GREThA
Groupe de Recherche en Économie Théorique et Appliquée

Unintended triadic closure in social networks: The strategic formation of research collaborations between French inventors.

Nicolas CARAYOL

GREThA, CNRS, UMR 5113

Université de Bordeaux

Observatoire des Sciences Techniques

Paris

&

Lorenzo CASSI

CES - Université Paris 1 Panthéon Sorbonne

Observatoire des Sciences et Techniques, Paris

&

Pascale ROUX

GREThA, CNRS, UMR 5113

Université de Bordeaux

Cahiers du GREThA

n° 2014-13

July

Closure' triadique non intentionnelle dans les réseaux sociaux: La formation stratégique des collaborations de recherche entre inventeurs français

Résumé

Dans la plupart des études empiriques et théoriques qui cherchent à comprendre les comportements individuels conduisant à la formation de réseaux sociaux, il est avancé que si les réseaux réels sont clusterisés c'est parce que les individus valorisent la fermeture (« closure ») triadique de leurs relations (ils aiment que leurs amis soient amis entre eux). Dans ce papier, nous soutenons que le clustering peut aussi s'observer en dépit du fait que les individus n'aiment pas la redondance de leurs connexions. Nous proposons un modèle théorique de formation de nouvelles collaborations de recherche que nous estimons sur l'évolution longitudinale du réseau français de co-invention de brevets. Nous montrons que si ce réseau social est clusterisé, c'est parce qu'il est corrélé avec des métriques exogènes affectant les coûts de formation des liens directs, et non parce que les agents préfèrent, en soi, fermer des triangles. Ce résultat est obtenu grâce à la richesse de nos données qui nous permettent de contrôler pour les effets fixes des dyades ainsi que pour les divers coûts de formation des relations (la distance géographique, la spécialisation technologique, les frontières institutionnelles et leurs caractéristiques), omis dans les études précédentes.

Mots-clés : Formation stratégique des réseaux; Collaborations inter-individuelles; Closure; Clustering; Brevets.

Unintended triadic closure in social networks: The strategic formation of research collaborations between French inventors

Abstract

Most of the empirical and theoretical literature aimed at understanding the behavioral patterns that lead to the formation of social networks argue that such networks are clustered because agents like social closure, since it facilitates cooperation enforcement, for instance, or increases match quality. We argue that, in certain circumstances, network clustering may arise for other reasons, even though agents may actually not like redundancy in connections. We propose a theoretical model of the formation of new research collaboration that we estimate on the longitudinal evolution of the French co-invention network. We show that if this type of social network is closed it is because it correlates with exogenous metrics affecting the costs of direct link formation, not because agents prefer to close triangles per se. This result is obtained thanks to the richness of our dataset, allowing us to control for dyadic fixed-effects and various costs of network formation (geographical distance, technological specialization, and institution boundaries and attributes) omitted in previous studies.

Keywords: Strategic network formation; Inter-individual collaborations; Closure; Clustering; Patents.

JEL: D85; C23; O31, Z13

Reference to this paper: CARAYOL Nicolas, CASSI Lorenzo, ROUX Pascale (2014) Unintended triadic closure in social networks: The strategic formation of research collaborations between French inventors, *Cahiers du GREThA*, n°2014-13.

http://ideas.repec.org/p/grt/wpegrt/2014-13.html.

GRETHA UMR CNRS 5113
Université Montesquieu Bordeaux IV
Avenue Léon Duguit - 33608 PESSAC - FRANCE
Tel : +33 (0)5.56.84.25.75 - Fax : +33 (0)5.56.84.86.47 - www.gretha.fr

Unintended triadic closure in social networks: The strategic formation of research collaborations between French inventors[1]

Nicolas Carayol [◊,†,2], Lorenzo Cassi[‡,†,3], Pascale Roux[◊,4]
[◊] GREThA, Université de Bordeaux - CNRS
[‡] CES - Université Paris 1 Pantheon Sorbonne
[†] Observatoire des Sciences et Techniques, Paris

July 16, 2014

[1]We would like to thank Matt Jackson, Tom Snijders, and participants in the 2012 EEA conference in Malaga, the 2012 Workshop on Economic Design and Institutions held at Facultes Universitaires Saint Louis in Brussels, the 2011 conference on Regional Innovation and Growth in Pecs, the 2011 First International NPR Conference at the Universita Cattolica del Sacre Cuore in Milano, and the ARS 2013 conference in Rome. We would also like to thank the Agence Nationale de la Recherche (grant ANR-06-JCJC-0076) and the Aquitaine Region (AAP program) for their financial support.

[2]Corresponding author. Nicolas Carayol, Université de Bordeaux, GREThA - CNRS, Avenue Leon Duguit, F-33608 Pessac Cedex. Tel: +33-556844051. Email: nicolas.carayol@u-bordeaux.fr

[3]Lorenzo Cassi, CES, Université Paris 1 Pantheon Sorbonne, 106 - 112 Boulevard de l'Hôpital, 75013 Paris. Email: lorenzo.cassi@univ-paris1.fr

[4]Pascale Roux, Université de Bordeaux, GREThA - CNRS, Avenue Leon Duguit, F-33608 Pessac Cedex. Tel: +33-556842558. Email: pascale.roux@u-bordeaux.fr

1 Introduction

In the quest shared by several scientific disciplines to understand how social and economic networks are formed, we have benefited from numerous empirical results characterizing their structural properties. We now know that large real networks, though complex and heterogeneous in many respects, do nonetheless share certain common structural features. Perhaps the most salient feature of most real social networks is that they are highly clustered, in the sense that the neighborhoods of neighbors tend to overlap, forming triangles. The frequency with which agents' neighbors are themselves neighbors is, on average, several thousand times higher than in similar (size and density) randomly built networks. This property has been observed in a variety of network contexts such as those involving Hollywood actors (Watts, 1999), corporate board members (Davis et al., 2003), Broadway musicians (Uzzi and Spiro, 2005), inventors (Fleming et al., 2007), scientists (Newman, 2001) or alliances between firms (Kogut and Walker, 2001; Baum et al., 2003). The literature thus clearly converges to acknowledge that social networks exhibit what Rapoport (1953) and Granovetter (1973) first called "triadic closure".

This strong tendency of individuals to cluster in networks initially puzzled researchers seeking to understand the factors affecting the formation of connections. One direct and natural explanation of the social bias for triangles is what we call the "love for triadic closure" hypothesis. This takes different forms in the literature. Following on from Simmel who, in the early twentieth century, was the first to systematically describe the sophisticated games at play between three persons, Heider (1958) and Newcomb (1961) explored the psychological motives of individuals to maintain a "cognitive balance" between their social relations. They argued that agents dislike it when their neighbors are not also friends, because this creates tensions and instability. Recently, Bearman and Moody (2004) have shown that network intransitivity (the proportion of two-step relations that are not closed) is associated with a higher rate of suicide among female teenagers. Taking this argument one step further, agents may be directly interested in their neighbors also being friends and, thus, may intentionally favor such connections. Along these lines, Karlan et al. (2009) argue that common neighbors can act as social collaterals in bilateral inter-individual lending agreements. For Fafchamps et al. (2010), agents play the role of referees between their neighbors, so that information on the quality of potential bilateral matches flows through existing connections. Though the authors do not explicitly model intermediaries' strategic behaviors, they do interpret the negative effect of social distance on the probability of two economists collaborating (writing a paper together) due to their common neighbors having served as referrals.

Rather than focusing on the role played by common neighbors in arranging for their friends to meet and then set up connections, several authors highlight the beneficial influence of a common third-party on the decision by agents to form a link. Granovetter (1973) suggests that two agents who independently spend time with a third one have a greater chance of meeting and, thus, of becoming linked. Coleman (1988) emphasizes that closure, by facilitating collective monitoring and sanctions, thereby prevents free-riding and enforces cooperative behavior. Recently, several authors (e.g. Raub and Weesie, 1990; Ali and Miller, 2012) have used repeated games theory to model this idea: assuming that information on bad behavior flows through network connections, overlapping neighborhoods speed its diffusion in the community (between the recipients of misbehavior and potential punishers) and therefore shorten the waiting time before punishment. Granovetter (1985) also argues that closure facilitates interpersonal trust, since it creates a reputation cost for individuals who misbehave. Buskens (2003) builds on this idea to develop a model with incomplete information about the utility an agent derives from abusing the trust of two trustees. In a situation of network embeddedness, in which two trustees can learn from each other's experience with the agent, s/he has greater incentives to build a reputation for honoring others' trust. Jackson et al. (2012) adopt a slightly different perspective. They show that, in a complete information setting, local cliques are sustained in the equilibrium not just because they induce cooperation, but especially because they minimize the social contagion of punishment.[1]

All these arguments highlight the advantages of triadic closure. An important line of research in sociology, however, claims that individuals benefit from non-redundant information obtained from different clusters, provided they are able to build "bridges" between those disconnected communities. According to this approach, closure produces informational homogeneity and redundancy or, even, isolation, since individuals share the same sources of information. These ideas, which were developed in the early work of Granovetter (1973), hark back to another original idea advanced by Simmel (1922). He argues that a third person,"the Tertius Gaudens", gains from acting as an intermediary between two disconnected persons. In Burt's theory of "structural holes" (1992), gaps in networks provide opportunities for (entrepreneurial) individuals to bridge separate clusters and, thus, to play such an intermediary role. Thanks to bridges between separate communities, those individuals can synthesize, select and broker the flow of information or ideas arising from disconnected parts of the network, and can even control projects from opposite sides of the hole (Burt, 2000). Recent empirical studies show

[1]Though the authors highlight the notion of "support" rather than clustering, the equilibrium networks still exhibit a high level of clustering.

that brokers have higher performances, for instance, in tracking job opportunities (Granovetter, 1974), generating good ideas (Burt, 2004), obtaining promotions (Burt, 1997), or enhancing firms' inventiveness (Ahuja, 2000; Baum et al., 2000).[2]

The opposition between those who highlight the advantages of bridging connections, and those who might favor the "love for triadic closure" hypothesis, has challenged those researchers attempting to develop a synthetic explanation of link formation. One way of reconciling these views is to introduce heterogeneity between agents, assuming that intrinsic abilities, personality traits, or specific incentives of agents could eventually lead them to specialize in one type of neighborhood. Burt (2000) noted, for instance, that agents are not equally well-off as brokers between communities. Most people (more risk adverse agents, in particular) would favor closure in their relations, whereas certain entrepreneurial-minded individuals would build bridges between disconnected clusters. Burt (2000) also suggests that agents can sustain intrinsically different types of link. They can, therefore, eventually specialize in building just one type.

