The strategic formation of inter-individual collaboration networks. Evidence from co-invention patterns

Nicolas Carayol and Pascale Roux

ADIS, Université Paris Sud

BETA, CNRS, Université Louis Pasteur

April 2007

This version : June 2007

Abstract

This paper contributes to an emerging literature aiming to understand the behavioral patterns that lead to the formation of social networks. We introduce a strategic model of inter-individual collaborations formation which is a variation of the Connections model. Heterogeneous agents benefit from knowledge spillovers flowing through the network and bear the costs of maintaining their direct links. Such costs increase with geographic distance over a ring on which agents are located. We show that this simple model generates emergent networks that share the main structural properties observed in most real social networks. Moreover, we bring the model to co-invention networks data and find that the model fits quite well the data through the various structural measures used. In particular, it provides a consistent explanation of the observed asymmetry in the distribution of neighborhood sizes and of the high concentration of connections in local areas while some distant connections are formed.

JEL classification : D85 ; C63 ; O31

Keywords : Strategic network formation ; Inter-individual collaborations ; Small worlds ; Geography

Corresponding author. Nicolas Carayol, Université Paris Sud, Faculté Jean Monnet (Office A102), 54 Bvd Desgranges, F-92331 Sceaux. Tel : +33-140911843. Fax : +33-141138273. Email : nicolas.carayol@u-psud.fr.

1. Introduction

There is an increasing consensus in the economic literature to recognize that network structures significantly influence the outcomes of many social and economic activities. As recently highlighted in the theoretical economics literature, such networks are also often strategically shaped by the decentralized behaviors of the participating agents.[1] Nevertheless, this literature has not yet dedicated much attention to the full characterization of the endogenous networks and, most importantly, to the strategic conditions that may lead to the emergence of networks which might resemble real networks. Indeed, such real networks are usually much more complex than the structures discussed in the theoretical literature.[2]

A series of contributions showed that most social networks, though they are complex and heterogeneous in many respects, share some common structural features. Real networks generally have a very large connected component that links directly or indirectly most agents while a restricted proportion of agents remain within disconnected subsets (De Castro and Grossman, 1999). These networks are also proven to be very short in the sense that, when counting the minimal number of inter-individual social connections (that is the social distance) between them, agents are found to be in average very close the one to the others (Milgram, 1967). In the meantime, real networks are highly clustered : there is a high probability that an agent's neighbors are also neighbors together (Newman, 2001 ; Kogut and Walker, 2001). It has also been shown that agents are highly unequal with respect to the size of their neighborhood (degree) : few agents have many connections while many agents have few links (Price, 1965 ; Redner, 1998). In addition to those structural properties, some authors started to study the spatial distribution of networks and to show that some of them tend to correlate with geography (e.g. Gastner and Newman, 2006). Indeed, earlier studies have long documented, in various contexts, the strong inverse relationship between geographical distance and social ties formation (e.g. Bossard, 1932 ; Zipf, 1946 ; Festinger et al., 1950 ; Caplow and Forman, 1950 ; Kono et al., 1998 ; Sorenson and Stuart, 2001).

Nevertheless, it is still a challenge on the agenda to understand the social and economic conditions that lead to the formation of networks that exhibit such properties, and beyond, to explain why some social networks exhibit more specific structural attributes. Actually, several models of network formation, which incorporate, in various extents, randomness and ad-hoc settings, have been introduced (Erdös and Rényi, 1960 ; Watts and Strogatz, 1998 ; Price, 1965 ; Barabasi and Albert, 1999). Nevertheless, these models only explain some of the structural features usually observed in social networks.[3] Moreover, none of them present networks formation

[1] Models of strategic network formation encompass various contexts such as imperfect information diffusion in networks (Jackson and Wolinski, 1996 ; Bala and Goyal, 2000), job-contact networks (Calvó-Armengol, 2004), oligopolies and R&D collaborations (Goyal and Moraga, 2001 ; Goyal and Joshi, 2003), buyer-seller networks (Kranton and Minehart, 2000), etc.

[2] Such as the empty network, the complete network or even the complete star network.

[3] With the exception of some "mixture" models. For instance, Jackson and Rogers (2007) have proposed a model in which agents are attached to other agents in two manners. Some neighbors are picked at random among all other agents while some others are picked by searching locally through the current structure of the network.

as the outcome of the explicit decentralized behavior of agents.

In this paper, we make a first step in this direction by confronting the strategic approach of network formation with social networks data. We try to appreciate the extent to which the structure of interpersonal research collaborations can result from a decentralized process of strategic network formation. So as to model in a simple and stylized manner the impact of inter-individual collaboration networks on individual payoffs, we rely upon a variation of the connections model (Jackson and Wolinsky, 1996). Moreover, the strategic formation of networks occurs in a dynamic process close to the one introduced by Jackson and Watts (2002). A numerical methodology, which has been first introduced in Carayol and Roux (2004, 2006), allows us to compute emergent stable networks in the long run.

In our model, agents benefit from positive externalities from other agents given that there exists at least a sequence of continuous inter-individual connections between them on the relational network. Moreover, the strength of the positive externality decays geometrically with social distance. Thus, this model seems particularly relevant if one aims to describe knowledge networks since the positive externality can account for knowledge spillovers and there is a decay parameter that tunes the quality of knowledge transmission through the network connections. The kind of knowledge the transmission and impact of which this model intends to capture, is all together sensitive, advanced and mainly tacit. Such knowledge is usually "naturally" disclosed to the selected direct individual partners in the research process : if a link exists, then it is (imperfectly) conducive to knowledge transmission.[4] These features of knowledge transmission in collaboration networks are consistent with the very recently available empirical evidence which shows that inter-individual connections are the support of knowledge spillovers and that the probability of observing a knowledge flow decreases sharply with social distance between individuals (e.g. Breschi and Lissoni, 2006a and Singh, 2006). Moreover, because agents do not benefit from someone else by being connected to him by multiple paths, the model stresses that agents have greater incentives to sustain links with agents to whom they are not indirectly connected. This is also consistent with an extensive empirical literature which shows that sustaining such "bridging links" are associated with higher incomes (a higher probability to find a job for Granovetter, 1973 ; a higher rate of good ideas generation for Burt, 1992 ; and obtaining faster promotions for Burt, 1997).

Our model further introduces geography in a very simple and stylized way. Agents are assumed to be located at equidistant intervals on a circle of a given dimension. As in Johnson and Gilles (2000) whose world is linear, we assume that link formation costs increase linearly with geographic distance. Our underlying assumption is that geographically distant research collaborations are as effective as close connections *per se* but do impose more monitoring and costly interactions to be achieved. This feature stresses a specific impact of geography on network formation, that traces back to Stouffer (1940) and Zipf (1946), who, thought they did not

[4] Thus our approach does not aim to model the strategic disclosure of knowledge on the top of the strategic formation of networks, a feature that would be more relevant to stress collaborations between organizations.

explicitly model the (individual) benefits and costs of connections formation, assume and interpret their results as geographical distance affecting positively connection costs. Another feature of our model lies in the introduction of agents' heterogeneity in their ability to sustain direct connections at given costs.

Very few data are available to track inter-individual research collaboration patterns in a reliable, systematic and quantitative manner over a long period of time. Our empirical evidence is based on the relational information contained in all European Patent Applications, at least one inventor of which declared an address in France and which occurred over the period 1977-2003.[5] We build the co-invention network by allocating a connection between two French individuals if they both appear among the inventors of the same patent, at least once. A connection in this network reveals a strong and deliberate collaboration between two persons. We believe that this is a restricted but acceptable manner to evidence research collaborations, which are the support of advanced and sensitive knowledge transmission. From a methodological point of view, the procedure is similar to the one performed for measuring scientific collaboration networks from data on co-authorships of scientific publications (Newman, 2001, Barabasi et al. 2002). Very recently, Goyal et al. (2006) also apply this method to study the scientific collaboration networks in economics from papers co-authorships relying upon Econlit database. Our dataset counts more than 114,000 patents, invented by nearly 98,000 French inventors among whom more than 76,000 have invented a patent with at least one other French inventor.

