Network proximity in the geography of research collaboration*

Laurent R. Bergé†

GREThA (UMR CNRS 5113), University of Bordeaux, France

Published Version:

Papers in Regional Science, 96(4), 2017, doi: https://doi.org/10.1111/pirs.12218
https://onlinelibrary.wiley.com/doi/abs/10.1111/pirs.12218

Abstract

This paper deals with the questions of how network proximity influences the structure of inter-regional collaborations and how it interacts with geography. I first introduce a new, theoretically grounded measure of inter-regional network proximity. Then, I use data on European scientific co-publications in the field of chemistry between 2001 and 2005 to assess those questions. The main findings reveal that inter-regional network proximity is important in determining future collaborations but its effect is mediated by geography. Most importantly, a clear substitution pattern is revealed showing that network proximity mainly benefits international collaborations.

Keywords: network proximity; gravity model; research collaboration; network formation; co-publication

JEL codes: D85, O31, R12

*I would like to thank Pascale Roux, Ernest Miguélez, Ulrich Ziehran and Francesco Lissoni who provided helpful comments; I also benefited from comments from participants at the GREThA seminars (Bordeaux, March 2013, April 2014) and at the 54^{th} European Regional Science Association Conference (Saint-Petersburg, August 2014). Three anonymous referees provided very constructive comments. I gratefully acknowledge the financial support of the French region of Aquitaine (Conseil régional d'Aquitaine) for the research project REGNET (Grant #20101402006).

† *e-mail*: laurent.berge@u-bordeaux.fr

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). 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 paper is a step toward a better understanding of the role of networks in the geo-

graphy 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., 2014; Balland et al., 2015), 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., 2007; 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 paper 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, 2012, 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

[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 C.

of EU policy.

The remainder of this paper is organized as follows: in Section 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 3 then presents the estimation methodology and describes the measure of network proximity used in this paper, along with the model from which it can be retrieved; in Section 4, the data set is presented, as well as the econometric strategy; the empirical findings are reported and discussed in Section 5; and Section 6 concludes the paper.

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.

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 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).

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 paper. 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.

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. (2007) 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 paper 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.

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.

3.1 Gravity model

The object of this paper 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

[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).

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., 2007; 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:

$$Interaction_{ij} = Mass_i^{\alpha_1}\ Mass_j^{\alpha_2} F\left(Distances_{ij}\right), \tag{1}$$

with $F(.)$ being a decreasing function of the distances. The distance functions are usually of the form $F(x) = 1/x^{\gamma}$ or $F(x) = \exp\left(-\gamma x\right)$, 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 paper, $Interaction_{ij}$ 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.

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, 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.

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.

[Figure 1 about here.]

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 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.

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:

$$ENB_{ij}^{k} = \frac{g_{ik}g_{jk}}{n_k}. \tag{2}$$

Proof. See Appendix A.

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 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:

$$TENB_{ij} = \sum_{k \in N_i \cap N_j} \frac{g_{ik}g_{jk}}{n_k} \tag{3}$$

The measure of network proximity that will be used in this paper is the total expected number of bridging paths (TENB). The link between the TENB and the notion of network proximity is discussed in Subsection 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.

[3]For instance, consider the network in Figure 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).

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, $\iota \in \{1, \dots, n\}$, 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 on the number of links an agent already has. Notably, as shown in Appendix B.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 (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:

[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.

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

Proof. See Appendix B.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.

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 microstructure.

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.

4 Data and methodology

This section first explains the construction of the data set and all the variables; Subsection 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.

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.

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.

[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'.

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.

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.

[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 4.2), where the empirical construction of the TENB is described.

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 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.

[Figure 2 about here.]

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 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

[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.

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 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 (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 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

[10] Using the fractional count instead of the full count methodology does not alter the results.

[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.

