Papers in Evolutionary Economic Geography

15.07

Network proximity in the geography of research collaboration

Laurent R. Bergé

Utrecht University
Urban & Regional research centre Utrecht

Network proximity in the geography of research collaboration*

Laurent R. Bergé†

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

25th February 2015

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 benefits mostly 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 Miguelez, Ulrich Ziehran and Francesco Lissoni who provided helpful comments; I also benefited from comments from participants at GREThA seminars (Bordeaux, March 2013, April 2014) and at the 54^{th} European Regional Science Association Conference (Saint-Petersburg, August 2014) .

†*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 then no wonder that it is a central issue for policy makers, at the regional, national or even supranational scale. This in turn puts at the forefront policies handling 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 (e.g. Wuchty et al., 2007; Jones et al., 2008; Singh and Fleming, 2010). In the European Union (EU), the political will towards knowledge creation is supported by the recent Horizon 2020 program which 'should be implemented primarily through transnational collaborative projects' (European commission, 2013, article 23). This policy tool aims at developing an European research area (ERA) where collaborations don't suffer from the impediments of distance or country borders so that EU researchers could act as if they were in one and the same country. Such policies are backed with large EU budget: But is funding large-distance collaboration efficient? To apprehend this issue, one needs to clearly understand the determinants of collaboration and particularly what helps to bypass geography.

Despite the announced 'death of distance' due to the development in means of communication and in transportation technologies (Castells, 1996), geography is still important in explaining collaborations. Co-location helps having face to face contacts, eases the sharing of tacit knowledge (e.g. Gertler, 1995; Storper and Venables, 2004) and also enhances the likelihood of serendipitous fruitful collaborations (Catalini, 2012). Furthermore, country borders, a by-product of geography, also play an important role as differences in national systems renders collaborations more difficult (Lundvall, 1992). A recent stream of literature has shown that indeed geographical distance and country borders are 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). Linking to the ERA, it seems like EU policies failed to develop an integrated area, where distant collaborations are eased. Yet,

geography is not the sole determinant of collaborations (Boschma, 2005; Torre and Rallet, 2005; Frenken et al., 2009a). 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 depart from including potential network effects influencing the collaboration process.

This paper is a step toward a better understanding of the role of networks in the geography of collaboration. Previous studies on this topic have mostly been descriptive, a-geographic or did not weld the network and the geography together (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 of geography and network has been set aside. Yet, the answer to this question is important policy-wise. If they are substitutes, then heightening the network proximity of distant agents would in turn help them in creating new distant links because network proximity would partly compensate the loss of geographic proximity. On the contrary, in the case of complementarity 'forcing' distant collaboration may be inefficient because distant agents would be those who benefit the less from network proximity. Only the former case would support EU policies and only if the network mattered 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 regional pair, it 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 of understanding 'the position of region[s] within the European and global economy' (European commission, 2012, p.18).

To assess empirically how network proximity affects collaborations, I then make use of European co-publication data. It consists of co-publications stemming from the five largest

European countries (France, Germany, Italy, Spain, the United Kingdom), for the field of chemistry between 2001 and 2005. The analysis consists in estimating the determinants of flows of collaboration between 8,515 regional dyads from 131 NUTS 2 regions by the means of a gravity model.[1] The results demonstrate an interplay between geography and network proximity: while being not or only weakly beneficial to regions located close by, *the importance of network proximity grows with distance*, reaching an elasticity of 0.23 for a distance of 1,000 km. In other words, network proximity benefits most international collaborations. Those results then supports the claims of EU policies.

The remaining of the paper is organized as follows: first 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 describe 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 while Section 6 concludes.

2 The determinants of inter-regional collaborations

In this section I describe the determinants of scientific collaborations. 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 analyse the determinants of collaboration, the concept of proximity prove to be a very useful framework (Boschma, 2005; Torre and Rallet, 2005; Kirat and Lung,

[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 B.

1999). By distinguishing several types of proximity between agents (such as geographical, institutional, cognitive or organizational), this framework allows analysing each of them and to easily assess their interplay. Also, one can distinguish two mechanisms through which proximity, whatever the form, favours collaboration: proximity augments the probability of potential partners to meet and reduces the costs involved in collaboration; thus rising at the same time its expected net benefits and the likelihood of success.

Geographical proximity can be decomposed in such a way. First, the context of collaborative production of knowledge may imply the partners to 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 contacts may be important in conducting effectively the research by overcoming the problem of sharing tacit knowledge (Gertler, 1995; Collins, 2001; Gertler, 2003). Moreover, face to face contacts allows direct feedback, eases communication and the litigation of problems and facilitates coordination (Beaver, 2001; Freeman et al., 2014). All these elements heightens the probability of success of a collaboration. Thus, geographical distance, by implying greater travel costs and lesser opportunities to exchange knowledge by the means of face to face contacts, 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. Indeed, social events where researchers meet to share ideas, such as conferences, seminars or even informal meetings, are linked to geographical distance; thus heightening the chances to find the research partner at a local scale. For instance, by analysing data on participants at the European congresses of the regional science association, van Dijk and Maier (2006) show that distance to the event affect negatively the likelihood to attain it. Also, the social embeddedness of researchers and inventors has been shown to decay with geographical distance (Breschi and Lissoni, 2009), such that they will have a better knowledge of the potential partners at a closer distance.

Consequently, the effect of geographical distance should be 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), or in the case of cooperation among firms and research institutions within the European Framework Program (Scherngell and Barber, 2009).

Another impeding effect related to geography is country borders. In the context of inter-regional collaboration, country 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.3). Indeed, several features affecting knowledge flows happen at the national level (Banchoff, 2002; Glänzel, 2001). For instance funding schemes are more likely to be at a national scale, thus facilitating the collaborations within country. In the same vein, workers are more mobile within than across countries, and as they may keep ties with their former partners, their social network appear to be more developed at the national level (Miguélez and Moreno, 2014). Also, norms, values and language are likely to be shared within a country, facilitating collaboration. As a consequence, the literature shows evidence that country borders reduces collaborations (e.g. Hoekman et al., 2010; Morescalchi et al., 2015).

2.2 The role of networks in the process of collaboration

A 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. Triadic closure may occur because triads, by opposition to dyads, have some advantages. By reducing the individual power, triads can help to mitigate conflicts and favours trust among the individuals (Krackhardt, 1999). Bad behaviour of one of the agents is more limited because 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 where the reliability of different partners may be difficult to assess. Then relying on the network and form a triad, that is to collaborate with a partner of

a partner, is interesting to limit opportunistic behaviours, thus reducing the risk associated to the sunk cost of engaging in a collaboration. In a recent study on the German biotechnology industry, Ter Wal (2013) shows that triadic closure among German researchers has been increasingly important over time, as the technological regime was changing and more trust was needed among the partners. Also, by examining the behaviour of Stanford's researchers, Dahlander and McFarland (2013) show that having an indirect partner rises significantly the probability to collaborate.

Another feature of social networks that may influence their evolution is homophily. Homophily can be seen as a compelling feature of social networks. It can be depicted 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 context, like in friendships at school or working relationships, and show that similarity among individuals is a force driving the creation of ties. As McPherson et al. (2001, p.429) puts it: 'Homophily characterizes network systems, and homogeneity characterizes personal networks' and science is no exception. For instance Blau (1974) studies the relationships among theoretical high energy physicists and show that the similarity of their specialized research interest 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, and there is no need of the network to benefit from it. Yet, once the problem is reversed, one can see that the network can influence new connections via homophily. Indeed, take two agents already connected, they are likely to share some similarities that helped them succeed in collaboration. For instance it can be sharing a similar research topic, having the same approach to research questions or simply being compatible in terms of team working (i.e. they are a good match with respect to idiosyncratic characteristics). Therefore, if two agents are connected to a same partner, they are likely to be in some way similar to their common collaborator and then share themselves some similarities. This similarity may in turn favour their future collaboration.