Our main results are the following. We show that, if the population size, the heterogeneity across agents and the geographical range are sufficiently large, the theoretical model generates networks that share all the standard structural properties of most real social networks (exposed above) for a large range of intermediary values of the decay parameter (that tunes the strength of the positive externalities). Next, we provide a first global analysis of the whole French co-invention network. Given its high level of disaggregation in components, we propose a within-components methodology to compare empirical with theoretical networks. We find that they exhibit close structural measures. Moreover, most of the predictions on the effects of the different parameters (such as population size or geographical dimension) on networks structure are corroborated in the data. The theoretical model also generates emergent networks whose degree distribution is highly right-asymmetric in a comparable extent as in empirical networks. Lastly, the theoretical model shows that most connections are formed in the local geographic environment while still a few distant connections are formed, a property shared by the empirical co-invention networks.

The paper is organized as follows. The next section presents the model of network formation. Section 3 presents a series of results on the structure of emergent networks as compared to the standard properties that real social networks share. Section 4, presents the data and compares the (theoretical) emergent networks with the empirical networks. The last section concludes.

[5]The data, nicely provided by Francesco Lissoni, are an extraction of the EP-INV database produced by CESPRI-Universita Bocconi.

2. The strategic formation of inter-individual research collaboration networks

In this section, we begin with basic notions on non directed graphs. We then introduce our model and discuss the individuals' incentives to form collaborative links. Finally, we turn to the dynamic perturbed process that leads to networks formation.

2.1 Networks

We consider a finite set of n agents, $N = \{1, 2, ..., n\}$ with $n \geq 3$. Let i and j be two members of this set. Agents are represented by the nodes of a non-directed graph the edges of which represent the links between them. A link between two distinct agents i and $j \in N$ is denoted ij. A graph g is a list of non ordered pairs of connected and distinct agents. Formally, $\{ij\} \in g$ means that ij exists in g. We define the complete graph $g^N = \{ij \mid i, j \in N\}$ as the set of all subsets of N of size 2, where each player is connected with all others. Let $g \subseteq g^N$ be an arbitrary collection of links on N. We define $G = \left\{g \subseteq g^N\right\}$ as the finite set of all possible graphs between the n agents. The empty graph, denoted $g^\emptyset$, is such that it does not contain any links.

Let $g' = g + ij = g \cup \{ij\}$ and $g'' = g - ij = g \backslash \{ij\}$ be respectively the graph obtained by adding ij and the one obtained by deleting ij from the existing graph g. The graphs g and g' are said to be *adjacent* as well as the graphs g and g''. Let a *path* in a non empty graph $g \in G$ connecting i to j, be a sequence of edges between distinct agents such that $\{i_1i_2, i_2i_3, ..., i_{k-1}i_k\} \subset g$ where $i_1 = i$, $i_k = j$. Let $i \longleftrightarrow_g j$ be the set of paths connecting i and j on graph g. The set of *shortest paths* between i and j on g noted $i\widetilde{\longleftrightarrow}_g j$ is such that $\forall k \in i\widetilde{\longleftrightarrow}_g j$, then $k \in i \longleftrightarrow_g j$ and $\#k = \min_{h \in i \longleftrightarrow_g j} \#h$. The *geodesic distance* between two agents i and j is the number of links of a shortest path between them : $d(i,j) = d_g(i,j) = \#k$, with $k \in i\widetilde{\longleftrightarrow}_g j$. When there is no path between i and j then their geodesic distance is conventionally infinite : $d(i,j) = \infty$.

For any g, we then define $N(g) = \{i \mid \exists j : ij \in g\}$, the set of agents who have at least one link in the network g. We also define $N_i(g)$ as the set of i's direct neighbors, that is : $N_i(g) = \{j \mid ij \in g\}$. The cardinal of that set $\eta_i(g) = \#N_i(g)$ is called the *degree* of node i. $N_i^2(g) = \{j \mid j \neq i,\ d(i,j) \leq 2\}$ is the set of agents who are either directly connected or indirectly connected at distance two to i. The total number of links in the graph g is $\eta(g) = \#g$. Let us denote $\hat{\eta}(g) \equiv \frac{1}{n}\sum_{i \in N} \eta_i(g) = 2\eta(g)/n$, the average degree of g. A network g is said to be connected if for any pair of distinct agents $i, j \in N(g), i \longleftrightarrow_g j \neq \emptyset$. A *component* C is a subset of $N(g)$ such that for any pair $i, j \in C$, $i \neq j$, $i \longleftrightarrow_g j \neq \emptyset$ and, for any $i \in C$ and $k \notin C$, $i \longleftrightarrow_g k = \emptyset$. It follows that the set of non isolated agents $N(g)$ is such that $N(g) = \cup_{k=1,...K} C_k$, with K the number of components C_k, $k = 1, ..., K$.

Finally, let us assume that agents are equidistantly located on a circle. Without loss of generality, agents are ordered according to their index, such that i is the immediate geographic neighbor of agent $i+1$ and agent $i-1$ but agents 1 and agent n who are neighbors since they close the ring. We then define a operator denoted $l(i,j)$ that simply counts the number of inter-individual

intervals on the ring separating i and j. It is given by $l(i,j) = \min\{|i-j|\,; n-|i-j|\}$. Now assume that the maximum geographical distance on the circle is given by S (the geographical dimension of the circle). Then, the geographic distance between i and j is simply given by $s_{ij} = l(i,j)S/\lceil n/2\rceil$,[6] with $\lceil n/2\rceil$ the smallest integer higher than or equal to $n/2$. $S/\lceil n/2\rceil$ can be viewed as an inverse measure of the density of the population.

2.2 The model

We now propose a simple model of strategic formation of inter-individual collaboration networks which intends to capture the impact of knowledge diffusion in networks on agents payoffs. We assume that interpersonal connections are conducive to knowledge or ideas diffusion. In the spirit of Jackson and Wolinski (1996), we consider that agents benefit from their costly direct connections but also from indirect links through the relational neighborhood of their partners. We propose the following specification for agent i's payoffs derived from the position he occupies in network g :

$$\pi_i(g) = \sum\nolimits_{j\in N\setminus i} \delta^{d(i,j)}\omega_{ij} - \sum\nolimits_{j\in N_i(g)} c_{ij}. \tag{1}$$

The first argument of the payoffs accounts for the gross payments one agent would gain from knowledge diffusion through its direct or indirect connections, assuming no time lag for simplicity. ω_{ij} denotes the "intrinsic value" of individual j knowledge to individual i. For simplicity, it is assumed that such value is fixed across agents and normalized to unity : $\forall i \neq j : \omega_{ij} = 1$. There is a decay parameter $\delta \in]0;1[$ which gives the share of knowledge effectively transmitted through connections. The externality is geometrically decreasing with geodesic distance since δ is less than the unity. The empirical evidence on knowledge spillovers from patent citations support this assumption. In particular, Singh (2005) and Breschi and Lissoni (2006a) showed that the probability of patents citations decreases sharply with social distance between patents inventors. The value of δ shall be associated with the characteristics of knowledge : communication quality is likely to decrease with the degree of tacitness of knowledge while it would increase with the codification of knowledge.

The second part of the right hand side of equation (1) describes the costs of sustaining direct links, with each direct link costing c_{ij} to i. The cost of each direct link is given by :

$$c_{ij} \equiv a_i s_{ij} = a_i \frac{l(i,j)}{\lceil n/2\rceil} S. \tag{2}$$

Agents are assumed to be heterogeneous in their link formation costs in the sense that the parameters a_i are independently distributed according to the uniform distribution $\forall i \in N(g), a_i \sim U[\underline{a},\overline{a}]$ with strictly positive support and so that its mean equals unity : $(\overline{a}-\underline{a})/2 = 1$. This implies that the costs supported by two agents to be linked together may differ due to agents' heterogeneous abilities ($c_{ij} \neq c_{ji}$). Moreover, we assume that the costs linearly increase with the geographic distance separating agents on the circle metric s_{ij}. The assumption according to

[6] In the circle metric, the maximum number of interindividual connections is given by : $\max_{i,j\in N} l(i,j) = \lceil n/2\rceil$.

which link costs increase with distance can be justified by the fact that closely located agents incur lower costs to establish communications and to coordinate. Indeed, when agents are geographically distant, face to face interactions imply higher transporting costs and time.[7] Moreover, geographic distance could also generate higher monitoring costs (e.g. Lerner, 1995).