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:

$$TENB_{ij} = \sum_{k \in N_i \cap N_j} \frac{bilateral\_copub_{ik}^{2001-2003} \times bilateral\_copub_{jk}^{2001-2003}}{total\_publications_k^{2001-2003}}. \quad (4)$$

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 (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 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]

[13] This footnote shows how to derive this result. Using the notations from Section 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 and j is then defined as $TENB_{ij} = \sum_{k \in N_i \cap N_j} \left(g_{ik} \times g_{jk}/n_k\right)$. 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 $TENB_{ij}^{new} = \sum_{k \in N_i \cap N_j} \left(g_{ik}^{new} \times g_{jk}^{new}/n_k\right) = \sum_{k \in N_i \cap N_j} \left((1.01 \times g_{ik}) \times (1.01 \times g_{ik})/n_k\right) \simeq 1.02 \times \sum_{k \in N_i \cap N_j} \left(g_{ik} \times g_{jk}/n_k\right) = 1.02 \times TENB_{ij}$.

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

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 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

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

[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.

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.

4.3 Model and estimation procedure

As the dependent variable $Copub_{ij}$ is a count variable, a natural way to estimate equation (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:

$$\begin{aligned} \mathrm{E}\left(Copub_{ij}|X_{ij}\right) &= d_i \times d_j \times (TENB_{ij}+0.01)^{\alpha_1} \times GeoDist_{ij}^{\alpha_2} \\ &\quad \times \exp\left(\alpha_3 notContig_{ij} + \alpha_4 CountryBorder_{ij} + \alpha_5 CogDist_{ij} + \alpha_6 TopRegions_{ij}\right), \end{aligned} \tag{5}$$

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.

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 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 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 (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 2 shows the correlations among the explanatory variables. The highest correlation is between the geographical distance and national border variables.

[Table 1 about here.]

[Table 2 about here.]

[Figure 3 about here.]

5 Results

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 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

[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.

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.

[Table 3 about here.]

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.

[Figure 4 about here.]

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 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 TENB strictly increases for distances in the range of those in the sample. The estimates indicate that the effect is even negative for regions

[18] From Section 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.

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.

[Table 4 about here.]

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 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 they are non-contiguous instead of contiguous.[19] Figure 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.

[Figure 5 about here.]

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,

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

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: 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 paper regarding network proximity are likely to be a lower bound.

6 Conclusion

This paper 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 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 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.

[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.

References

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.

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.

Autant-Bernard, C., Billand, P., Frachisse, D., Massard, N., 2007. Social distance versus spatial distance in R&D cooperation: Empirical evidence from European collaboration choices in micro and nanotechnologies. ***Papers in Regional Science*** 86(3): 495–519.

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., 2015. 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.

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

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.

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.

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.

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

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

Boschma, R., Balland, P.-A., de Vaan, M., 2014. 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., 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.

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.

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

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.

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).

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

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.

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.

European Commission, May 2012. Guide to research and innovation strategies for smart specialisations (RIS 3). ***Joint Research Centre***Luxembourg: Publications Office of the European Union.

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.

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

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

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: Pyka, A., Scharnhorst, A. (Eds.), ***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.

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

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

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.

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.

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.

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.

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.

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.

Kirat, T., Lung, Y., 1999. Innovation and proximity territories as loci of collective learning processes. ***European Urban and Regional Studies*** 6(1): 27–38.

Krackhardt, D., 1999. The ties that torture: Simmelian tie analysis in organizations. ***Research in the Sociology of Organizations*** 16(1): 183–210.

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.

McPherson, M., Smith-Lovin, L., Cook, J. M., 2001. Birds of a feather: Homophily in social networks. ***Annual Review of Sociology*** 27: 415–444.

Miguélez, E., Moreno, R., 2014. What attracts knowledge workers? the role of space and social networks. ***Journal of Regional Science*** 54(1): 33–60.

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.

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.

Newman, M. E., 2001. The structure of scientific collaboration networks. ***Proceedings of the National Academy of Sciences*** 98(2): 404–409.

OST, 2010. Indicateurs de sciences et de technologies.

Picci, L., 2010. The internationalization of inventive activity: A gravity model using patent data. ***Research Policy*** 39(8): 1070–1081.

Ponds, R., van Oort, F., Frenken, K., 2007. The geographical and institutional proximity of research collaboration. ***Papers in Regional Science*** 86(3): 423 – 443.

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.

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.

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., Marx, M., 2013. Geographic constraints on knowledge spillovers: political borders vs. spatial proximity. ***Management Science*** 59(9): 2056–2078.