Finally, the network can be seen as a provider of externalities of information, thus being critical in determining future collaborations. Indeed, as the need for collaboration is getting more and more acute (Jones, 2009), finding the right partners is critical; but may be time consuming. Katz and Martin (1997) points out that time is one of the most important resource 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 can then be best viewed by analogy to optimization problems: though not giving the global best match, the network helps to get the local best 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 to select the best one. Then 'picking' the partner in the network vicinity may be a rational and efficient choice.

To summarize, the network is swathed in various mechanisms favouring collaboration, thus affecting its evolution. This yields the following hypothesis:

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

Yet, although it may influence the formation of new collaborations, can its effect be regulated by other factors, like geographical distance or country borders? Or is the effect of network proximity merely independent of other determinants? This question needs to be investigated to unravel the precise mechanism shaping the landscape of knowledge networks. 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, geographic proximity is

here intended to be a shorthand for non-network forms of proximity.

If both network and geographic proximity influence the creation of new collaborations, what would be the net outcome of these two effects? The first case one could consider is that network proximity benefits homogeneously to all prospective partners, meaning an independence between the effects of both network and geographical proximity. That is, the more network proximity, the higher the likelihood of a collaboration, in a magnitude independent from geography. But this independence could only occur if geographical proximity and network proximity played on two complete different grounds. That is, if the very mechanism through which they affect collaboration were unrelated. As soon as they influence collaboration by the same common mechanisms (like enhancing trust, or facilitating the search for prospective partners), their interplay should not be independent. So if one departs from the hypothesis of independence then two opposing standpoints compete.

On the one hand, network proximity can reinforce the benefits of being geographically close. Particularly in the case where agents have a taste for similarity, then network proximity can foster collaborations in situations where agents already benefits from geographical proximity. This taste for similarity can be seen as a need to be close in different dimensions in order to convey research. For instance, in the case where the research is highly subject to opportunistic behaviour, then several forms of proximity may be complementary in mitigating it.

On the other hand, the benefits of proximity may suffer from decreasing returns. In that case, network proximity would be a substitute to other forms of proximity. Indeed, take the case where two prospective partners are far apart. For them, network proximity will be crucial to engage in a successful collaboration as it would be their sole source of proximity. On the contrary, if they are already close to each other, because of the decreasing returns, having network proximity would matter less and would then not be decisive in triggering collaboration. These effects depicts a pattern of substitutability. Another possible interpretation yielding the same conclusion is that the net rewards of collaborations may increase with distance (this view is supported by e.g. Narin et al., 1991; Glänzel, 2001;

Frenken et al., 2010). In that 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 also picturing 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 to 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 network and the role of geography but not unravelling their interplay. For instance, Maggioni et al. (2007), compare 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 firm collaborations at the EU 6^{th} framework program; they asses the effect of network proximity and geographical proximity on the probability of collaborations. Both studies find a positive effect of both geography and the network.

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

This paper departs from the previous literature by specifically focusing on network proximity and, more importantly, its relation with geography. In line with previous studies, the focus will be on inter-regional flows of collaborations in Europe (e.g. Scherngell and Barber, 2009; Morescalchi et al., 2015; Hoekman et al., 2009). But before detailing the data, I will first present the modelling strategy and will spend some time describing the measure of network proximity used 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. It will be shown that 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 interactions in various contexts such as trade flows or migration flows (Roy and Thill, 2004; Anderson, 2011), and is also applied to the context of collaborations (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 by the combinations of centripetal and centrifugal forces; while the masses of the regional entities act as attractors, the distance separating them hampers the attraction. It can be written as:

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

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 destination; also in the case where $Interaction_{ij}$ and the distance variables are undirected, as in collaboration networks, α_1 is constrained to be equal to α_2. In the context

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

of this paper, $Interaction_{ij}$ will represent collaboration flows. Within the gravity framework, the network proximity should act 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 applied network analysis tools to assess the position of regions within a network. Some cope with the position of regions within the network by making use of centrality measures (see e.g. Sebestyén and Varga (2013b,a) or Wanzenböck et al. (2014, 2013)). Others use the network by linking the performance of a given region to the performance of the regions that have connections with it, in a fashion similar to spatial econometrics (see e.g. Maggioni et al. (2007) or Hazir et al. (2014)).

To fit in the gravity model framework and later in the econometric analysis, a measure of inter-regional network proximity should have two properties. First, it should be defined for each pair of region. Second, for the sake of coping with potential endogeneity problem, it should be independent from direct collaboration. 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. First the measure and its main components is introduced. Then I show that it actually reflects a notion at the micro level, the notion of 'bridging path', and give the model from which it can be formally derived. Last, a variation in the model's assumption is analysed.

3.2.1 The measure and its idea

The measure is aimed to fit the gravity model and thus relates the intensity of network proximity between a pair of regions. It is defined by the following formula:

$$\sum_{k \neq i,j} \frac{Collab_{ik} Collab_{jk}}{Researchers_k}, \tag{2}$$

where $Collab_{ik}$ is the number of collaborations between researchers of regions i and k, and $Researchers_k$ is the number of researchers of region k. Thus, it is based on concepts defined at the regional level: collaboration flows and regional size. This measure can be seen as conservative as it is based only on indirect connections and neglects the form of network proximity that may arise from direct connections.

The main ideas underlying this measure can be summarized by looking at changes in its parameters all else being equal. First of all, if two regions have not any common partner, they will have a measure of zero, even though they could have direct links. Now suppose two regions, i and j, are tied to a third one, k. In this case, collaborating with large regions (i.e. high $Researchers_k$) yields less proximity than with regions of smaller size. The main idea is that in the latter case, the links are more concentrated so that the agents from the two regions i and j will be closer in the social space. For instance, take the opposite case, where the size of region k is very high (e.g. $Researchers_k$ tends to infinity), then despite the positive number of collaborations with k, they are diluted among so many researchers from k that it will be very unlikely that agents from i and j know each other thanks to agents from k. So there is a negative effect of the size of the common region k.

When analysing inter-regional networks, one has to keep in mind that they are the aggregated view of micro-economic decisions. Thus, it may be difficult to consider regions simply as individual agents and to apply them the same concepts as the ones used at the micro level (see e.g. Ter Wal, 2011; Brenner and Broekel, 2011). Indeed, regions do not collaborate with each other, only the agents within them do. So the logic of applying the same concepts may be irrelevant. Yet, it would also be inadequate to consider that the aggregate flows of collaboration do not convey any information about their micro structure.

Following this line of thought, this measure 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. Those indirect connections at the micro level are coined 'bridging paths' and are precisely defined in the next subsection while the section 3.2.3 provides a simple model from which the measure can be derived.

3.2.2 The notion of bridging path and some notations

[Figure 1 about here.]

First some notations, as they will be useful to define the concept of bridging path and will be used in the model of next subsection. Consider N regions, each populated with n_i researchers. A link between two regions is defined as a collaboration occurring between two researchers, one of each of those regions. Let g_{ij} to 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_{ij}^a to represent the a^{th} link, $a \in \{1, \ldots, g_{ij}\}$, between agents from regions i and j.

Using these notations, a bridging path between region i and j via the bridging regions k is defined as a set of two links (L_{ik}^a, L_{jk}^b) 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_{ik}^1, L_{jk}^1) forms a bridging path, while others like the pair (L_{ik}^1, L_{jk}^1) 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.