2.3 Incentives to form bilateral connections

We now turn toward the formation of networks. It is assumed that the relational network emerges from the willingness of agents to form links in order to benefit from knowledge flows. As a consequence, agents try to maximize the value generated from direct and indirect connections, avoiding superfluous connections. Nevertheless, agents are myopic in the sense that they take their decisions on the basis of the immediate perceived impact of such decisions on their payoffs.[8] We also assume that, when agents take their decisions to form links, they only consider knowledge flows at geodesic distance less or equal to two ($d(i,j) \leq 2$). In other words, they ignore (or neglect) the indirect connections at distance greater or equal to three.

Let us examine the individuals' incentives to form direct connections that are the expected reward an agent i would get in sustaining a connection with some other agent j, $\Delta_i(g, ij)$. For that purpose, we define $\widehat{N}_{ij}(g) = N_i^2(g) \cap (N_j(g) \cup \{j\})$ the set of agents who are both at a distance less or equal to two from agent i and in the direct neighborhood of agent j, including j himself. Notice that $\widehat{N}_{ij}(g) \neq \widehat{N}_{ji}(g)$. Let $\widehat{\eta}_{ij}(g)$ be the cardinal of this set. We then define the function $\Delta_i(\cdot,\cdot)$ as follows :

$$\Delta_i(g, ij) = \delta + \delta^2 \left(\eta_j(g) - \widehat{\eta}_{ij}(g)\right) - a_i s_{ij}. \tag{3}$$

The incentive of agent i to form a link with j increases with the decay parameter δ, with the size of j's relational neighborhood η_j and decreases with $\widehat{\eta}_{ij}(g)$ which accounts for the overlapping of the two agents' neighborhoods (from i's point of view, since $\widehat{\eta}_{ij}(g) \neq \widehat{\eta}_{ji}(g)$). Thus there is a disincentive to form a link with an agent who is already connected to some other agents whom one already benefits from (and thus to form triangles). In other words, the model stresses higher incentives to form relations with agents with whom they are not indirectly connected (through one intermediate agent here), an assumption which is consistent with the empirical literature on "structural holes" (Burt, 1992). Link formation incentives also decrease with agent i's idiosyncratic costs to sustain connections a_i, and with the geographic distance separating i and j on the circle metric s_{ij}. Through the geographic distance (reminding that $s_{ij} = l(i,j)S/\lceil n/2 \rceil$), S plays negatively on incentives since n is held constant, it decreases the density of agents on the geographical space and thus increases the link formation costs. On the

[7] This seems to still hold despite the introduction of Internet technologies since they are complementary to face to face interactions (Gaspar and Glaeser, 1998).

[8] This assumption is standard in the literature on network formation (e.g. Jackson and Wolinski, 1996; Jackson and Watts, 2002). It is well designed for "not too small" networks analyses since introducing agents anticipations on others' further behaviors requires enormous computation capabilities that grow exponentially with the size of the population. It is only when the population is rather small that one can reasonably assume agents are farsighted.

opposite, incentives increase with n since, S being held constant, the larger the population, the greater the agglomeration of agents (agents are closer the one to the others on the circle metric) and so the lower the costs.

2.4 Dynamic network formation

Following Jackson and Watts (2002), we assume that, at each period, two agents $i, j \in N$ are randomly chosen with a given constant non null probability p. If these agents are already connected, they consider whether they may unilaterally severe the link or bilaterally keep it. If they are not directly connected, they consider whether they should add this connection or stay disconnected. As in Jackson and Wolinski (1996), we assume that the formation of a link between two agents requires the consent of both of them, but not its deletion, which can unilaterally emanate from one of them. Formally, writing g_t the network at discrete time period t, the dynamic process can be described as follows :

$$\begin{array}{lll} \text{i) if } ij & \in & g_t : ij \in g_{t+1} \text{ iff } \Delta_i (g_t - ij, ij) \geq 0 \text{ and } \Delta_j (g_t - ij, ij) \geq 0, \\ \text{ii) if } ij & \notin & g_t : ij \in g_{t+1} \text{ iff } \Delta_i (g_t, ij) \geq 0 \text{ and } \Delta_j (g_t, ij) > 0, \text{ or } \Delta_i (g_t, ij) > 0 \text{ and } \Delta_j (g_t, ij) \geq 0. \end{array} \tag{4}$$

Small but non vanishing random perturbations affect agents' decisions in creating, maintaining or deleting links. These perturbations may be understood as mistakes or as random experiments. We propose to let such an error term decrease in time according to the following simple rule :

$$\varepsilon_t = 1/(t+1) + \overline{\varepsilon}. \tag{5}$$

This rule ensures that a significant noise affects the dynamics in the beginning while it decreases monotonically with time down to a small strictly positive limit : $\lim_{t \to \infty} \varepsilon_t = \overline{\varepsilon}$. Agents are likely to make less and less errors through time though still a very small error probability persists in the long run.

The evolution of the system at any time t only depends on the present state of the system given by the graph structure g_t. The stochastic process is thus Markovian. The evolution of the system $\{g_t, t > 0\}$ can be described by the time-varying probability matrix $(\mathrm{P}(\varepsilon_t))$ describing the one-step transition probabilities at each period t between all possible states of the finite state space G. According to Robles (1998), the long run equilibrium $\psi(\overline{\varepsilon})$ of such time-inhomogeneous Markov chain exists, is unique and is equal to the equilibrium of the Markov chain perturbed by the constant error $\overline{\varepsilon}$.[9] It is then ergodic. This property is interesting since it renders numerical experiments more tractable in order to examine with good confidence the long run behavior of the system (Vega-Redondo, 2006). We label the networks on which the process stabilizes in the long run as *emergent networks*. The definition follows.[10]

[9] See Proposition 3.1 of Robles (1998, p. 211).

[10] Notice that the set of emergent networks is broader than the set of stochastically stable networks (Young, 1993) which is included in $\widehat{G}$ (cf. Definition 1). If we label $\mu(\overline{\varepsilon})$ the (unique) stationary distribution of the time-

Definition 1 *A network $g \in G$ is emergent if its probability of occurrence ($\psi_g(\overline{\varepsilon})$) in the long run equilibrium of the stochastic process described by the transition matrix $P(\varepsilon_t)$ is strictly positive. The set of emergent networks is* $\widehat{G} = \left\{ g \in G \,|\, \psi_g(\overline{\varepsilon}) > 0 \right\}$.

3. The structure of emergent networks

In this section, we analyze the structure of the emergent networks formed by the stochastic process presented in (4) and (5). We first provide some measures of the network structures that have been mainly used to characterize empirical social networks. These statistics are then used to characterize the architecture of the (theoretical) emergent networks for various values of the model parameters. Of particular interest to us, are the conditions under which the networks that formed would share the properties of many real social networks.

3.1 Measuring networks

To study the structural properties of emergent networks, we compute several dedicated statistics. The first index is the *average distance* (or average path length) between any two (directly or indirectly) connected agents of the network. It is given by :

$$d(g) = \frac{\sum_i \sum_{j \neq i} d(i,j) \times 1\{i \leftrightarrow_g j \neq \emptyset\}}{\#\{i,j \,|\, i \neq j \in N, i \leftrightarrow_g j \neq \emptyset\}}, \tag{6}$$

if $\eta(g) > 0$, with $\#\{\cdot\}$ denoting the cardinal of the set defined into brackets and $1\{\cdot\}$, the indicator function that is equal to unity if the condition into brackets is verified and zero otherwise. This index allows us to appreciate the extent to which directly or indirectly connected agents are distant in the relational network.

