature on social networks is combined to the broad literature on the determinants of innovation and on the sociology, and determinants, of collaboration. The social network under scrutiny is the network based on formal collaborations for knowledge production, be it patents or scientific articles.

Although this thesis focuses on a specific determinant of knowledge creation, social networks, it also considers the role of geography. Indeed, since the actors of knowledge production are highly concentrated in space, we then try to further understand the interplay between geography and social networks. In this vein, social networks are a means to get into the mechanisms involved in "knowledge spillovers".

We argue that social networks are a vector of diffusion of information and ideas, and a strong factor influencing knowledge creation and the formation of new collaborations. Tools originating from social networks theory are drawn to help to define new measures and to assess the role of social networks as drivers of knowledge creation.

The three chapters of this thesis contribute to: *i*) better understand the determinants of collaboration in knowledge production, *ii*) provide a thorough assessment of the network-position of regions in knowledge networks, and *iii*) investigate the link between an inventor's position within a network and the production of new knowledge.

Each of these chapters provide methodological contributions, and two include an empirical study to cope with these questions. We first detail the contribution, limitation and follow-up research contained in each chapter. In the end of this conclusion, some future paths of research relating to this thesis are sketched.

Main results, limitations and follow-up research

Chapter 1. The first chapter assesses the role of the social network as a determinant of inter-regional collaborations. As geography is also a competing element driving collaborations, this chapter evaluates the interplay between the social network and geography in driving new collaborations. This chapter is a contribution to the literature on the determinants of collaborations (Maggioni et al., 2007; Maggioni and Uberti, 2009). In particular, it is anchored in the debate opposing geography versus social networks (e.g.,

Bathelt et al., 2004; Boschma, 2005; Rodríguez-Pose and Crescenzi, 2008). To empirically evaluate the relation between geography and social networks as determinants of new collaborations, we employ data on scientific collaboration in chemistry in the period 2000–2005. To answer to this question, we first introduce a new measure of inter-regional network proximity. This measure is used to approximate the level of social proximity between regions. We then estimate the number of collaborations between regions with respect to geographical factors and this measure of social proximity. First of all, the results show that social proximity positively affects the creation of new collaborations. Furthermore, the results depict a clear substitutability pattern between social proximity and geographical proximity. This means that the benefits to social proximity do increase with geographical distance. Concretely, the elasticity of the social proximity variable increases with the geographical distance separating the regions. This is evidence that social proximity helps to bypass the barrier of geography.

This study has been limited in scope. *First*, it has focused on scientific collaborations in the specific field of chemistry. Therefore, it is an open question to know if the results hold for other kind of scientific fields. However, the pattern of substitutability seems strong enough to be robust to other fields, provided the collaboration behaviour in these fields is not too different from the one in chemistry. Therefore, a natural extension would be to apply the methodology of this study to other scientific fields. *Second*, the empirical study was geographically circumscribed to 5 countries of the European Union. Natural extension of this study includes the application of the methodology to larger geographical zones. In particular, it can be worthwhile to extend the analysis to the US, to assess the interplay between geographical distance and social networks in a large area which does not suffer from country border effect. *Third*, the empirical analysis is based on one point in time. Therefore, it does not cope with the evolution of the substitutability/complementarity pattern between social and geographic proximity. Further research can use the same methodology to investigate whether the substitutability between network and geography has evolved over time due to the development in the means of communication.

Chapter 2. The second chapter intends to clarify the notion of network position of regions in social networks. There is a growing interest to characterize the position of regions in inter-regional networks (Maggioni and Uberti, 2011; Scherngell, 2013). However, the question of how to measure the relative position of regions in networks remains unclear. Therefore, this chapter critically assesses the question of the measurement of regional network centrality in the context of R&D networks. It demonstrates the drawbacks of existing centrality measures to cope in a meaningful way with inter-regional R&D networks, after describing how to apply them in this context. Mainly, there are two ways to apply existing network centrality measures to regional R&D networks. Either consider the inter-regional network as a weighted network (i.e. a network in which links between the nodes can be valued more than one) in which the regions are the actors of the network, and then apply existing centrality measures suited for weighted networks. Or compute

the network centrality of each micro-level agent (e.g., the firms or the researchers) and then aggregate these centralities at the regional level. These two methodologies suffer from the drawback driven by the links internal to the regions. We further evidence the problem of the duality between the meso-level, for which the centrality is to be computed, and the micro-level, in which are the actors of the network.

After this discussion, a new measure, the *bridging centrality*, is introduced. This measure is based on how much a region does indirectly connects other regions. In the context of R&D networks, this kind of position is important for two reasons. For the region itself, as it is the sign that the agents from this region are connected to a variety of other regions which can diversify the regional pool of knowledge (Bathelt et al., 2004; Berliant and Fujita, 2012). Secondly, in a network formation perspective, regions which are in the middle of others can be seen as some repository of information and can facilitate future connections between these regions. Another advantage of the measure we introduce is that it was specifically designed for regional R&D networks.

We then use data on co-patents in the EU to compare the new centrality measure with the existing ones. Despite some similar patterns between the various centrality measures, the bridging centrality shows some differences. For instance, the region "Île de France" has a lower rank with the bridging centrality than with other measures because the share of connections which are internal to the region is very high, thus providing less interconnections between other regions.

This new measure is an attempt to incorporate meaningfully the notion of centrality in this context. However, it is only a first step towards integrating a multi-dimensional approach to characterize the position of regions in knowledge networks. A next step would be to complement this approach by including other regional characteristics, such as, for instance, to define how much a region is a bridge between other national and non-national regions.

Chapter 3. The third chapter investigates the link between the production of knowledge and the position of inventors in the co-patent network. Following the literature on knowledge production, the network structure should have an influence on the diffusion of ideas and on the productivity of inventors (Cowan and Jonard, 2004; Singh and Fleming, 2010). We introduce a simple model in which the productivity of inventors is assumed to be dependent on their collaborator's behaviour. This model contains three elements: *connectivity*, *complementarity* and *rivalry*. The first element, *connectivity*, simply states that there is a positive relation between an inventor's productivity and her/his network of collaborators (i.e. without connectivity, the productivity of inventors is independent of their network). The two other elements relate to *how* the inventor's network affect their productivity. Inventors are assumed to exert efforts to produce new knowledge. The *complementarity* states that an inventor can be more productive if his collaborators exert more efforts. Finally, *rivalry* means that an inventor may benefit less from very connected collaborators, as they would have less time to devote to their collaboration. The model

predicts that at equilibrium, the network-related production of the agents would be dependent on a measure of their position in the network. This measure of network position is a new form of network centrality which depends on the three elements of connectivity, complementarity and rivalry.

To empirically assess if the network influences regional innovation, we use data on the French co-patent network for the period 1981–2003. We estimate a region's production of innovation with respect to regional determinants, a various set of fixed effects and the average network centrality of the agents in the region. The aim is to estimate the parameters of the network centrality, these parameters tell whether the network has any influence on regional production, and what kind of structure favours most innovation.