3.2.3 Deriving the measure from a model of random matching

This subsection shows how, by assuming that collaborations 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 can be either a lone researcher or a team of researchers, as teams can be fairly considered as behaving like a unique entities (see e.g. Beaver, 2001; Dahlander and McFarland, 2013). The second assumption is in line with the intuition as it simply implies that for two regions, say i and j, the more observed collaborations between i and j, the more likely a randomly picked agent from i has 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, called 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:

$$ENB_{ij}^{k} = \frac{g_{ik}g_{jk}}{n_k}. \tag{3}$$

Proof. Let L_{ik}^{a} to represent the a^{th} link, $a \in \{1, \ldots, g_{ik}\}$, between agents from regions i and k, and L_{jk}^{b} to be the b^{th} link, $b \in \left\{1, \ldots, g_{jk}\right\}$, between agents from regions j and k. By definition, the pair of links (L_{ik}^{a}, L_{jk}^{b}) forms a bridging path if and only if they are both connected to the same agent in region k (as depicted by figure 1). Let the Greek letter ι, $\iota \in \{1, \ldots, n_k\}$, to designate agent ι from region k. Hence, from the random matching process, we know that the probability that agent ι is connected to any incoming link is $p_\iota = 1/n_k$. Thus, the probability that agent ι is connected to both links L_{ik}^{a} and L_{jk}^{b} is $p_\iota^2 = 1/n_k^2$. Then the pair (L_{ik}^{a}, L_{jk}^{b}) is a bridging path with probability $p = \sum_{\iota=1}^{n_k} p_\iota^2 = 1/n_k$ (summing over all the agents of region k, because each agent can be

[3]For instance, consider the network of figure 1: if one selects randomly one agent from region i, it is more likely that she/he has collaborated with one from j than one from k (because there are two links with the former and only one with the later).

connected to both links). Let X_{ab} to 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 $\{1, \ldots, g_{ik}\}$ and $\left\{1, \ldots, g_{jk}\right\}$, 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: $ENB^k_{ij} = \sum_{a=1}^{g_{ik}} \sum_{b=1}^{g_{jk}} E(X_{ab}) = \sum_{a=1}^{g_{ik}} \sum_{b=1}^{g_{jk}} p = \sum_{a=1}^{g_{ik}} \sum_{b=1}^{g_{jk}} (1/n_k) = (g_{ik} g_{jk})/n_k$. □

Proposition 1 relates to the expected number of bridging paths stemming from a specific bridging region. But two regions have more than just one common neighbour. The total expected number of bridging paths between two regions i and j is then the sum of the bridging paths stemming from all other regions to which i and j are both connected:

$$TENB_{ij} = \sum_{k \neq N_i \cap N_j} \frac{g_{ik} g_{jk}}{n_k} \tag{4}$$

This equation, which reflects a concept at the micro level, is the same as equation (2) first describing the measure. So the measure of network proximity that will be used in this paper is the total expected number of bridging paths (TENB).

Next subsection elaborates on the consequence of a variation on the random matching assumption and show that it would imply only trivial variation. The description of the data and the empirical details are in section 4.

3.2.4 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. But 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. Actually, 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 to the previously defined random matching process, I investigate the case of a matching process with preferential attachment.

A form of preferential attachment. The matching is not done at random anymore, but some nodes (the 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 $\iota \in \{1, \ldots, n\}$ to 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 with 10 agents, the probability to be tied to an incoming link is 20% for agent 1, 14% for agent 2, etc, and 6% for agent 10. This is to be compared to the random matching process where each agent had the same likelihood to be 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 A.1, the expected distribution of agents' degree along 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 along such a process, and analyse the difference with the measure obtained with the random matching process in equation (3).

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

[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 to have a degree k is equal to $f(k) = c \times k^{-\gamma}$ with c a constant.

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

Proof. See appendix A.2.

This result implies that, even using a more complex matching mechanism, the result is very similar to proposition 1. Indeed, $ENB_{ij}^{k,Pref}$ is merely an inflation of ENB_{ij}^{k}. Surely there are some variation as $\log(n_k)$ varies, but the logarithmic form flatten most of them, so that the correlation between $ENB_{ij}^{k,Pref}$ and ENB_{ij}^{k} is very high. This ends to show that the measure is robust to such a variation in its assumptions.

4 Data and methodology

This section first explains the construction of the data set and of all the variables, then subsection 4.3 presents the full model to be estimated as well as the estimation procedure. Some descriptive statistics are given last.

4.1 Data

To measure the intensity of collaborations between two regions, I will make use of co-publication data.[5] Collaborations are 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 papers published in most scientific journals, with, for each article, the list of all the participating authors along with their institutions.

The data is extracted for a time period ranging from 2001 to 2005 and the geographical scale is restricted to the five largest European countries: Italy, France, Germany, Spain and

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

the United Kingdom, as in Maggioni et al. (2007). Also, to avoid the problems that can arise when mixing several disciplines because of researchers' behaviour and publishing schemes that may differ between fields, I restrict the analysis to one specific field: Chemistry.

For each paper, this database reports the institution of each author's by-line. As there is an address assigned to each institution, it is possible to geographically pinpoint each of them. The 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 with the postcodes. The remaining 15% were located using an online map service with the information 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 is assigned to a NUTS 2 region with respect to their latitude/longitude coordinates. The data then consists of 131 NUTS 2 regions having at least one publication in the field of chemistry.

To sum up, the database consists of papers from Chemistry journals, with at least one author affiliated to an institution from the selected countries, for a total of 125,075 publications distributed along 131 NUTS 2 regions and over 5 years.

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) mentions, 'international refereed journals [in chemistry] play an important role in communicating results' so that most of chemistry scientific activities, including collaborations, that provide any result should leave a paper trail. Thus, scientific articles in this field allow tracking down the bulk of collaborations.

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 could

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

[7] Despite its simplicity, the accuracy of the location with only the name of the city and the country is quite high. Indeed, I located all addresses with both methods: city/country or postcodes. When comparing the two methods, one can see that only less than 1.5% of the NUTS 3 codes differ between the two methodologies. This number falls to less than 0.4% when considering the NUTS 2 codes.

be explained by the behaviour of existing researchers rather than by the actions of newly active ones. Put differently, as the focus 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 5 in the period 2001-2005, which seems high enough to fit this purpose.[8]

Last, most of inter-regional papers involves researchers from only two regions. As figure 2 shows, 2-regions papers account for 85% of the sample while 3-regions papers represent a share of 13%. This propensity for 2-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 separate the construction of the explanatory and the dependent variables, to avoid any simultaneity bias. The period used to construct the explanatory variables is 2001-2003. This three-year span is used to have enough information on collaboration patterns. Then the period 2004-2005 is 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, in the time period 2004-2005. Several methods could have been used to build this variable. Mainly there are the 'full count' or the 'fractional count' methodology. The former gives a unitary value for any dyad participating to 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

[8] In order to infer some statistics on the number of publications per researcher, I considered only the researchers having published an article in the year 2001 and then counted their publications in the range 2001-2005. Yet, to be sure these researchers were from the 5 studied countries, I only selected the ones that had at least an article whose institutions were exclusively within these 5 countries. Last, the researchers were identified using their surnames and the initial of their first names. Despite the rough identification of the researchers, this methodology allows to have an insight into the question of researchers' productivity in chemistry.

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 as it relates to the idea of participation to knowledge production instead of net contribution to knowledge production (OST, 2010, p.541).[9]

The masses. The mass of a region represents a natural force of attraction in the perspective of the gravity model. It is defined as the total number of articles produced from researchers of a given region. More precisely, $Mass_i$ is the number of articles published between 2001 and 2003 that have at least one author who is affiliated to an institution of region i. In the context of the gravity model, when the attraction between regions i and j is analysed, $Mass_i$ is called *mass of origin* and $Mass_j$ is called *mass of destination*.

Network proximity. The main explanatory variable catches 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 inventors of different regions. This variable is expected to positively influence future collaborations.