The second index is the *average clustering*. It indicates the extent to which neighborhoods of connected agents overlap or, in other words, the propensity with which the neighbors of an agent are also neighbors together is high. It is given by :

$$c(g) = \frac{1}{n} \sum_{i \in N; \eta_i(g) > 1} \frac{\#\{jl \in g \,|\, j \neq l \in N_i(g)\}}{\#\{j,l \,|\, l \neq j \in N_i(g)\}}. \tag{7}$$

The two indicators presented above are affected by the average degree of the network ($\hat{\eta}(g)$) that is likely to vary with δ and the other parameters of the model. Therefore, these indicators are somehow biased and we must find an efficient control for average degree.[11] In the spirit of Watts

homogeneous Markov chain associated with transition matrix $P(\overline{\varepsilon})$, this claim is formally : for all g such that, if $\lim_{\overline{\varepsilon} \to 0} \mu_g(\overline{\varepsilon}) > 0$ then $\psi_g(\overline{\varepsilon}) > 0$. This can be easily proved by recalling the Freidlin and Wentzell (1984) theorem that states : $\forall g$, $\psi_g(\overline{\varepsilon}) = \mu_g(\overline{\varepsilon})$ is of the form $\psi_g(\overline{\varepsilon}) = v_g(\overline{\varepsilon}) / \sum_{g'} v_{g'}(\overline{\varepsilon})$ with $v_g(\overline{\varepsilon})$ a polynomial in $\overline{\varepsilon}$.

[11] For instance, it is easy to see that the density of network affects crucially the average distance of the networks : more dense networks are likely to exhibit a shorter average path length. Without any control for density, one cannot know whether a network is shorter thanks to the structural allocation of links or simply thanks to a higher density. This is also true for the average clustering.

and Strogatz (1998), we associate *control random graphs* to emergent networks characterized by the same number of agents and links (and thus by the same average degree). Such random networks are simply built by allocating a given number of edges to randomly chosen pairs of agents (Erdös and Rényi, 1960). For each given number of edges of emergent networks, the average distance and the average clustering are numerically computed and averaged over $1,000$ of such random graphs. Thus, instead of looking at $c(g)$, where g is an emergent network, we compute the ratio $\widetilde{c}(g) = c(g)/c\left(g^{rd}\right)$, where $c\left(g^{rd}\right)$ denotes the mean average clustering of the $1,000$ random networks that have exactly the same average degree as g. Similarly we compute $\widetilde{d}(g) = d(g)/d\left(g^{rd}\right)$.

Those two ratios can be used to define the small world property *à la* Watts and Strogatz (1998).[12] Such a structure is characterized by the two following properties :

$$\widetilde{c}(g) \gg 1 \quad \text{and} \quad \widetilde{d}(g) \approx 1. \tag{8}$$

This means that small world networks are simultaneously highly clustered as compared to random graphs and that their average distance is close to the one of random graphs which are known to exhibit very short average path length.

Though it has been proven that most real social networks are small worlds *à la* Watts and Strogatz, two additional properties have also been emphasized in the literature as commonly observed in social networks. The former is the existence of a very large component, the size of which is much larger than the size of second largest component. More its size is often approaching the size of the whole population (Newman, 2001). The last property is a high inequality in the size of agents' neighborhoods : few agents have many connections while many agents have few links. Such a property can be studied using the distribution of the nodes' degrees in a graph g, $\rho(k)$, defined as the fraction of agents having k links in g. It is defined as :

$$\rho(k) = \frac{1}{n}\sum_{i \in N} 1\{\eta_i = k\}, \tag{9}$$

for all $k = 0, ..., n-1$. Again, if links where allocated fully randomly, we know, since Erdös and Rényi (1960), that this distribution would be Poisson and consequently its variance would be equal to its mean, which is the average degree $\hat{\eta}(g)$. On the contrary, most real networks exhibit a long right tail that is a very unequal distribution of links among agents.[13] A usual manner to compute the asymmetry of a distribution is by relying on its Gini coefficient. Assume agents are given new indexes given the relative size of their neighborhoods so that $i < j$ iff $\eta_i < \eta_j$. Then, the Gini coefficient is given by :

$$Gi(g) = 1 - \sum_{i=1,...,n} \frac{\rho(\eta_i(g))\,(\theta_i(g) + \theta_{i-1}(g))}{\theta_i(g)}, \tag{10}$$

[12] Watts and Strogatz (1998) presented an experiment allowing to build small world networks by rewiring a given proportion of edges from an initial regular lattice and reallocating them randomly. Notice that their aim was not to model how such networks are formed by the agents' decentralized behaviors.

[13] As a matter of fact, since Barabasi and Albert (1999), many scholars have demonstrated that the degree distribution of social networks can be approximated by a power law. See the next section and Table 2 for such an application on co-invention empirical data.

with $\theta_i(g) = \sum_{j=1,\ldots,i} \rho\big(\eta_j(g)\big)\,\eta_j(g)$ and $\theta_0(g) = 0$. It is equal to twice the area between the Lorenz curve and the 45° straight line crossing the horizontal axis at zero. Notice that, if all agents have the same number of nodes, that is agents are all equal, then the Gini coefficient is null. When inequality is maximal, then the Gini coefficient equals one.

3.2 Results

The various statistics are computed on networks that are on the support of the unique limiting stationary distribution $\psi(\overline{\varepsilon})$, obtained through a series of numerical experiments.[14] In order to provide some intuition on the impacts of the various parameters of the model, we consider three population sizes ($n^1 = 20$, $n^2 = 50$ and $n^3 = 100$), as well as two geographic sizes ($S^1 = 1$ and $S^2 = 2$) and two amplitudes for agents' heterogeneity ($a_i \sim U[0.5, 1.5]$ and $a_i \sim U[0.1, 1.9]$). Notice that, since the means of the a_i are by definition always equal to unity, the two distributions only differ in their variances, and so we will refer to the two heterogeneity distributions with respect to their standard deviation σ^1 and σ^2 (with $\sigma^2 > \sigma^1$). It would be cumbersome to present the twelve configurations of the different values of the parameters. Since it is likely that when n increases, both the heterogeneity of the population and the size increase, and since the remaining configurations do not bring more relevant information, we will limit the exposition of our results to the following four configurations of parameters values : $\big(n^1, S^1, \sigma^1\big)$, $\big(n^2, S^1, \sigma^1\big)$, $\big(n^2, S^2, \sigma^2\big)$ and $\big(n^3, S^2, \sigma^2\big)$. Finally, for each of these configurations, 1,000 simulations are performed with randomly drawn values of $\delta \in]0; 1[$ so as to fully explore the impact of δ on the structure of emergent networks.

The ratios ($\widetilde{c}\,(g)$ and $\widetilde{d}\,(g)$) as well as the Gini coefficient $Gi(g)$ and the average degree ($\hat{\eta}(g)$) are plotted in Figure 1. We can observe that the architecture of emergent networks strongly vary with the parameters value.

When δ is very close to zero, emergent networks are empty and thus their average degree $\hat{\eta}(g)$ is null. A very high value of the Gini coefficient is obtained when δ is small and both heterogeneity and the geographical dimension of the ring (S) are high because then many agents remain not connected while a few have some connections. Inequality is then very high since a few agents accumulated all the graph connectivity. It is only in these configurations that the model generates isolated agents and eventually several components. Otherwise, all agents always belong to a unique component. As δ increases, the incentives to form collaboration links also increase and so does the average degree. The average distance ratio reaches a maximum value when $\delta \approx 0.1$. Networks are then only locally connected since the incentives to form distant connections are not sufficient. As a consequence, it takes in average many inter-individual connections to reach some other agent potentially located far in the geographical ring. This also explains why a very high clustering ratio is then reached (despite the disincentive to form triangles) : agents connect

[14] We also fixed the limit error term $\overline{\varepsilon}$ to 10^{-4}. Moreover, all experiments are stopped at $T = 20,000$, period after which the process is proven to have surely stabilized on a given pairwise stable state with this model. A network is said to be *pairwise stable* if no incentive exists for any two agents to form a new link or for any agent to break one of his existing links (Jackson and Wolinski, 1996).

to their nearest geographical neighbors and so clustering is achieved in the local geographical areas.