An alternative, and somewhat less ad-hoc point of departure, is to be found in the observation that it is not necessary to assume any "love for triadic closure" ingredient, or specific individual abilities or link attributes when modeling link formation, in order to obtain clustered equilibrium networks. Let us consider, for instance, the well-known connections model introduced by Jackson and Wolinsky (1996). In this model, agents benefit from positive externalities from other agents with whom they are indirectly connected, and the strength of the externality decays with social distance. As a consequence, a specific feature of the model is that agents do not benefit from multiple paths to some other agent: only the length of one of the shortest paths matters. There is a replacement effect. In particular, when two agents already have at least one common friend, they have fewer incentives to form a link with each other, *ceteris paribus*, because, thanks to this common neighbor, they already benefit from each other. Does this trait disqualify the model from explaining the formation of the particular kind of social network which does exhibit triadic closure? Carayol and Roux (2009) have shown that this is not the case in a simple extension of the connections model. In fact, introducing an exogenous structure that affects the direct link costs is a sufficient condition to ensure that agents form triadic connections in all pairwise stable networks for non-extreme values of the strength of the externalities. Agents do so simply because, though the gross returns of these connections are limited, their costs are also very small. Many triangles can therefore be observed in real social networks, even though agents do not like closing them. Interestingly, in the

[2]For a survey, see Burt (2000).

long run, some agents do strategically sustain long-distance (very costly) connections. There are only a few such social bridges because, once long-distance connections are formed, the neighbors' incentives to form similar long-distance connections disappear. Neighbors free-ride on these connections to gain access to distant agents. Therefore, the strategically formed long-run equilibrium networks are closed locally, though a few bridges between separate communities are formed, thereby constituting small worlds, in the sense of Watts and Strogatz (1998). Since, in that model, all agents are *ex ante* identical, and are not even allowed to broker knowledge, this result provides a simple and neat explanation of the puzzle, going against the "love for triadic closure" hypothesis.

The main goal of the present paper is to develop a strategic model of network formation compatible with both hypotheses, and to estimate that model, using large real network data so as to obtain empirical validation or rejection of the "love for triadic closure" hypothesis. Admittedly, the results are limited to the context of our empirical application: the formation of research collaborations between inventors. However, knowledge networks are good candidates for challenging the "love for triadic closure" hypothesis. On the one hand, Zander and Kogut (1995) argue that knowledge transfer and the speed of its diffusion are facilitated by closure, in particular when that knowledge is highly tacit or complex. Reagans and McEvily (2003) note that closure favors individuals' willingness to make efforts to share knowledge. On the other hand, social bridges have been associated with greater ability to generate knowledge (Burt, 1992, 2004; Ahuja, 2000; Baum et al., 2000).

We introduce a simple theoretical model in which agents (individual researchers), at any point in time, may consider the formation of a bilateral collaboration that will produce an expected direct return and generate various costs. If this research collaboration is undertaken, a new social tie is formed within the larger social network generated by preexisting collaborations. These connections are conducive to positive externalities. Though the model has been voluntarily kept as simple as possible, it is sufficiently general to encompass opposing assumptions about triadic closure: either multiple paths to some agent are complementary and agents typically like forming triangles, or multiple paths are partial or perfect substitutes, which implies that agents do not care about forming triangles, or even dislike doing so. This model leads to a simple expression of the incentive schemes for forming collaborations at any point in time, either in or out of equilibrium, which we estimate by using co-invention data.

Our empirical evidence is based on the relational information contained in all European patent applications over the period 1978-2004, for which at least one inventor has

declared a personal address in France. For each year concerned, we build the co-invention network by allocating a connection between two individuals if they have both previously appeared among the inventors of the same patent application. From a methodological point of view, the procedure is similar to that performed for measuring scientific collaboration networks from data on the co-authorship of scientific publications (Newman, 2001, Barabasi et al. 2002). Individual information, such as inventors' geographic location, is available. We also build various covariates accounting for the costs of link formation. Some of these covariates rely upon the identification of patent applicants. These are subsequently matched against the list of companies, whose mandatory annual survey provides us with detailed information on applicants, including their R&D investments and research personnel.

If we now turn to the estimation of our theoretical model, the above discussion of the "love for triadic closure" hypothesis highlights the need for the proper consideration of the link formation costs when estimating models of network formation. Should such costs not be accounted for, the omitted variables (the unobserved costs) that affect the decision to form links would be correlated with closure. Such regressions are likely to provide spurious estimates of having friends in common, leading to the mistaken conclusion that agents have incentives to form triadic connections. Typically, if one intends to estimate the impact of having a common friend on the probability to form a link, but ignores some (potentially time-varying) costs, such as geographical distance, institutional barriers or common interests, one may find positive and significant estimates of closure whereas, in fact, it plays absolutely no role. This is likely to occur if, in the true data generation process, the costs affect in the same way both the probability of connecting and the probability of having common friends.

Our main result is that, when the connection costs are not properly accounted for, the estimations lead to the conclusion that agents like closing triangles, whereas when we account for the costs, we find that agents do not like closing triangles. We also observe that, according to certain specifications, they even dislike triadic closure. This result thus speaks against the "love for closure hypothesis".

This paper is also related to the literature on knowledge spillovers and invention. Abundant empirical evidence highlights the role of knowledge spillovers, which are proven to be spatially and technologically bounded (Jaffe, 1986), as important determinants of inventiveness and productivity growth (e.g. Griliches 1992). Several recent studies have also shown that interpersonal networks are crucial determinants of their transmission (Singh, 2005; Breschi and Lissoni, 2006). Nevertheless, we still have little research on

the formation of knowledge networks since, in most studies, network ties are taken as exogenous. Here, we consider individuals' decisions to form new bonds in order to take advantage of the resulting network. A better understanding of such individual strategies should help to provide stronger foundations for policies aimed at supporting innovation.

The following section introduces our simple theoretical model of strategic research collaboration formation, and our empirical strategy. The third section presents the data. The fourth section exposes our findings and discusses their robustness. The last section concludes.

2 The strategic formation of inter-individual research collaboration networks

In this section, we present the different building blocks of the theoretical model and discuss our empirical strategy.

2.1 The game

At each period t of the discrete time, we consider a finite set of n^t agents, $N^t = \{1, 2, ..., n^t\}$. New agents may enter the population at the beginning of any period and, for the sake of simplicity in the exposure, agents are assumed never to retire or die,[3] so that $N^t \in N^{t+1}$. A (non-directed) link between two distinct agents i and $j \in N^t$ is denoted ij. Let g^t denote the relational network in place at the beginning of period t, that is the collection of all existing links at that point in time. We also assume that agents never consider link deletion, so that $g^t \in g^{t+1}$. At the beginning of each period, all pairs of unconnected agents simultaneously meet with some given uniform small probability p. They may then decide to establish research collaboration or not on the basis of the perceived impact of the new link on the discounted net present value of their payoffs. Agents can bargain bilaterally when they consider forming a link together, so that a link will be formed between two agents who meet, if their expected joint payoffs are greater when the project is launched. We do not consider the precise way in which agents bargain, but just assume that the bilateral transfers are such that the link is always formed when the two agents find it jointly profitable to do so. Finally, agents are myopic, in the sense that they do not anticipate the impact of their present moves on subsequent moves: they consider that the network formed in the present period is a permanent one. This standard assumption is usually considered as relevant when one considers large

[3] This assumption could be easily relaxed without changing any of the predictions.

networks in which forward looking computations become extremely complex (Jackson, 2009).

2.2 Individual payoffs

A research project generates immediate (pair specific) net payoffs, and brings a social connection into the web of already existing connections, which also generates per period returns.

The expected (net) returns for i of a shared research project with j formed at period t is given simply by:

$$r^t(i,j) = \theta_{ij} + \varepsilon^t_{ij} - \zeta c^t_{ij}, \tag{1}$$

where the gross returns of the research collaboration are composed of θ_{ij}, an idiosyncratic, pair-specific and time-invariant parameter, and of ε^t_{ij}, its time-variant counterpart, interpreted as the opportunities of research collaboration between i and j that particular year. The variable c^t_{ij} captures all the (sunk) costs, supported by i, for running the research collaboration with j at period t. ζ is a non-null and positive parameter. To simplify the exposure of the model, though that is not necessary for our results, we will further assume that $\theta_{ij} = \theta_{ji}$, $\varepsilon^t_{ij} = \varepsilon^t_{ji}$ and $c^t_{ij} = c^t_{ji}$ so that $r^t(i,j) = r^t(j,i)$, namely the net primary payoffs of a research collaboration, are identical for the two agents involved.

Research collaboration between two agents who are not already connected consists in a bilateral social connection that is assumed to be permanent. The complex of bilateral social connections is also assumed to be the support of positive externalities at each period. We propose the following simple specification of these "network" per period payoffs:

$$\pi_i\left(g^t\right) = \sum_{j \neq i} \left(\alpha \eta_{ij}\left(g^t\right) + \beta \eta^2_{ij}\left(g^t\right) + \gamma \overset{\triangle}{\eta}_{ij}\left(g^t\right) \right), \tag{2}$$

with $\eta_{ij}\left(g^t\right)$ the number of direct links between i and j on g^t (equal to 0 or 1), $\eta^2_{ij}\left(g^t\right)$ the number of paths of length two between i and j on g^t, provided there is no direct link, between i and j, and $\overset{\triangle}{\eta}_{ij}\left(g^t\right)$ the number of triangles on g^t having i and j as summits.[4] The two positive parameters, α and β, capture the imperfect knowledge spillovers that flow through local connections: α scales the knowledge spillover from a direct neighbor, β gives the spillover that flows through any path of length two from some other, provided there is no direct link to that agent. Parameter γ scales the positive externality captured

[4] That is also the number of common neighbors of i and j, or the number of paths of length 2 between i and j, provided there is a direct link between i and j.

by i for each indirect connection of length two to any direct neighbor. It captures both the knowledge spillover flowing on such a path, and the closure effect.