Let $TENB_{ij}$ to be the empirical counterpart of the TENB defined in equation (4). As the measure from the theoretical model is transposed to real data, two comments are in order. First, the theoretical model assumes that each collaboration involves only two agents. Yet, in the data, some papers involve more than two regions. To stick to the philosophy of the model, I then use only bilateral co-publications, i.e. 2-regions articles, to construct $TENB_{ij}$. 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 researchers of each region, yet this information is not directly available in the data. As an alternative, I chose the total number of publications of a given region to approximate its number of researchers.[10] Indeed, by the law of large numbers and for large enough regions, these two values should be proportional. Thus, in

[9] Using the fractional count instead of the full count methodology do not alter the results. The results with fractional counting are given by table 5.

[10] I lack precise information on researchers' affiliation. Indeed, within the data set, institutions are not pinpointed to the affiliated researchers.

the case where the number of researchers is proportional to the number of publications, we have $Researchers_k = a \times Mass_k$. This approximation allows to circumvent the problem of researchers' identification and will still yield a reliable measure of the TENB as it should only be proportional to the theoretical value.

Finally, the empirical variable can be defined. Let $copub_{ij}^{bilateral}$ to be the number of bilateral publications, i.e. 2-regions articles, between regions i and j occurring between 2001 and 2003. The empirical TENB is then defined as:

$$TENB_{ij} = \sum_{k \neq N_i \cap N_j} \frac{copub_{ik}^{bilateral} copub_{jk}^{bilateral}}{Mass_k}. \tag{5}$$

Because of the empirical specification, this variable is then 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 to the variable $TENB_{ij}$ in the econometric analysis is done in terms of elasticity, so that it is unaffected by a, the coefficient of proportionality.

Furthermore, an advantage of the the TENB is that its variation can easily be interpreted. Take the case of two region, i and j, from equation (5), we can see that the increase of 1% of the number of collaborations of the two regions *with their common neighbours* leads to an increase of 2% of the TENB measure.[11] Conversely, an increase of 1% of the TENB can then be interpreted by an increase of 0.5% of the collaborations with the common neighbours.

Other covariates. The variable $GeoDist_{ij}$ is 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 one when the regions i and j are from different countries and zero otherwise. To further take into account the notion of geographical proximity, a variable of regional

[11] Using the notations from section 3.2.2, the increase of 1% of the collaborations of i and j with their common neighbour k lead to a new level of collaboration with region k of $1.01g_{ik}$ and $1.01g_{jk}$. This in turn leads to an increase of the TENB of: $\sum_{k \neq i,j} (1.01g_{ik}) \times (1.01g_{jk}) / n_k \simeq 1.02 \sum_{k \neq i,j} g_{ik} \times g_{jk} / n_k = 1.02 \times TENB_{ij}$.

contiguity is created. This variable is aimed at capturing the effects of geography that are not seized by the geographical distance alone. The variable $notContig_{ij}$ is then of value one when two regions are not contiguous and of value zero otherwise. Additionally, to control for unobservable effects specific to the regional level, I include regional dummies in the model. Those dummies are specific to each region, whenever they appear as origin or destination.

4.3 Model and estimation procedure

As the dependent variable $Copub_{ij}$ is a count variable, a natural way to estimate equation (6) is via a Poisson regression as in other recent studies (e.g. Agrawal et al., 2014 or 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, more dispersion is allowed as the conditional mean gets larger, then 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 for instance. Particularly, they showed, by using simulations, that the estimates obtained in Poisson regression suffer from less bias than those obtained using other methods.

The structure of the data set, not differently from 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. are either the origin or the destination of the flow. When it comes to properly estimate the standard errors of the estimators, this dyadic structure is problematic. Indeed, in most econometric models, not controlling for the structure of correlation lead to erroneous standard errors that greatly overstate the precision of the estimator (Cameron and Miller, 2015a). As Cameron and Miller (2015b) show, the problem is even more acute for dyadic data. By the means of a Monte-Carlo study, they demonstrate that using the simple White heteroskedasticity-robust covariance matrix is unreliable as leading to standards errors several times lower than the dyadic-robust one. Moreover, this effect is only scarcely limited by using one-way or two-way clustering. Thus, in this econometric analysis, I use the methodo-

logy described by Cameron and Miller (2015b) to compute the dyadic-robust standard errors of the estimators.

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

$$E\left(Copub_{ij}|X_{ij}\right) = \alpha d_i d_j \left(TENB_{ij}+1\right)^{\beta_1} Mass_i^{\beta_2} Mass_j^{\beta_2} GeoDist_{ij}^{\beta_3} \times \exp\left(\beta_4 notContig_{ij} + \beta_5 CountryBorder_{ij}\right), \tag{6}$$

where X_{ij} represent the set of all explanatory variables, while d_i and d_j are the regional dummies of regions i and j. Note that one is added to the variable $TENB_{ij}$ as its value may be equal to zero. Further, as the relation $Copub_{ij}$ is undirected, the coefficient (β_2) associated to $Mass_i$ is the same as the one associated to $Mass_j$.

4.4 Descriptive statistics

The data set is composed of all the bilateral relations among 131 NUTS 2 regions, which leads to 8,515 $(= 131 \times 130/2)$ 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 3.3. The maximum is 156 and is between regions Île de France and Rhône-Alpes. The TENB defined by equation (5), is also unevenly distributed, but less than the number of co-publications, with a coefficient of variation of 2.3. Its maximum value, 16.6, is attained 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 9.26. Table 2 shows the correlations among the explanatory variables. The highest correlation is between the geographical distance and the country border variable.

[Table 1 about here.]

[Table 2 about here.]

5 Results

First, I focus on model (1), the standard gravity model where only size and geographical factors are taken into account. Consistent with previous literature (e.g. Hoekman et al., 2009, 2010; Scherngell and Barber, 2009), geography greatly affects collaboration. The most impeding effect is the country border effect. All else being equal, if two regions are from different countries, their collaboration flows will suffer a decrease of 82% $(1-\exp(-1.707))$. Though the effect of country borders is very high, the order of magnitude is in line with other estimates in the literature (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 collaborations by 0.35%. Seen with a larger variation, when the geographic distance doubles, collaborations decrease by 22% $(1-2^{-.35})$. Turning to the contiguity effect, as other distances, it has a non negligible effect on collaborations: being non contiguous instead of contiguous reduces the expected number of collaborations by 20%.

[Table 3 about here.]

Now I 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 0.01, positive although not significant. At first sight, network proximity, captured by the TENB variable, does not seem to influence network formation. One can conclude from this model that there is no homogeneous benefit from the network. Yet, it does not mean that network proximity has no effect at all as it could be mediated by geography.

[Figure 3 about here.]

To see 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) depicts significant estimates for both network proximity and its interaction with geographical distance, with a positive sign for the interaction. These estimates seem to imply a growing effect of network proximity with distance. Yet, 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 3 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. In fact, the estimates indicate that the effect is even negative (but not significant) for regions located at a distance lower than 114 km while the elasticity of the TENB is positive for regions farther apart. For instance, the effect starts to be significantly positive at the 5% level for regions at a distance of 537 km. For regions separated by the median distance, 1,000 km, the elasticity is 0.23, meaning that a 10% increase of the TENB would lead to an increase of co-publications of 2.3%.[12] This result is in line with the hypothesis of substitution between network proximity and geographical proximity. Finally, adding the interaction with the squared distance, in model (4), does not improve the estimation so that the effect of distance on network proximity is only monotonous.

[Table 4 about here.]

As geographical distance *per se* does not seize all forms of proximity, I decompose the effect of the TENB with respect to the country border dummy and the contiguity dummy. The first dummy will capture whether regions from different countries benefits more from network proximity, along with the substitution hypothesis. The second dummy captures a form of geographic proximity that is not directly seized by geographic distance. 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.

[12] From section 4.2, a 10% increase in the TENB between two regions can be interpreted by a 5% increase of collaboration flows of these two regions with their common neighbours.