When δ increases from 0.15, the average distance ratio straightly decreases to become equal or below unity at values of δ which range from 0.2 to 0.35 depending on the parameters. For these values of δ, the clustering ratio also decreases, but in a much smoother slope. The average clustering of emergent networks still ranges there from two to six times greater than the clustering of their random graphs controls. Thus, in this region of δ, the property expressed in (8) is clearly verified. A small world property *à la* Watts and Strogatz is obtained since some agents find sufficient incentives to form (geographically) distant connections. These agents, who are also much more densely connected than others, are likely to be precisely the ones who bear the lower individual costs of forming links (a_i). It is interesting to notice that increasing simultaneously agents' heterogeneity and S preserves a high clustering. Nevertheless, as Figure 2 illustrates, this high clustering tends then to be achieved in different manners depending on these parameters. When σ and S are high, it is more the intermediary of some central agents that allows the overlapping of neighborhoods since local agents are now only minimally locally connected (only connected to their two nearest neighbors). When σ and S are lower, the high clustering is rather due to the overlapping of local connections. In all configurations of n, σ and S, it is in this region of δ that the Gini coefficient is the largest if we do not consider the region $\delta \lesssim 0.15$.

When δ increases again from 0.35, the average distance ratio remains below unity while the average clustering ratio still decreases. Indeed, the more δ approaches the unity, the more δ^2 tends to δ, and so the stronger the disincentive to form triangles. This also explains why the average degree tends to stabilize or even to decrease either when the population is not large, or when the heterogeneity and the geographical dimension are low. Indeed, as δ increases, the incentives to form connections and the disincentives to form triangles increase simultaneously, each affecting average degree in an opposite manner. When either n, or σ and S are high, the disincentives to form triangles are less effective because the formation of connections is then more specific to individuals and less sensitive to variations in δ. As a consequence, the region of δ for which the four properties that characterize a small world are obtained tend to be larger when the geographical dimension and agents' heterogeneity are greater. For instance, when $S = S^2$, and $\sigma = \sigma^2$, not only the Gini coefficient remains high (thanks to the high heterogeneity) but also does the average clustering ratio up to $\delta = 0.7$. Lastly, as expected from the incentives to form connections given in equation (3), the density is greater when the population is larger and when the geographical size is smaller. This simply derives from the fact that costs of link formation depend crucially on the density of the population on the geographical space : either if one reduces the geographical size of the ring, or if one increases the number of agents, agents are geographically more proximate the one to the others and thus link formation costs are reduced.

Therefore we can conclude that when δ ranges from 0.2 to 0.35, the standard structural properties that characterize most real social networks (small average distance ratio, large clustering ratio, existence of a very large component and a high inequality in degrees) are verified for all explored values of the other parameters (n, S and σ). Moreover, these properties are also

obtained in a much larger range of δ (when δ is between 0.2 and 0.8), when n, S and σ are sufficiently large.

4. Are inter-individual collaborations strategically formed ? Evidence from co-invention networks

The aim of this section is to investigate whether our model of inter-individual collaborations formation offers acceptable predictions of some observed empirical patterns. Few data are available to realize such an experiment. Our application concerns the network of French co-inventors, built from patent data over a 25 years period. We will consider inventors as connected nodes, i.e. as collaborators, in a non directed relational graph if they have co-invented a patent together.[15]

The detailed study of the structural properties of co-invention networks, which remain ignored, should allow us to appreciate whether collaboration networks may have been strategically built by agents as described in the theoretical model presented in Section 2. For that purpose, the global structural properties of the (empirical) co-invention networks are first examined. Next, a methodology is developed to compare emergent (or theoretical) networks to empirical ones. This methodology can be designated as a within-components approach according to which agents' strategies can be, to some extent (and given a selection of -non active anymore- components), understood independently of the agents that do not belong to their component as they are in the theoretical model. A specific focus is made on within-components degree distribution as well as the relation between networks and geography. Moreover, we restrict our attention to the value of the delta parameter for which realistic properties of emergent networks are obtained.

4.1 Co-invention networks

Our empirical evidence is built upon all European Patents Applications, at least one inventor of which declared an address in France, and the priority date of which is between January 1977 and August 2003 included.[16] All non French inventors of these patents have been dropped so that our evidence is limited to France. The dataset counts 97,966 French inventors of more than 114,000 patents. Among these inventors, 76,612 have invented a patent with at least one other French inventor. The co-invention network is constructed by allocating a link between any pair of agents who have invented at least one patent together. Thus we make the assumption that any

[15] Of course, not all collaborations lead to inventions and not all inventions are patented. Nevertheless, we can suppose that if a research collaboration is strong, if it lasts over a significant period of time and if research occurs in a patenting domain, then it nearly always leads to a patent.

[16] These data are an extraction of the EP-INV database produced by CESPRI-Universita Bocconi. These data have been treated for dealing with the homonymy of inventors' names. The procedure was performed in three successive steps : 1) standardization of names and addresses and attribution of an inventor code for a unique surname, name and address ; 2) computation of similarity scores for all possible pairs of inventors with the same names and surnames but different addresses, 3) identification of a threshold value of the score over which two inventors are considered as the same inventor. For more details on the data see Lissoni et al. (2006b).

pair of agents who co-invent a patent are personally acquainted. This assumption is standard in the literature on co-authorship networks (see e.g. Newman, 2004; Moody, 2004; Goyal et al., 2006). Moreover, the assumption according to which co-inventors collaborate together is particularly acceptable in the framework of our study since co-invented patents (with at least two inventors) mostly involve small teams of collaborators : the average and median number of inventors of co-invented patents are respectively 2.7 and 2 with a standard deviation only equal to 1.17.

Table 1 provides some basic statistics of this network. Beyond such global statistics, a standard way to grasp a network structural properties relies on a detailed analysis of its degree distribution. Remember that social networks are proved to exhibit a high inequality in neighborhood sizes and that heavy tail degree distributions such as power law distribution are usually found (Barabasi and Albert, 1999). Figure 3 presents the histogram of the distribution and Table 2 the power law estimates of degree distribution of the population of non isolated inventors.

Patent data mention the personal addresses of inventors. We were thus able to locate inventors on the metropolitan French area thanks to the matching of the post codes mentioned in their addresses with their corresponding latitude and longitude coordinates[17]. Nevertheless, inventors may have different locations : up to 11,970 among the connected inventors have declared at least two different addresses. Note that this number might be overestimated since an inventor may be abusively recorded as being located in two different places just because two different writings were used for the same address. Most geographically mobile inventors remain in the same area : nearly 79% (86%) of mobile inventors have a maximal distance between their different locations which is less than 20 km (50 km).

The Euclidian 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 can be associated to a pair of connected agents : some agents invent several times with the same co-inventor while at least one of the two changed address in between. Thus, behind the 134,224 direct connections between agents, one effectively records nearly 150,000 distances between co-inventors. We apply the distributional approach to study the geographic organization of connections, that is the extent to which direct relations correlate with geography. We observe that the distribution of connections according to the geographic distance between connected agents is very skewed. More than 75% of the connections are achieved between inventors that live at less than 50 km from each other while less than 4% of the connections are formed between agents who live at more than 550 km from each other. Figure 3 presents the (properly weighted for multiple counting) histogram of the distribution while power law estimates of the distribution are exposed in Table 2.

It should also be noticed that the largest component of the network counts 43.92% of the population[18] (34.35% if we include isolated individuals) which is a relatively low proportion as

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

[18] Breshi and Lissoni (2006a,b) find that the largest component of the co-invention networks is 16% of the Italian

compared to other social networks. As a point of comparison, the largest component of scientific co-authorship networks rarely include less than 70% of the population (see for instance Newman, 2001 and Barabasi et al. 2002). Nevertheless, these studies always focused on given scientific domains and disciplines (for instance medicine, condensed matter physics or computer science). Such a low proportion can also be partially explained by a lower density of the network.[19] One can also argue that technological knowledge may be more fragmented as compared to scientific knowledge. Furthermore, the institutional configuration could generate a higher fragmentation of the population of inventors as compared to authors who evolve in a more open scientific mode of knowledge production. When applying the distributional approach to components, we find that the distribution of components with respect to their population size can be fitted by a power law distribution (see estimates presented in Table 2). The following subsection investigates the possibility that, though the global analysis remains relevant, there might be some interest in having a component based approach.