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.

Torre, A., Rallet, A., 2005. Proximity and localization. ***Regional studies*** 39(1): 47–59.

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.

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

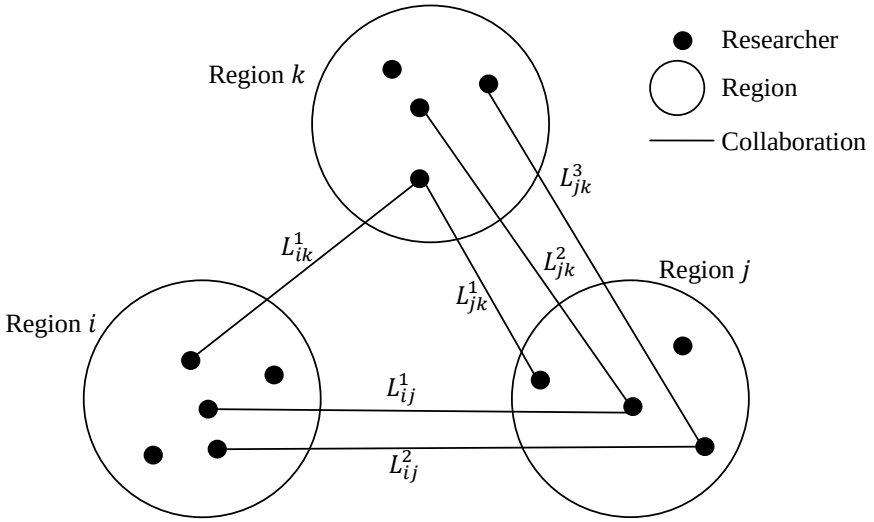

Figure 1: Illustration of a regional network of collaboration and of the notion of bridging paths.
Notes: The figure depicts three bridging paths formed by the following pairs of links: (L^1_{ik}, L^1_{jk}), (L^1_{ij}, L^2_{jk}) and (L^2_{ij}, L^3_{jk}). So the regional dyads (i,j), (i,k) and (j,k) have 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.

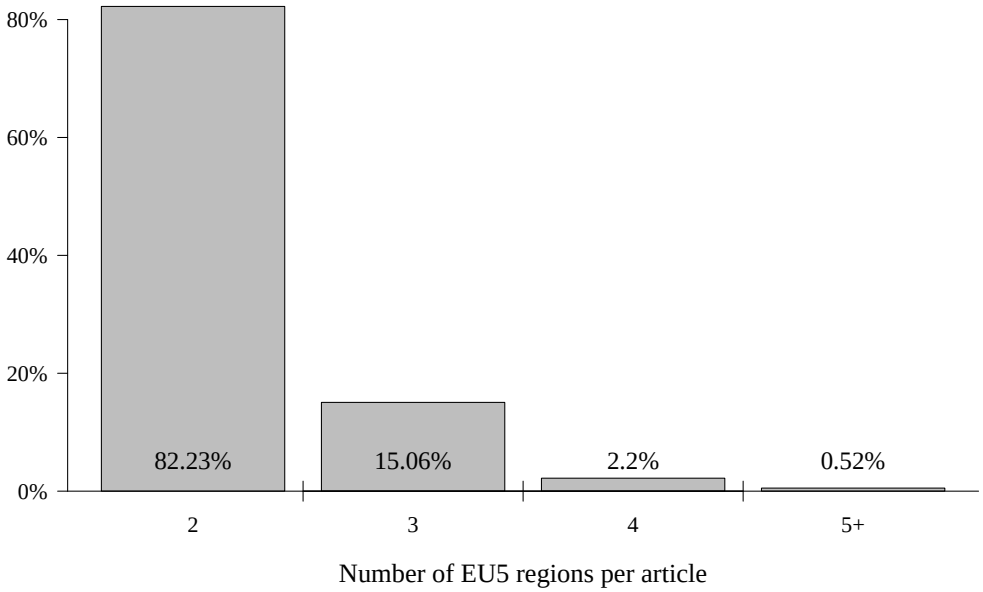

Figure 2: Distribution of the number of regions (from the EU5 countries) per inter-regional article in chemistry.