It should be noted that, according to this payoff specification, agents are assumed to only consider social network externalities at geodesic distance less than or equal to two ($d(i,j) \leq 2$). This is a natural assumption for the closure effect, but needs to be justified for the knowledge spillover effect. One convincing justification for not considering knowledge flows at distances strictly greater than two is provided by Singh (2005) and Breschi and Lissoni (2006), who show that the probability of patent citations decreases sharply in function of the social distance between patent inventors, and that these spillovers are null or nearly null at a social distance equal to or greater than three. It should also be noted that externalities are here associated with paths, and not agents.[5] Therefore, one agent may benefit from another agent via different paths, the total gain from that second agent being additive to the gain from each path.

2.3 Bilateral incentives to form connections

We now focus on the bilateral incentives to form connections, once two unconnected agents have just met. Agents can bargain bilaterally when they consider forming a link together and operate sunk bilateral transfers, in particular when these transfers are necessary to convince one of the two agents. This assumption is consistent with the idea that, in research collaboration, not all agents contribute equally: often more peripheral agents accept to contribute more to a project, which materializes here as a bilateral transfer. Since agents do not consider the further moves induced by their present collaboration, the transfers are rationally limited to the private returns generated by the link. Agents are, however, not allowed to subsidize the formation of a link they are not directly involved in, which is also a reasonable behavioral assumption. Therefore, the total variation of expected worth for the two agents, due to the creation of a new link between them, constitutes the (dyadic) incentives to form connections, whatever the effective bilateral transfers they operate. Let $\Delta(g^t, ij)$ denote that variation of agents i and j discounted payoffs, if link ij is created while the network g^t is in place (with $ij \notin g^t$). Using Equations 1 and 2, this is given by:

$$\begin{aligned}\Delta(g^t, ij) = {} & 2\left(\theta_{ij} + \varepsilon_{ij}^t - \zeta c_{ij}^t\right) \\ & + \frac{1}{1-\delta}\left(2\alpha + \beta\bar{\eta}_{ij}\left(g^t\right) + (4\gamma - 2\beta)\overset{\triangle}{\eta}_{ij}\left(g^t + ij\right)\right),\end{aligned} \tag{3}$$

[5] As in the connections model (Jackson and Wolinski, 1996), for instance.

with $\bar{\eta}_{ij}\left(g^t\right)$ the number of non-common neighbors of i and j on g^t,[6] and with $\overset{\triangle}{\eta}_{ij}(g^t)$ the number of common neighbors of i and j defined above.[7] The first component of the right-hand side of Equation 3 is related to the per period average joint gain of the research collaboration. The second one captures the net present value of the variation in the flow of network payoffs, due to the new link ij having been added to g^t. All the agents discount time by factor δ. The variation in the per period payoffs is composed of the payoffs obtained thanks to: two new direct relations, $\bar{\eta}_{ij}\left(g^t\right)$ new indirect relations between agents not having a direct link, $4\overset{\triangle}{\eta}_{ij}\left(g^t+ij\right)$ new indirect relations between agents having a direct link, and $2\overset{\triangle}{\eta}_{ij}\left(g^t+ij\right)$ less indirect relations between agents having no direct link on g^t. The following example illustrates how exactly these computations are made.

Example 1 *Let us consider the network* $g=\{ix,jx,iv,iu,iy,yj,js\}$ *depicted in Figure 1, and let us focus on the potential formation of a new link between agents* i *and* j *that does not exist in* g*. It should be noted that here,* $\bar{\eta}_{ij}(g)=3, \overset{\triangle}{\eta}_{ij}(g+ij)=2$*. Thus, according to Equation 3, the new link* ij *would bring to the dyad an expected average net payoff of* $(1-\delta)\left[\Delta(g,ij)\right]=2(1-\delta)\left[\theta_{ij}+\varepsilon^t_{ij}-\zeta c_{ij}\right]+\left[2\alpha+3\beta+2(4\gamma-2\beta)\right]$*. Let us explain the second term of the right-hand side of this equation, which corresponds to the variation in the per period network payoffs. The dyad first enjoys the returns of two new direct connections (*j *with* i *and* i *with* j*), each providing an* α*. Thanks to link* ij*,* i *benefits from the returns of four indirect connections, provided there is a direct link: two that point to* j *(*$\{ix,xj\}$*, and* $\{iy,yj\}$*), one that goes to* x *(*$\{ij,jx\}$*), and one to* y *(*$\{ij,jy\}$*). Simultaneously, two indirect connections, provided there is no direct link, have disappeared (*$\{ix,xj\}$*, and* $\{iy,yj\}$*) on* $g+ij$*. The same occurs for* j*, which explains the multiplication by two.*

2.4 Network evolution and empirical strategy

The relational network emerges gradually from the uniform meeting process exposed above and the willingness of agents to form links. The social network is not assumed to be at equilibrium but in some, possibly transient, state. In our context, new agents enter the population at all periods, and connection costs evolve over time. Therefore, to assume that the network would be at equilibrium would amount to considering that

[6] That is, agents in the direct neighborhood of i (j), but from which the other agent j (i) does not already benefit (at a social distance strictly greater than two).

[7] It should be noted that, by definition: $\bar{\eta}_{ij}\left(g^t\right)+2\overset{\triangle}{\eta}_{ij}(g^t+ij)=\eta_i\left(g^t\right)+\eta_j\left(g^t\right)$, where $\eta_i\left(g^t\right)$ denotes the number of neighbors of agent i in g^t.

agents could rearrange all their collaborations at each period, which would obviously not be consistent.[8]

At each period of time, with the network g^t being in place, and provided that the link between i and j does not already exist, the probability of a dyadic connection being established between the two agents is written $\Pr(g_{ij}^{t+1} = 1|g^t, g_{ij}^t = 0)$. As explained above, at each period, any pair of unconnected agents $i, j \in N^t$ is randomly chosen with a given constant and a non-null probability p and, provided that two agents i and j meet, a link will be formed between them, if $\Delta(g^t, ij) > 0$. We further assume that $\varepsilon_{ij}^t \sim$ Logit, and we denote $F(.)$ its associated cumulative distribution function. The probability of i and j forming a collaboration in period t is thus given by:

$$\Pr\left(ij \in g^{t+1} \left| ij \notin g^t\right.\right) = \Pr(\Delta(g^t, ij) > 0) \times p \propto F(\bar{\Delta}(g^t, ij)), \tag{4}$$

with $\bar{\Delta}(g^t, ij) \equiv \Delta(g^t, ij) - \varepsilon_{ij}^t$. We propose to estimate Equation 4 by relying on the following specification of the incentives to form a bilateral collaboration:

$$\frac{1}{2}\Delta(g^t, ij) = \beta_1 + \beta_2 \overline{\eta}_{ij}\left(g^t\right) + \beta_3 \overset{\triangle}{\eta}_{ij}\left(g^t + ij\right) + \theta_{ij} + \varepsilon_{ij}^t + \beta_4 c_{ij}^t, \tag{5}$$

where θ_{ij} is the time-invariant fixed effect, and ε_{ij}^t the error term. This expression is directly derived from our specification of the bilateral payoffs of a link formation exposed in Equation 3, with $\beta_1 = \frac{\alpha}{(1-\delta)}$, $\beta_2 = \frac{\beta}{2(1-\delta)}$, $\beta_3 = \frac{2\gamma-\beta}{1-\delta}$ and $\beta_4 = -\zeta$. If $2\gamma < \beta$, then $\beta_3 > 0$, and thus the number of triangles should impact positively the incentives for forming links. Though the significance and the sign of β_3 is our primary interest here, we will also be concerned with β_2 being positive and significant, which would provide support for the idea that the collaboration network is a vehicle for knowledge spillovers.

3 Data and variables

Our primary empirical evidence is built upon all European patent applications, of which at least one inventor has declared an address in France, and the priority date of which is between January 1978 and December 2004 included. All non-French inventors of these patents have been excluded. Before describing the co-invention network and the various explanatory variables, we first describe the procedure we developed to disambiguate inventors, a major issue when tackling large network data based on administrative files.

[8] Carayol and Roux (2008) adopt the alternative perspective by studying inert components, assuming that they reach some stable state.

3.1 A Bayesian methodology to disambiguate inventors' names

For each inventor listed in a patent document, her/his name, first name and personal address information are available, but a unique identification is not. This raises a disambiguation issue, or a "name game", according to Trajtenberg et al. (2006), due to the homonymy of inventors and to spelling errors. Most often, such errors should not be neglected, since an accumulation of small identity errors could easily trigger great changes in the network data. For instance, a positive error of homonymy would lead to considering that different persons are the same, thereby mistakenly generating some apparently extremely connected agents who would abusively link different communities. A negative error of homonymy would lead to ignoring the role of bridging agents. As is well known in the literature on networks, many network statistics are very sensitive to such errors. Therefore, the use of the information on patent inventors necessitates the correct identification of individual identities in patent data through some reliable, systematic and reproducible methodology.

Though a growing literature tackling this issue is emerging,[9] a widely accepted standard has not yet been fixed, and a whole range of more or less *ad hoc* techniques can be seen. Any disambiguation procedure needs, in particular, to have a filtering step, in which different observable attributes, already listed in the patent dataset, are used to provide similarity scores to determine whether two homonyms refer to one and the same person. Most of the methodologies used at this stage are not based on clear theoretical grounds, and thus encounter two major drawbacks. First, they arbitrarily assign a given increase in their similarity scores when they record that two homonyms have the same modality for variable. But should, for instance, information about the city of residence contribute more or less to the similarity scores than information about the technological classes? Second, the relative frequencies of each variable modality are not taken into account. Clearly, it is not as informative to observe that two homonyms live in Paris rather than in a small town, and we would like to know what difference this makes exactly.

In line with Carayol and Cassi (2009), we adopt a Bayesian methodology for estimating the probability that two homonyms are the same person, given a series of observables provided by the data. This methodology, further detailed in Appendix A, completely overcomes the two drawbacks encountered by other studies, as stressed above. Out of 133,764 patents considered, we find 262,186 patent×inventor occurences that correspond to an address in France. We use the following list of observable attributes of individuals:

[9] For a review, see Miguélez and Gomez-Miguélez (2011).

name and first name, address (the full string and the extracted name of the city), technological class, patent citation, applicant (at company and group level). Our methodology also makes use of an empirical benchmark of nearly five thousand reliable (positive or negative) matches. We thus know that we were able to reach ninety-eight percent correct inferences out of a linear combination of positive and negative errors in the benchmark. Out of a total initial population of 126,887 agents, we obtain 103,309 French inventors.