4.2 A methodology for comparing theory with empirical data

A within-components approach

We now address the issue of the structural comparison between empirical and theoretical emergent networks. To be compared with emergent networks, we are in need of co-invention networks that might have been generated in a similar context. The main problem faced then is the following. In the theory, a population of n agents is given. The agents deal at one time or the other, and probably several times, with the eventuality to connect with any of the $n-1$ other agents. In the empirical data, since the co-invention network is not restricted *ex ante* to a specific domain or institution, it is clear that not any agent may have been connected with any other. Agents belong to subsets within which they may be connected the one to the others whereas establishing connections between subsets is very unlikely. Nevertheless, it is very difficult to clearly identify separate subsets of agents from the rest of the population. If this was fully observable, agents' behaviors could be explained by only considering the eventuality of forming connections with other people within a given subset.

To deal with this problem, an obvious way is to rely on the observed allocation of agents in components, assuming that agents of a same component are, for some unobserved factors, more likely to be isolated from people outside the component. But a typical error can then be made : some agents belonging to different components, though they did not yet, might have been connected to each other and will probably be connected in a soon future. If this happens, agents' strategies would be incorrectly perceived. In order to reduce as much as possible this potential measurement error, we only analyze components that exhibit inertia in the last

inventors of European patents recorded over 1978-1995 and 46% of the US inventors of European patents recorded over 1978-1999.

[19] It is a well known property of random as well as scale free networks that increasing network density non linearly leads to the emergence of a "giant component" which tends to encompass nearly all the population (Erdös and Rényi,1960).

years of observation. Namely, we keep only components that experienced no link creation in the last two years of observation (2002-2003). These components are assumed to have reached an equilibria.[20,21]

Components of small population size are not relevant for a detailed within-component structural analysis. Thus all components of population lower than 15 are dropped. Among the 194 components of size $n \geq 15$, 129 exhibited inertia over 2002-2003 (the procedure drops out the largest component). After these elimination procedures, only one component is of population size greater than 70 inventors. We thus restrict to components of size around $n^1 = 20$ agents ($15 \leq n \leq 25$) and around $n^2 = 50$ ($30 \leq n \leq 70$)[22]. We are also interested in the potential impact of geography on network structure which, in the model, is assumed to increase links formation costs. For that purpose, a maximal geographic distance between any two agents of each component is computed. When agents have several addresses, one of them is picked at random. Such a measure can be considered as the Euclidian equivalent to the parameter S which appears in the theoretical model. Inventors, are split into two subsets according to the fact that their components have S lower or greater than the median $\overline{S}$.[23] Notice that these two subsets of components do not differ significantly in their number of agents : mean number of agents is 44.63 when $S < \overline{S}$ and 43.79 when $S > \overline{S}$.

Setting parameters values of the theoretical model

First, we consider theoretical networks whose population sizes correspond to the ones of the selected empirical components ($n^1 = 20$, $n^2 = 50$). We recall that, as we have seen in the previous section, all agents belong to a single connected component if δ is not too close to 0. Next, the two geographical sizes ($S^1 = 1$ and $S^2 = 2$) and the two standard deviations σ^1 and σ^2 of agents' heterogeneity (still with $\sigma^2 > \sigma^1$) proposed in the previous section are also retained. Again, for the ease of the exposure, we only examine the following combinations of the parameters : $\left(n^1, S^1, \sigma^1\right)$, $\left(n^2, S^1, \sigma^1\right)$ and $\left(n^2, S^2, \sigma^2\right)$.

Lastly, we restrict our attention to the emergent networks obtained with values of δ for which we have seen that the emergent networks exhibit the standard structural properties of most real social networks whatever the other parameters of the model ($\delta \in [0.2; 0; 35]$). So as to confront our choice with some external validation, we rely upon the empirical results of Breschi and Lissoni (2004). They report empirical estimates of the probability that a patent receives a citation from

[20] Such observed stability is not fully equivalent to the long run equilibrium that theoretical emergent networks reached. Nevertheless, the long run observations are not available in the data.

[21] It should also be acknowledged that two or more components might persistently remain separated though no barrier of any kind would prevent connection between agents belonging to different components. Even though it is not possible to completely avoid this eventuality, theoretical results however tend to suggest that it is quite unlikely since separate components are rarely found in the emergent networks and only for extreme values of the parameters (when δ is very close to zero).

[22] The mean numbers of agents in the components are slightly different than n^1 and n^2 : 19.20 and 44.25.

[23] The median maximal geographic distance for all components of size around 50 is $\overline{S} = 640$ kilometers. This means that half of the population belong to a component whose maximal geographical distance found among all pairs agents in this component is less than 640 kilometers.

another patent given the social distance between their inventors benchmarked by the same probability when agents are not connected. Such estimates have been obtained on European Patent data (with Italian inventors) which are thus quite close to ours. Assuming a geometric decrease of the citation probability (a proxy for the externality) with social distance, we use their results to estimate δ which is precisely found within our selected region ($\hat{\delta} \in [0.2; 0; 35]$).[24] This result tends to confirm our initial choice. After an iterative process, it appears that the best fit between our empirical and theoretical networks for our various measures are obtained for $\delta = 0.25$. Consequently, only the theoretical results obtained with this value of δ are presented.

4.3 Results

The indexes proposed in the preceding section can be quite easily modified to characterize components. Table 3 exposes some basic network computations performed for both theoretical and empirical networks. There are two main manners to read these results. One may first compare directly numerical values computed on the two types of networks while the second consists, more interestingly, in comparing the effects of the parameters (S, n) on each of them. The first approach essentially aims to verify that the two types of networks exhibit comparable structures while the second aims to check whether the expected effects of external factors on the strategic formation of networks are corroborated by the data.

As regards the former approach, we find that the average degree of theoretical and empirical networks are quite close when $n = 20$ while theoretical networks are denser when $n = 50$. The average distances are also very similar when $n = 20$ while agents are in average closer the one to the others in theoretical networks when $n = 50$. This difference can be explained as a straightforward consequence of a lower density of empirical networks. The Gini coefficient of the degree distributions are again comparable though degree inequality is always higher in empirical networks. The difference is reduced when agents' heterogeneity and geographical size are large suggesting that agents heterogeneity is higher in real networks. Furthermore, though the average clustering of empirical networks is always higher, theoretical and empirical networks are very similar in this respect.

According to the second approach, we find that when n increases, the average degree of empirical and theoretical networks also increases. There is no clear impact of n on the inequality of the degree distribution in the theory and in the empirics. Population size affects negatively the geodesic distance of theoretical networks because it increases the density of agents on a given geographical space and thus reduces link formation costs. Such a phenomenon is not observed in the empirical networks since n impacts positively the geodesic distance of real networks. This

[24] Breschi and Lissoni (2004) report estimations of $y = \Pr(citation\,|d)/\Pr(citation\,|d=\infty)$. We assume that the citation probability (the externality) decreases geometrically with social distance : $\Pr(citation\,|d) = \omega\delta^d$, with $\omega \leq 1$ the externality potential. We can thus write $y = a\delta^d$, with $a = \omega/\Pr(citation\,|d=\infty)$. Taking the log on both sides, we obtain : $Y = C + \Delta \times d$ with $Y = \ln y$, $C = \ln a$ and $\Delta = \ln \delta$. Breschi and Lissoni provide values for y when $d = 1, ..., 6$. Since their estimates (excluding self-citations) are all significant up to a distance equal to 4, we limit our investigation to such values. By simply regressing Y on d, we obtain $\hat{\Delta} = -1.19$ (with a $R^2 = 0.92$), the exponential of which gives us $\hat{\delta} = 0.30$.

could be explained by a less homogeneous geographical space in the empirics or by the effect of unobserved exogenous factors (institutional or cognitive) eventually correlated with geography. For instance, if agents were located in geographical clusters on the ring, then increasing n would increase geodesic distance as soon as agents do not find more incentives to form distant (between clusters) connections. This seems to indicate that, when the population is larger, theoretical networks are even more small worlds à la Watts and Strogatz than empirical nets.