The results first show that the connectivity of inventors affects positively the regional production of innovation. Moreover, we find a significant sign of complementary, meaning that connections to inventors better positioned in the network increase the performance of agents. Finally, the results depict, consistently across all estimations, no sign of rivalry occurring in the network. This absence of rivalry means that new network connections are always beneficial.

The main policy recommendation stemming from these results aim at increasing the network connectivity of inventors. One way to attain such a goal is to enable/facilitate more movement of inventors across firms, since they will therefore be able to access new sets of possible partners to collaborate with. This recommendation is very in line with recent works on non-compete agreements in the US which demonstrate the counter-productive effect of non-compete agreements on regional innovation (e.g., Marx et al., 2009, 2015).

Although this study is a step towards a better understanding of how the social network can affect innovation performance, it suffers from limitations. The first limitation is its scope. Indeed, the empirical evidence is only based on data on French inventors. Therefore the implications of the results are not universal and must be taken with care. A next step is to implement this methodology to other geographical areas. In particular, the US could provide a nice point of comparison, which the network of inventors has been the subject of many (if not most) innovation studies.

Another limitation is that the network is taken as given and the question of network formation is not dealt with. Indeed, the formation of the inventors network may be endogenous, so that collaborations are not random but may rather be the outcome of a selection (see e.g., Carayol et al., 2015). In this case, the fact that there is consistently no sign of rivalry in the results may be in part driven by this selection process. An avenue of investigation would be to integrate the notion of network formation to the study of the network-related determinants of innovation.

Research paths

One essential mechanism taking part in the production of new knowledge is the diffusion of information and ideas, in which the social network plays a critical role. However, the causal effect of *diffusion* on the creation of innovation is challenging to single out. The evidence on this issue are mostly indirect: some articles document that inventors tend to cite other patents based on their social network (Breschi and Lissoni, 2005, 2009; Singh, 2005), but what remains unclear is whether the knowledge diffusion mechanism is pivotal to the generation of innovations.

An avenue for further research concerns the integration of the notion of causality to the question of diffusion in innovation networks, based on the very recent work of Athey et al. (2015). Athey et al. (2015) describe how to implement tests of causality in networks. The type of question for which it is suited is: How the access of one agent to a treatment (e.g., some specific information) can causally affect its partners behaviour (e.g., the collaborators)? This kind of methodology can then be used to infer the causal effect between the access of some agent to some knowledge and how it can propagate through its network.

Another promising area of research on innovation lies on the combination of different levels of innovation networks by taking advantage of the *two-modes network* nature of innovation networks. For instance, in the case of patent data, one patent provides information on the authors, the technological class, the assignee, etc, which are possible nodes of the network. These nodes can be connected in a network perspective in which the 'network-link' is identified with the patent. Since the breakthrough research of Hidalgo and Hausmann (2009), there is growing research on this issue, with a large set of applications. For instance, recently Balland and Rigby (2014) apply Hidalgo and Hausmann's methodology to characterize the technological specialization of cities. There is room for interesting research on innovation using this perspective.

Adams, J., 2013. Collaborations: The fourth age of research. Nature 497 (7451), 557–560.

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.

Aghion, P., Howitt, P., March 1992. A model of growth through creative destruction. Econometrica 60 (2), 323–351.

Agrawal, A., Cockburn, I., Galasso, A., Oettl, A., 2014. Why are some regions more innovative than others? the role of small firms in the presence of large labs. Journal of Urban Economics 81, 149–165.

Agrawal, A., Cockburn, I., McHale, J., 2006. Gone but not forgotten: knowledge flows, labor mobility, and enduring social relationships. Journal of Economic Geography 6 (5), 571–591.

Ahuja, G., 2000. Collaboration networks, structural holes and innovation: A longitudinal study. Administrative Science Quarterly 45 (3), 425–455.

Almeida, P., Kogut, B., 1997. The exploration of technological diversity and geographic localization in innovation: start-up firms in the semiconductor industry. Small Business Economics 9 (1), 21–31.

Almendral, J. A., Oliveira, J. G., López, L., Mendes, J., Sanjuán, M. A., 2007. The network of scientific collaborations within the European framework programme. Physica A: Statistical Mechanics and its Applications 384 (2), 675–683.

Anderson, J. E., 2011. The gravity model. Annual Review of Economics 3 (1), 133–160.

Anselin, L., Varga, A., Acs, Z., 1997. Local geographic spillovers between university research and high technology innovations. Journal of urban economics 42 (3), 422–448.

Arora, A., Gambardella, A., 1994. The changing technology of technological change: general and abstract knowledge and the division of innovative labour. Research policy 23 (5), 523–532.

Arthur, W. B., 1989. Competing technologies, increasing returns, and lock-in by historical events. Economic Journal 99, 116–131.

Athey, S., Eckles, D., Imbens, G. W., 2015. Exact p-values for network interference. Working paper.
URL http://arxiv.org/abs/1506.02084v1

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

Audretsch, D. B., Feldman, M. P., 2004. Knowledge spillovers and the geography of innovation. In: Henderson, J. V., Thisse, J.-F. (Eds.), Cities and Geography. Vol. 4 of Handbook of Regional and Urban Economics. Elsevier, Ch. 61, pp. 2713 – 2739.

Autant-Bernard, C., 2001. Science and knowledge flows: evidence from the French case. Research Policy 30 (7), 1069–1078.

Autant-Bernard, C., Billand, P., Frachisse, D., Massard, N., 2007a. 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.

Autant-Bernard, C., Mairesse, J., Massard, N., 2007b. Spatial knowledge diffusion through collaborative networks. Papers in Regional Science 86 (3), 341–350.

Baker, A., Bollobás, B., Hajnal, A., 1990. A Tribute to Paul Erdos. Cambridge University Press.

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.

Balland, P.-A., Boschma, R., Frenken, K., 2015b. Proximity and innovation: from statics to dynamics. Regional Studies 49 (6), 907–920.

Balland, P.-A., De Vaan, M., Boschma, R., 2013. The dynamics of interfirm networks along the industry life cycle: The case of the global video game industry, 1987–2007. Journal of Economic Geography 13 (5), 741–765.

Balland, P.-A., Rigby, D., 2014. The geography and evolution of complex knowledge. Working Paper.

Ballester, C., Calvó-Armengol, A., Zenou, Y., 2006. Who's who in networks. Wanted: the key player. Econometrica 74 (5), 1403–1417.

Banchoff, T., 2002. Institutions, inertia and European Union research policy. Journal of Common Market Studies 40 (1), 1–21.

Banerjee, A., Chandrasekhar, A. G., Duflo, E., Jackson, M. O., 2013. The diffusion of microfinance. Science 341 (6144).

