Unintended Triadic Closure In Social Networks: The Strategic Formation Of Research Collaborations Between French Inventors[1]

Nicolas Carayol [2], Laurent Bergé[3], Lorenzo Cassi[4], Pascale Roux[5]

August 10, 2018

[1]We thank Matt Jackson, Tom Snijders, Robin Cowan, and participants in the ARS Conference held in Rome, the EEA Conference, the Workshop on Economic Design and Institutions held at Facultes Universitaires Saint Louis in Brussels, the Pecs conference on Regional Innovation and Growth, and the First International NPR Conference at the Universita Cattolica del Sacre Cuore in Milano. The suggestions of two anonymous referees significantly helped us improving the paper. We are grateful to the Agence Nationale de la Recherche (grant ANR-06-JCJC-0076) and the Aquitaine Region (AAP program, no 20101402006) for their financial supports, as well as the French Ministry of Higher Education and Research for providing us with the administrative R&D surveys used in this study.

[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]Laurent Bergé, Université du Luxembourg, CREA, 162A, avenue de la Faïencerie, L-1511 Luxembourg. Tel: +352 46 66 44 6596. Email: laurent.berge@uni.lu

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

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

Abstract

Observing that most social networks are clustered, the literature often argues that agents are more willing to form links that close triangles. We challenge this idea by proposing a simple model of new collaboration formation that shows why network clustering may arise even though agents do not "like" network closure. We address empirically this question on the longitudinal evolution of the French co-invention network, and find that two inventors are less likely to form a first research collaboration when they have common partners. Our findings further reveal the preferences of inventors towards forming non-redundant connections.

Keywords: Social networks; Link formation; Closure; Patents; Conditional logit; Monte Carlo simulations.

JEL codes: D85, O31.

1 Introduction

One of the main salient features of most real social and economic networks is that they are highly clustered, in the sense that the neighborhoods of neighbors tend to overlap, exhibiting what Rapoport (1953) and Granovetter (1973) first called "triadic closure". In other words, agents having a common friend or partner are highly likely to be friends or partners themselves, so that the whole network typically incorporates far more triangles than would be obtained by chance. 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; Carayol and Roux, 2008), scientists (Newman, 2001) and alliances between firms (Kogut and Walker, 2001; Baum *et al.*, 2003).

A direct and natural explanation of the social tendency for closure is based on a individuals' presumed preference for closing triangles, what we call the *love-for-triadic-closure* hypothesis. This idea was first suggested by Simmel (1922), then followed by Heider (1958) and Newcomb (1961), who explored the psychological motives of individuals to maintain a "cognitive balance" between their social relations. Later, sociologists, management scholars and more recently economists, have emphasized the benefits of network closure in social relations. Coleman (1988) argues that closure, by facilitating collective monitoring and sanctions, prevents free-riding and enforces cooperative behavior. Granovetter (1985) highlights the fact that closure facilitates interpersonal trust, since it creates a reputation cost for individuals who misbehave. Common relations play the role of social collateral which favors the formation of relations in a high-value exchange environment (Karlan *et al.*, 2009). A common partner may also act as a referee between his or her acquaintances (e.g. Granovetter, 1973; Burt and Knez, 1995; Fafchamps *et al.*, 2010). Accordingly, the perceived gains of a new relationship between two agents will be higher if they have common partners.

The empirical literature challenges this rationale, however, as it documents mixed effects of closure on economic performances or social achievements. In particular, whereas they generally expect a positive effect of closure on the inventive performances of various actors, studies on research collaboration networks are inconclusive as these effects are either positive, non significant or even negative (e.g. Reagans and Mc Evily, 2003; Schilling and Phelps, 2007; Fleming *et al.,* 2007; Bettencourt *et al.*, 2007; Breschi and Lenzi, 2016). A lively debate also concerns the effect

of (generational or inter-generational) closure on children's school performances or well-being. Famous studies are those of Coleman *et al.* (1982), Carbonaro (1999) and Morgan and Sorensen (1999) who obtain an opposite effect of parents' closure on school achievements. Similarly, contradictory effects are obtained regarding the effect of closure on diffusion. For instance, the findings of Ugander *et al.* (2012) suggest that closure in virtual networks is a negative predictor of Facebook diffusion, while the opposite effect is found by Centola (2010) on the adoption of a health forum. A number of studies also report in various contexts an inverse-U shape relation between closure and, among others, the probability of artistic success of Broadway musicals (Uzzi and Spiro, 2005), team effectiveness (Oh *et al.*, 2004), or new technology-based venture performance (Wang and Chen, 2016).

Rather than making the assumption that agents "like" closure, an alternative point of departure would be that contextual factors or homophily may influence network formation and ultimately lead to closed networks. It has often been claimed in the sociological literature that people tend to form links with others who are similar to them (in terms of age, education, ethnicity, religious beliefs for instance) or share some common neighborhoods, such as a school or company (see McPherson *et al.*, 2001 for an extensive review). In fact, such factors could raise the level of clustering in real networks as people with similar or close characteristics to others may form cliques together (Easley and Kleinberg, 2010). Therefore, it is not necessary to assume *love-for-triadic-closure* to explain clustered networks. In theory, it is even possible to simultaneously reject the *love-for-triadic-closure* hypothesis, introduce a *love-for-non-redundant-connections* ingredient and still observe a high level of triadic closure.

To see this point, consider for instance the 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 declines with social distance. As only the length of the shortest paths matters, agents do not benefit from multiple paths to some other agent. Therefore, 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. In this network formation model, equilibrium networks will typically not exhibit a high level of clustering. However, introducing an exogenous structure (geography for instance) that affects direct link costs in this model is sufficient to ensure that agents form triadic connections (Carayol and Roux, 2009). Though the gross returns of those redun-

dant connections are limited, agents form and maintain them simply because their costs are also very low. Many triangles could therefore be observed in real social networks, even though agents do not particularly like them. Socially distant (non-redundant) connections are more valuable but in the same time more costly because they need to be formed with (geographically) distant agents. There are only a few such social bridges because their formation dissipates the incentives to further form similar connections.[1] This view is consistent with the literature highlighting that individuals who bridge separate clusters (usually named brokers) experience higher performance, for instance in tracking job opportunities (Granovetter, 1974), obtaining promotions (Burt, 1997), generating good ideas (Burt, 2004), or enhancing firms' performances (Zaheer and Bell, 2005) or inventiveness (Ahuja, 2000; Baum *et al.*, 2000).[2] Yet, bridges are expected to provide opportunities for individuals or organizations to benefit from new information or ideas arising from disconnected parts of the network (Granovetter, 1973; Burt, 2000; Letterie *et al.*, 2008).

In this paper, we empirically challenge the ideas that *love-for-triadic-closure* and/or *love-for-non-redundant-connections* may drive the formation of new collaborations. These hypotheses are tested on the formation of collaboration ties between individual inventors. More specifically, we estimate the formation of links in a large co-invention network.[3] Such networks are particularly well suited to test our hypotheses for several reasons. First, we were able to build a dataset on a large scale networks which we observe longitudinally over a sufficiently long period of time. Further, this dataset has been matched with companies data sets in order to use interesting covariates on top of other covariates built using information contained in the patents. The second reason is that the formation of collaborations between inventors remains understudied since previous studies in the field have mainly considered network ties as exogenous. Moreover, the results obtained concerning network effects on inventive productivity remain contradictory in the literature in particular regarding the effects of closure (Schilling and Phelps, 2007; Fleming *et al.,* 2007; Bettencourt *et al.*, 2007; Breschi and Lenzi, 2016). A deeper understanding of how collaborative ties are formed should provide new insights for empirical research to

[1] Carayol and Roux (2009) show that 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).