Turning to the impacts of the geographical dimension S, it seems to have a very similar influence on both theoretical and empirical networks. When S increases (n held constant) then the average degree is significantly reduced in both theoretical and empirical networks. This confirms that geographical distance plays positively on links formation costs. An expected mechanical consequence of a significant reduction of average degree resides in an increase of the average geodesic distance. Nevertheless, no significant change of average geodesic distance is observed in both the theory and in the empirics. Even more, it slightly decreases in empirical networks. We also observe that clustering increases in both cases. When S increases, there are less numerous distant connections because these connections have a much lower chance to be payoffs increasing given that costs are then higher. As geography increases, agents then tend to collaborate proportionally even more with their nearest geographic neighbors and even less with agents located far away. Finally, we observe that the Gini of the degree distribution increases : there are more central agents when the geographical space is greater. This explains why, despite a much lower connectivity, the average distance of emergent and empirical networks remains somehow constant : more central agents play a very important role in reducing the distance between their numerous neighbors.

Within-components degree asymmetry

We now turn to a deeper analysis of the within-components asymmetry in agents' neighborhoods size in a component. For that purpose, we define $\rho_C(k)$, the relative within component C degree distribution, as follows :

$$\rho_C(k/\left(\#C-1\right)) = \frac{1}{\#C}\sum\nolimits_{i\in C} 1\left\{\eta_i(g)=k\right\}, \tag{11}$$

for all $k = 0, ..., (\#C-1)$ with $\#C$ the number of agents in component C. Such a measure allows us to directly compare the degree distribution of components that may not have exactly the same population size while inequality might to some extent be influenced by the size of the component (rather than the opposite) : within a component of size n, one can not be linked with more than $n-1$ other agents.

The results are to be found in Figure 4 which plots simple histograms of averaged ρ_C distributions computed on empirical components and theoretical networks for the different population size and geography as defined above. We find that the degree distribution of empirical components is right-asymmetric, even when the population remains limited around 20 agents. The theoretical networks also exhibit a right-asymmetric degree distribution, though inequality

is higher in empirical distributions. Let's notice that, in the theoretical networks, the right-asymmetry of the distribution is an emergent property of the decentralized process of network formation since the distribution of agents-specific costs is uniform. Nevertheless, it is found that there are larger star agents in the empirics as compared to the theory when the population is small (n^1). This difference might be explained by more heterogeneity even when population size is still small. Moreover, it appears that increasing S generates larger (global) stars in the theory whereas it generates more intermediary (local) stars in the empirics. This might again be explained by a difference in the distribution of agents in space between the model and the data. If agents are already clustered in space, increasing the geographical dimension stimulates more the emergence of local stars rather than the emergence of more global stars.

How do networks relate to geography ?

We are also interested in the way in which network links are distributed in space. In order to provide a systematic analysis of the within-components correlation of social connections with geography, we propose to study $\phi_C(\cdot)$, the density distribution of direct connections according to the geographic distance between two linked agents corrected by the maximal distance (S) between any two agents of the component C. It is formally given by :

$$\phi_C\left(h/S\right) = \frac{1}{\#\left\{ij \,|ij \in g \text{ st } i,j \in C\right\}} \sum\nolimits_{ij \in g \text{ st } i,j \in C} 1\left\{s_{ij} = h\right\}, \tag{12}$$

for all $h = 1, ..., S$.[25]

How do the network links relate to the geographical space in theory and empirics ? The averaged ϕ_C distributions computed on empirical and theoretical components for the different parameters as defined above are presented in Figure 5.

As regards the empirics, we find that agents preferentially form links with other agents located in their geographical neighborhood. A very strong asymmetry is preserved in the within-component approach though it is reduced as compared to the across components distribution (Figure 3). Indeed, 54% (80%) of the connections formed between agents within components of population size around 20 (50) are at a geographical distance that is less than 10% of the maximal distance between any two agents of the component. In the meantime 20% (12%) of the connections formed between agents within components of population size around 20 (50) are of a geographical distance greater than 50% of the maximal distance. When n increases, there are proportionally less long distance connections. This corroborates our intuition on the effect of n on the average geodesic distance (cf. Table 3) provided above : when n increases, the average geodesic distance increases in empirical networks because links are more frequently formed within clusters and distant connections are proportionally more scarce. Finally, S seems to have no significant impact on the distribution.

Theoretical networks do also share a strong asymmetry in the distribution. In all configurations of the parameters, the mode of the distribution is obtained for the lowest geographic

[25] The components for which all agents do live in the same town are not included so that the relative distributional measure keeps consistency.

distance. However, such asymmetry does not reach the same extent as empirical networks. This difference may be explained by the absence, in the theoretical model, of any *ex ante* agglomeration of agents in space, whereas in reality agents are located within territories.

5. Conclusion

In this paper, we have introduced a model of network formation in which heterogeneous agents balance the benefits of forming links against their costs which increase with the geographic distance over a ring on which agents are located at equidistant intervals. Networks are strategically formed in a decentralized manner by rational but myopic agents through a dynamic meeting process.

We show that, if the population, the heterogeneity and the geographical range are sufficiently large, the theoretical model generates networks that share the standard structural properties of most real social networks (the existence of a large connected component that links directly or indirectly most agents, a short average path length, a high average clustering and a high inequality in degree distribution) for a very large range of intermediary values of the decay parameter (that tunes the strength of the positive externalities). We can also find a more restricted region of the decay parameter, for which these properties are also found for all examined alternatives of the other parameters values. Clustering is achieved in the local space though networks also slightly discorrelates with geography thanks to some agents who establish connections with distant partners. These agents are likely to be the ones that have the greater abilities to sustain connections at low costs and who also tend to play the role of local intermediaries.

More, the model simultaneously provides a synthetical rationale for the emergence of such real networks. Agents have incentives to form connections with agents from whom they do not already indirectly benefit. Nevertheless, local (in space) relations are less costly and are thus naturally established, implying the overlapping of neighborhoods. Then, bridging connections can only be formed between distant agents and are thus more costly. To be formed and maintained, such distant relations need to be compensated by a significant surplus in wealth brought about by a sufficient number of indirect connections. The model thus reconciliates the works that emphasize the (gross) returns to forming bridging connections with the observed clustered structures : some "short cut" relations are endogenously formed while most agents remain embedded in overlapping neighborhoods. When a bridging connection is formed between separated and distant communities, it dissipates the incentives to form further of such relations. The ex ante heterogeneity between agents is not neutral on the selection of the (entrepreneurial) agents who establish such bridging relations.

Moreover, we bring the model to network data concerning all inter-individual collaborations between French inventors of all European patents over a 25 years period. A first global analysis of the whole French co-invention network is provided. Given its high level of disaggregation in components, we propose a within-components methodology to compare empirical networks with

theoretical networks. We find that theoretical and empirical networks present close structural measures. More, most of the predictions on the impacts of several parameters effects on networks structure are corroborated in the data. In particular, the dimension of the geographical space on which agents are located reduces the density of the network because this increases links formation costs, and thus both the relational distance between agents and the propensity to form local connections (i.e. local clustering) increase.

The theoretical model also generates networks whose degree distribution is highly right-asymmetric in a comparable extent as empirical networks. This property is a consequence of the decentralized network formation process since agents are only assumed to be *ex ante* uniformly heterogeneous in the link costs they bear. It is thus not necessary to assume that agents' abilities are asymmetrically distributed to fit the data : It is the network formation process that tends to enhance inequalities as measured through agents' degrees. Lastly, the theoretical model succeeds in generating networks, most connections of which are formed in the local environment while still a few distant connections are formed, a property shared by the empirical co-invention networks.

Among the limits of the model, we should notice that agents are obviously not located uniformly in space but in clusters, a difference which is observed in the results. Nevertheless, this simple strategic model of network formation succeeds in predicting most of the structural properties of co-invention networks and provides explanations or at least some intuitions for the remaining features.