Barabási, A.-L., Albert, R., 1999. Emergence of scaling in random networks. Science 286 (5439), 509–512.

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.

Basalla, G., 1988. The evolution of technology. Cambridge University Press, Cambridge, MA.

Bathelt, H., Malmberg, A., Maskell, P., 2004. Clusters and knowledge: local buzz, global pipelines and the process of knowledge creation. Progress in Human Geography 28 (1), 31–56.

Bavelas, A., 1948. A mathematical model for group structures. Human organization 7 (3), 16–30.

Bavelas, A., 1950. Communication patterns in task-oriented groups. The Journal of the Acoustical Society of America 22, 271–282.

Beauchamp, M. A., 1965. An improved index of centrality. Behavioral Science 10 (2), 161–163.

Beaver, D. d., 2001. Reflections on scientific collaboration (and its study): past, present, and future. Scientometrics 52 (3), 365–377.

Belderbos, R., Cassiman, B., Faems, D., Leten, B., van Looy, B., 2014. Co-ownership of intellectual property: Exploring the value-appropriation and value-creation implications of co-patenting with different partners. Research Policy 43 (5), 841 – 852.

Bell, D. C., Atkinson, J. S., Carlson, J. W., 1999. Centrality measures for disease transmission networks. Social Networks 21 (1), 1–21.

Bercovitz, J., Feldman, M., 2011. The mechanisms of collaboration in inventive teams: Composition, social networks, and geography. Research Policy 40 (1), 81–93.

Bergé, L. R., 2015. Network proximity in the geography of research collaboration. Papers in Evolutionary Economic Geography, Number 15.07 Working paper.

Berliant, M., Fujita, M., 2008. Knowledge creation as a square dance on the hilbert cube. International Economic Review 49 (4), 1251–1295.

Berliant, M., Fujita, M., 2012. Culture and diversity in knowledge creation. Regional Science & Urban Economics 42 (4), 648 – 662.

Bettencourt, L. M., Lobo, J., Strumsky, D., 2007. Invention in the city: Increasing returns to patenting as a scaling function of metropolitan size. Research Policy 36 (1), 107 – 120.

Blau, J. R., September 1974. Patterns of communication among theoretical high energy physicists. Sociometry 37 (3), 391–406.

Bloom, N., Schankerman, M., Van Reenen, J., 2013. Identifying technology spillovers and product market rivalry. Econometrica 81 (4), 1347–1393.

Bonacich, P., 1972. Factoring and weighting approaches to status scores and clique identification. Journal of Mathematical Sociology 2 (1), 113–120.

Bonacich, P., 1987. Power and centrality: A family of measures. American journal of sociology 92, 1170–1182.

Borgatti, S. P., 1995. Centrality and aids. Connections 18 (1), 112–114.

Borgatti, S. P., 2005. Centrality and network flow. Social Networks 27 (1), 55–71.

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

Boschma, R., Balland, P.-A., de Vaan, M., 2014a. The formation of economic networks: a proximity approach. In: Torre, A., Wallet, F. (Eds.), Regional Development and Proximity Relations. New Horizon in Regional Science. Edward Elgar Publishing, pp. 243–266.

Brambor, T., Clark, W. R., Golder, M., 2006. Understanding interaction models: Improving empirical analyses. Political Analysis 14 (1), 63–82.

Brenner, T., Broekel, T., 2011. Methodological issues in measuring innovation performance of spatial units. Industry and Innovation 18 (1), 7–37.

Breschi, S., Lenzi, C., 2012. Net city: how co-invention networks shape inventive productivity in US cities. Unpublished.

Breschi, S., Lenzi, C., 2015. The role of external linkages and gatekeepers for the renewal and expansion of US cities knowledge base, 1990-2004. Regional Studies 49 (5), 782–797.

Breschi, S., Lissoni, F., 2001. Knowledge spillovers and local innovation systems: a critical survey. Industrial and Corporate Change 10 (4), 975–1005.

Breschi, S., Lissoni, F., 2003. Mobility and social networks: Localized knowledge spillovers revisited. Working paper. Bocconi University.

Breschi, S., Lissoni, F., 2005. "Cross-firm" inventors and social networks: Localized knowledge spillovers revisited. Annales d'Économie et de Statistique (79/80), 189–209.

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.

Brin, S., Page, L., 1998. The anatomy of a large-scale hypertextual Web search engine. Computer networks and ISDN systems 30 (1-7), 107–117.

Broekel, T., Balland, P.-A., Burger, M., van Oort, F., 2013. Modeling knowledge networks in economic geography: A discussion of four empirical strategies. Annals of Regional Science 53 (2), 423–452.

Burgess, R. L., 1969. Communication networks and behavioral consequences. Human Relations 22 (2), 137–159.

Burt, R. S., 1992. Structural holes: The social structure of competition. Harvard University Press, Cambridge, MA.

Burt, R. S., 2005. Brokerage and Closure: An Introduction to Social Capital. OUP Oxford.

Buzard, K., Carlino, G., 2013. The geography of research and development activity in the US. In: Giarratani, F., Hewings, G., McCann, P. (Eds.), Handbook of Industry Studies and Economic Geography. Edward Elgar Publishing, p. 389.

Calvó-Armengol, A., Patacchini, E., Zenou, Y., 2009. Peer effects and social networks in education. Review of Economic Studies 76 (4), 1239–1267.

Calvó-Armengol, A., Zenou, Y., 2004. Social networks and crime decisions: The role of social structure in facilitating delinquent behavior. International Economic Review 45 (3), 939–958.

Cameron, A. C., Gelbach, J. B., Miller, D. L., 2011. Robust inference with multiway clustering. Journal of Business & Economic Statistics 29 (2), 138–249.

Cameron, A. C., Miller, D. L., 2015. A practitioner's guide to cluster-robust inference. Journal of Human Resources 50 (2), 317–372.

Carayol, N., Cassi, L., 2009. Who's who in patents. a Bayesian approach. In: Les Cahiers du GREThA.

Carayol, N., Cassi, L., Roux, P., 2014. Unintended triadic closure in social networks: The strategic formation of research collaborations between French inventors. Working paper GREThA 2014-13.

Carayol, N., Cassi, L., Roux, P., 2015. Unintended closure in social networks. the strategic formation of research collaborations among French inventors (1983-2005). mimeo.

Carayol, N., Roux, P., 2007. The strategic formation of inter-individual collaboration networks. evidence from co-invention patterns. Annales d'Economie et de Statistique, 275–301.

Carlino, G. A., Hunt, R. M., Carr, J. K., Smith, T. E., 2012. The agglomeration of R&D labs. FRB of Philadelphia Working Paper (12-22).

Carlino, G. A., Kerr, W. R. K., 2015. Agglomeration and innovation. In: Duranton, G., Henderson, J. V., Strange, W. C. (Eds.), Handbook of Regional and Urban Economics. Vol. 5. Elsevier, Ch. 6, pp. 349–404.