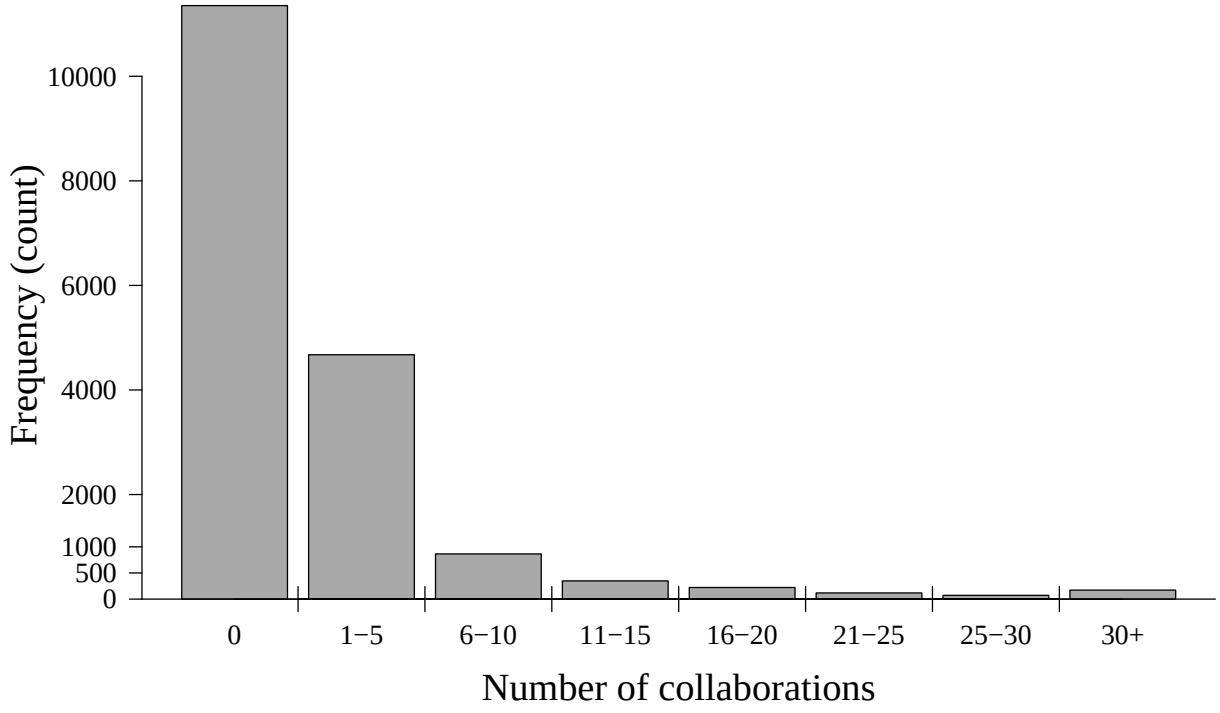

Figure 3: Distribution of EU5 inter-regional collaborations in chemistry for the period 2004-2005.