Model (5) considers the sole decomposition with respect to country borders, it shows that network proximity influences international collaborations with an elasticity of 0.50 (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 is strictly increasing with the loss of other forms of proximity.[13] Figure 4 represents these estimates with their 95% confidence interval. For the most favourable case, that is when two regions are from the same country and are contiguous, the estimated elasticity is negative (-0.17) but not statistically different from 0 due to a large standard error. When the two regions lose the benefits of contiguity, the elasticity of the TENB becomes positive, rising to 0.16, while becoming significant at the 10% level. When they lose the benefits of belonging to the same country, the coefficient jumps to 0.46, and even reaches 0.54 when the regions are neither contiguous. These results confirms the hypothesis 2.b of substitution. Although the pattern is clear, it is worth to mention that only the elasticity of the TENB for international non-contiguous pairs of regions is significantly different from zero at the 0.001 level.

[Figure 4 about here.]

The main conclusions of the results are then twofold. First, the estimates show that network proximity has not an overall homogeneous effect but rather acts as a substitute to other forms of proximity: the effect of network proximity gets stronger with distance, either pure geographic distance or other forms of distance, namely country borders or non-contiguity. This fact validates the hypothesis 2.b of substitution. Second, for the regional pairs that benefit the most from non-network forms of proximity, the effect is non significant: network proximity is not always beneficial, so hypothesis 1 is only partially validated. Finally, as the TENB is a measure of network proximity that is rather conservative, the effects found in this paper are likely to be a lower bound.

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

6 Conclusion

This paper has investigated the role of networks in the formation of inter-regional research collaborations and 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, named TENB, is proposed in section 3.2. This measure fits well the gravity framework as it is independent from direct linkage, to prevent any endogeneity issue, and is defined for each dyad of regions. The measure ends out to be a conservative measure of network proximity as it is based only on indirect connections, neglecting any potential network benefit that would arise from direct linkages. 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 indirect connection at the micro level.

Next, I empirically assessed the influence of network proximity on network formation using data on co-publications over 131 NUTS 2 regions in the field of chemistry. To that purpose, the TENB variable was embedded in a gravity model estimated using Poisson regressions. Consistent with the existing literature, I find a significant, negative effect of separation variables such as geographical distance or country borders.

Notably, a clear substitutability pattern is revealed: the strength of network proximity rises when the benefits of geographic proximity, or of other non-network forms of proximity, wane. This suggests that network proximity alleviates the impeding effects of distances. Particularly, this result underscores the importance of network related effects in international collaborations. This fact bears an important significance in the context of policy making. Indeed, an important characteristic of distant collaborations, such as international ones, is that they provide higher quality of research production (see e.g. Adams et al., 2005; Narin et al., 1991). From this view point, the EU policies aiming at fostering international collaborations can have a sustained positive effect on knowledge production and ease future knowledge flows. As new international connections arise, the network proximity of regions

from different countries increases.[14] This in turn may trigger new international collaborations thanks to network effects, implying that more distant/more yielding collaborations are more likely to be created.

Natural extensions of this study could consider other fields of science which may have different collaboration patterns, as well as the extension to larger geographical areas. Particularly, a comparison with US data may be worthwhile to further 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? It can also be interesting to see 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.

[14]Consider two regions from different countries: i and j. If these two have a new collaboration, in consequence it rises the indirect connections (measured with the TENB) between i and all regions connected to j from j's country, and vice versa. Then new international collaborations indeed increase the network proximity between regions from the two countries.

References

Adams, J. D., Black, G. C., Clemmons, J. R. and Stephan, P. E. (2005). Scientific teams and institutional collaborations: Evidence from U.S. universities, 1981-1999. *Research Policy* 34: 259 – 285. 6

Aghion, P. and Howitt, P. (1992). A model of growth through creative destruction. *Econometrica* 60: 323–351. 1

Agrawal, A., Cockburn, I., Galasso, A. and 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. 4.3

Almendral, J. A., Oliveira, J. G., López, L., Mendes, J. and Sanjuán, M. A. (2007). The network of scientific collaborations within the European framework programme. *Physica A: Statistical Mechanics and its Applications* 384: 675–683. 1

Anderson, J. E. (2011). The gravity model. *Annual Review of Economics* 3: 133–160. 3.1

Autant-Bernard, C., Billand, P., Frachisse, D. and 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: 495–519. 1, 2.3, 3.1

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: 741–756. 1

Banchoff, T. (2002). Institutions, inertia and European Union research policy. *Journal of Common Market Studies* 40: 1–21. 2.1

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science* 286: 509–512. 3.2.4, 3.2.4, A

Barabâsi, A.-L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A. and Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A* 311: 590–614. 1

Beaver, D. d. (2001). Reflections on scientific collaboration (and its study): past, present, and future. *Scientometrics* 52: 365–377. 2.1, 3.2.3

Belderbos, R., Cassiman, B., Faems, D., Leten, B. and Looy, B. van (2014). Co-ownership of intellectual property: Exploring the value-appropriation and value-creation implications of co-patenting with different partners. *Research Policy* 43: 841 – 852. 4.3

Blau, J. R. (1974). Patterns of communication among theoretical high energy physicists. *Sociometry* 37: 391–406. 2.2

Boschma, R. (2005). Proximity and innovation: A critical assessment. *Regional Studies* 39: 61 – 74. 1, 2.1

Brambor, T., Clark, W. R. and Golder, M. (2006). Understanding interaction models: Improving empirical analyses. *Political Analysis* 14: 63–82. 5

Brenner, T. and Broekel, T. (2011). Methodological issues in measuring innovation performance of spatial units. *Industry and Innovation* 18: 7–37. 3.2.1

Breschi, S. and Lissoni, F. (2009). Mobility of skilled workers and co-invention networks: an anatomy of localized knowledge flows. *Journal of Economic Geography* 9: 439 – 468. 2.1

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

Cameron, A. C. and Miller, D. L. (2015a). A practitioner's guide to cluster-robust inference. *Journal of Human Resources* forthcoming. 4.3

Cameron, A. C. and Miller, D. L. (2015b). Robust inference for dyadic data. *Unpublished* . 4.3, 3, 4, 5

Castells, M. (1996). *The Rise of the Network Society*. Blackwell Publishers, Oxford. 1

Catalini, C. (2012). Microgeography and the direction of inventive activity. *Rotman School of Management Working Paper* . 1

Collins, H. M. (2001). Tacit knowledge, trust and the Q of sapphire. *Social Studies of Science* 31: 71–85. 2.1

Dahlander, L. and McFarland, D. A. (2013). Ties that last: Tie formation and persistence in research collaborations over time. *Administrative Science Quarterly* 58: 69 – 110. 2.2, 3.2.3, 5

Defazio, D., Lockett, A. and Wright, M. (2009). Funding incentives, collaborative dynamics and scientific productivity: Evidence from the EU framework program. *Research Policy* 38: 293–305. 4.1

d'Este, P., Guy, F. and Iammarino, S. (2013). Shaping the formation of university–industry research collaborations: what type of proximity does really matter? *Journal of Economic Geography* 13: 537–558. 2.3

Dijk, J. van and Maier, G. (2006). ERSA conference participation: does location matter? *Papers in Regional Science* 85: 483 – 504. 2.1

European commission (2012). Guide to research and innovation strategies for smart specialisations (RIS 3). *Joint Research Centre* . 1

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. 1

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

Freeman, R. B., Ganguli, I. and Murciano-Goroff, R. (2014). Why and wherefore of increased scientific collaboration. *National Bureau of Economic Research* . 1, 2.1

Frenken, K., Hardeman, S. and Hoekman, J. (2009a). Spatial scientometrics: Towards a cumulative research program. *Journal of Informetrics* 3: 222 – 232. 1

Frenken, K., Hoekman, J., Kok, S., Ponds, R., Oort, F. van and Vliet, J. van (2009b). Death of distance in science? A gravity approach to research collaboration. In *Innovation networks*. Springer Berlin Heidelberg, 43–57. 2.1, 4.2