Carrincazeaux, C., Lung, Y., Rallet, A., 2001. Proximity and localisation of corporate R&D activities. Research Policy 30 (5), 777–789.

Cassi, L., Morrison, A., Rabellotti, R., 2015. Proximity and scientific collaboration: Evidence from the global wine industry. Tijdschrift voor Economische en Sociale Geografie 106 (2), 205–219.

Cassi, L., Plunket, A., 2015. Research collaboration in co-inventor networks: Combining closure, bridging and proximities. Regional Studies 49 (6), 936–954.

Cassiman, B., Veugelers, R., 2006. In search of complementarity in innovation strategy: Internal R&D and external knowledge acquisition. Management Science 52 (1), 68–82.

Castells, M., 1996. The Rise of the Network Society. Blackwell Publishers, Oxford.

Catalini, C., 2012. Microgeography and the direction of inventive activity. Rotman School of Management Working Paper (2126890).

Coffano, M., Tarasconi, G., 2014. CRIOS-Patstat Database: Sources, contents and access rule. Center for Research on Innovation, Organization and Strategy, CRIOS working paper 1.

Collins, H. M., 2001. Tacit knowledge, trust and the Q of sapphire. Social Studies of Science 31 (1), 71–85.

Cowan, R., David, P. A., Foray, D., 2000. The explicit economics of knowledge codification and tacitness. Industrial and Corporate Change 9 (2), 211–253.

Cowan, R., Foray, D., 1997. The economics of codification and the diffusion of knowledge. Industrial and Corporate Change 6 (3), 595–622.

Cowan, R., Jonard, N., 2004. Network structure and the diffusion of knowledge. Journal of economic Dynamics and Control 28 (8), 1557–1575.

Crépon, B., Duguet, E., 1997. Estimating the innovation function from patent numbers: GMM on count panel data. Journal of Applied Econometrics 12 (3), 243–263.

Criscuolo, P., Verspagen, B., 2008. Does it matter where patent citations come from? inventor vs. examiner citations in European patents. Research Policy 37 (10), 1892–1908.

Czepiel, J. A., 1974. Word-of-mouth processes in the diffusion of a major technological innovation. Journal of Marketing Research 11, 172–180.

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.

Dasgupta, P., David, P. A., 1994. Toward a new economics of science. Research Policy 23 (5), 487–521.

David, P. A., 1985. Clio and the economics of QWERTY. American Economic Review 72 (2), 332–337.

Debreu, G., Herstein, I. N., 1953. Nonnegative square matrices. Econometrica: Journal of the Econometric Society, 597–607.

Defazio, D., Lockett, A., Wright, M., 2009. Funding incentives, collaborative dynamics and scientific productivity: Evidence from the EU framework program. Research Policy 38 (2), 293–305.

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.

Duranton, G., Puga, D., 2004. Micro foundation of urban agglomeration economies. In: Henderson, J. V., Thisse, J.-F. (Eds.), Handbook of Regional and Urban Economics. Vol. 4. North-Holland, Ch. 48, pp. 2063–2117.

European Commission, 2012b. Guide to research and innovation strategies for smart specialisations (RIS 3). Joint Research Centre.

European Commission, 2013. Regulation (EU) no 1291/2013 of the European parliament and of the council of 11 December 2013. Official Journal of the European Union, L347.

European Commission, 2014. HORIZON 2020 in brief. the EU framework programme for research & innovation. Directorate-General for Research and Innovation.

Everett, M. G., Borgatti, S. P., 1999. The centrality of groups and classes. The Journal of Mathematical Sociology 23 (3), 181–201.

Fafchamps, M., van der Leij, M. J., Goyal, S., 2010. Matching and network effects. Journal of the European Economic Association 8 (1), 203 – 231.

Faucheux, C., Moscovici, S., 1960. Etudes sur la créativité des groupes tâche, structure des communications et réussite. Bulletin du C.E.R.P. 9, 11–22.

Feldman, M. P., 1999. The new economics of innovation, spillovers and agglomeration: A review of empirical studies. Economics of Innovation & New Technology 8 (1-2), 5–25.

Feldman, M. P., Kogler, D. F., 2010. Stylized facts in the geography of innovation. Handbook of the Economics of Innovation 1, 381–410.

Fleming, L., 2001. Recombinant uncertainty in technological search. Management science 47 (1), 117–132.

Fleming, L., 2002. Finding the organizational sources of technological breakthroughs: the story of Hewlett-Packard's thermal ink-jet. Industrial and Corporate Change 11 (5), 1059–1084.

Fleming, L., King III, C., Juda, A. I., 2007. Small worlds and regional innovation. Organization Science 18 (6), 938–954.

Fleming, L., Marx, M., 2006. Managing creativity in small worlds. California Management Review 48 (4), 6.

Freeman, L. C., 1977. A set of measures of centrality based on betweenness. Sociometry 40 (1), 35–41.

Freeman, L. C., 1979. Centrality in social networks conceptual clarification. Social Networks 1 (3), 215–239.

Freeman, R. B., Ganguli, I., Murciano-Goroff, R., 2014. Why and wherefore of increased scientific collaboration. National Bureau of Economic Research.

Frenken, K., Hardeman, S., Hoekman, J., 2009a. Spatial scientometrics: Towards a cumulative research program. Journal of Informetrics 3 (3), 222 – 232.

Frenken, K., Hoekman, J., Kok, S., Ponds, R., van Oort, F., van Vliet, J., 2009b. Death of distance in science? A gravity approach to research collaboration. In: Innovation networks. Springer Berlin Heidelberg, pp. 43–57.

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

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.

Giuliani, E., Morrison, A., Pietrobelli, C., Rabellotti, R., 2010. Who are the researchers that are collaborating with industry? an analysis of the wine sectors in chile, south africa and italy. Research Policy 39 (6), 748–761.

Glaeser, E. L., 2011. Triumph of the city. Penguin Press, New York.

Glänzel, W., 2001. National characteristics in international scientific co-authorship relations. Scientometrics 51 (1), 69–115.

Graf, H., 2011. Gatekeepers in regional networks of innovators. Cambridge Journal of Economics 35 (1), 173–198.

Granovetter, M., 1995. Getting a job: A study of contacts and careers. University of Chicago Press.

Granovetter, M. S., 1973. The strength of weak ties. American Journal of Sociology 78 (6), 1360–1380.

Guimera, R., Uzzi, B., Spiro, J., Amaral, L. A. N., 2005. Team assembly mechanisms determine collaboration network structure and team performance. Science 308 (5722), 697–702.

Gulati, R., Gargiulo, M., 1999. Where do interorganizational networks come from? American Journal of Sociology 104 (5), 1439 – 1493.

Hall, B. H., Jaffe, A., Trajtenberg, M., Spring 2005. Market value and patent citations. RAND Journal of economics 36 (1), 16–38.