3.2 The French co-invention network

Of those 103,309 French inventors, 82,994 invented a patent with at least one other French inventor over the period 1978-2004. In the evolving co-invention network, connection exists if two persons have already invented at least one patent together. Implicitly, we assume that all inventors of a patent are personally acquainted. This assumption, which is standard in the literature on co-authorship networks (see e.g. Newman, 2004; Moody, 2004; Goyal et al., 2006), is even more acceptable in the co-invention context, since co-invented patents (with at least two inventors) mostly involve small teams of collaborators: the average and median numbers of inventors of co-invented patents are respectively 2.8 and 2, with a standard deviation equal to only 1.19. Different assumptions can be made about the duration of a link. As is usually done in the literature (e.g. Singh, 2005), we will here mostly rely upon a five-year backward-moving window. However, we have also computed network data on an alternative ten-year moving window, for the cumulated network. Table 1 provides some basic statistics for each of these three networks in the last year of our sample (2004). As could be expected, the number of connected agents changes in function of the assumption made about link duration. Note that the largest component of the cumulated network represents 50% of the whole population (62% of the connected agents). As a point of comparison, the largest component of scientific co-authorship networks rarely includes less than 70% of the population[10]. Such a discrepancy may be explained by a greater density of the co-authorship networks.[11] One could also argue that technological knowledge may be more fragmented than scientific knowledge, or that the institutional configuration could generate a higher fragmentation of the population of inventors than authors, who evolve in a more open scientific mode of knowledge production. A very interesting statistic for our study is

[10]See, for instance, Newman (2001) where a 5-year window is taken into account, and Barabasi et al. (2002) where the data cover a 8-year period.

[11]It is a well-known property of both random and scale-free networks that increasing network density non-linearly leads to the emergence of a "giant component" tending to encompass almost all the population (Erdos and Renyi, 1960).

the average clustering coefficient. This gives the (averaged among all connected agents) number of triangles to be found in agents' neighborhoods, divided by the number of all the triangles that could be built between these neighbors. We find high values of the average clustering (between 53 and 59%), a result which is very close to those usually found in large social networks.

3.3 Variables

We now present the variables that will be used in the regressions. They include network variables that are reliable, thanks to the disambiguation of inventors' names, geographical and technological distances directly extracted from patent data, and applicant data that rely both on the cleaning of the applicant field of patent data and the match of patent applicants with companies names in mandatory national surveys of companies.

Descriptive statistics on all variables are presented in Table 2. Since the fixed effect approach we use in our econometric estimations deletes all dyads with only null-dependent variables (no link is ever formed), the data are limited to the yearly observations of the dyads that are eventually formed. Moreover, since the variability in the explanatory (cost) variables comes from the observation of at least one previous patent, we only consider the $97,551$ dyads in which an inventor has already invented at least one patent. Each one of these dyads is observed starting from the first year the two concerned agents are considered to be part of the population of inventors,[12] until the link is formed (that year being included). These dyads involve $54,886$ distinct inventors. All in all, there are $407,001$ dyad$\times$year observations.

3.3.1 Network variables

The dependent variable g_{ij}^{t+1} is a dummy, equal to one if the link between the two active agents i and j is formed in year $t+1$, and zero otherwise. This concerns the period 1983-2004. For each year t during the period 1982-2003,[13] we calculated the two explaining variables of major interest on the 5-year window network g^t, namely the number of non-common neighbors $\overline{\eta}_{ij}\left(g^t\right)$ and the number of common neighbors $\overset{\triangle}{\eta}_{ij}\left(g^t\right)$.[14] We

[12] In order to build the unbalanced panel data set, we had to formulate some assumption about the entry of inventors in the population. An inventor is considered as active three years before his first patent application year.

[13] There is a one-year lag for all explanatory variables compared with the dependent variable, as suggested by the theoretical model.

[14] The five-year window was used, since it is the one most commonly employed in similar network empirical studies. However, a larger ten-year window will also be used to build the right-hand side

also computed a series of network controls (noted net_controls$_{ij}^t$) that concerns the time-variant network attributes of inventors of each focal dyad: the average number of patents per year of the two agents, the rate of difference (absolute value of the difference divided by the mean) of agents' degree and the rate of difference in their average number of patents.

3.3.2 Geographical and technological distances

As patent data mention the personal addresses of inventors, we were able to locate inventors in the metropolitan French area by matching the post codes mentioned in their addresses with their corresponding latitude and longitude coordinates.[15] By means of name disambiguation, we were able to identify inventors who changed location: as many as 11,970 of the connected inventors declared at least two different addresses. Most geographically mobile inventors remain in the same area: nearly 79% (86%) of mobile inventors have a maximal distance between their different locations of less than 20 km (50 km).

The Euclidean geographical distance can be computed for any pair of addresses, given their coordinates (latitude and longitude). Since some agents change location, more than one distance may be associated with a pair of connected agents: some pairs of agents invent together on several occasions, while at least one of the two changes addresses in the meanwhile. If we restrict ourselves to our data set of dyads, matters are much simpler. Overall, we have identified more than 145,000 distances (in kilometers) between co-inventors. If we just consider the distance for the year in which the link is formed, we observe that the distribution of connections, according to the geographic distance between agents, is very skewed. More than 63% of the connections are achieved between inventors that live at less than 50 km from each other, while fewer than 6.2% of the connections are formed between agents who live at more than 550 km from each other. Figure 2 presents the histogram of the geographic distance between inventors, restricted to the dyadic observations of the year when the link is formed. The variable geo$_{ij}^t$, which is equal to twice the geographic distance (as suggested by the theoretical model) between agents i and j at period t, accounts for some of the connection costs.

For each pair of inventors, in each year, we have also computed the technological distance. This has been defined using the similarity measurement proposed by Jaffe

network variables in the robustness check analyses. By doing so, we lose five years of observation and the covered period is 1988-2004 for the dependent variable and 1987-2003 for the explanatory variables.

[15] Those coordinates were kindly provided to us by the IGN (Institut Géographique National).

(1988), i.e. un-centered correlation measurement of two inventors' distribution vector of patents over 30 technological IPC classes defined by OST (2010). It is given by :

$$\text{jaffe}_{ij}^{t} = 1 - \frac{\sum_k n_i^{k,t} n_j^{k,t}}{\left(\left(\sum_k \left(n_i^{k,t}\right)^2\right)\sum_k \left(n_j^{k,t}\right)^2\right)^{1/2}},$$

with $n_i^{k,t}$ the number of patents i invented in technological class k before year t. Alternative measurements of technological distance, such as the Euclidean or the Manhattan distance, have also been computed (see robustness check in the following section).

3.3.3 Applicants

The association of inventors to applicants on a yearly basis has been based on the two following principles: *i*) the inventor is associated with her/his first applicant and permanently if she/he does not switch to another applicant; *ii*) if the inventor switches to a new applicant, she/he is associated with that new applicant from the year of the application of the new patent.

To account for the institutional costs of collaboration, a dummy variable, app_{ij}^{t} was created: it is equal to unity if i and j have ever been associated with the same applicant. We also identify public research institutions (universities and other public bodies) among all the applicants of our database. This variable is of interest for us since we hypothesize that, when inventors are in the academic sphere, they may follow different behavioral patterns, somewhat reducing the perceived costs of collaborations. The dummy variable acad_{ij}^{t} captures this: it is equal to unity if one of the two agents has already invented a patent for which the applicant is a public research institution.

A final step in the enrichment of our data proceeded as follows. We matched the patent dataset with the French R&D surveys conducted annually by the French Ministry of Research, using the name and location of applicants-companies as matching key. These surveys, exhaustive for all the companies employing at least one full-time researcher (whatever their size), provided us with annual data on companies' R&D internal and external expenditures, number of researchers, as well as more general information, such as total number of employees. We then deflated R&D internal and external expenditures by a national investment price index. Information concerning the applicants associated with the inventors has been used to build dyadic variables (denoted $\text{app\_controls}_{ij}^{t}$), both by summing for the two agents and by calculating the rate of difference (the absolute values of the difference within the dyad divided by the sum). Though the surveys on

companies used are exceptionally extensive, it was not possible to obtain this information for every applicant in the dyad, or for every year. This results from a sharp decrease in the number of observations available: approximately 130 thousand observations for all dyadic sums, and approximately ninety-three thousand observations for all the dyadic sums and all the differences within the dyad.

4 Estimations and results

The direct empirical counterpart of the theoretical model described in Equations 4 and 5, is given in the following equation:

$$\begin{aligned}\Pr\left(ij \in g^{t+1} \left| ij \notin g^t\right.\right) = F\Big(&\beta_1 + \beta_2 \overline{\eta}_{ij}\left(g^t\right)) + \beta_3 \overset{\triangle}{\eta}_{ij}\left(g^t\right) \\ &+ \beta_4^1 \text{geo}_{ij}^t + \beta_4^2 \text{jaffe}_{ij}^t + \beta_4^3 \text{app}_{ij}^t + \beta_4^4 \text{acad}_{ij}^t \\ &+ \beta_5 \text{app\_controls}_{ij}^t + \beta_6 \text{net\_controls}_{ij}^t + \theta_{ij}\Big).\end{aligned} \tag{6}$$

The significance and sign of parameters β_2 and, in particular, β_3 are our main interest, provided that we properly control for the benefits and costs of link formation. To do so, we have introduced a fixed effect θ_{ij} that accounts for the time-invariant matching quality of the two agents which, we hypothesize, corresponds to the returns of the research collaboration between i and j captured by these two agents. Controlling for time-invariant fixed effect is, however, not sufficient since there may be some time-variant factors that affect the probability of collaboration. Four variables introduced above account for the costs, geographic distance (geo_{ij}^t) between agents, technological distance (Jaffe_{ij}^t), having already invented for the same applicant (app_{ij}^t), which we interpret as being associated with the same institution, and having already invented for an academic institution (acad_{ij}^t) that may generate more collaborative research patterns.