Frenken, K., Ponds, R. and Oort, F. van (2010). The citation impact of research collaboration in science-based industries: A spatial-institutional analysis. *Papers in regional science* 89: 351–271. 2.3

Gertler, M. S. (1995). 'Being There': Proximity, organization, and culture in the development and adoption of advanced manufacturing technologies. *Economic Geography* 71: 1–26. 1, 2.1

Gertler, M. S. (2003). Tacit knowledge and the economic geography of context, or the undefinable tacitness of being (there). *Journal of Economic Geography* 3: 75–99. 2.1

Glänzel, W. (2001). National characteristics in international scientific co-authorship relations. *Scientometrics* 51: 69–115. 2.1, 2.3

Gulati, R. and Gargiulo, M. (1999). Where do interorganizational networks come from? *American Journal of Sociology* 104: 1439 – 1493. 2.2

Hazir, C. S., Lesage, J. and Autant-Bernard, C. (2014). The role of R&D collaboration networks on regional innovation performance. *Working paper GATE* 2014-26. 3.1

Hoekman, J., Frenken, K. and Oort, F. van (2009). The geography of collaborative knowledge production in Europe. *Annals of Regional Science* 43: 721 – 738. 1, 2.1, 2.3, 4.1, 5

Hoekman, J., Frenken, K. and Tijssen, R. J. (2010). Research collaboration at a distance: Changing spatial patterns of scientific collaboration within Europe. *Research Policy* 39: 662 – 673. 1, 2.1, 4.2, 5

Hoekman, J., Scherngell, T., Frenken, K. and Tijssen, R. (2013). Acquisition of European research funds and its effect on international scientific collaboration. *Journal of Economic Geography* 13: 23–52. 3.1

Jackson, M. O. and Rogers, B. W. (2007). Meeting strangers and friends of friends: How random are social networks? *American Economic Review* 97: 890–915. 1

Jones, B. F. (2009). The burden of knowledge and the 'death of the renaissance man': is innovation getting harder? *Review of Economic Studies* 76: 283–317. 1, 2.2

Jones, B. F., Wuchty, S. and Uzzi, B. (2008). Multi-university research teams: shifting impact, geography, and stratification in science. *Science* 322: 1259–1262. 1

Jones, C. I. (1995). R&D-based models of economic growth. *Journal of Political Economy* : 759–784. 1

Katz, J. S. (1994). Geographical proximity and scientific collaboration. *Scientometrics* 31: 31–43. 2.1

Katz, J. S. and Martin, B. R. (1997). What is research collaboration? *Research Policy* 26: 1–18. 1, 2.1, 2.2

Kirat, T. and Lung, Y. (1999). Innovation and proximity territories as loci of collective learning processes. *European Urban and Regional Studies* 6: 27–38. 2.1

Krackhardt, D. (1999). The ties that torture: Simmelian tie analysis in organizations. *Research in the Sociology of Organizations* 16: 183–210. 2.2

Lundvall, B.-Å. (1992). *National systems of innovation: Toward a theory of innovation and interactive learning*. London: Pinter, vol. 2. 1

Maggioni, M. A., Nosvelli, M. and Uberti, T. E. (2007). Space versus networks in the geography of innovation: A European analysis. *Papers in Regional Science* 86: 471 – 493. 1, 2.1, 2.3, 3.1, 3.1, 4.1, 5

Maggioni, M. A. and Uberti, T. E. (2009). Knowledge networks across Europe: which distance matters? *Annals of Regional Science* 43: 691 – 720. 3.1

McPherson, M., Smith-Lovin, L. and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 27: 415–444. 2.2

Miguélez, E. and Moreno, R. (2014). What attracts knowledge workers? the role of space and social networks. *Journal of Regional Science* 54: 33–60. 2.1

Morescalchi, A., Pammolli, F., Penner, O., Petersen, A. M. and Riccaboni, M. (2015). The evolution of networks of innovators within and across borders: Evidence from patent data. *Research Policy* 44: 651–668. 1, 2.1, 2.3

Narin, F., Stevens, K. and Whitlow, E. S. (1991). Scientific co-operation in Europe and the citation of multinationally authored papers. *Scientometrics* 21: 313–323. 2.3, 6

Newman, M. E. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences* 98: 404–409. 1

OST (2010). Indicateurs de sciences et de technologies. 4.2

Ponds, R., Oort, F. van and Frenken, K. (2007). The geographical and institutional proximity of research collaboration. *Papers in Regional Science* 86: 423 – 443. 2.3, 4.1

Roy, J. R. and Thill, J.-C. (2004). Spatial interaction modelling. *Papers in Regional Science* 83: 339–361. 3.1, 3.1

Santos Silva, J. a. M. C. and Tenreyro, S. (2006). The log of gravity. *Review of Economics and Statistics* 88: 641–658. 4.3

Scherngell, T. and 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: 531 – 546. 1, 2.1, 2.3, 5

Sebestyén, T. and 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, 71–97. 3.1

Sebestyén, T. and Varga, A. (2013b). Research productivity and the quality of interregional knowledge networks. *Annals of Regional Science* 51: 155–189. 3.1

Singh, J. and Fleming, L. (2010). Lone inventors as sources of breakthroughs: Myth or reality? *Management Science* 56: 41–56. 1

Singh, J. and Marx, M. (2013). Geographic constraints on knowledge spillovers: political borders vs. spatial proximity. *Management Science* 59: 2056–2078. 1

Storper, M. and Venables, A. J. (2004). Buzz: face-to-face contact and the urban economy. *Journal of Economic Geography* 4: 351–370. 1

Ter Wal, A. L. J. (2011). Networks and geography in the economics of knowledge flows: a commentary. *Quality & Quantity* 45: 1059–1063. 3.2.1

Ter Wal, A. L. J. (2013). The dynamics of the inventor network in German biotechnology: geographic proximity versus triadic closure. *Journal of Economic Geography* : In press. 2.2

Torre, A. and Rallet, A. (2005). Proximity and localization. *Regional studies* 39: 47–59. 1, 2.1

Wagner, C. S. and Leydesdorff, L. (2005). Network structure, self-organization, and the growth of international collaboration in science. *Research policy* 34: 1608–1618. 1

Wanzenböck, I., Scherngell, T. and Brenner, T. (2013). 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* : 1–32. 3.1

Wanzenböck, I., Scherngell, T. and Lata, R. (2014). Embeddedness of European regions in European Union-funded research and development (R&D) networks: A spatial econometric perspective. *Regional Studies* : 1–21. 3.1

Wuchty, S., Jones, B. F. and Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science* 316: 1036–1039. 1

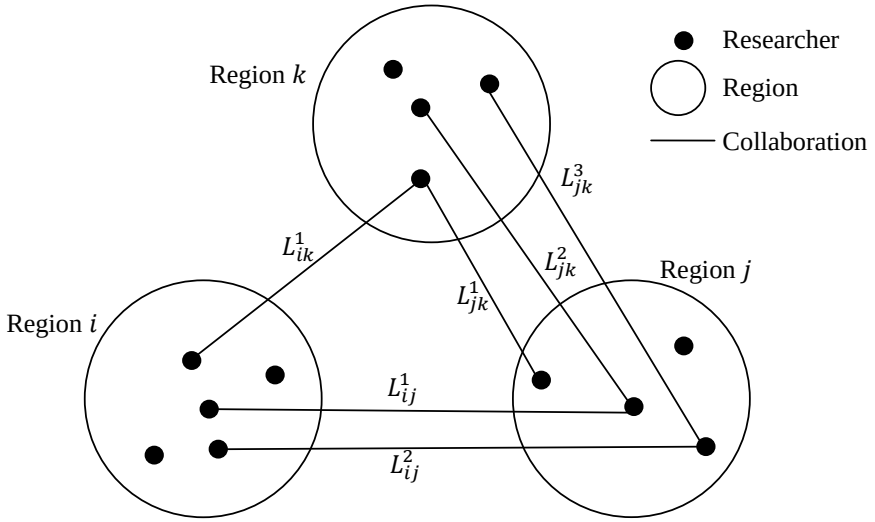

Figure 1: Illustration of a regional network of collaboration and of the notion of bridging path.
Notes: The figure depicts three bridging paths formed by the following pairs of links: (L^1_{ik}, L^1_{jk}), (L^1_{ij}, L^2_{jk}) and (L^2_{ij}, L^3_{jk}). So the regional dyads (i,j), (i,k) and (j,k) have 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 because they have no agent indirectly connected.