Hanaki, N., Nakajima, R., Ogura, Y., 2010. The dynamics of R&D network in the IT industry. Research policy 39 (3), 386–399.

Harhoff, D., Narin, F., Scherer, F. M., Vopel, K., 1999. Citation frequency and the value of patented inventions. Review of Economics and statistics 81 (3), 511–515.

Hazir, C. S., Lesage, J., Autant-Bernard, C., 2014. The role of R&D collaboration networks on regional innovation performance. Working paper GATE 2014-26.

Helsley, R. W., Strange, W. C., 2004. Knowledge barter in cities. Journal of Urban Economics 56 (2), 327–345.

Henderson, R. M., Clark, K. B., 1990. Architectural innovation: The reconfiguration of existing product technologies and the failure of established firms. Administrative Science Quarterly, 9–30.

Hidalgo, C. A., Hausmann, R., 2009. The building blocks of economic complexity. Proceedings of the National Academy of Sciences 106 (26), 10570–10575.

Hoekman, J., Frenken, K., Tijssen, R. J., 2010. Research collaboration at a distance: Changing spatial patterns of scientific collaboration within Europe. Research Policy 39 (5), 662 – 673.

Hoekman, J., Frenken, K., van Oort, F., 2009. The geography of collaborative knowledge production in Europe. Annals of Regional Science 43 (3), 721 – 738.

Hoekman, J., Scherngell, T., Frenken, K., Tijssen, R., 2013. Acquisition of European research funds and its effect on international scientific collaboration. Journal of Economic Geography 13 (1), 23–52.

INSEE, 2010. Atlas des zones d'emploi 2010.

Jackson, M., Wolinsky, A., 1996. A strategic model of social and economic networks. Journal of Economic Theory 71 (1), 44–74.

Jackson, M. O., 2010. Social and economic networks. Princeton University Press.

Jackson, M. O., Rogers, B. W., 2007. Meeting strangers and friends of friends: How random are social networks? American Economic Review 97 (3), 890–915.

Jacobs, J., 1961. The death and life of great American cities. Vintage.

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

Jaffe, A. B., 1986. Technological opportunity and spillovers of r & d: Evidence from firms' patents, profits, and market value. American Economic Review 76, 984–1001.

Jayet, H., 1985. Les zones d'emploi et l'analyse locale des marchés du travail. Économie et statistique 182 (1), 37–44.

Jones, B. F., 2009. The burden of knowledge and the 'death of the renaissance man': is innovation getting harder? Review of Economic Studies 76 (1), 283–317.

Jones, B. F., Wuchty, S., Uzzi, B., 2008. Multi-university research teams: shifting impact, geography, and stratification in science. Science 322 (5905), 1259–1262.

Jones, C. I., 1995. R&D-based models of economic growth. Journal of Political Economy 103 (4), 759–784.

Kaiser, U., Kongsted, H. C., Rønde, T., 2015b. Does the mobility of R&D labor increase innovation? Journal of Economic Behavior & Organization 110, 91 – 105.

Katz, J. S., 1994. Geographical proximity and scientific collaboration. Scientometrics 31 (1), 31–43.

Katz, J. S., Martin, B. R., 1997. What is research collaboration? Research Policy 26 (1), 1–18.

Katz, L., 1953. A new status index derived from sociometric analysis. Psychometrika 18 (1), 39–43.

Kirat, T., Lung, Y., 1999. Innovation and proximity territories as loci of collective learning processes. European Urban and Regional Studies 6 (1), 27–38.

Kogut, B., Zander, U., 1992. Knowledge of the firm, combinative capabilities, and the replication of technology. Organization Science 3 (3), 383–397.

Krackhardt, D., 1999. The ties that torture: Simmelian tie analysis in organizations. Research in the Sociology of Organizations 16 (1), 183–210.

Krugman, P. R., 1991b. Geography and Trade. The MIT Press.

Lanjouw, J. O., Pakes, A., Putnam, J., 1998. How to count patents and value intellectual property: The uses of patent renewal and application data. The Journal of Industrial Economics 46 (4), 405–432.

Lata, R., Scherngell, T., Brenner, T., 2015. Integration processes in European research and development: A comparative spatial interaction approach using project based research and development networks, co-patent networks and co-publication networks. Geographical Analysis, 1–27.

Leavitt, H. J., 1951. Some effects of certain communication patterns on group performance. The Journal of Abnormal and Social Psychology 46 (1), 38–50.

Lee, Y.-N., Walsh, J. P., Wang, J., 2015. Creativity in scientific teams: Unpacking novelty and impact. Research Policy 44 (3), 684–697.

Lobo, J., Strumsky, D., 2008. Metropolitan patenting, inventor agglomeration and social networks: A tale of two effects. Journal of Urban Economics 63 (3), 871–884.

Lucas, R. E., 1988. On the mechanics of economic development. Journal of Monetary Economics 22 (1), 3–42.

Lundvall, B.-Å., 1992. National systems of innovation: Toward a theory of innovation and interactive learning. Pinter, London, vol. 2.

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.

Maggioni, M. A., Uberti, T. E., 2009. Knowledge networks across Europe: which distance matters? Annals of Regional Science 43 (3), 691 – 720.

Maggioni, M. A., Uberti, T. E., 2011. Networks and geography in the economics of knowledge flows. Quality & Quantity 45, 1031 – –1051.

Maraut, S., Dernis, H., Webb, C., Spiezia, V., Guellec, D., 2008. The OECD REGPAT Database. OECD Science, Technology and Industry Working Papers.

Marshall, A., 1890. Principle of Economics, 8th Edition. London, Macmillan and Co.

Marx, M., Singh, J., Fleming, L., 2015. Regional disadvantage? employee non-compete agreements and brain drain. Research Policy 44 (2), 394–404.

Marx, M., Strumsky, D., Fleming, L., 2009. Mobility, skills, and the Michigan non-compete experiment. Management Science 55 (6), 875–889.

McPherson, M., Smith-Lovin, L., Cook, J. M., 2001. Birds of a feather: Homophily in social networks. Annual Review of Sociology 27, 415–444.

Menon, C., 2015. Spreading big ideas? the effect of top inventing companies on local inventors. Journal of Economic Geography 15 (4), 743–768.

Miguélez, E., Gòmez-Miguélez, I., 2011. Singling out individual inventors from patent data. Institut de Recerca en Economia Aplicada Regional i Pública Working Papers 5.

Miguélez, E., Moreno, R., 2014. What attracts knowledge workers? the role of space and social networks. Journal of Regional Science 54 (1), 33–60.

Montgomery, J. D., 1991. Social networks and labor-market outcomes: Toward an economic analysis. American Economic Review 81 (5), 1408–1418.

Montobbio, F., Primi, A., Sterzi, V., 2015. IPRs and international knowledge flows: Evidence from six large emerging countries. Tijdschrift voor Economische en Sociale Geografie 106 (2), 187–204.