Lastly, we include two series of controls: $\text{app\_controls}_{ij}^t$ and $\text{net\_controls}_{ij}^t$. The former refers to the research capacity of the applicant(s) associated with the inventors of the dyad. This series includes the total internal and external R&D expenditures, the number of researchers and the number of employees, as well as the difference rates of these three variables between the two agents of the dyad. The second series of controls concerns inventors' time-variant network attributes: the average number of patents per year of the two agents, the rate of difference of agents' degree, and the rate of difference in their average number of patents. These variables allow us to control, in particular, for the time-varying individual propensities to patent (which may affect the probability

p that is assumed to be uniform across dyads in the model).[16]

Our estimation strategy immediately raises a first issue. Allisson and Christakis (2006) recently showed that the estimation of such a fixed effect logit model leads to spurious estimates when the explanatory variables are trended. Therefore, as suggested in Fafchamps et al. (2010), all quantitative explanatory variables are first detrended by assuming a linear trend (we later relax this assumption by assuming other forms of trend).

Table 3 synthesizes our main results. We find first that, as expected, the number of neighbors that the two agents do not have in common always positively and significantly impacts the probability of connection. This supports the idea that agents benefit from indirect connections (at least at distance two), and that a natural justification for this is network-based knowledge spillovers. Our second and most important result is the following. When we do not properly control for the costs of network formation (no control of costs in Column 1) or only account for technological and geographic distance (in Column 2), the number of common neighbors is positively and significantly associated with the probability to create a link. This result seems to indicate that agents like closing triangles. However, when we extensively control for the connection costs, in particular when we control for all applicant variables (Columns 3 to 5), it appears clearly that this effect disappears and, even, that the number of common neighbors now significantly decreases the probability of becoming connected.

We now raise a second estimation issue: our observations may not be independently distributed, since all observations corresponding to dyads related to the same agent may be correlated. The omission of correlation between observations may lead to an incorrect assessment of the significance of some coefficients. We are in particular worried about a possible overestimation of coefficient significance, but the contrary may also occur. One simple way to correct for such a bias is clustering. However, regressions clustered on several variables are still not available, to the best of our knowledge. Of course, it would be possible to cluster on the dyads, but this is not what we would like to do, preferring instead to cluster on each member of the dyad. To do so, in all the estimations presented, we cluster observations on the identity of the agents "on the left" of the dyad (included in all regressions, among which those in Table 3, discussed above). In order to control for a potential correlation of dyadic observations related to the same agent listed "on the right" of the dyads, we also run all estimations on restricted samples in which only one dyad of each agent "on the right" is randomly selected, and all the

[16] Fafchamps et al. (2010) use similar network controls.

dyads involving one agent "on the right" who is also present on the list of agents "on the left", are discarded. Since this strongly reduces the size of the dataset (down to sixty-four thousand observations), such a restriction is likely to seriously underestimate coefficient significance. However, the estimation results on the restricted sample (see Table 4) remain qualitatively the same as those obtained on the full sample as regards the impact of the number of non-common neighbors. Only the impact of the number of common neighbors becomes non-significant in the last two regressions (Columns 4 and 5), which more fully control for the costs of link formation. It should be noted that these coefficients remain negative. However, significance here is much likely to be lost, due to the massive reduction of the number of observations (down to fewer than 20 thousand dyads, which represents more than 20 times fewer observations than in the complete sample).

The two main results appear to be globally robust to a long list of robustness checks. All the supplementary regressions we discuss below are reported in Appendix B. The results are, in particular, robust to different assumptions about the trend of the quantitative explaining variables: the results do not change when we assume either a logarithmic or a quadratic trend (Tables B1 and B2). When a larger time-window (ten years) is retained to build the right-hand side networks variables (Table B3), the results on the impact of the number of common neighbors remain the same. However, it appears then that the number of non-common neighbors negatively affects the probability of connecting. This could be explained by the lower influence of (older) non-common neighbors on the probability of connecting. It could also be due to the assumption that agents never retire or die: older agents are those most likely to have larger neighborhoods, while they are also more likely to be no longer active. This last remark could be interpreted as a reinforcement of our positive result obtained for the five-year moving window networks. The divergence with the main regressions is, however, limited since this coefficient remains positive in the most complete specification (Column 6).

We were also concerned by a possible bias due to the simultaneous arrival of all connections in the research teams inventing a patent. Closed groups could have been formed before the patent is filed, and all closed links may have arrived simultaneously. We therefore ran regressions on those dyads that become connected via a two-inventor research team, which reduces the sample to approximately sixty-two thousand observations. The results (Table B4) show that the coefficients of the number of common neighbors and of the numbers of non-common neighbors tend to lose significance in the complete models, but also that the number of non-common neighbors remains significant when we restrict the sample to one unclustered dyad per agent (Table B5). Other, more

minor and unreported robustness checks (the use of other forms of indexes of technological distance, various lists of controls, non-linear effects), show that the main results remain unchanged.[17]

The impacts of the cost variables are also interesting in themselves. We find that all the cost variables are always significant, and that the coefficients always have the expected signs: geographic distance and technological distance significantly decrease the probability of becoming connected. Having previously had one common applicant strongly and positively affects that probability. Interestingly, if one inventor has already invented for an academic applicant in the dyad, the probability of forming a connection is higher, which highlights the role of academics in creating connections among inventors.

5 Conclusion

The main result of our paper is that inventors do not like closing triangles *per se*. Some econometric specifications tend, even, to support the idea that they dislike doing so. This result is obtained thanks to the availability of very precise data that account for the time-varying costs of network formation, and the possibility of controlling for the dyadic time-invariant fixed effect in our longitudinal framework. This result may be limited to our particular context, in which the institutions (mostly companies) have incentives to create the conditions for the enforcement of cooperative behaviors between their employees. Even though different conclusions may be obtained, for instance, in friendship networks or in risk-sharing networks, our results call, however, for properly controlling for the costs or network formation in any context, before assessing whether agents like triadic closure.

If we now examine the consequences of our findings for innovation policy, we should first stress that our results are consistent with the idea that professional social networks of inventors are conducive to positive externalities due, for example, to knowledge spillovers. Since, in addition, the transfers between agents are limited,[18] inventor networks are likely to be less than optimally dense. This provides a rationale for innovation policy to focus on the community of inventors (rather than on companies) and to sustain the formation of connections among inventors. To be more specific, we also highlight the finding that the costs of link formation across institutions, in particular, are higher. Since these links mediate knowledge spillovers, their social value is likely to be very high, because they are non-redundant, much higher than their private returns for the directly concerned

[17] All results can be obtained from the authors.

[18] In particular, between agents that are not connected; see Bloch and Jackson (2007) on this point.

agents. It is these links that need to be targeted by innovation policy. Two factors could contribute significantly to building these bridging links: the number of academics who are likely to collaborate with different companies, and an active labor market of mobile inventors.

References

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

Ali, S.N., Miller, D.A., 2012, Cooperation and collective enforcement in networked societies, mimeo.

Allison, P.D., Christakis, N.A., 2006, Fixed effects methods for the analysis of non-repeated events. Sociological Methodology 36(1), 155-172.

Baum, J.A.C., Shipilov, A.V., Rowley, T.J., 2003,Where do small worlds come from? Industrial and Corporate Change 12(4), 697-725.

Baum, J.A.C., Calabrese, T., Silverman, B.S., 2000, Don't go it alone: Alliance networks and startups' performance in Canadian biotechnology, 1991-97. Strategic Management Journal 21, 267-294.

Barabasi, A.L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., Vicsek, T., 2002, Evolution of the social network of scientific collaborations. Physica A 311, 590-614.

Bearman, P.S., Moody, J., 2004, Suicide and friendships among American adolescents. American Journal of Public Health 29(4), 89-95

Bloch, F., Jackson, M.O., 2007, The formation of networks with transfers among players. Journal of Economic Theory 133, 83-110.

Breschi, S., Lissoni, F., 2006, Cross-firm inventors and social networks: Localised knowledge spillovers revisited. Annales d'Economie et de Statistique 79-80, 189-209.

Burt, R.S., 2004, Structural holes and good ideas. American Journal of Sociology 110, 349–399.

Burt, R.S., 2000, The network structure of social capital. In Staw B.M., Sutton R.I., Research in Organizational Behavior. Amsterdam; London and New York: Elsevier Science JAI, 345-423

Burt, R.S., 1997, The contingent value of social capital. Administrative Science Quarterly 42(2), 339–365.

Burt, R.S., 1992, Structural holes. Cambridge, MA: Harvard University Press

Buskens, V., 2003, Trust in triads: Effects of exit, control, and learning. Games and Economic Behavior 42, 235-252.

Carayol, N., Cassi, L., 2009, Who's who in patents. A Bayesian approach. GREThA working paper 2009-07.

Carayol, N., Roux, P., 2009, Knowledge flows and the geography of networks. A strategic model of small worlds formation. Journal of Economic Behavior and Organization 71, 414-427.

Carayol, N., Roux, P., 2008, The strategic formation of inter-individual collaboration networks. Evidence from co-invention patterns. Annales d'Economie et de Statistique 89/90, 275-302.

Coleman, J.S., 1988, Social capital in the creation of human capital. American Journal of Sociology 94, S95-S120.

Davis, G.F., Yoo, M., Baker, W.E., 2003, The small world of the American corporate elite, 1982-2001. Strategic Organization 1(3), 301-326.

Erdos, P., Renyi, A., 1960, On the evolution of random graphs. Publications of the Mathematical Institute of the Hungarian Academy of Sciences 5, 290-297.

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

Fleming, L., King, C., Juda, A., 2007, Small worlds and regional innovation. Organization Science 8(2), 938-954.

Goyal, S., Moraga, J.L., Van Der Leij, M., 2006, Economics: An emerging small world. Journal of Political Economy 114, 403-412.

Granovetter, M.S., 1985, Economic action and social structure: The problem of embeddedness. American Journal of Sociology 91(3), 481-510.

Granovetter, M.S., 1974, Getting a job: A study of contacts and careers. Cambridge, Mass., Harvard University Press.

Granovetter, M.S., 1973, The strength of weak ties. American Journal of Sociology 78, 1360-1380.

Griliches, Z., 1992, The Search for R&D spillovers. NBER Working Papers 3768, National Bureau of Economic Research.

Heider, F., 1958, The psychology of interpersonal relations, New York: Wiley.