[2] Note that these networks, like other social networks, typically exhibit a high level of clustering (i.e. triadic closure).

[3] A link is drawn between two individuals if they have previously co-invented at least one patent. The procedure is similar to that for drawing scientific collaboration networks from data on the co-authorship of scientific publications (Newman, 2001; Barabási *et al.*, 2002; Fafchamps *et al.*, 2010).

consider those effects[4] and should ultimately better ground policy recommendations.

We first design a very simple heuristic model which clarifies the relation between network gains and link formation. In our model, 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.[5] 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 externalities. Though the model has intentionally been kept as simple as possible, it is sufficiently general to encompass opposing assumptions about triadic closure. Either direct and indirect links to an agent are complementary and then agents typically like forming triangles, or they are partial or perfect substitutes, which implies that agents do not care about forming triangles, or even dislike doing so. Simultaneously, the model allows us to test whether two agents are more likely to become connected when that new link grants access to agents they do no yet benefit from directly (non-redundant links). This leads to a simple and generic expression of the incentive schemes for forming collaborations at any point in time, either in or out of equilibrium. As such, this expression can be applied to different contexts of network formation.

We use 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. Once we have disambiguated inventor identities thanks to an original Bayesian methodology, we obtain a population of about one hundred thousand inventors and reliable individual information on them, such as their precise geographic location, technological specialization or patent applicant identities (mostly companies in the EPO system). These applicants are matched against the list of French companies in mandatory annual surveys to gather detailed information on applicants, including their yearly R&D investments and research personnel. We suspect that the omission of some (potentially time-varying) variables could lead to the mistaken conclusion that agents have incentives to form triadic connections. This is likely to occur if, in the true data generation process, these covariates affect both

[4]For instance, though we would expect non-redundant connections to be very important to access fresh ideas in such professional networks, Lee (2010) find that individual fixed effect may in fact explain both the formation of distant connections and inventor performance. He then shows that controlling for inventor fixed effects, the position-performance correlation disappears.

[5]In our model, each agent can create a new collaboration with any other existing agent. This network formation mechanism is then different from the ones where new connections can only be ensured by new agents entering the network, as in Barabási and Albert (1999), Jackson and Rogers (2007) or more recently König (2016).

the probability of connecting and the probability of having common friends in the same way. The main variables we have in mind are the geographical distances between inventors, the institutional barriers they may face or the existence of common research interests, which should affect the opportunities and costs of forming links and be simultaneously correlated with closure. Their influence on knowledge flows between workers, scientists or inventors has already been evidenced in a number of papers (e.g. Kogut and Zander, 1992; Jaffe *et al.*, 1993; Fafchamps *et al.*, 2010 ; Lee, 2010).

We rely on conditional logit regressions which allow us to control for such dyadic fixed effects to estimate the probability of two as yet unconnected agents forming a first connection at any period of time. The identification of the main explaining variable effects thus comes from their variation in time. As this may bias estimates, explaining variables are detrended (Allisson and Christakis, 2006; Fafchamps *et al.*, 2010). To account for the interdependence between dyads emanating from the same person, inference is based on dyadic-robust standard errors (see Fafchamps and Gubert, 2007 and Cameron and Miller 2014). Importantly, we introduce a dyadic fixed effect that accounts for the time-invariant matching quality of the two agents, as well as for the invariant individual abilities of the two inventors.[6]

When the connection costs are not properly accounted for, our estimations lead to the conclusion that agents like closing triangles. But since we account for these costs, the closure effect disappears. According to our preferred specification, agents are even 23% less likely to collaborate when they have one more common previous collaborator. This speaks against the *love-for-triadic-closure hypothesis*. Our second result is that inventors are more likely to collaborate when they each have more partners with whom the other is not already connected to. If agents of a given dyad have one standard deviation of non-common partners more, they are 20% more likely to become connected. This supports the *love-for-non-redundant connections hypothesis* in the context of knowledge creation. Such findings appear to be globally robust to different assumptions about the trend of the quantitative explaining variables (logarithmic or quadratic vs. linear), to a larger time-window used for building network variables (ten-year vs. five-year), and to different methods for clustering standard errors (two-way vs. dyadic).

[6]In a recent paper, Graham (2016) introduces a methodology to disentangle closure effects from other homophily effects in network formation. The identification strategy hinges on dyads creating or severing a link together while all their other connections as well as their neighbors' connections remain stable across time. Unfortunately this method is very demanding in terms of network data and cannot be used on our dataset since only a handful of stable dyads are present.

Finally, we conduct a series of Monte Carlo simulations to investigate the consequences of departing from implicit assumptions of the econometric model. In particular, as meetings can not be observed, we do not know if some dyads remain unconnected because agents have met but decided not to collaborate, or if they did not meet. We show that this may lead to a downward bias in magnitude for our network variables. We also explore the consequences of assuming a different meeting process, continuous updating of the network, and mistakes in the identification of agents. Monte Carlo simulations overall reinforce the conclusion that no *love-for-triadic-closure* is at play in our data and even further suggest that the "true" negative effect on triadic closure in network formation may even be underestimated.

The following section introduces the heuristic theoretical model of strategic research collaboration formation, and our empirical strategy. The third section presents the data. The fourth section describes our findings and investigates several robustness checks. In Section 5, we discuss the results of the Monte Carlo experiments. The last section concludes and outlines some managerial and policy implications of our findings.

2 The formation of inter-individual research collaboration networks

In this section, we present the different building blocks of a simple theoretical model which illustrates how individual expected returns from collaboration shape link-forming strategies. This model leads to a reduced form equation specifying agents' incentives to bilaterally form collaborations. Lastly, we show how this equation can be tested empirically and how it relates to our hypotheses on triadic closure and non-redundant connections.

2.1 The setup

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 exposition, agents are assumed never to retire or die,[7] so that $N^t \subseteq N^{t+1}$. 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

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

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 bilaterally 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. 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 networks in which forward-looking computations become extremely complex.

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. Let us assume that the expected (net) returns for i of a shared research project with j formed at period t is given simply by:

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

where the net returns of the research collaboration are composed of: i) θ_{ij}, an idiosyncratic, pair-specific and time-invariant parameter, of ii) ε_{ij}^t a noise interpreted as the opportunities of research collaboration between i and j that particular year, and of c_{ij}^t which captures the (sunk) time-variant costs and benefits, supported by i, for running the research collaboration with j at period t. ζ is a non-null and positive parameter so that variable c_{ij}^t is interpreted as a net cost. To simplify the exposition of the model, though that is not necessary for our results, we will further assume that $\theta_{ij} = \theta_{ji}$, $\varepsilon_{ij}^t = \varepsilon_{ji}^t$ and $c_{ij}^t = c_{ji}^t$ 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 for reasons of simplicity. The complex of bilateral social connections is also assumed to support positive externalities at each period. We propose the following simple specification of these "network" (per unit of time) payoffs:

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

with $\eta_{ij}(g^t)$ the number of direct links between i and j on g^t (equal to 0 or 1), $\eta^2_{ij}(g^t)$ 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 $\overset{\triangle}{\eta}_{ij}(g^t)$ the number of triangles on g^t having i and j as summits.[8]

The two parameters α and β are likely to be positive. A standard interpretation for these parameters would be that they 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 externality captured by i for each indirect connection of length two to any direct neighbor. It captures both possible knowledge spillover flowing on such a path, and a closure effect. As such, it is also expected to be positive.