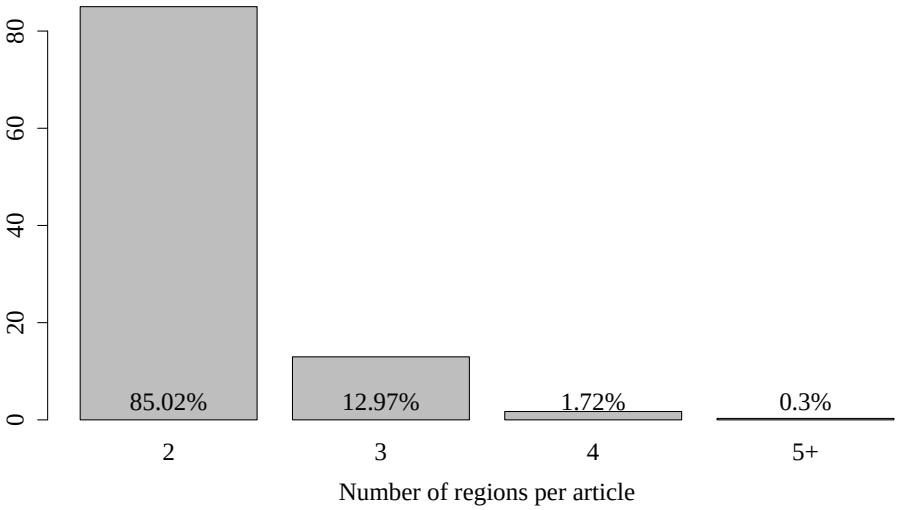

Figure 2: Distribution of the number of regions involved in inter-regional articles.

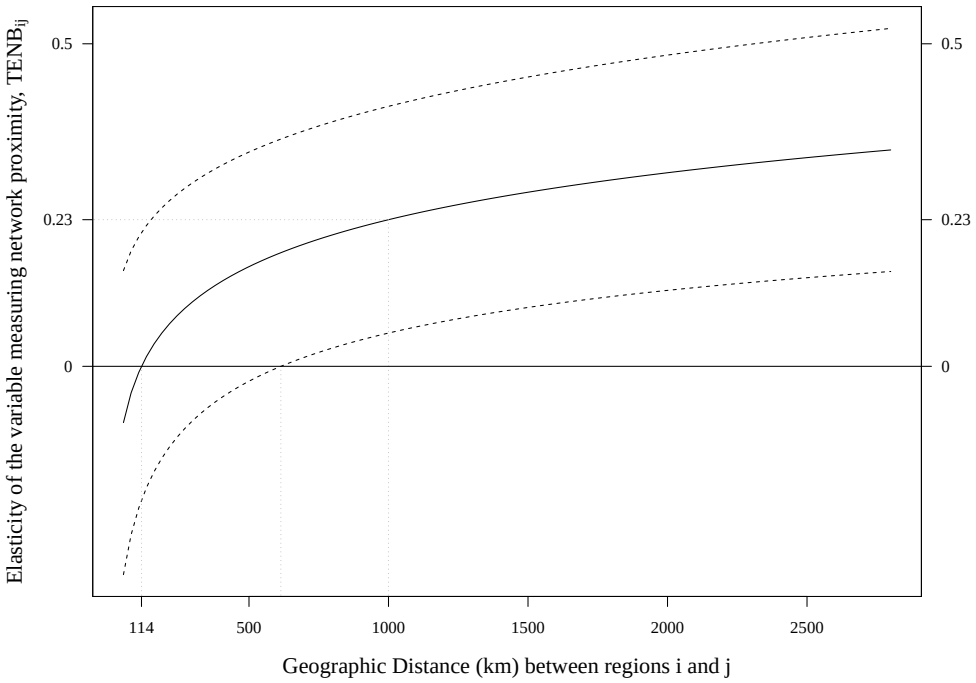

Figure 3: Graph of the interaction between network proximity and geographical distance.
Notes: The graph represents the estimated elasticity of the TENB on co-publications with respect to geographical distance (solid line) along with its 95% confidence interval (dashed lines). It was made using the estimates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

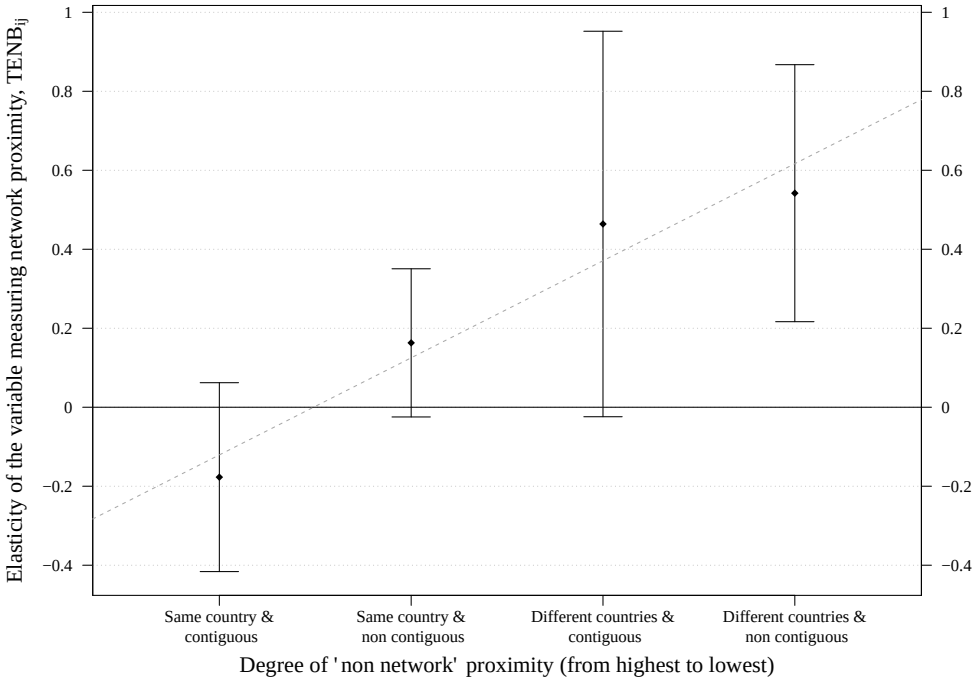

Figure 4: Graph of the link between network proximity and other forms of proximity.
Notes: The graph relates the elasticity of the TENB on co-publications with respect to different degrees of 'non-network' proximity. Both the estimates of the elasticity and their 95% confidence intervals are represented. It uses 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 of the collaboration network.

	Min	Median	75^{th} percentile	Max	Mean	SD
Co-publications	0	0	1	156	1.587	5.239
TENB	0	0.090	0.323	16.65	0.337	0.790
Mass origin	2	507	862	4859	671.0	693.1
Mass destination	2	410	803	4859	598.6	655.7
Geographical distance	5.48	1002.5	1368.4	2626.9	999.2	526.7
Non-Contiguity	0	1	1	1	0.967	0.177
Different Country	0	1	1	1	0.784	0.411

Table 2: Correlation matrix of the covariates.