Jackson, M.O., 2009, Social and economic networks. Princeton University Press.

Jackson, M.O., Rodriguez-Barraquer T., Tan X., 2012, Social capital and social quilts: Network patterns of favor exchange. American Economic Review 102(5), 1857-1897.

Jackson, M.O., Wolinsly, A., 1996, A strategic model of social and economic networks. Journal of Economic Theory 71, 44-74.

Jaffe, A.B., 1988, Demand and supply influences in R&D intensity and productivity

growth. Review of Economics Statistics 70(3), 431-437.

Jaffe, A.B., 1986, Technological opportunity and spillovers of R&D: Evidence from firms' patents profits and market value. American Economic Review 76(5), 984-1001.

Karlan, D., Mobius, M., Rosenblat, T., Szeidl, A., 2009, Trust and social collateral. The Quarterly Journal of Economics 124, 1307-1361.

Kogut, B., Walker, G., 2001, The small world of Germany and the durability of national networks. American Sociological Review 66, 317-335.

Li, G-C., Lai, R., D'Amour, A., Doolin, D.M, Sun, Y., Torvik, V.I. , Yu, A.Z., Fleming, L., 2014, Disambiguation and co-authorship networks of the U.S. patent inventor database (1975-2010). Research Policy 43(6), 941-955.

Miguélez, E., Gomez-Miguélez, I., 2011, Singling out individual inventors from patent data. IREARP working paper 2011/05.

Moody J., 2004, The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. American Sociological Review 69(2), 213-238.

Newcomb, T.M., 1961, The acquaintance process, New York: Holt, Rinehart and Winston.

Newman, M.E.J., 2004, Who is the best connected scientist? A study of scientific coauthorship networks. In: Ben-Naim E., Frauenfelder H., Toroczkai, Z. (Eds.). Complex Networks, Springer, Berlin, 337-370.

Newman, M.E.J., 2001, The structure of scientific collaborations. Proceedings of the National Acadademy of Science USA 98, 404-409.

OST, 2010, Indicateurs de Sciences et de Technologies. Rapport de l'Observatoire des Sciences et des Techniques, Paris, Economica.

Pezzoni, M., Lissoni, F., Tarasconi, G., 2012. How to kill inventors: Testing the Massacrator© algorithm for inventor disambiguation. Cahiers du GREThA 2012-29.

Raffo, J., Lhuillery, S., 2009. How to play the "Names Game": Patent retrieval comparing different heuristics. Research Policy 38, 1617-1627.

Rapoport, A., 1953, Spread of information through a population with socio-structural bias: I. Assumption of transitivity. Bulletin of Mathematical Biophysics 15(4), 523-533.

Raub, W., Weesie, J., 1990, Reputation and efficiency in social interactions: An example of network effects. American Journal of Sociology 96(3), 626-654.

Reagans, R., McEvily B., 2003, Network structure and knowledge transfer: The effects of cohesion and range. Administrative Science Quarterly 48(2), 240-267.

Simmel, G., 1922 [1955], Conflict and the Web of group affiliations, translated and edited by Kurt Wolff, Glencoe, IL: Free Press.

Singh, J., 2005, Collaborative networks as determinants of knowledge diffusion patterns. Management Science 51(5), 756-770.

Trajtenberg, M., Shiff, G., Melamed, R., 2006, The "Names Game": Harnessing inventors' patent data for economic research. NBER WP 12479.

Uzzi, B., Spiro J., 2005, Collaboration and creativity: The small world problem. American Journal of Sociology 111(2), 447-504

Watts, D.J., 1999, Networks, dynamics, and the small world phenomenon. American Journal of Sociology 105(2), 493-527

Watts, D.J., Strogatz, S.H., 1998, Collective dynamics of 'small worlds' networks. Nature 393, 440-442.

Zander, U., Kogut, B., 1995, Knowledge and the speed of the transfer and imitation of organisational capabilities: An empirical test. Organisation Science 6(1), 76-92.

Tables and Figures

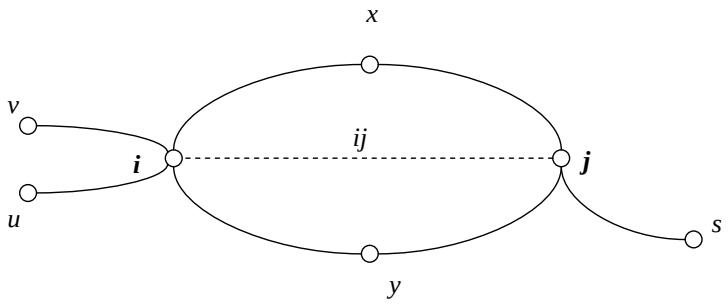

Figure 1: Example 1.

Table 1: Descriptive statistics of the 2004 co-invention networks built with different assumptions about the duration of links.

	cumulated	10-year window	5-year window
# isolated agents	$20,315$	$53,555$	$73,063$
# connected agents	$82,994$	$49,754$	$30,246$
# links	$161,724$	$92,756$	$51,763$
# of components	$10,198$	$7,104$	$5,586$
largest component size	$51,761$	$24,744$	$7,357$
2nd largest component size	82	153	130
av. degree (all agents)	3.13	1.80	1.00
av. degree (connected agent)	3.89	3.74	3.42
av. clustering	0.53	0.57	0.59

Table 2: Descriptive statistics of the dyads that are formed at some point, observed from the year the two agents are considered as active, until the year the link is formed (included).

	Variable	**Mean**	**Std. Dev.**	**N**
network	common	0.11	0.73	407,001
variables	non-common	4.73	6.79	407,001
	geo distance	265.91	395.83	407,001
cost	Jaffe tech distance	0.49	0.46	407,001
variables	public research	0.11	0.32	407,001
	common applicant	0.16	0.37	407,001
	cpny researchers	819.66	1635.14	154,535
	RD dpt size	1912.3	3355.97	154,535
applicant controls	internal RD	216,785.65	588,788.08	154,535
(sum in the dyad)	external RD	52,692.32	139,987.39	147,743
	cpny size	19,212.2	55,975.64	154,535
	cpny turnover	1,994,384.71	8,439,402	154,535
	diff. cpny researchers	0.18	0.35	129,133
	diff. RD dpt size	0.18	0.36	129,133
applicant controls	diff. internal RD	0.18	0.35	129,133
(difference in the dyad)	diff. external RD	0.21	0.38	111,472
	diff. cpny size	0.18	0.35	129,120
	diff. cpny turnover	0.2	0.37	119,898
	av. productivity	0.38	0.44	407,001
net controls	diff. in degree	0.64	0.45	407,001
	diff. in productivity	0.71	0.39	407,001

Figure 2: Distribution of first connections according to the geographic distance (in km) between the connected agents.

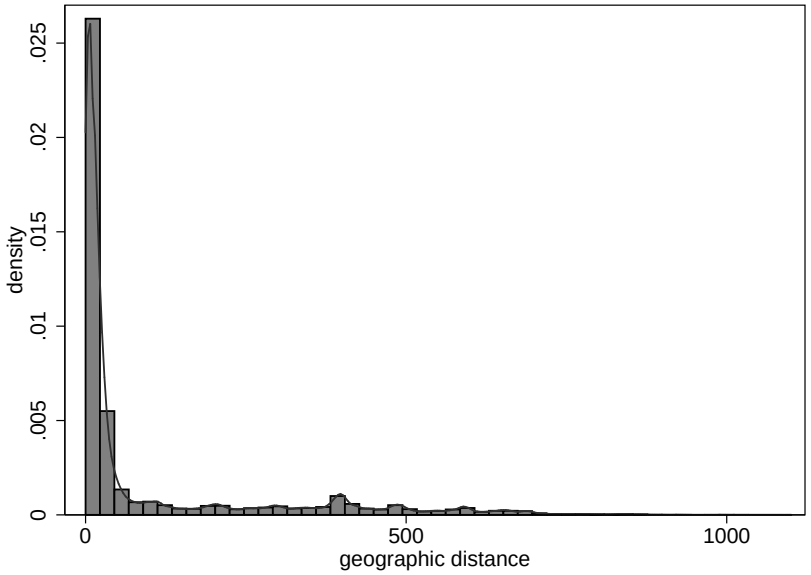

Table 3: Conditional logit on the occurrence of the first connection, All sample, five-year window network, linear detrending.

	1	2	3	4	5
non-common	0.0321***	0.0328***	0.0276***	0.0313***	0.0220**
	(7.65)	(7.83)	(4.73)	(3.74)	(2.66)
common	0.199***	0.204***	-0.529***	-0.255***	-0.138**
	(12.50)	(12.75)	(-6.36)	(-3.77)	(-2.62)
geo distance		-0.00132***	-0.00143***	-0.000621***	-0.000764***
		(-33.90)	(-33.41)	(-8.25)	(-8.90)
Jaffe tech distance		-0.339***	-0.258***	-0.507***	-0.580***
		(-6.14)	(-3.74)	(-4.41)	(-4.76)
public research			23.13***	21.83***	19.30***
			(129.55)	(56.44)	(44.56)
common applicant			31.71***	24.75***	22.75***
			(11.90)	(178.41)	(266.21)
network controls	yes	yes	yes	yes	yes
applicant controls sum	no	no	no	yes	yes
applicant controls diff.	no	no	no	no	yes
observations	407,001	407,001	407,001	129,924	93,215

t statistics in parentheses

$^{*}p < 0.05,^{**}p < 0.01,^{***}p < 0.001$

Clustered standard errors in parentheses on the identity of the agent "on the left" of the dyad.

Table 4: Conditional logit on the occurrence of the first connection, sample restricted to one randomly chosen dyad per inventor "on the right", five-year window network, linear detrending.