It should be noted that, according to this payoff specification, agents are assumed only to consider social network externalities at a social distance less than or equal to two. 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. 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 (which is supposed to be random for simplicity). For simplicity, agents are assumed to be able to 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.[9] Therefore, the total

[8] 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.

[9] 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. This assumption allows us to consider the formation of a link as a joint dyadic decision,

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 the variation in the discounted payoffs of agents i and j 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:

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

with $\bar{\eta}_{ij}\left(g^t\right)$ the number of non-common neighbors of i and j on g^t,[10] and with $\overset{\triangle}{\eta}_{ij}(g^t)$ the number of common neighbors of i and j defined above.[11] 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* $(1-\delta)\left[\Delta(g, ij)\right] = 2(1-\delta)\left[\theta_{ij} + \varepsilon_{ij}^t - \zeta c_{ij}\right] + [2\alpha + 3\beta + 2(4\gamma - 2\beta)]$*. 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.*

instead of two separate decisions. It is consistent with the idea that in research collaboration, not all agents contribute equally: more peripheral agents often 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.

[10] 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).

[11] 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.

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.[12] 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 chosen randomly 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:

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

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 $\beta_1 = \frac{\alpha}{(1-\delta)}$, $\beta^{nc} = \frac{\beta}{2(1-\delta)}$, $\beta^c = \frac{2\gamma-\beta}{1-\delta}$ and $\beta^{cost} = -\zeta$. If $2\gamma < \beta$, then $\beta^c > 0$, and thus the number of triangles impact positively the

[12] 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 is at equilibrium would amount to considering that agents could rearrange all their collaborations at each period, which would obviously not be consistent. Carayol and Roux (2008) adopt the alternative perspective by studying inert components, assuming that they reach some stable state.

incentives to form a link. If it turns out to be positive and significant, this would mean that agents like closing triangles, thus supporting the love-for-triadic-closure hypothesis. We will also be concerned with β^{nc} being positive and significant, which would provide support for the love-for-non-redundant-connections hypothesis, in line with the idea that the collaboration network is a vehicle for knowledge spillovers.

Turning to the costs of research collaboration formation, we will identify several factors that may impact the probability of forming a connection. Several forms of homophily affect such costs in terms of uncertainty, time and effort to form a link.

3 Data and variables

Our primary empirical evidence is built upon all European patent applications in 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.

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 Type 1 error of homonymy would lead to considering that different persons are the same, thereby mistakenly generating some apparently extremely connected agents creating unjustified links between different communities. A Type 2 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,[13] a widely accepted

[13]For an overview see Miguélez and Gomez-Miguélez (2011), Pezzoni *et al.* (2014) and Li *et al.*

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. Two main issues need to be addressed. First, how much should similarity scores increase when two homonyms have the same modality for some given variable? Should, for instance, information about the city of residence contribute more or less to the similarity scores than information about the technological classes? Second, how should the relative frequencies of each variable modality be taken into account? Clearly, it is not as informative to observe that two homonyms live in Paris as in a small town, and we would like to know what difference this makes exactly.

We have thus developed 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,[14] addresses the two main issues stressed above. Out of 133,764 patents considered, we find 262,186 patent×inventor occurrences that correspond to an address in France. We use the following list of observable attributes of individuals: 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 Type 1 and Type 2 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

(2014).

[14]Even more details are available in a technical note written by two of us (Carayol and Cassi, 2009), on the same data set, but less updated.

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 mostly rely here upon a five-year backward-moving window. However, we have also computed network data on an alternative ten-year moving window, and 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 according to 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.[15] Such a discrepancy may be explained by a greater density of the co-authorship networks.[16] 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 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 for 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, 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 in mandatory national surveys.

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

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

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

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 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,[17] 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,[18] we calculated the two explaining variables of major interest on the 5-year window network g^t, namely the number of non-common neighbors $\overline{\eta}_{ij}(g^t)$ and the number of common neighbors $\overset{\triangle}{\eta}_{ij}(g^t)$.[19] We also computed a series of network controls (noted $\text{net\_controls}_{ij}^t$) 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 France area by matching the post codes mentioned in their addresses with their corresponding latitude and longitude coordinates.[20] 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

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

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

[19] The five-year window was used, since it is the one most commonly employed in similar empirical network studies. However, a larger ten-year window will also be used to build the right-hand side network variables in the robustness check analyses. By doing so, we lose five years of observation and the period covered is 1988-2004 for the dependent variable and 1987-2003 for the explanatory variables.

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

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 less than 50 km from each other, while fewer than 6.2% of the connections are formed between agents who live 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 also computed the technological distance. This has been defined using the similarity measurement proposed by Jaffe (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. Our results are invariant when we use alternative measurements of technological distance, such as the Euclidean or Manhattan distance which are very correlated together and with Jaffe distance (coefficients above .97).[21]

[21]Estimations are not included due to space constraints but are available from the authors.

3.3.3 Applicants

The association of inventors to applicants on a yearly basis was 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. A simple justification would be that academics are less likely to perceive each other as competitors and are therefore more likely to be willing to engage in joint research projects. The dummy variable acad_{ij}^{t} captures this: it is equal to unity if the two agents have already invented a patent for which the applicant is a public research institution. About eleven percent of all dyads formed are between academics.

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 the matching key. These surveys are exhaustive for all the companies employing at least one full-time researcher (whatever their size) and provided us with annual data on companies' internal and external R&D expenditure, number of researchers, as well as more general information, such as total number of employees. We then deflated internal and external R&D expenditure by a national investment price index. Information concerning the applicants associated with the inventors was 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 company surveys we 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 130,000 observations for all dyadic sums, and approximately 93,000 observations for all the relative differences within the dyad (which can not be calculated when the sum is null). Since such a massive

reduction in the number of observations sharply decreases coefficient significance, we decided to present only the results in which applicant controls are limited to dyadic sums.[22]

4 Estimations and results

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

$$\Pr\left(ij \in g^{t+1} \left| ij \notin g^t\right.\right) = F\Big(\beta_1 + \beta^{nc}\overline{\eta}_{ij}\left(g^t\right) + \beta^{c}\overset{\triangle}{\eta}_{ij}\left(g^t\right) + \beta_1^{cost}\text{geo}_{ij}^t + \beta_2^{cost}\text{jaffe}_{ij}^t + \beta_3^{cost}\text{app}_{ij}^t + \beta_4^{cost}\text{acad}_{ij}^t + \beta_5^{cost}\text{app\_controls}_{ij}^t + \beta_6^{cost}\text{net\_controls}_{ij}^t + \theta_{ij}\Big). \tag{6}$$

The significance and signs of parameters β^{nc} and β^{c} are our main interest, provided that we control properly for the direct benefits and costs of collaboration formation. To do so, we 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. That term accounts for any time-invariant effect on the probability of connecting, such as the individual abilities of i and j. 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 capture more collaborative research patterns.

Lastly, we included two series of controls: $\text{app\_controls}_{ij}^t$ and $\text{net\_controls}_{ij}^t$. The former refer 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 expenditure, 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. This information is however available only for one subset of the whole sample.[23] The second series of

[22] We have found that the decrease in significance can be attributed mainly to the sample reduction and not to the inclusion of more controls by running all model regressions on the only observations that are fully informed.

[23] For the dyads in which link formation is assessed thanks to a patent for which the applicant is found in the company data.

controls concerns inventors' time-variant network attributes: the average number of patents per year of the two agents, the rate of difference in the degree of the agents, 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 meeting probability p that is assumed to be uniform across dyads in the model).[24]

Our estimation strategy immediately raises a first issue. Allisson and Christakis (2006) show that the estimation of such a fixed effect logit model leads to spurious estimates when the explanatory variables are trended. This is due to the fact that it is not possible (by design) to observe link deletion in such data. All dyadic time series take the form of a series of zeros followed by a one. Therefore, as suggested in Fafchamps *et al.* (2010), all quantitative explanatory variables are first detrended by assuming a linear trend (in the robustness checks, we relax this assumption by considering other forms of trend).