		1	2	3	4	5	6
1	TENB (ln)	1.000					
2	Mass Origin (ln)	0.393*	1.000				
3	Mass Destination (ln)	0.436*	-0.002	1.000			
4	Geographical distance (ln)	-0.276*	0.059*	-0.034*	1.000		
5	Non-Contiguity (ln)	-0.155*	0.002	-0.014	0.298*	1.000	
6	Different Country (ln)	-0.366*	0.064*	-0.019	0.641*	0.328*	1.000

*: 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]		0.0173	-1.329***	-1.772***
		(0.0755)	(0.299)	(0.555)
TENB (ln) $*$ Geo. Distance (ln)			0.248***	0.401**
			(0.0521)	(0.159)
TENB (ln) $*$ Squared Geo. Distance (ln)				-0.0134
				(0.0119)
Mass Origin (ln), Mass Destination (ln)	0.810***	0.806***	0.734***	0.730***
	(0.137)	(0.140)	(0.0324)	(0.0317)
Geographical Distance (ln)	-0.346***	-0.346***	-0.546***	-0.564***
	(0.0350)	(0.0349)	(0.0775)	(0.0802)
$\mathbb{1}_{\{Not\ Contiguous\}}$	-0.225**	-0.225**	-0.296***	-0.291***
	(0.0692)	(0.0694)	(0.105)	(0.101)
$\mathbb{1}_{\{DifferentCountries\}}$	-1.707***	-1.695***	-1.540***	-1.520***
	(0.0470)	(0.0736)	(0.152)	(0.157)
Constant	-6.942***	-6.888***	-4.762***	-4.616***
	(1.811)	(1.833)	(0.562)	(0.614)
Regional dummies (Origin & Destination)	yes	yes	yes	yes
Number of Observations	8515	8515	8515	8515
Pseudo-R^2	0.679	0.679	0.679	0.681
BIC	19682.2	19691.1	19669.9	19597.6

Notes: The model estimated is depicted by equation (6). 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 e.g. section 3.2). Dyadic-robust standard errors in parenthesis (see Cameron and Miller, 2015b). Level of statistical significance: * 10%, ** 5%, *** 1%.

Table 4: Results of the Poisson regression where the TENB is decomposed with respect to country borders and contiguity.

Model:	(5)	(6)
Dependent variable:	Co-publications	Co-publications
TENB (ln) $* \mathbb{1}_{\{SameCountry\}}$	0.0571	
	(0.110)	
TENB (ln) $* \mathbb{1}_{\{DifferentCountries\}}$	0.509***	
	(0.172)	
TENB (ln) $* \mathbb{1}_{\{SameCountry\}} * \mathbb{1}_{\{Contiguous\}}$		-0.177
		(0.122)
TENB (ln) $* \mathbb{1}_{\{SameCountry\}} * \mathbb{1}_{\{Not\ Contiguous\}}$		0.163*
		(0.0957)
TENB (ln) $* \mathbb{1}_{\{DifferentCountries\}} * \mathbb{1}_{\{Contiguous\}}$		0.464*
		(0.249)
TENB (ln) $* \mathbb{1}_{\{DifferentCountries\}} * \mathbb{1}_{\{Not\ Contiguous\}}$		0.542***
		(0.166)
Mass Origin (ln), Mass Destination (ln)	0.770***	0.754***
	(0.0332)	(0.0319)
Geographical Distance (ln)	-0.320***	-0.301***
	(0.0617)	(0.0575)
$\mathbb{1}_{\{Not\ Contiguous\}}$	-0.249**	-0.651***
	(0.103)	(0.128)
$\mathbb{1}_{\{DifferentCountries\}}$	-1.950***	-1.870***
	(0.144)	(0.132)
Constant	-6.530***	-6.118***
	(0.439)	(0.405)
Regional dummies (Origin & Destination)	yes	yes
Number of Observations	8515	8515
Pseudo-R^2	0.682	0.683
BIC	19547.4	19480.8

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). Dyadic-robust standard errors in parenthesis (see e.g. Cameron and Miller, 2015b). Level of statistical significance: * 10%, ** 5%, *** 1%.

Table 5: Results of the Poisson regression when the dependant variable is the number of collaborations in fractional count.

Dependent variable:	Co-publications (fractional)	Co-publications (fractional)	Co-publications (fractional)
TENB (ln) [proxy for network proximity]	0.0259	-1.334***	
	(0.140)	(0.321)	
TENB (ln) $*$ Geo. Distance (ln)		0.250***	
		(0.0551)	
TENB (ln) $* \mathbb{1}_{\{SameCountry\}} * \mathbb{1}_{\{Contiguous\}}$			-0.188
			(0.127)
TENB (ln) $* \mathbb{1}_{\{SameCountry\}} * \mathbb{1}_{\{Not\ Contiguous\}}$			0.180*
			(0.0978)
TENB (ln) $* \mathbb{1}_{\{DifferentCountries\}} * \mathbb{1}_{\{Contiguous\}}$			0.451*
			(0.254)
TENB (ln) $* \mathbb{1}_{\{DifferentCountries\}} * \mathbb{1}_{\{Not\ Contiguous\}}$			0.542***
			(0.155)
Mass Origin (ln), Mass Destination (ln)	0.795***	0.724***	0.743***
	(0.0368)	(0.0337)	(0.0323)
Geographical Distance (ln)	-0.367***	-0.571***	-0.322***
	(0.0764)	(0.0797)	(0.0582)
$\mathbb{1}_{\{Not\ Contiguous\}}$	-0.227*	-0.295***	-0.682***
	(0.116)	(0.109)	(0.136)
$\mathbb{1}_{\{DifferentCountries\}}$	-1.723***	-1.568***	-1.885***
	(0.177)	(0.152)	(0.132)
Constant	-7.516***	-5.378***	-6.734***
	(0.544)	(0.568)	(0.398)
Regional dummies (Origin & Destination)	yes	yes	yes
Number of Observations	8515	8515	8515
Pseudo-R^2	0.619	0.621	0.622
BIC	10843.9	10791.1	10758.4

Notes: The dependent variable is the number of co-publications *in fractional count* 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). Dyadic-robust standard errors in parenthesis (see e.g. Cameron and Miller, 2015b). Level of statistical significance: * 10%, ** 5%, *** 1%.

A Preferential attachment

In this section I consider the matching mechanism described in section 3.2.4. 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.

A.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:

$$\begin{aligned}\left(\iota^{-0.5}/\Gamma\right)\times L &< \mathtt{k}\\ \iota^{-0.5} &< \frac{\mathtt{k}\Gamma}{L}\\ \iota &> \left(\frac{L}{\mathtt{k}\Gamma}\right)^2. \end{aligned} \tag{7}$$

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 (7). The share of agents having a degree lesser than $\mathtt{k}$ is then:[15]

$$\begin{aligned}F(\mathtt{k}) &= \frac{1}{n}\left(n-\iota(\mathtt{k})\right)\\ &= 1-\frac{1}{n}\left(\frac{L}{\Gamma}\right)^2\mathtt{k}^{-2}. \end{aligned} \tag{8}$$

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:

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

[15] 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 $n-\iota(\mathtt{k})$ but, as this number cannot be negative, its value is $\max(n-\iota(\mathtt{k}),0)$. Now let $\mathtt{k}^*$ to 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 (8) is defined only for $\mathtt{k} \geq \mathtt{k}^*$ and is zero 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.

A.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 3.2.3. Consider a region k with n_k agents. The number of links between k and regions i and j are g_{ik} and g_{jk} respectively.

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

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

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

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

Further, notice that $\Gamma = \sum_{\iota=1}^{n_k} \iota^{-0.5} \simeq \int_1^{n_k} x^{-0.5} \mathrm{d}x = 2\left(\sqrt{n_k} - 1\right)$, 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. □

B List of the 131 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

CODE	NAME	CODE	NAME
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
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

CODE	NAME	CODE	NAME
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		