	1	2	3	4	5
non-common	0.123***	0.125***	0.105***	0.162***	0.150***
	(10.88)	(10.93)	(8.47)	(5.28)	(6.07)
common	0.306***	0.311***	-26.51**	-4.209	-0.879
	(4.34)	(4.46)	(-3.14)	(-0.26)	(-1.57)
geo distance		-0.00148***	-0.00131***	-0.000726***	-0.000908***
		(-18.96)	(-12.32)	(-4.42)	(-5.28)
Jaffe tech distance		-0.455***	-0.506***	-0.730*	-0.783*
		(-3.44)	(-3.89)	(-2.32)	(-2.48)
public research			19.94***	20.74***	20.12***
			(56.10)	(15.89)	(24.34)
common applicant			242.2***	35.83	22.63***
			(3.61)	(0.57)	(21.05)
network controls	yes	yes	yes	yes	yes
applicant controls sum	no	no	no	yes	yes
applicant controls diff.	no	no	no	no	yes
observations	63,820	63,820	63,820	19,907	12,856

t statistics in parentheses

$^{*}p < 0.05,^{**} p < 0.01,^{***} p < 0.001$

Clustered standard errors in parentheses on the identity of the agent "on the left" of the dyad.

Appendix A: A Bayesian methodology to disambiguate inventors' names.

Introduction

In this Appendix, we present the basic features of a Bayesian methodology for estimating the probability that two *ex ante* different identities correspond to the same person, given a series of observables provided by the data.[19] This methodology has been presented much more extensively in a technical note wroten by two of us (Carayol and Cassi, 2009). In the second section of this appendix, we show how this methodology applies to the disambiguation of patent inventors. We also briefly present the results we obtain on an actualized data set of French inventors. The disambiguated data produced are basic inputs for the article to which this Appendix is attached (Carayol et al., 2014). Other methodologies have been developed and applied on patent data in recent works, such as Trajtenberg et al (2006), Pezzoni et al. (2014) or Li et al. (2014).

Methodology

According to Raffo and Lhuillery (2009), any procedure of disambiguation should be performed in three stages *i*) a parsing stage, finalized in the standardization and cleaning of different data set fields; *ii*) a matching stage, where different algorithms could be used to group homonyms; *iii*) a filtering stage, where different sets of information (i.e. observable attributes already listed in patent data sets such as, for instance, technological class) are used to give a similarity score in order to determine whether homonyms refer to the same person. If the two first steps are essentially technical, the third one requires non-trivial methodological issues to be solved. Here, we focus on this third step, since the first two have already been treated. Basically, it consists in establishing criteria for assigning similarity scores between homonyms.

Let us first consider a list I of *ex ante* agents i defined in the most disaggregated way possible. Each *ex ante* agent i is characterized by a series of K variables[20] labeled X^k, with $k = 1, ..., K$. The main goal of the methodology proposed here is to provide an estimation of the probability that any *ex ante* agent i is the same person as some other *ex ante* agent j. In short, we are in search of the partition $\pi = \{C_1, ..., C_m\}$ of I,

[19] Though this methodology has been developed for inventors in patent data, it can be applied to other similarly structured data.

[20] See Table A1 below to have the list of variables we use here.

the m elements of which should correspond to the correct *ex post* identities. We note $\{i,j\} \subset C_h, \forall h = 1,...,m$ by writing "$i = j$".

In order to assess the probability of that event, we must rely on observables of agents i and j, that is the observed realizations x_i^k and x_j^k of random variables X_i^k and X_j^k for all $k = 1,...,K$ and their respective frequencies of occurrence. All these variables are independent and $\forall k = 1,...,K; \forall i,j$, X_i^k and X_j^k have the same support. Without loss of generality, let us assume that we observe $x_i^k = x_j^k$, for all $k = 1,...,\bar{k}-1$ and $x_i^{k'} \neq x_j^{k'}$ for all $k' = \bar{k},...,K$. One may think of j as an identity which first appears in the data, and then a new identity i appears and one wants to check whether i and j identities correspond to the same person. We have some information on j and i from which we can use. In short, we would like to estimate the following conditional probability:

$$\Pr\left(i = j \left| X_i^k = x_j^k, \forall k = 1,...,\bar{k}-1 \text{ and } X_i^{k'} \neq x_j^{k'}, \forall k' = \bar{k},...,K \right.\right). \tag{7}$$

In principle, it could be possible to apply Bayes' rule to calculate (7). However, it is not possible to compute $\Pr(i = j)$, and thus it is not possible to compute the conditional probability using Bayes' rule. One way to avoiding this difficulty is to focus on the similarity score $\Delta(i,j)$, defined as follows. It is the probability that $i = j$, knowing that indeed $X_i^k = x_j^k, \forall k = 1,...,\bar{k}-1$, divided by the probability that $i = j$, knowing that $X_i^{k'} \neq x_j^{k'}, \forall k' = 1,...,\bar{k}-1$, other things remaining the same. This differentiates out $\Pr(i = j)$ and it can be shown that the similarity score is equal to:

$$\Delta(i,j) = \Pi_{k=1,...\bar{k}-1} \quad \frac{\left(1-\varepsilon^k\right)}{\varepsilon^k} \times \Omega^k(i,j), \tag{8}$$

where $\varepsilon^k \equiv \Pr\left(X_i^k \neq x_j^k \,|i = j\right)$ is the probability that any individual changes k^{th}, observable between two invention occurrences, and where $\Omega^k(i,j) \equiv \left(1 - Pr(X_i^k = x_j^k)\right) \Big/ Pr(X_i^k = x_j^k)$, that is the probability that the two *ex ante* agents i and j have a different k^{th} observable divided by the reverse (irrespective of the fact that they are or are not the same persons *ex post*). The latter term accounts for the frequency of occurrence of the observables (x_j^k through $\Omega^k(i,j)$ in Equation (8)). As will be shown in the next section, these two probabilities ε^k and Ω^k can be estimated iteratively.

At this point, let us assume that we know the relevant threshold value $\bar{\Delta}$ for the similarity score below which two *ex ante* agents should be considered as different agents, and above which they should be considered as being the same person.[21] Then, a

[21] We show in the next section how we make use of a benchmark sample to compute this threshold.

transitivity issue arises. For instance, consider three *ex ante* agents z, w and h and $\Delta(h,w) < \bar{\Delta} < \Delta(z,h) < \Delta(z,w)$. In this situation, *ex ante* agents z and w will *ex post* be considered as referring to the same person. The same applies to z and h. If these two statements hold true, h and w should also be the same person *ex post* by transitivity, even though their similarity score is below the threshold value. We thus need to modify the values of $\Delta(h,w)$ so as to take into account the transitivity of identities. To do so, an algorithm is proposed in order to modify the values of $\Delta(i,j)$.

Algorithm

For all considered pairs of distinct ex ante agents i and j, we apply:

$$\Delta(i,j) \hookleftarrow \max\left(\Delta(i,j); \max_{k \in I \setminus \{i,j\}} \min(\Delta(i,h); \Delta(j,h))\right)$$

recursively until one can not find any triplet of distinct ex ante agents $h,i,j \in I$, such that:

$$\Delta(i,j) < \min\left(\Delta(i,h); \Delta(j,h)\right).$$

Data, estimation and results

Our empirical evidence is built upon all European Patent Applications for which at least one inventor has declared an address in France, with a patent priority date between January 1978 and December 2005. All non-French inventors of these patents have been deleted. The data set counts $136,285$ patents and $266,724$ inventor$\times$patent occurrences. At this stage, the total number of *ex ante* agents corresponds to all the inventor$\times$patent occurrences that can be observed (for instance Pierre_Dupont/Patent_X; Pierre_Dupont/ Patent_Y; Olivier_Dupuy/Patent_Y and so on). This represents our list of *ex-ante* inventors, *I.* The variables used for computing similarity scores are presented in Table 1.

Table A.1: The variables used to build the similarity scores.

Variables
X_1 : name & first name
X_2 : assignee
X_3 : city
X_4 : IPC (6 digits)
X_5 : citation link

The name, first name and full address information are initially used to obtain a starting partition of agents noted π^0. Since full address information is used, we certainly minimize incorrect aggregations.[22] This partitioning generates an initial evolution of the set of agents, reduced to 126,887 inventors. This evolution allows us to compute initial conditional probabilities ε^k. However, the ε^k are underestimated here since the identities are not yet sufficiently aggregated, and we thus encounter the risk of abusive aggregations of agents. Therefore, we propose to process identities recursively, which allows us to progressively determine both the identities and the ε^k. The first similarity scores are computed for the 1,074,946 couples of agents, taken from the previous step, with the same and first name.[23] A precautionary conservative rule is at this step arbitrarily adopted: a high value is given to the threshold $\bar{\Delta}$ which defines, after having applied the transitivity algorithm, a new partition π^1. Then, at each stage $t \geq 2$, the partition obtained from the previous period π^t is considered, and new conditional probabilities ε^k, new similarity scores and a new threshold $\bar{\Delta}$ are computed. All pairs of agents whose similarity score is above the threshold are aggregated within the elements of π^{t+1}. That partition defines a new population of agents taken as an input in the next iteration. This process is repeated until it converges to an equilibrium partition, π^*, which will constitute the final set of inventors.

Fixing the value of the threshold, $\bar{\Delta}$, is obviously a key issue, which deserves careful attention. In order to determine this threshold, we rely on a benchmark data set. A list of French faculty members was matched with the patent data set on the basis of the name and first name of their inventors. Checks on the internet and phone calls to the faculty

[22]The full string, reporting the city and street address, is considered. The probability of two different persons having exactly the same name and first name have the same address (i.e. living in the same building) can reasonably be assumed to be equal to zero. However, it may happen that the company address is reported as the inventor's personal address. Such cases were checked in the data and treated separately.

[23]Without relying on the location data at this stage, because addresses were used in defining the identities so that the probability of moving is here null by assumption.

members were made in order to verify that they are the inventors of patents when their first name and name are mentioned therein. In all, reliable information was collected on 445 French scholars.[24] Their positive and negative declarations have been transformed into assertions on the fact that an *ex ante* agent i and another agent j who have the same name and first name refer to the same person. In all, we have $4,989$ assertions, $4,567$ of which are positive and 422 are negative. This sample of positive and negative couples of agents identities is used as a reliable benchmark to select the appropriate value of the threshold in the interim stages. For each threshold value chosen, the share of positive errors ϵ_1 and the share of negative errors ϵ_2 (in the benchmark) are computed, as well as any linear combinations of these two values: $\phi(\theta) = \theta\epsilon_1 + (1-\theta)\epsilon_2$, with $\theta \in [0,1]$, which accounts for any given weighting schemes of the two types of errors. A threshold that would minimize $\phi(\theta)$ for some θ is noted $\bar{\Delta}(\theta)$. On our data set, it appears that fixing the threshold equal to $\exp(12.49)$ minimizes $\phi(\theta)$ for a wide range of θ, between 0.09 and 0.64. Our preferred value is $\theta = .1$ because it avoids abusive aggregation of agents.