A second issue concerns inference. Our observations are not independently distributed since all the observations corresponding to dyads involving the same agent are likely to be correlated. Ignoring such correlation between observations may lead to an inference problem. We are particularly worried about a potential overestimation of coefficient significance, although the contrary may also occur. Clustering observations on the dyads is not satisfactory since clustering should be performed on the identities of the two members of each dyad. Cameron and Miller (2015) suggest a two-way clustering approach by which observations are clustered on the identity of the two persons involved. However it does not fully account for all the correlation between observables. In particular, it does not take into consideration the correlation between dyads that share the same agent on different "sides" of the dyad (on the right and on the left). In this paper, we will follow previous work by Fafchamps and Gubert (2007) and Cameron and Miller (2014)[25] to provide dyadic-robust standard errors estimation. We have adapted the sandwich variance estimator of Cameron and Miller (2014) to the conditional logit model which is estimated through maximum likelihood.[26]

Table 3 synthesizes our baseline regression results. We find first that, as ex-

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

[25]Previous work by Snijders and Borgatti (1999) should also be mentioned.

[26]Adaptation of their Stata code "regdyad2.ado". In the robustness checks, we show that this method is, as expected, more conservative (t statistics are closer to zero) than the two-way clustering approach suggested by Cameron and Miller (2015).

pected, the number of neighbors that the two agents do not have in common always positively and significantly impacts the probability of connection. This result is significant at the 0.1% level for the three regressions for which all 407,001 observations are available. Significance at the 5% level is obtained when we also control for the applicant (company) characteristics, imposing a significant reduction in sample size down to one third of its initial size. This result supports the idea that agents benefit from indirect connections (at least at distance two) which is consistent with the existence of network-based knowledge spillovers. It supports the *love-for-non-redundant-connections hypothesis*. This effect may seem limited as having one more non-common friend increases the probability to connect by approximately 3% across all specifications. However, note that agents in the dyads under investigation have in average four-to-five non-common partners. Moreover, the standard deviation of that variable is even slightly larger (reported in Table 1), so that a one-standard deviation of the number of non-common friends actually increases the probability to connect by more than 20%. Remind that such probability increase is obtained within the dyad and while controlling for all other time-varying covariates.

Our second and main result is the following. When we do not control for the costs of network formation (no cost variable is introduced in the model of Column 1, Table 3) or only account for technological and geographic distances (Column 2), the number of common neighbors is positively and significantly associated with the probability of creating a link. This seems to indicate that agents like closing triangles. However, when we properly control for omitted variables such as the connection costs, in particular when we control for all applicant variables (Columns 3 and 4), it appears clearly that this effect disappears. Therefore, if inventors close triangles, this is totally explained by the controls. Moreover, it turns out that the number of common neighbors now significantly decreases the probability of becoming connected. These results are significant at the 0.1% level for the models of columns 1 to 3, and at the 1% level for the full model of Column 4 - on the reduced sample. This effect is strong as, according to the last specification, having one more friend in common decreases the probability to connect by 23%. In terms of within-sample variation, a one standard deviation in the number of non-common partners raises the probability to connect by 16%. These results clearly speak against the *love-for-triadic-closure hypothesis*.

The results appear to be globally robust to a list of alternative specifications. All the supplementary regressions we discuss below are reported in Appendix B. The results are robust in particular 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 then appears 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 is never significant and very close to zero.[27] Lastly, in Table B4 we report the same regression coefficients as in Table 3, but the t statistics are obtained with two-way clustered standard errors instead of dyadic clustered standard errors. These standard errors were computed as suggested in Cameron and Miller (2015). As expected, these results are less conservative in terms of inference as compared to our main results presented in Table 3.

The impacts of the cost variables are also interesting in themselves. All the cost variables are always significant at the 1% level,[28] and 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, and while controlling for individual fixed effects, the probability of forming a connection is higher, which highlights the role of academics in creating connections among inventors.

5 Monte Carlo experiments

In this section, we explore the validity of the estimated negative effect of common neighbors on link formation, relying on a series of Monte Carlo simulations. We suspect that our estimates may be biased by underlying assumptions, in particular

[27] Moreover, this coefficient remains positive in a more complete specification in which difference variables (between applicants in the dyad) are included. Results are not reported, but are available from the authors.

[28] In fact, all the cost variables in all models but one (Jaffe technological distance in Model 3) are significant at the 0.1% level.

concerning the unobserved meeting process between agents which we have assumed to be uniform.

The Monte Carlo model, further detailed in Appendix C,[29] simulates network formation according to the model presented in Section 2. At each period, non-connected dyads meet with probability p and decide to collaborate if their dyadic net return is positive, calculated as follows:

$$\Delta_{ij}^{t} \equiv \theta_{ij} + \beta^{c} \overset{\Delta}{\eta}_{ij} \left(g^{t} + ij\right) + \beta^{nc} \bar{\eta}_{ij} \left(g^{t}\right) - c_{ij}^{t} + \varepsilon_{ij}^{t}, \tag{7}$$

which is a simplified version of Equation 3. β^{c} and β^{nc} are generative coefficients chosen on purpose.[30] Only dyads of agents who eventually collaborate are considered for inclusion in the data table, up to their first connection, so that the dependent variable is a time series of 0 followed by a single 1 for each dyad (as for the empirical data estimation in Section 4). Then, we estimate the occurrence of a first connection in a conditional Logit model on linearly detrended explanatory variables. The whole network data generation process and the estimation are repeated 100 times so that a distribution of estimated coefficients is obtained for each single set of generative coefficients.

In a first set of Monte Carlo simulations, the generative coefficients β^{c} and β^{nc} are calibrated using the estimated coefficients of Column 3–Table 3. The regression results obtained on those data are presented in Model 1–Table 4. The mean estimated coefficient of common neighbors is -0.25 (SD is only 0.07) and are significant for 98 out of 100 runs. The sign is thus correctly estimated but the coefficient is about half of its associated generative coefficient (-0.505). We suspect this downward bias in magnitude may be due to the systematic inclusion in the sample of agent dyads for all the periods before their first collaboration, even when they actually do not meet. Those dyad-periods have been included to match the very nature of the empirical data. Meetings between inventors are not directly observed as in most real network data. Meetings are however observable in the Monte Carlo generated data allowing us to appreciate to what extent using this information would reduce the estimation bias. In Model 2 agents dyads are excluded for all periods they do not meet. This significantly alleviates the downward bias as the average estimated coefficient of the number of common neighbors is now -0.41, closer to its generative value (-0.505).

[29] Note that the results of this section are robust to a variation in the generating parameters, as shown in Appendix D.

[30] The generative coefficient of costs is normalized to -1.

If the underestimation of this coefficient by 50% is roughly preserved for different generative values of β^c, it is pretty straightforward to find out the generative value which leads to estimates matching the ones obtained on real data. Setting the generative β^c to -1 (nearly twice the estimated β^c on the real data), we obtain estimated coefficients that are indeed very close to the coefficients obtained from real network data regressions (-0.53 in Model 3, vs. -0.505). Further, network generation free of any network determinants (i.e., generative $\beta^c = \beta^{nc} = 0$) leads to positive estimated coefficients (Model 5) that are significant in 71% of the regressions. Therefore, in the absence of any network effect, spurious positive (not negative) effects of common neighbors may be obtained. Such bias is mainly due to not observing meetings as excluding non-meeting pairs almost completely solves the problem (the fraction of regressions for which we obtain significant coefficients falls to 7% in Model 6, and the mean point estimates are close to 0).