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.

Murata, Y., Nakajima, R., Okamoto, R., Tamura, R., 2014. Localized knowledge spillovers and patent citations: A distance-based approach. Review of Economics and Statistics 96 (5), 967–985.

Narin, F., Stevens, K., Whitlow, E. S., 1991. Scientific co-operation in Europe and the citation of multinationally authored papers. Scientometrics 21 (3), 313–323.

Nelson, R., Winter, S., 1982. An Evolutionary Theory of Economic Change. Belknap Press, Cambridge, MA.

Newman, M. E., 2001. The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences 98 (2), 404–409.

Opsahl, T., Agneessens, F., Skvoretz, J., 2010. Node centrality in weighted networks: Generalizing degree and shortest paths. Social Networks 32 (3), 245–251.

Orlando, M. J., 2000. On the importance of geographic and technological proximity for R&D spillovers: An empirical investigation. FRB of Kansas City Research Working Paper No. 00-02.

OST, 2010. Indicateurs de sciences et de technologies.

Owen-Smith, J., Powell, W. W., 2004. Knowledge networks as channels and conduits: The effects of spillovers in the Boston biotechnology community. Organization Science 15 (1), 5–21.

Pezzoni, M., Lissoni, F., Tarasconi, G., 2014. How to kill inventors: Testing the Massacrator[I] algorithm for inventor disambiguation. Scientometrics 101 (1), 477–504.

Picci, L., 2010. The internationalization of inventive activity: A gravity model using patent data. Research Policy 39 (8), 1070–1081.

Pitts, F. R., 1965. A graph theoretic approach to historical geography. The Professional Geographer 17 (5), 15–20.

Ponds, R., Oort, F. v., Frenken, K., 2010. Innovation, spillovers and university-industry collaboration: an extended knowledge production function approach. Journal of Economic Geography 10 (2), 231–255.

Ponds, R., van Oort, F., Frenken, K., 2007. The geographical and institutional proximity of research collaboration. Papers in Regional Science 86 (3), 423 – 443.

Porter, M., November-December 1998. Clusters and the new economics of competition. In: Harvard Business Review. Harvard University.

Porter, M. E., 1990. The competitive advantage of nations. Harvard Business Review 68 (2), 73–93.

Powell, W. W., Grodal, S., 2005. Networks of innovators. In: The Oxford Handbook of Innovation. Oxford University Press, Oxford, pp. 56–85.

Rodríguez-Pose, A., Crescenzi, R., 2008. Mountains in a flat world: why proximity still matters for the location of economic activity. Cambridge Journal of Regions, Economy and Society 1 (3), 371–388.

Romer, P. M., 1990. Endogenous technological change. Journal of Political Economy 98 (5), 71–102.

Roy, J. R., Thill, J.-C., 2004. Spatial interaction modelling. Papers in Regional Science 83 (1), 339–361.

Royal Society Science Policy Centre, 2011. Knowledge, Networks and Nations: Global Scientific Collaboration in the 21st Century. Royal Society, London.

Santos Silva, J. a. M. C., Tenreyro, S., 2006. The log of gravity. Review of Economics and Statistics 88 (4), 641–658.

Saxenian, A., 1991. The origins and dynamics of production networks in Silicon Valley. Research Policy 20 (5), 423–437.

Saxenian, A., 1996. Regional advantage: Culture and competition in Silicon Valley and Route 128. Harvard University Press.

Scherngell, T. (Ed.), 2013. The geography of networks and R&D collaborations. Springer-Physica Verlag, Berlin-Heidelberg-New York.

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.

Schumpeter, J. A., 1943. Capitalism, socialism and democracy. Reprint 2010. Routledge.

Sebestyén, T., Varga, A., 2013a. A novel comprehensive index of network position and node characteristics in knowledge networks: Ego network quality. In: Scherngell, T. (Ed.), The Geography of Networks and R&D Collaborations. Advances in Spatial Science. Springer International Publishing, pp. 71–97.

Sebestyén, T., Varga, A., 2013b. Research productivity and the quality of interregional knowledge networks. Annals of Regional Science 51 (1), 155–189.

Singh, J., 2005. Collaborative networks as determinants of knowledge diffusion patterns. Management Science 51 (5), 756 – 770.

Singh, J., Fleming, L., 2010. Lone inventors as sources of breakthroughs: Myth or reality? Management Science 56 (1), 41–56.

Singh, J., Marx, M., 2013. Geographic constraints on knowledge spillovers: political borders vs. spatial proximity. Management Science 59 (9), 2056–2078.

Smith, R., 2006. The network of collaboration among rappers and its community structure. Journal of Statistical Mechanics, PO2006.

Sorenson, O., Fleming, L., 2004. Science and the diffusion of knowledge. Research policy 33 (10), 1615–1634.

Sorenson, O., Rivkin, J. W., Fleming, L., 2006. Complexity, networks and knowledge flow. Research Policy 35 (7), 994–1017.

Sorenson, O., Stuart, T., 2001. Syndication networks and the spatial distribution of venture capital investments. American Journal of Sociology 106 (6), 1546–588.

Sorenson, O., Stuart, T. E., 2008. Bringing the context back in: Settings and the search for syndicate partners in venture capital investment networks. Administrative Science Quarterly 53 (2), 266 – 294.

Storper, M., Venables, A. J., 2004. Buzz: face-to-face contact and the urban economy. Journal of Economic Geography 4 (4), 351–370.

ter Wal, A. L. J., 2011. Networks and geography in the economics of knowledge flows: a commentary. Quality & Quantity 45, 1059–1063.

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, 589–620.

Thompson, P., Fox-Kean, M., 2005. Patent citations and the geography of knowledge spillovers: A reassessment. American Economic Review 95 (1), 450–460.

Torre, A., Rallet, A., 2005. Proximity and localization. Regional studies 39 (1), 47–59.

Trajtenberg, M., Spring 1990. A penny for your quotes: Patent citations and the value of innovations. The Rand Journal of Economics 21 (1), 172–187.

Uzzi, B., Spiro, J., 2005. Collaboration and creativity: The small world problem. American journal of sociology 111 (2), 447–504.

van Dijk, J., Maier, G., 2006. ERSA conference participation: does location matter? Papers in Regional Science 85 (4), 483 – 504.

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

Wanzenböck, I., Scherngell, T., Brenner, T., 2014. Embeddedness of regions in European knowledge networks: a comparative analysis of inter-regional R&D collaborations, co-patents and co-publications. Annals of Regional Science 53 (2), 337–368.

Wanzenböck, I., Scherngell, T., Lata, R., 2015. Embeddedness of European regions in European Union-funded research and development (R&D) networks: A spatial econometric perspective. Regional Studies 49 (10), 1685–1705.

Watts, D. J., Strogatz, S. H., 1998. Collective dynamics of 'small-world' networks. nature 393 (6684), 440–442.