Finally, the algorithm converges after four iterations towards a final population of 105,086 French inventors. Restricting ourselves to the period 1978-2004, we have 103,309 inventors.[25] The benchmark can also be used to assess the quality of the terminal results. If the most appropriate weighting scheme is $\theta = .1$, the weighted share of errors obtained is $\phi(.1) = 1.81\%$, which remains very low.

[24] We are indebted to the KEINS project and BETA at the University of Strasbourg for kindly allowing us to use these data.

[25] It should be worth noting that, in order to solve the issue of homonomy betwen inventors, we make use of all the data available to us (i.e. 1978-2005). Nevertheless, since the data for the last year (2005) is not complete, we exclude it from our analysis of the co-invention network in the article.

Appendix B: Robustness check regressions for the article: "Unintended triadic closure in social networks.

Table B.1: Conditional logit on the occurence of the first connection, all sample, five-year window network, log detrending.

	1	2	3	4	5
non-common	0.0322***	0.0329***	0.0277***	0.0314***	0.0222**
	(7.68)	(7.85)	(4.75)	(3.76)	(2.67)
common	0.199***	0.204***	-0.527***	-0.254***	-0.137**
	(12.54)	(12.78)	(-6.36)	(-3.76)	(-2.61)
geo distance		-0.00132***	-0.00143***	-0.000623***	-0.000765***
		(-33.92)	(-33.43)	(-8.27)	(-8.92)
Jaffe tech distance		-0.340***	-0.260***	-0.509***	-0.583***
		(-6.17)	(-3.76)	(-4.43)	(-4.78)
public research			22.63***	21.83***	19.29***
			(126.78)	(56.55)	(44.59)
common applicant			31.02***	24.75***	22.75***
			(11.70)	(179.34)	(266.78)
network controls	yes	yes	yes	yes	yes
applicant controls sum	no	no	no	yes	yes
applicant controls dif	no	no	no	no	yes
observations	407,001	407,001	407,001	129,924	93,215

t statistics in parentheses

$^{*}p < 0.05, ^{**}p < 0.01, ^{***}p < 0.001$

Clustered standard errors in parentheses on the identity of the agent "on the left" of the dyad.

Table B.2: Conditional logit on the occurence of the first connection, all sample, five-year window network, quadratic detrending.

	1	2	3	4	5
non-common	0.0320***	0.0327***	0.0275***	0.0312***	0.0219**
	(7.63)	(7.80)	(4.71)	(3.73)	(2.64)
common	0.198***	0.203***	-0.531***	-0.256***	-0.139**
	(12.47)	(12.71)	(-6.36)	(-3.78)	(-2.63)
geo distance		-0.00132***	-0.00143***	-0.000620***	-0.000762***
		(-33.87)	(-33.39)	(-8.23)	(-8.89)
Jaffe tech distance		-0.338***	-0.256***	-0.505***	-0.578***
		(-6.12)	(-3.71)	(-4.38)	(-4.74)
public research			23.13***	21.83***	19.30***
			(129.52)	(56.34)	(44.52)
common applicant			31.76***	24.76***	22.75***
			(11.87)	(177.49)	(265.64)
network controls	yes	yes	yes	yes	yes
applicant controls sum	no	no	no	yes	yes
applicant controls dif	no	no	no	no	yes
observations	407,001	407,001	407,001	129,924	93,215

t statistics in parentheses

$^{*}p < 0.05,^{**}p < 0.01,^{***}p < 0.001$

Clustered standard errors in parentheses on the identity of the agent "on the left" of the dyad.

Table B.3: Conditional logit on the occurence of the first connection, all sample, ten-year window network, linear detrending.

non-common	-0.0178**	-0.0161*	-0.0185*	-0.0105	0.00904
	(-2.62)	(-2.39)	(-2.06)	(-0.98)	(0.93)
common	0.179***	0.184***	-0.470***	-0.310***	-0.157**
	(10.48)	(10.73)	(-6.16)	(-3.95)	(-2.72)
geo distance		-0.00134***	-0.00144***	-0.000647***	-0.000770***
		(-32.85)	(-32.19)	(-8.48)	(-8.88)
Jaffe tech distance		-0.420***	-0.336***	-0.544***	-0.542***
		(-6.76)	(-4.34)	(-4.44)	(-4.18)
public research			21.97***	20.81***	20.35***
			(109.36)	(60.84)	(45.38)
common applicant			27.95***	23.47***	23.46***
			(13.29)	(145.10)	(255.74)
network controls	yes	yes	yes	yes	yes
applicant controls sum	no	no	no	yes	yes
applicant controls dif	no	no	no	no	yes
observations	347707	347707	347707	118839	86950

t statistics in parentheses

$^{*}p < 0.05,^{**}p < 0.01,^{***}p < 0.001$

Clustered standard errors in parentheses on the identity of the agent "on the left" of the dyad.

The observations before year 1988 were dropped to obtain a consistent ten-year window for each year considered.

Table B.4: Conditional logit on the occurence of the first connection, sample limited to the formation of inventor teams of size 2, five-year window network, linear detrending.

	1	2	3	4	5
non-common	0.0533***	0.0536***	0.0288*	0.0371*	0.0257
	(5.79)	(5.82)	(2.30)	(1.99)	(1.34)
common	0.257***	0.269***	-0.513**	0.0137	0.0870
	(5.89)	(6.12)	(-2.84)	(0.13)	(0.87)
geo distance		-0.00153***	-0.00172***	-0.000634***	-0.000790***
		(-19.59)	(-18.89)	(-3.85)	(-4.40)
Jaffe tech distance		-0.215	-0.0992	-0.333	-0.480
		(-1.94)	(-0.69)	(-1.30)	(-1.73)
public research			20.54***	21.82***	17.56***
			(46.34)	(22.06)	(24.38)
common applicant			23.95***	22.53***	21.47***
			(89.56)	(126.59)	(122.40)
network controls	yes	yes	yes	yes	yes
applicant controls sum	no	no	no	yes	yes
applicant controls dif	no	no	no	no	yes
observations	62,690	62,690	62,690	18,994	13,965

t statistics in parentheses

$^{*}p < 0.05,^{**}p < 0.01,^{***}p < 0.001$

Clustered standard errors in parentheses on the identity of the agent "on the left" of the dyad.

Table B.5: Conditional logit on the occurence of the first connection, sample limited to the formation of inventor teams of size 2 and sample restricted to one randomly chosen dyad per inventor "on the right", five-year window network, linear detrending.

	1	2	3	4	5
non-common	0.132***	0.130***	0.120***	0.158***	0.150***
	(7.19)	(7.11)	(5.24)	(4.38)	(3.80)
common	0.465***	0.478***	-1.163	-0.220	0.108
	(4.72)	(4.79)	(-1.07)	(-0.43)	(0.29)
geo distance		-0.00171***	-0.00189***	-0.000816***	-0.00106***
		(-14.15)	(-13.55)	(-3.43)	(-3.87)
Jaffe tech distance		-0.343	-0.469*	-0.884*	-0.873
		(-1.87)	(-2.13)	(-2.12)	(-1.82)
public research			18.13***	17.63***	17.30***
			(44.69)	(28.13)	(17.33)
common applicant			23.65***	21.96***	20.76***
			(3.88)	(57.42)	(72.31)
network controls	yes	yes	yes	yes	yes
applicant controls sum	no	no	no	yes	yes
applicant controls dif	no	no	no	no	yes
observations	26,669	26,669	26,669	7,337	5,101

t statistics in parentheses

$^{*}p < 0.05,^{**}p < 0.01,^{***}p < 0.001$

Clustered standard errors in parentheses on the identity of the agent "on the left" of the dyad.

Cahiers du GREThA
Working papers of GREThA

GREThA UMR CNRS 5113

Université Montesquieu Bordeaux IV
Avenue Léon Duguit
33608 PESSAC - FRANCE
Tel : +33 (0)5.56.84.25.75
Fax : +33 (0)5.56.84.86.47

http://gretha.u-bordeaux4.fr/

Cahiers du GREThA (derniers numéros – last issues)

2014-01 : BLANCHETON Bertrand, PASTUREAU Guillaume, *Le Mont-de-Piété à Bordeaux, les raisons d'un succès (1802-1913)*

2014-02 : FRIGANT Vincent, MIOLLAN Stéphane, *La restructuration de la géographie de l'industrie automobile en Europe durant les années 2000*

2014-03 : BLANCHETON Bertrand, *Les improvisations financière de la guerre de 1914-1918 en France. Les enjeux de la liquidité*.

2014-04 : ARNAUD Brice, *Extended Producer Responsibility and Green Marketing: an Application to Packaging*

2014-05 : CARAYOL Nicolas, DELILLE Rémi, VANNETELBOSCH Vincent, *Allocating value among farsighted players in network formation*

2014-06 : ARIFOVIC Jasmina, YILDIZOGLU Murat, *Learning the Ramsey outcome in a Kydland & Prescott economy*

2014-07 : BECUWE Stéphane, BLANCHETON Bertrand, *Relations internationales et discriminations tarifaires : le cas de la France (1850-1913)*

2014-08 : BECUWE Stéphane, BLANCHETON Bertrand, *La politique commerciale de la France et les filières sucrières de ses vieilles colonies sous le Second Empire*

2014-09 : FRIGANT Vincent, ZUMPE Martin, *Are automotive Global Production Networks becoming more global? Comparison of regional and global integration processes based on auto parts trade data*

2014-10 : BEN OTHMEN Asma, *De la mise à contribution des bénéficiaires au financement de la préservation des espaces naturels : tarification de l'accès ou augmentation de taxe?*

2014-11 : MENON Carlo, *Spreading Big Ideas? The effect of Top Inventing Companies on Local Inventors*

2014-12 : MIGUELEZ Ernest, *Inventor diasporas and the internationalization of technology*

2014-13 : CARAYOL Nicolas, CASSI Lorenzo, ROUX Pascale, *Unintended triadic closure in social networks: The strategic formation of research collaborations between French inventors*

La coordination scientifique des Cahiers du GREThA est assurée par Emmanuel PETIT. La mise en page est assurée par Anne-Laure MERLETTE.