In Appendix E, the Monte Carlo procedure is used to investigate three other potential biases. We first explore the consequence of miss identifying agents in the data. We find that artificially adding *Type 1 or Type 2* errors[31] into the data has limited consequences on the estimates. Large type 1 errors only lead to a downward bias in magnitude of the estimated coefficients. Such errors are however controlled and minimized by the disambiguation algorithm (see Appendix A). Secondly, we consider a non uniform meeting process, assuming that meetings are more likely when agents have common friends. This alternative meeting process does not bias estimates. Thirdly, we assume agents take their decisions in continuous time, observing the current state of the network whereas estimations are performed on a discrete time basis. The bias is also limited, decreasing slightly the magnitude of the common neighbors estimates.

In a nutshell, none of those Monte Carlo experiments suggests that the estimated negative effect of common neighbors could be spurious. If biased, the negative effect of common neighbors may have only been underestimated. This reinforces our main conclusion leading to the rejection of the *love-for-triadic-closure* hypothesis in our context.

[31] Type 1 errors are false positives while Type 2 errors are false negatives.

6 Conclusion

While a large number of empirical studies examine how social networks shape aggregate or individual outcomes, the individual strategies that drive their evolution are often not considered. At the same time, a growing theoretical literature explores the properties of networks that emerge when self-interested agents strategically form their connections (Jackson, 2009). In this paper, we contribute to a recent empirical literature that aims to fill the gap between these two approaches by studying the formation of links using panel data on large social networks (e.g. Fafchamps *et al.*, 2010; Snijders, 2017). Specifically, we examine the incentives of inventors to form new research collaborations, with a special focus on the effect of triadic closure and non-redundant connections which have been identified as some of the main drivers of social network formation. We estimate the proposed incentive equation to bilaterally form new links using precise data on patents produced by French inventors over the period 1978-2004 that allow us to account for time-varying co-variates and to control for dyadic as well as individual fixed effects.

We find that two inventors are significantly less likely to form a first research collaboration when they have more common partners, once all the potential confounding factors are properly controlled for. Our heuristic model leads us to interpret this result as meaning that inventors are not willing to close triangles *per se*, and even that they dislike doing so. A series of Monte Carlo experiments even suggest we may underestimate this effect. It may (or may not) 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. As we control for patent applicants, we also control for the positive effect of being employed by the same organization. In any context of application, however, this result does urge us to control for the various potential costs and constraints borne by the agents when testing individuals' preferences toward forming links as a function of the existing network connections. A second finding reveals the preferences of inventors towards non-redundant connections. Taken together, these results indicate that connections among socially closed agents might provide lower benefits but that the lower costs or constraints faced to form such connections strongly encourages their formation.

This paper also contributes to the recent literature on knowledge spillovers, networks and invention. Several studies have shown that interpersonal networks are crucial determinants of knowledge transmission (e.g. Singh, 2005; Breschi and Lissoni, 2006). Our evidence further shows that inventors preferentially form links with

partners with whom they are not already indirectly connected. One interpretation would be that agents prefer to form non-redundant connections to benefit from fresh ideas and gather information from socially distant sources. Nevertheless, our results also suggest that inventors mostly build those less fruitful collaborations within clustered communities because they are less costly. In terms of innovation policy, an implication of our findings is that the focus should be more on the communities of individuals (rather than on companies or spatial clusters) and on subsidizing the formation of those supposedly more efficient but costly connections that span institutional and other boundaries. Their social value is likely to be very high because they are non-redundant, much higher than their private returns for the directly concerned agents.

References

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

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

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

Barabási, 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.

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.

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

Breschi, S., Lenzi, C., 2016. Co-invention networks and inventive productivity in US cities. Journal of Urban Economics 92: 66-75.

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 BM, Sutton RI, 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., Knez, M., 1995. Kinds of third-party effects on trust. Rationality and Society 7(3): 255-292.

Cameron, A.C., Miller, D.L., 2015. A Practitioner's Guide to Cluster-Robust Inference, Journal of Human Resources 50(2): 317-372.

Cameron, A.C., Miller, D.L., 2014. Robust Inference for Dyadic Data, mimeo.

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.

Carbonaro, W., 1999. Opening the Debate on Closure and Schooling Outcomes. American Sociological Review 64: 682-686.

Centola, D., 2010. The Spread of Behavior in an Online Social Network Experiment. Science 329: 1195-1197.

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

Coleman, J.S., Hoffer, T., Kilgore, S., 1982. High School Achievement: Public, Catholic, and Private Schools Compared. New York: Basic Books.

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.

Easley, D., Kleinberg, J., 2010. Networks, crowds, and markets: Reasoning about a highly connected world. Cambridge University Press.

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., Gubert, F., 2007. Risk sharing and network formation. The

American Economic Review 97(2): 75-79.

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.

Graham, B.S., 2016. Homophily and transitivity in dynamic network formation. NBER working paper series (22186).

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.

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., Rogers, B.W., 2007. Meeting strangers and friends of friends: How random are social networks? American Economic Review 97 (3): 890–915.

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., Trajtenberg, M., Henderson, R., 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. The Quarterly Journal of Economics 108(3): 577–598.

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

Kogut, B., Zander, U., 1992. Knowledge of the firm, combinative capabilities, and the replication of technology. Organization Science 3(3): 383–397.

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

König, M.D., 2016. The formation of networks with local spillovers and limited observability. Theoretical Economics 11 (3): 813–863.

Lee, J., 2010. Heterogeneity, brokerage, and innovative performance: Endoge-

nous formation of collaborative inventor networks. Organization Science 21(4): 804-822.

Letterie, W., Hagedoorn, J., van Kranenburg, H., Palm, F., 2008. Information gathering through alliances. Journal of Economic Behavior and Organization 66: 176-194.

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.

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

Miguélez, E., 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.

Morgan, S.L., Sørensen, A.B., 1999. A test of Coleman's social capital explanation of school effects. American Sociological Review 64: 661-681.

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 Academy of Science 98: 404-409.

Oh, H., Chung, MH., Labianca, G., 2004. Group social capital and group effectiveness: The role of informal socializing ties. The Academy of Management Journal 47(6): 860-875.

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., 2014. How to kill inventors: Testing the Massacrator© algorithm for inventor disambiguation. Scientometrics 101(1): 477-504.

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.

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

Schilling, M.A., Phelps, C.C., 2007. Interfirm collaboration networks: The impact of large-scale network structure on firm innovation. Management Science 53(7): 1113-1126.

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.

Snijders, T.A.B., 2017. Stochastic Actor-Oriented Models for Network Dynamics. *Annual Review of Statistics and its Application* 4: 343-363.

Snijders, T.A.B., Borgatti, S.B., 1999. Non-Parametric Standard Errors and Tests for Network, Connections, 22: 161-70.

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

Ugander, J., Backstrom, L., Marlow, C., Kleinberg, J., 2012. Structural Diversity in Social Contagion. Proceedings of the National Academy of Sciences 109(16): 5962-5966.

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

Wang, M.C., Chen, M.H., 2016. The more, the better? The impact of closure collaboration network and network structures on technology-based new ventures' performance. R&D Management 46: 174-192.

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.

Zaheer, A., Bell, G.G., 2005. Benefiting from network position: Firm capabilities, structural holes, and performance, Strategic Management Journal 26: 809-825.