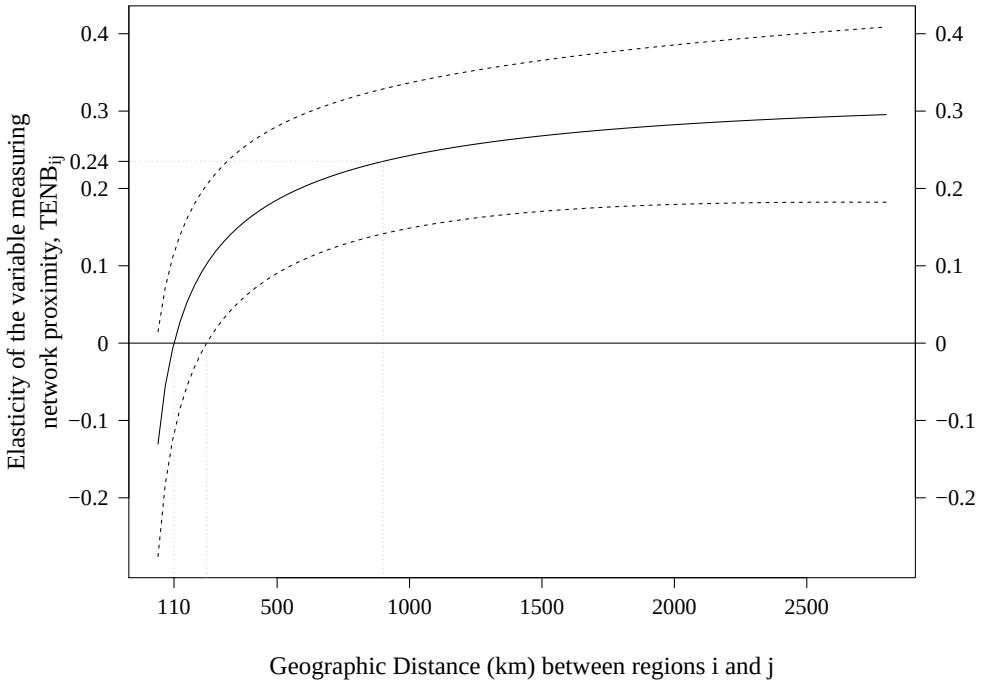

Figure 4: Graph of the interaction between network proximity and geographical distance.
Notes: The graph represents the estimated elasticity of the TENB on co-publications with respect to geographical distance (solid line) along with its 95% confidence interval (dashed lines). This graph is based on the estimates from model (4) of Table 3.

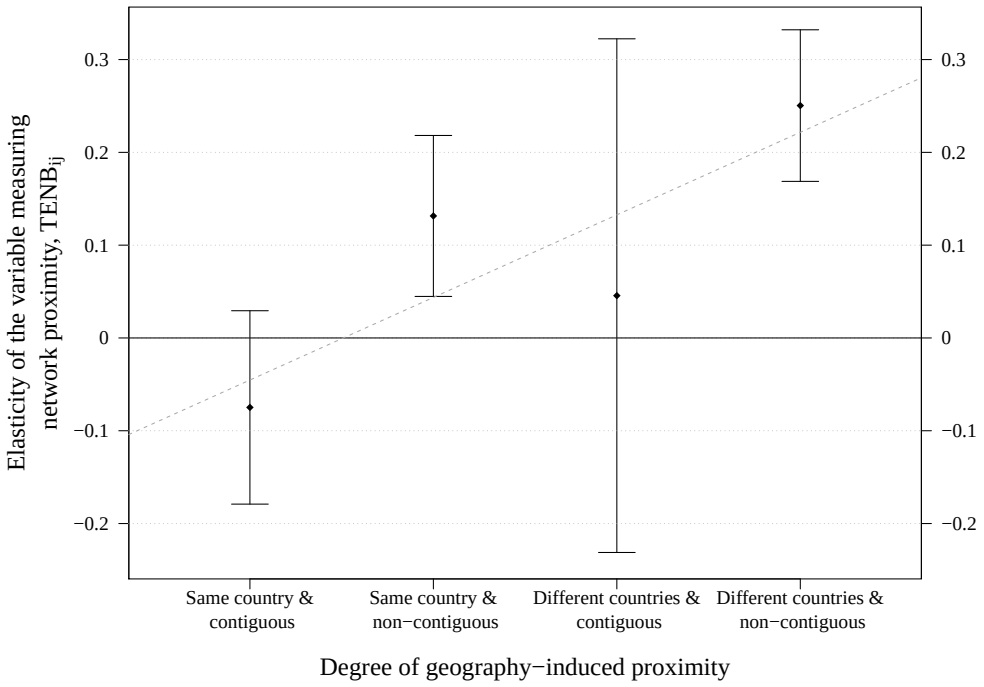

Figure 5: Graph of the link between network proximity and geography-induced proximity.
Notes: The graph reports the elasticity of the TENB on co-publications with respect to different degrees of geography-induced proximity. Both the estimates of the elasticities, as well as their 95% confidence intervals, are represented. This graph is based on the estimates from model (6) of Table 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.

Table 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 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).