Weitzman, M. L., 1998. Recombinant growth. Quarterly Journal of Economics, 331–360.

Windmeijer, F., 2008. GMM for panel count data models. In: Laszlo, M., Sevestre, P. (Eds.), The Econometrics of Panel Data, 3rd Edition. Springer.

Wooldridge, J. M., 2010. Econometric analysis of cross section and panel data, 2nd Edition. The MIT Press, Cambridge, Massachussets, London, England.

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

Zucker, L. G., Darby, M. R., Armstrong, J., 1998. Geographically localized knowledge: spillovers or markets? Economic Inquiry 36 (1), 65–86.

Summary in French – Résumé en français

Le but principal de cette thèse est de clarifier les mécanismes impliqués dans la création de connaissance, en s'intéressant particulièrement au rôle des réseaux sociaux dans ce processus. A cette fin, la littérature sur les réseaux sociaux est combinée à la littérature sur les déterminants des collaborations, et sur la sociologie, et déterminants, des collaborations. Le réseau social considéré ici est celui basé sur le réseau formel de collaborations pour la production de connaissance (par exemple les brevets ou les articles scientifiques).

Bien que cette thèse porte spécifiquement sur un déterminant de la création de connaissance, le réseau social, elle prend aussi en compte le rôle de la géographie. En effet, comme les acteurs de la création de connaissance sont très concentrés dans l'espace, cette thèse propose de mieux comprendre l'interaction entre la géographie et les réseaux sociaux. Ainsi, le réseau social est un moyen d'investiguer les les mécanismes impliqués dans les "externalités de connaissances localisées".

Le point de vue porté par cette thèse est que les réseaux sont un vecteur de diffusion d'information et d'idées, et est un vecteur important de la création de connaissance et de la formation de collaborations. La question de recherche requiert d'évaluer précisément la position des agents dans le réseau. De fait, des outils issus de la théorie des réseaux sociaux sont utilisés pour définir de nouvelles mesures et afin d'évaluer le rôle du réseau social comme un vecteur de création de connaissance.

Les trois chapitres de cette thèse contribuent à : *i*) mieux comprendre les déterminants des collaborations dans la création de connaissances en mettant en avant la relation entre réseau social et géographie, *ii*) apporter une évaluation critique de la position des régions dans les réseaux de R&D, et *iii*) étudier le lien entre la position des inventeurs dans le réseau et la production régionale de connaissance.

Chacun de ces chapitres apporte des contributions théoriques et méthodologiques, et deux incluent une étude empirique. La suite décrit les questions posées par chaque chapitre ainsi que leurs méthodologies et résultats.

Chapitre 1. Dans le premier chapitre il est question d'évaluer le rôle conjoint du réseau social et de la géographie dans la formation des collaborations scientifiques. Ce chapitre

s'inscrit à la fois dans la littérature sur les déterminants de collaborations (Maggioni et al., 2007; Hoekman et al., 2010; Morescalchi et al., 2015) et dans le débat opposant la géographie au réseau social (Bathelt et al., 2004; Boschma, 2005; Rodríguez-Pose and Crescenzi, 2008). Dans un premier temps, ce chapitre discute de façon théorique des causes affectant la collaboration entre chercheurs, en focalisant principalement sur les mécanismes impliqués par le réseau social. En particulier, les possibles conséquences de l'interaction entre la proximité de réseau et la proximité géographique sont discutées, en mettant en avant dans quelles conditions elles peuvent être complémentaires ou substituts. Dans ce chapitre, nous soutenons l'idée que la proximité de réseau affecte positivement les collaborations. Par contre, nous restons agnostiques sur la question de la complémentarité / substituabilité de l'effet des proximités géographique et de réseau.

Dans un deuxième temps, une analyse empirique est mise en place afin de tester les hypothèses développées dans la discussion théorique. À cette fin, nous employons des données sur les collaborations scientifiques dans le domaine de la chimie, ayant lieu dans cinq pays Européens (France, Allemagne, Royaume-Uni, Espagne et Italie), entre 2000 et 2005. Un modèle gravitaire est employé pour évaluer les déterminants des flux de collaborations inter-régionales. De par l'absence de mesure existante pour évaluer la proximité sociale entre régions, ce chapitre fait aussi une contribution méthodologique en introduisant une telle mesure. Cette mesure de proximité sociale inter-régionale se base sur les connexions indirectes entre chercheurs et s'appuie sur des fondements micro-économiques.

Les résultats démontrent un effet positif de la proximité sociale comme déterminant des collaborations. De plus, la substituabilité entre les proximités sociale et géographique est clairement mise en évidence. Cela veut dire que la probabilité de collaboration due à la proximité sociale croît avec la distance géographique séparant les régions. C'est une mise en avant que la proximité sociale permet de passer les barrières liées à la géographie.

Chapitre 2. Le deuxième chapitre se propose de caractériser le positionnement des régions dans les réseaux sociaux, en particulier les réseaux d'innovation. En effet, il y a une importance croissante est donnée à l'évaluation la position des régions dans les réseaux inter-régionaux (Maggioni and Uberti, 2011; Scherngell, 2013). Néanmoins, la question de la mesure de la position relative des régions dans les réseaux n'a toujours pas de réponse claire. Le but de ce chapitre est donc d'évaluer de façon critique de la notion de centralité de réseau dans le contexte de réseau inter-régional de R&D.

Tout d'abord, ce chapitre met en avant les possibilités d'appliquer à ce contexte les centralité de réseau existantes. Principalement, il existe deux approches. La première possibilité est de considérer le réseau inter-régional comme un réseau pondéré. Les nœuds composants le réseau représentent alors chaque région et les liens entre chaque nœud mesurent l'intensité des flux de collaborations entre régions. Dans ce cas, il suffit alors d'employer les versions pondérées des mesures de centralité existantes. La deuxième approche consiste à considérer la centralité régionale comme la somme des centralités de ses

agents. Dans ce cas, il faut calculer les centralités au niveau individuel (i.e. au niveau du réseau des chercheurs ou des firmes), puis agréger les centralités au niveau régional. La discussion théorique portant sur ces méthodes met en avant des limites claires en termes d'interprétation. En effet, les mesures de centralité sont en premier lieu définies à l'échelle des réseaux individuels, or nous montrons dans ce chapitre que ces dernières perdent l'essentiel de le pouvoir explicatif lorsque l'on passe à un niveau agrégé. Un problème majeur réside dans la dualité entre le niveau meso (i.e., régional) pour lequel la centralité doit être calculée, et le niveau micro, où interviennent les vrais acteurs du réseaux (i.e., les chercheurs ou les firmes).