Tables and Figures

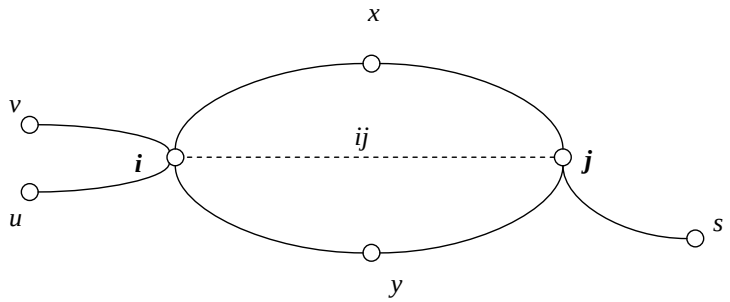


Figure 1: Example 1.

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.

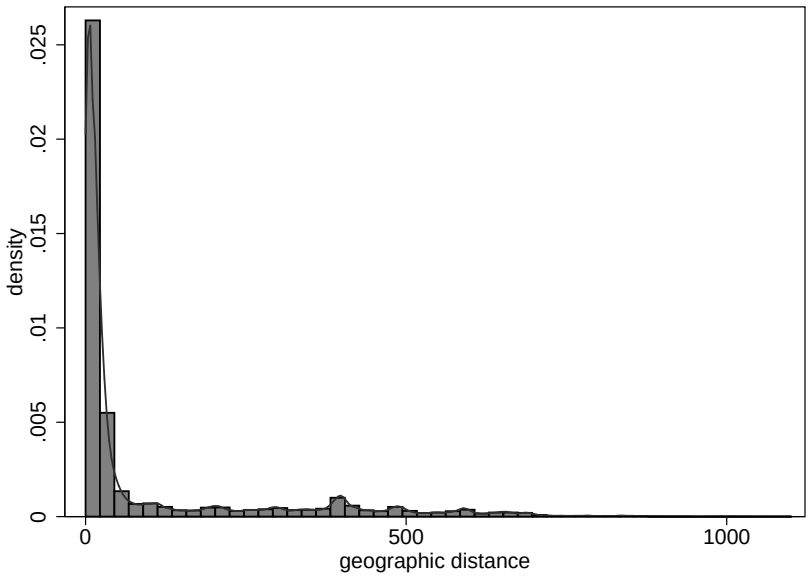


Table 3: Conditional logit on the occurrence of the first connection, all sample, five-year window network, linear detrending.

	1	2	3	4
non-common	0.0321***	0.0328***	0.0276**	0.0313*
	(4.26)	(4.33)	(2.35)	(2.01)
common	0.199***	0.204***	-0.529***	-0.255**
	(10.34)	(9.47)	(-4.31)	(-2.79)
geo distance		-0.00132***	-0.00143***	-0.000621***
		(-24.70)	(-25.15)	(-6.52)
Jaffe tech distance		-0.339***	-0.258**	-0.507***
		(-5.18)	(-2.77)	(-3.30)
public research			23.13***	21.83***
			(84.67)	(1312.63)
common applicant			31.71***	24.75***
			(8.30)	(133.39)
network controls	yes	yes	yes	yes
applicant controls	no	no	no	yes
observations	407,001	407,001	407,001	129,924

Notes: Dyadic clustered standard errors (t statistics in parentheses).
Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 4: Estimated coefficients of the number of non common neighbors on the probability to collaborate from given Monte Carlo simulations.

Generative Values	*Estimated coefficients*					
	Full sample			*Excl. Non-Meeting Pairs*		
	Model 1			Model 2		
$\beta^c = -0.53$	Mean	S.D.	# signif.	Mean	S.D.	# signif.
	-0.25	0.073	98	-0.41	0.204	48
	Model 3			Model 4		
$\beta^c = -1$	Mean	S.D.	# signif.	Mean	S.D.	# signif.
	-0.656	0.112	100	-0.989	0.241	99
	Model 5			Model 6		
$\beta^c = 0$	Mean	S.D.	# signif.	Mean	S.D.	# signif.
	0.189	0.0887	71	0.0849	0.347	7

Notes: The table reports the means and standard-deviations of the coefficient estimates β^c from 100 conditional logit estimations obtained. The column *# signif.* reports the number of times the coefficient is significant at the 5% level. In Models 1 to 4, the generative coefficient for β^{nc} is fixed to 0.028. In Models 5 and 6, it is kept to 0, as β^c. In Model 2, 4 and 6, the non meeting dyads a given year are excluded from the sample. All explaining variables are linearly detrended.

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

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.[32] This methodology has been presented much more extensively in a technical note authored 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. 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[33] 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, 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

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

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

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. These variables are assumed to be independent across agents ($\forall k, k' = 1, ..., K, \forall i \neq j : X_i^k \perp X_j^k$) and $\forall k = 1, ..., K; \forall i, j$, X_i^k and X_j^k have the same support (by construction). 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:

$$\Pr\left(i = j \left| X_i^k = x_j^k, \forall k = 1, ..., \bar{k} - 1 \text{ and } X_i^{k'} \neq x_j^{k'}, \forall k' = \bar{k}, ..., K \right.\right). \tag{8}$$

In principle, it could be possible to apply Bayes' rule to calculate (8). Applying Bayes rule, the probability in (8) is equal to

$$\frac{\Pr(i = j) \times \Pr\left(X_i^k = x_j^k, \forall k = 1, ..., \bar{k} - 1 \text{ and } X_i^{k'} \neq x_j^{k'}, \forall k' = \bar{k}, ..., K \,|i = j\right)}{\Pr\left(X_i^k = x_j^k, \forall k = 1, ..., \bar{k} - 1 \text{ and } X_i^{k'} \neq x_j^{k'}, \forall k' = \bar{k}, ..., K\right)}. \tag{9}$$

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 avoid 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 $X_i^{k'} \neq x_j^{k'}, \forall k' = 1, ..., \bar{k} - 1$, all other things remaining the same. This differentiates out $\Pr(i = j)$. If we make the additional simplifying assumption that the different variables are independent for any given agent across agents ($\forall k, k' = 1, ..., K, \forall i : X_i^k \perp X_j^{k'}$),[34] it can be shown that the similarity score is equal to:

$$\Delta(i, j) = \Pi_{k=1,...\bar{k}-1} \quad \frac{\left(1 - \varepsilon^k\right)}{\varepsilon^k} \times \Omega^k(i, j), \tag{10}$$

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 $\Omega^k(i, j) \equiv \left(1 - Pr\left(X_i^k = x_j^k\right)\right) / Pr\left(X_i^k = x_j^k\right)$, that is the probability that the two *ex ante* agents i and j have a

[34] This assumption is made for simplifying the exposition only. If, for instance, the correlation of the X_i^k with $X_i^{k'}$ variables (for all i) had to be considered, we would just have to consider the probability of jointly observing $X_i^k = x_j^k$ and $X_i^{k'} = x_j^{k'}$. Then, it is more a matter of computation. As it is shown below, in this disambiguation exercise, there is no need to introduce this since the results are already very good.

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 (10)). 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.[35] Then, a 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×patent occurrences. At this stage, the total number of *ex ante* agents corresponds to all the inventor×patent occurrences that can be observed (for instance

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

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 A1.

Table A1: 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.[36] 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 name and first name.[37] A precautionary conservative rule is arbitrarily adopted at this step: 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

[36]The full string, reporting the city and street address, is considered. The probability of two different persons with the same name and first name having 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.