Table 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 (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 4.2. The variable TENB approximates network proximity and is defined as a measure of the strength of indirect connections between regions (see Section 3.2). Two-way clustered standard errors in parenthesis (see Cameron et al., 2011). Level of statistical significance: * 10%, ** 5%, *** 1%.

Table 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 4.2. The variable TENB approximates network proximity and is defined as a measure of the strength of indirect connections between regions (see Section 3.2). Two-way clustered standard errors in parenthesis (see e.g. Cameron et al., 2011). Level of statistical significance: * 10%, ** 5%, *** 1%.

A Proof of proposition 1

Let L^a_{ik} represent the a^{th} link, $a \in \{1,\dots,g_{ik}\}$, between agents from regions i and k, and L^b_{jk} 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^a_{ik}, L^b_{jk}) forms a bridging path if and only if they are both connected to the same agent in region k (as depicted by figure 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^a_{ik} and L^b_{jk} is $p^2_\iota = 1/n^2_k$. Therefore, the pair (L^a_{ik}, L^b_{jk}) is a bridging path with probability $p = \sum_{\iota=1}^{n_k} p^2_\iota = 1/n_k$ (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^a_{ik}, L^b_{jk}) 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 $ENB^k_{ij} = E(\sum_{a=1}^{g_{ik}} \sum_{b=1}^{g_{jk}} X_{ab})$. From the property of the mean operator, it can be rewritten as: $ENB^k_{ij} = \sum_{a=1}^{g_{ik}} \sum_{b=1}^{g_{jk}} E(X_{ab}) = \sum_{a=1}^{g_{ik}} \sum_{b=1}^{g_{jk}} p = \sum_{a=1}^{g_{ik}} \sum_{b=1}^{g_{jk}} (1/n_k) = (g_{ik} g_{jk})/n_k$. □

B Preferential attachment

In this section I consider the matching mechanism described in section 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.

B.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:

$$\begin{aligned} \left(\iota^{-0.5}/\Gamma\right) \times L &< \mathtt{k} \\ \iota^{-0.5} &< \frac{\mathtt{k}\Gamma}{L} \\ \iota &> \left(\frac{L}{\mathtt{k}\Gamma}\right)^2. \end{aligned} \tag{6}$$

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 (6). The share of agents having a degree lesser than $\mathtt{k}$ is then:[21]

$$\begin{aligned} F(\mathtt{k}) &= \frac{1}{n}\left(n - \iota(\mathtt{k})\right) \\ &= 1 - \frac{1}{n}\left(\frac{L}{\Gamma}\right)^2 \mathtt{k}^{-2}. \end{aligned} \tag{7}$$

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

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

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:

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

[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 (7) is defined only for $\mathtt{k} \geq \mathtt{k}^*$ and is 0 otherwise. All these details were skipped for readability.

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:

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

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

B.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 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^a_{ik} be the a^{th} link, $a \in \{1, \ldots, g_{ik}\}$, between agents from regions i and k, and L^b_{jk} 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^a_{ik}, L^b_{jk}) 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^a_{ik} and L^b_{jk} are both connected to agent ι is $p^2_\iota = \left(\iota^{-0.5}/\Gamma\right)^2$. Then the pair (L^a_{ik}, L^b_{jk}) is a bridging path with probability $p = \sum_{\iota=1}^{n_k} p^2_\iota$. Let X_{ab} be the binary random variable relating whether the pair (L^a_{ik}, L^b_{jk}) 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 $ENB^{k,Pref}_{ij} = E(\sum_{a=1}^{g_{ik}} \sum_{b=1}^{g_{jk}} X_{ab})$. From the property of the mean, it can be rewritten as:

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

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

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

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. □

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

CODE	NAME	CODE	NAME
DE11	Stuttgart	FR52	Bretagne
DE12	Karlsruhe	FR53	Poitou-Charentes
DE13	Freiburg	FR61	Aquitaine
DE14	Tübingen	FR62	Midi-Pyrénées
DE21	Oberbayern	FR63	Limousin
DE22	Niederbayern	FR71	Rhône-Alpes
DE23	Oberpfalz	FR72	Auvergne
DE24	Oberfranken	FR81	Languedoc-Roussillon
DE25	Mittelfranken	FR82	Provence-Alpes-Côte d'Azur
DE26	Unterfranken	FR83	Corse
DE27	Schwaben	ITC1	Piemonte
DE30	Berlin	ITC3	Liguria
DE40	Brandenburg	ITC4	Lombardia
DE50	Bremen	ITF1	Abruzzo
DE60	Hamburg	ITF2	Molise
DE71	Darmstadt	ITF3	Campania
DE72	Gießen	ITF4	Puglia
DE73	Kassel	ITF5	Basilicata
DE80	Mecklenburg-Vorpommern	ITF6	Calabria
DE91	Braunschweig	ITG1	Sicilia
DE92	Hannover	ITG2	Sardegna
DE93	Lüneburg	ITH1	Provincia Autonoma di Bolzano/Bozen

CODE	NAME	CODE	NAME
DE94	Weser-Ems	ITH2	Provincia Autonoma di Trento
DEA1	Düsseldorf	ITH3	Veneto
DEA2	Köln	ITH4	Friuli-Venezia Giulia
DEA3	Münster	ITH5	Emilia-Romagna
DEA4	Detmold	ITI1	Toscana
DEA5	Arnsberg	ITI2	Umbria
DEB1	Koblenz	ITI3	Marche
DEB2	Trier	ITI4	Lazio
DEB3	Rheinhessen-Pfalz	UKC1	Tees Valley and Durham
DEC0	Saarland	UKC2	Northumberland and Tyne and Wear
DED2	Dresden	UKD1	Cumbria
DED4	Chemnitz	UKD3	Greater Manchester
DED5	Leipzig	UKD4	Lancashire
DEE0	Sachsen-Anhalt	UKD6	Cheshire
DEF0	Schleswig-Holstein	UKD7	Merseyside
DEG0	Thüringen	UKE1	East Yorkshire and Northern Lincolnshire
ES11	Galicia	UKE2	North Yorkshire
ES12	Principado de Asturias	UKE3	South Yorkshire
ES13	Cantabria	UKE4	West Yorkshire
ES21	País Vasco	UKF1	Derbyshire and Nottinghamshire
ES22	Comunidad Foral de Navarra	UKF2	Leicestershire, Rutland and Northamptonshire
ES23	La Rioja	UKF3	Lincolnshire
ES24	Aragón	UKG1	Herefordshire, Worcestershire and Warwickshire
ES30	Comunidad de Madrid	UKG2	Shropshire and Staffordshire
ES41	Castilla y León	UKG3	West Midlands
ES42	Castilla-La Mancha	UKH1	East Anglia
ES43	Extremadura	UKH2	Bedfordshire and Hertfordshire
ES51	Cataluña	UKH3	Essex
ES52	Comunidad Valenciana	UKI1	Inner London
ES53	Illes Balears	UKI2	Outer London
ES61	Andalucía	UKJ1	Berkshire, Buckinghamshire and Oxfordshire
ES62	Región de Murcia	UKJ2	Surrey, East and West Sussex
FR10	Île de France	UKJ3	Hampshire and Isle of Wight
FR21	Champagne-Ardenne	UKJ4	Kent
FR22	Picardie	UKK1	Gloucestershire, Wiltshire and Bristol/Bath area
FR23	Haute-Normandie	UKK2	Dorset and Somerset
FR24	Centre	UKK3	Cornwall and Isles of Scilly
FR25	Basse-Normandie	UKK4	Devon
FR26	Bourgogne	UKL1	West Wales and The Valleys
FR30	Nord - Pas-de-Calais	UKL2	East Wales
FR41	Lorraine	UKM2	Eastern Scotland
FR42	Alsace	UKM3	South Western Scotland
FR43	Franche-Comté	UKM5	North Eastern Scotland
FR51	Pays de la Loire	UKM6	Highlands and Islands

D 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').

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

Keyword	Count	Keyword	Count
Computer Science, Interdisciplinary Applications	384	Materials Science, Paper & Wood	2
Nutrition & Dietetics	289	Fisheries	1
Archaeology	268	Marine & Freshwater Biology	1
Engineering, Environmental	236		