A la suite de cette discussion, une nouvelle mesure, que l'on nommera *bridging-centrality*, est introduite. Cette dernière est basée sur la propension des agents d'une région à être des "ponts" entre agents d'autres régions, i.e. à être des agents connectés à plusieurs régions différentes. Dans le contexte des réseaux de R&D, ce type de position revêt un rôle important pour deux raisons. Dans un premier temps, cela joue un rôle important pour la région en soi. En effet, il s'agit d'un signe que ses agents sont connectés à une variété d'autre régions, ce qui permet de diversifier le fonds de connaissance régional (e.g., Bathelt et al., 2004; Berliant and Fujita, 2012) Ensuite, cette position est aussi bénéfique pour les autres régions. En effet, dans un contexte de formation de réseau, une régions qui en interconnecte d'autres peut être vue comme un répertoire d'information, et ainsi peut faciliter les connexions futures entre ces autres régions. Finalement, un autre avantage de cette mesure est qu'elle a été créée spécifiquement pour les réseaux régionaux et prend en compte la dualité des niveaux micro/meso.

Pour illustrer cette mesure et la comparer avec d'autres mesures existantes, des données sur les dépôts de brevet Européens ont été utilisées. Malgré des similarités entre les différentes mesures de centralité, la *bridging-centrality* montre des différences significatives avec les autres. Par exemple, pour le cas de la région Île de France, le rang que lui donne la *bridging-centrality* est inférieur à celui calculé par les mesures standard car la part des connexions internes à la région est si grande qu'au final les agents interconnectent peu les autres régions.

Chapitre 3. Le troisième chapitre s'intéresse à la relation existante entre la position des agents dans le réseau d'inventeur et la production régionale d'innovation. Selon la littérature sur la production de connaissance, la structure du réseau devrait influencer la diffusion d'idée et la productivité des inventeurs (e.g., Cowan and Jonard, 2004; Fleming et al., 2007; Singh and Fleming, 2010). Ce chapitre propose de tester cette assertion avec une méthodologie originale. Pour cela, un modèle est introduit dans lequel la productivité des inventeurs est reliée au comportement de leur collaborateurs. Ce modèle contient trois éléments : connectivité, complémentarité et rivalité. Le premier élément, la connectivité, relate simplement qu'il y a un lien entre la productivité d'un inventeur et son réseau. Les deux autres éléments rendent compte de *comment* le réseau affecte la production des inventeurs. La complémentarité stipule que la productivité d'un inventeur sera augmentée

s'il collaborent avec d'autres inventeurs qui sont plus productifs. Enfin, la rivalité relate l'idée qu'un inventeur peut bénéficier moins d'un collaborateur qui est très connecté, car il pourra moins consacrer de temps à la collaboration (see e.g., Jackson and Wolinsky, 1996). Ce simple modèle stylisé prédit qu'à l'équilibre, la production des inventeurs est dépendante de leur position dans le réseau. Cette position dans le réseau correspond à une nouvelle forme de centralité de réseau qui est basée sur les valeur de : connectivité, complémentarité et rivalité.

Pour évaluer empiriquement si l'influence du réseau sur la production régionale d'innovation, des données sur le réseau de brevets français sont utilisées, de 1981 à 2003. La production régionale d'innovation est estimée en fonction de déterminants régionaux, d'effets fixes pour contrôler pour les variations temporelles et inter-régionales, et de la centralité de réseau des agents de ces régions françaises. Le but étant d'estimer les paramètres du modèle. Ces derniers permettent d'appréhender quelle type de structure de réseau est la plus favorable à l'innovation régionale.

Dans un premier temps, les résultats montrent que le réseau d'inventeur influence positivement la production d'innovations (i.e., la connectivité est positive). Il y a aussi des signes de complémentarité dans le réseau, ce qui signifie que les inventeurs sont plus productifs quand ils collaborent avec des individus biens placés dans le réseau. Ce résultat corrobore l'importance accordée par la littérature au "star-inventors" dans la diffusion des idées au sein d'un réseau (Zucker et al., 1998; Menon, 2015). Enfin, les résultats ne permettent pas de montrer des signes de rivalité au sein du réseau. Ainsi, chaque nouvelle connexion entre deux inventeurs qui n'étaient pas déjà connectés reste toujours bénéfique pour l'innovation régionale.

Le rôle des réseaux sociaux dans la géographie de l'innovation et de la collaboration: Trois essais

Resumé : Cette thèse porte sur la création de connaissances scientifiques et technologiques, et son lien avec la géographie et le réseau social. En ce sens la thèse s'attache à mieux identifier le rôle du réseau social dans la production de connaissance, et à éclairer le lien entre réseau social et géographie dans la formation des collaborations, en mettant en avant dans quelles conditions le réseau permet de s'affranchir de cette dernière. A cet égard, cette thèse apporte plusieurs contributions théoriques, méthodologiques et empiriques. L'essentiel de la thèse s'applique à assembler les mécanismes qui lient le réseau social à la production de connaissances. La discussion théorique est ensuite appuyée par une analyse empirique dans deux contextes liés la création de connaissances. D'une part la thèse analyse la formation du réseau des collaborations scientifiques en Europe dans le domaine de la chimie, mettant en avant l'interaction réseau *versus* géographie dans la formation des collaborations. D'autre part, elle évalue le rôle du réseau d'inventeur dans la performance des zones d'emploi françaises en termes de production d'innovation, en se focalisant sur le type de structure de réseau qui favorise le plus l'innovation. Les résultats principaux sont que l'expansion du réseau social – mesuré par la connectivité des inventeurs – a un effet bénéfique sur l'innovation. De plus, il est montré que le réseau social permet en partie de s'affranchir de la barrière géographique pour collaborer. Ces résultats apportent des éclairages sur le rôle du réseau dans l'ogranisation spatiale des activités scientifiques et technologiques.

Mots-clefs : innovation ; collaboration ; formation de réseau ; économie géographique

---

Social networks and the geography of innovation and research collaboration: Three essays

This thesis pertains to understanding how social networks and geography affect the creation of new knowledge. More precisely, this thesis will question how the social network of collaboration can influence the production of knowledge, how do geography and the social network interact, and whether the social network can help to bypass geography. Answering these questions required to make some theoretical, methodological and empirical contributions. One part of the thesis gathers the mechanisms linking the social network to knowledge creation, while another focuses on the interplay of geography and the network into the collaboration process. Following this theoretical discussion, two empirical studies are laid out. First, it assesses the formation of scientific collaborations in Europe in the field of chemistry. This study focus on the competing role between the social network and geography to shaping new collaborations. Then, the thesis comes to evaluate how the network of inventors influence the innovation performance of French employment areas. In particular, a specific methodology is set up to address what kind of network structure favours the most collaboration. The main results of this thesis are that an increase in the connectedness of inventors is always beneficial to urban innovation performance. We also show that social network act as a substitute to geographic distance, so that social network allows to alleviate the burden of distance. These results shed light on the role of the network in shaping the spatial distribution of the scientific and technological activity.

Keywords: innovation; research collaboration; network formation; economic geography