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

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 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.[38] 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 agent 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 Type 1 (false positive) errors ϵ_1 and the share of Type 2 (false 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 $\bar{\Delta} = exp(12.49)$ minimizes $\phi(\theta)$ for a wide range of θ, between 0.09 and 0.64, and thus this is the chosen value for the threshold.

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.[39] The benchmark can also be used to assess the quality of the terminal results. If the most appropriate weighting scheme is $\theta = .1$,[40] the weighted

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

[39] It should be worth noting that, in order to solve the issue of homonymy between 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.

[40] It is indeed our preferred value because it avoids abusive aggregation of agents. Note that this preference does not constraint the disambiguation algorithm since the chosen threshold value for the similarity score is minimizing errors for a large and reasonable set of values of θ.

share of errors obtained is $\phi(.1) = 1.81\%$, which remains very low.

Appendix B: Robustness check regressions

Table B1: Conditional logit on the occurrence of the first connection, five-year window network, log detrending.

	1	2	3	4
non-common	0.0322***	0.0329***	0.0277**	0.0314*
	(4.28)	(4.34)	(2.36)	(2.01)
common	0.199***	0.204***	-0.527***	-0.254**
	(10.37)	(9.50)	(-4.31)	(-2.78)
geo distance		-0.00132***	-0.00143***	-0.000623***
		(-24.72)	(-25.17)	(-6.45)
Jaffe tech distance		-0.340***	-0.260**	-0.509***
		(-5.20)	(-2.79)	(-3.32)
public research			22.63***	21.83***
			(27.39)	(33.16)
common applicant			31.02***	24.75***
			(8.15)	(135.85)
network controls	yes	yes	yes	yes
applicant controls	no	no	no	yes
observations	407,001	407,001	407,001	129,924

Notes: Dyadic clustered standard errors (t statistics in parentheses).
Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table B2: Conditional logit on the occurrence of the first connection, five-year window network, quadratic detrending.

	1	2	3	4
non-common	0.0320***	0.0327***	0.0275**	0.0312*
	(4.24)	(4.31)	(2.34)	(1.98)
common	0.198***	0.203***	-0.531***	-0.256**
	(10.31)	(9.44)	(-4.31)	(-2.76)
geo distance		-0.00132***	-0.00143***	-0.000620***
		(-24.68)	(-25.14)	(-6.47)
Jaffe tech distance		-0.338***	-0.256**	-0.505***
		(-5.15)	(-2.75)	(-3.29)
public research			23.13***	21.83***
			(22.95)	(725.74)
common applicant			31.76***	24.76***
			(8.26)	(138.67)
network controls	yes	yes	yes	yes
applicant controls	no	no	no	yes
observations	407,001	407,001	407,001	129,924

Notes: Dyadic clustered standard errors (t statistics in parentheses).
Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table B3: Conditional logit on the occurrence of the first connection, ten-year window network, linear detrending.

	1	2	3	4
non-common	-0.00998	-0.00841	-0.00991	-0.00398
	(-0.68)	(-0.57)	(-0.46)	(-0.16)
common	0.115***	0.121***	-0.686***	-0.418***
	(5.33)	(4.92)	(-4.29)	(-3.29)
geo distance		-0.00122***	-0.00132***	-0.000536***
		(-21.54)	(-22.47)	(-5.43)
Jaffe tech distance		-0.00122***	-0.00132***	-0.000536***
		(-5.11)	(-2.90)	(-3.03)
public research			23.22***	22.15***
			(438.72)	(1470.51)
common applicant			35.82***	25.50***
			(7.02)	(74.01)
network controls	yes	yes	yes	yes
applicant controls	no	no	no	yes
observations	347,707	347,707	347,707	118,839

Notes: Dyadic clustered standard errors (t statistics in parentheses).
Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.
The observations before year 1988 were dropped to obtain a consistent ten-year window for each year considered.

Table B4: Conditional logit on the occurrence of the first connection, five-year window network, linear detrending, with the two-way clustered standard errors

	1	2	3	4
non-common	0.0321***	0.0328***	0.0276**	0.0313**
	(5.04)	(5.15)	(2.73)	(2.40)
common	0.199***	0.204***	-0.529***	-0.255***
	(10.67)	(10.86)	(-5.09)	(-3.19)
geo distance		-0.00132***	-0.00143***	-0.000621***
		(-28.14)	(-28.22)	(-7.24)
Jaffe tech distance		-0.339***	-0.258***	-0.507***
		(-5.28)	(-3.16)	(-3.79)
public research			23.13***	21.83***
			(113.25)	(54.27)
common applicant			31.71***	24.75***
			(9.72)	(147.829)
network controls	yes	yes	yes	yes
applicant controls	no	no	no	yes
observations	407,001	407,001	407,001	129,924

Notes: Dyadic clustered standard errors (t statistics in parentheses).
Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Appendix C: Network generation for the Monte Carlo experiments

Data are generated according to the theoretical model over 30 periods. One thousand agents are initially introduced and fifty new agents are added each period. Agents meet at random with a probability $p = 0.05$, and decide to form a link if $\Delta_{ij}(t) > 0$, with $\Delta_{ij}(t)$ defined in Equation 3. The network statistics are computed on the links that have been created between periods $t-5$ and $t-1$, consistently with the real data as network statistics are computed over the five previous years. The random term ε_{ij}^t is drawn from a centered Logistic distribution of unitary scale. The pair fixed-effect is defined as $\theta_{ij} = \gamma_i + \gamma_j + \xi_{ij}$, where γ_i and γ_j are specific to the agents and drawn from a Poisson law of mean $\lambda^\gamma = 2$, and where ξ_{ij} is specific to the pair and drawn from a Poisson law of mean $\lambda^\xi = 11$. Each time varying cost variable c_{ij}^t is drawn at from a Poisson law of parameter $\lambda^c = 23$. λ^c is greater than $\lambda^\xi + 2\lambda^\gamma$ so that on average the costs are higher than the benefits (i.e. $E(\Delta_{ij}^t) < 0$), meaning that it requires several meetings (on average) before getting connected.

At each period, the network evolves as follows:

1. Meeting step: Each non-connected pair of agents is selected with probability p.

2. Decision step: For each selected pair of agents, we

   - Compute the network statistics $\bar{\eta}_{ij}(g^t)$ and $\overset{\triangle}{\eta}_{ij}(g^t)$,
   - Generate the cost value c_{ij}^t and the error ε_{ij}^t,
   - Compute $\Delta_{ij}(t)$,
   - Create a link if $\Delta_{ij}(t) > 0$.

3. Entry step: 50 new agents enter the network.

Appendix D: Monte Carlo simulations with varying network generation parameters

We replicate the Monte Carlo simulations reported in Section 5 for different values of the generating parameters λ^{ξ} and λ^{c}. More precisely, we let the mean cost and the mean dyad fixed effect vary across a grid. We reduce and increase by 1 the values of the two parameters, thus replicating the MC simulations nine times. The different parameters lead to different expected probabilities to collaborate upon meeting. In Table 5, we report these expected probabilities and number of meetings prior collaboration, when two agents have no common nor non-common neighbors. In the baseline, the expected probability to collaborate upon meeting equals 11%. This means agents need to meet 9.5 times on average before a collaboration occurs. Both reducing the mean cost and augmenting the mean dyad fixed-effect by one increases collaboration probability upon meeting up to 18%, that is agents need to meet only 5.7 in average times before collaborating.

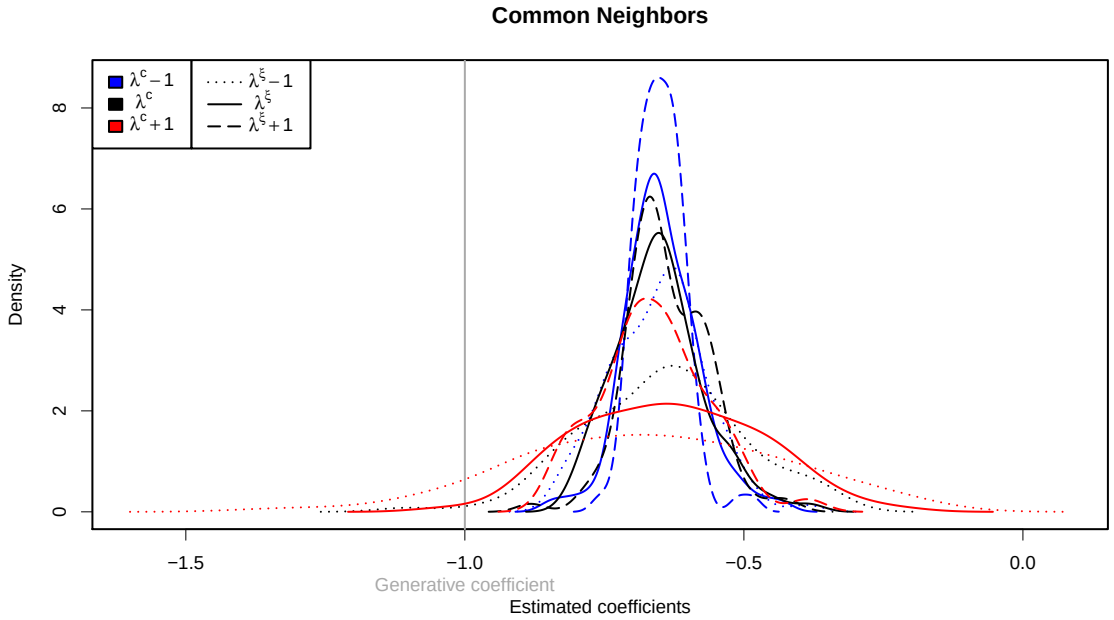
Table 5: Consequence of parameters variation: Expected probability to collaborate and expected number of meetings before collaboration.

Expected probability to collaborate				*Expected number of meetings before collaboration*			
	$\lambda^{\xi}-1$	λ^{ξ}	$\lambda^{\xi}+1$		$\lambda^{\xi}-1$	λ^{ξ}	$\lambda^{\xi}+1$
$\lambda^{c}-1$	0.1	0.13	0.18	$\lambda^{c}-1$	10	7.4	5.7
λ^{c}	0.076	0.11	0.14	λ^{c}	13.2	9.5	7.1
$\lambda^{c}+1$	0.059	0.082	0.11	$\lambda^{c}+1$	17	12.2	8.9

Notes: The expected number of meetings before a collaboration takes place is computed as $1+(1-p)/p$ with p the expected probability to collaborate. To obtain these probabilities, we generated 100,000 times the value of $\Delta_{ij} = (\gamma_i + \gamma_j + \xi_{ij}) - c_{ij}^{t} + \epsilon_{ij}^{t}$ and computed the average of $\mathbf{1}\{\Delta_{ij} > 0\}$. With: $\gamma_i \sim Poisson(2)$, $\gamma_j \sim Poisson(2)$, $\xi_{ij} \sim Poisson(x)$, $x \in \{\lambda^{\xi}-1, \lambda^{\xi}, \lambda^{\xi}+1\}$, $cost_{ij} \sim Poisson(t)$, $t \in \{\lambda^{c}-1, \lambda^{c}, \lambda^{c}+1\}$, $\lambda^{\xi} = 11$, $\lambda^{c} = 23$ and $\varepsilon_{ij}^{t} \sim Logit(1)$.

The distribution of estimated β^{c} from 100 conditional logit estimations for each variant of network costs and benefits, and with the same generative values of the network determinants than in Model 3-Table 4 are reported in Figure 3. Overall, patterns are very close to what is obtained in Section 5, exhibiting a similar downward bias in magnitude for β^{c}.

Figure 3: Distribution of estimates obtained with different values of the parameters.



Notes: The figure reports the distribution of the coefficient estimates β^c from 100 conditional logit estimations obtained, when $\gamma_i \sim Poisson\,(2)$, $\gamma_j \sim Poisson\,(2)$, $\xi_{ij} \sim Poisson\,(x)$, $x \in \{\lambda^\xi - 1, \lambda^\xi, \lambda^\xi + 1\}$, $cost_{ij} \sim Poisson\,(t)$, $t \in \{\lambda^c - 1, \lambda^c, \lambda^c + 1\}$, $\lambda^\xi = 11$, $\lambda^c = 23$ and $\varepsilon_{ij}^t \sim Logit\,(1)$. The generative coefficient for β^{nc} is fixed to 0.028 and the generative coefficient for β^c is -1.

Appendix E: Further Monte Carlo simulations

Errors in individuals identification

Dealing with individual data extracted from name registers, we had to play a "name game" described in Appendix A. Though we know identification errors are very limited, still the data suffer from some lack of precision (due to Type 1 errors) and recall (due to Type 2 errors). To appreciate the impact of such errors, we inject both types of errors in the simulated data before estimation. Type 1 errors are introduced by selecting randomly a fraction of individual IDs and merging each with another randomly drawn ID. Type 2 errors are introduced by splitting randomly chosen agents in two and each of their collaboration is then randomly assigned to one of the two "fake" agents. Though the network forms on the basis of real identities, network statistics and regressions are computed on data where IDs errors have been included. The results are reported in Table 6. All estimated common neighbors coefficients are very close to the ones of Model 3–Table 4. The largest difference is observed when a large fraction of Type 1 errors are introduced. Here the average of the estimated coefficients of common neighbors is significantly reduced in magnitude to a value of -0.40.

Table 6: Estimated coefficients obtained from 100 Monte Carlo simulations: Different ways to generate the data.

Generative		*Estimated Coefficients*					
Value		Type 1 errors		Type 2 errors		Alternative	Continuous
		1%	10%	1%	10%	Meeting	Updating
		Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
$\beta^c = -1$	Mean	-0.619	-0.397	-0.657	-0.656	-0.646	-0.564
	S.D.	0.112	0.0896	0.113	0.112	0.082	0.0903

Notes: The table reports the means and standard-deviations of the coefficient estimates from 100 conditional logit estimations. The data for each of these estimations consist in 100 different networks. The generative coefficient for β^{nc} is fixed to 0.028. The coefficient is significant at the 5% level in all regressions.

Alternative meeting process

In the baseline model, agents meet at random. We investigate the consequence of introducing an alternative meeting process in which ex ante connected agents in a "friendship network" are more likely to meet. When connected in this friendship network, agents are four times more likely to meet. The friendship network is assumed to be a small world created from a L1 lattice with degree 6, and a 10% rewiring probability. The results of the MC simulations are reported in Model 5–Table 6 and show estimated coefficients similar to the ones of Model 3-Table 4.

Continuous time

In the baseline model, all individuals take decisions at discrete points of time. To investigate the consequences of introducing continuous time, we assume that instead of all agents making their decisions simultaneously in one given period, now dyads make collaboration decisions sequentially as in a continuous time updating of the network. MC results are reported in Model 6–Table 6 where we see that the coefficient of common neighbors is slightly lower in magnitude than the coefficient of Model 3-Table 4. This implies that in the presence of continuous updating, the empirical methodology would still be able to detect the negative effect of common neighbors. Further, the continuous updating acts simply as adding further "noise" in the estimation.