References

Bala, V., Goyal, S., 2000. A non-cooperative model of network formation. Econometrica 68, 1181-1229.

Barabási, A.L., Albert, R., 1999. Emergence of scaling in random networks. Science 286, 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.

Bossard, J., 1932. Residential propinquity as a factor in marriage selection. American Journal of Sociology 38, 219-222.

Breschi S., Lissoni, F., 2006a. Mobility and Social Networks : Localised Knowledge Spillovers Revisited. Annales d'Economie et de Statistique, forthcoming.

Breschi S., Lissoni, F., 2006b. Mobility of inventors and the geography of knowledge spillovers. New evidence on US data. ADRES Conference "Networks of Innovation and Spatial Analysis of Knowledge Diffusion", CREUSET, Saint-Etienne, Sept 14-15.

Breschi S., Lissoni, F., 2004. Knowledge networks from patent data : methodological issues and research targets. In : H. Moed, W. Glänzel, U. Schmoch (Eds.). Handbook of Quantitative Science and Technology Research : The Use of Publication and Patent Statistics in Studies of S&T Systems. Springer, Berlin, 613-643.

Burt, R., 1992. Structural Holes : The Social Structure of Competition. Harvard University Press, Cambridge (MA).

Burt, R., 2004. Structural holes and good ideas. Amrical Journal of Sociology 110, 349-399.

Caplow, T., R. Forman, 1950. Neighborhood interaction in a homogenous community. American Sociological Review 15, 357-366.

Calvó-Armengol, A., 2004. Job contact networks. Journal of economic Theory 115, 191-206.

Carayol, N., Roux, P., 2006 Knowledge flows and the geography of networks. A strategic model of small worlds formation, BETA working Paper 2006-16.

Carayol, N., Roux, P., 2004. Behavioral foundations and equilibrium notions for social network formation processes. Advances in Complex Systems 7 (1), 77-92.

De Castro, R., Grossman, J., 1999. Famour trails to Paul Erdös. Mathematical Intelligencer 21, 51-63.

Egghe, L., Rousseau, R. , 1990. Introduction to Informetrics. Amsterdam : Elsevier.

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

Festinger, L., Schachter, S., Back, K., 1950. Social Pressures in Informal Groups. Harper, New York.

Freidlin, M., Wentzell, A., 1984. Random perturbations of dynamical systems. Springer Verlag, New York.

Gaspar, J., Glaeser, E.L., 1998. Information technology and the spatial structure of cities. Journal of Urban Economics 43, 136-156.

Gastner, M.T., Newman, M.E.J., 2006. The spatial structure of networks. European Physical Journal B 49, 247-252.

Goyal, S., Joshi, S., 2003. Networks of collaboration in oligopoly. Games and Economic Behavior 43(1), 57-85.

Goyal, S., Moraga, J.L., 2001. R&D networks. Rand Journal of Economics 32, 686-707.

Goyal, S., Moraga, J.L., van der Leij, M., 2006. Economics : an emerging small world. Journal of Political Economy, 114, 403-412.

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

Jackson, M.O., Rogers, B.W., 2007. Meeting Strangers and Friends of Friends : How Random are Socially Generated Networks ? American Economic Review, *forthcoming*.

Jackson, M.O., Watts, A., 2002. The evolution of social and economic networks. Journal of Economic Theory 106, 265-295.

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

Johnson, C., Gilles, R.P., 2000. Spatial social networks. Review of Economic Design 5, 273-299.

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

Kono, C., Palmer, D., Friedland, R., Zafonte, M., 1998, Lost in space : The geography of corporate interlocking directorates. American Journal of Sociology 103, 863-911.

Kranton, R., Minehart, D., 2000. Competition for goods in buyer-seller networks. Review of Economic Design 5, 301-332.

Lerner, J., 1995. Venture capitalists and the oversight of private firms. Journal of Finance 50, 301-318.

Lissoni F., Sanditov B. and Tarasconi G., 2006. The Keins Database on Academic Inventors : Methodology and Contents. WP cespri #181.

Moody, J., 2004, The Structure of a Social Science Collaboration Network : Disciplinary Cohesion from 1963 to 1999, American Sociological Review 69(2), 213-238.

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

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

Nicholls, P.T., 1986. Empirical validation of Lotka's law. Information Processing and Management 22, 417-419.

Price, D.J., de S., 1965. Networks of scientific papers. Science 149, 510-515.

Redner, S., 1998. How popular is your paper ? An empirical study of citation distribution. European Physical Journal B 4, 131-134.

Robles, J., 1998. Evolution with Changing Mutation Rates. Journal of Economic Theory 79, 207-223.

Singh, J., 2005. Collaborative Networks as Determinants of Knowledge Diffusion Patterns. Management Science 51(5), 756-770.

Sorenson, O., Stuart, T., 2001, Syndication networks and the spatial distribution of venture capital investments. American Journal of Sociology 106, 1546-1588.

Stouffer, S., 1940, Intervening opportunities : A theory relating mobility and distance. American Sociological Review 5, 845-867.

Vega-Redondo, F., 2006. Building up social capital in a changing world. Journal of Economic Dynamics and Control, *forthcoming*.

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

Young, H.P., 1993. The evolution of conventions. Econometrica 61, 57-84.

Zipf, G. (1946) "The P1P2/D hypothesis : On the intercity movement of persons."American Sociological Review, 11 : 677-686.

# isolated agents	21,354
# connected agents $\#N(g)$	76,612
# links $\eta(g)$	134,224
# of components	12,515
Size of the largest component	33,650
Size of the 2nd largest component	143
Average degree $\hat{\eta}(g)$ (over all agents)	2.74
Average degree $\hat{\eta}(g)$ (over connected agents)	3.50
Highest degree $\max_{i \in N} \eta(g)$	202
Average clustering $c(g)$	0.54
Average geographic distance of direct connections$^{\diamond}$	89.23 km

$^{\diamond}$ *Agents' location is considered as declared in the patent which evidences each connection. This implies the multiple counting of links (possibly associated to different distances) which connect a mobile inventor who has invented several times with the same co-inventor.*

Table 1. Descriptive statistics on the co-invention network.

	$\hat{C}$	$\hat{\gamma}$
Agents degree distribution	0.50	1.73
Link geographical distance distribution$^{\diamond,\star}$	0.39	1.51
Component size distribution	0.46	1.66

Table 2. Estimates of the parameters C and γ of the power law distribution on three empirical distributions. A power law distribution is defined as follows : $f(k) = Ck^{-\gamma}$, $\forall k = 1, ..., \infty$, such that $\sum_{k=1,...,\infty} Ck^{-\gamma} = 1$, with $f(k)$ the frequency of observations having value k. We used a maximum likelihood approach (introduced by Nicholls, 1986) which, as suggested by Egghe and Rousseau (1990), performs much better than a least square approach.

$^{\diamond}$ *Agents' location is considered as declared in the patent which evidences the connection. As a consequence, a connection between two persons is counted twice if at least one of the two declared a different address in the two patents they invented together.*

* *Since distance is per se a continuous variable, the estimated distance variable is the number of tens of kilometers initiated*= $\lceil$(# of kilometers)/10$\rceil$.

	Empirics			**Theory**		
	No recent link formation			$\delta = 0.25$		
	$15 \leq n \leq 25$	$30 \leq n \leq 70$		$n^1 = 20$	$n^2 = 50$	
	all S	$S < \overline{S}$	$S \geq \overline{S}$	σ^1, S^1	σ^1, S^1	σ^2, S^2
Average degree $\hat{\eta}(g)$	3.74	8.11	4.18	3.78	10.71	5.67
Average distance $d(g)$	2.51	3.22	3.14	2.43	2.02	2.25
Average clustering $c(g)$	0.73	0.71	0.75	0.63	0.63	0.70
Gini coefficient $Gi(g)$	0.31	0.33	0.37	0.22	0.16	0.33
Total # of agents	(1,671)	(504)	(496)	(10,000)	(10,000)	(10,000)

Table 3. Various structural measures computed on empirical and theoretical emergent networks. $\overline{S}$ is the median geographical size of components of population $35 \leq n \leq 65$.

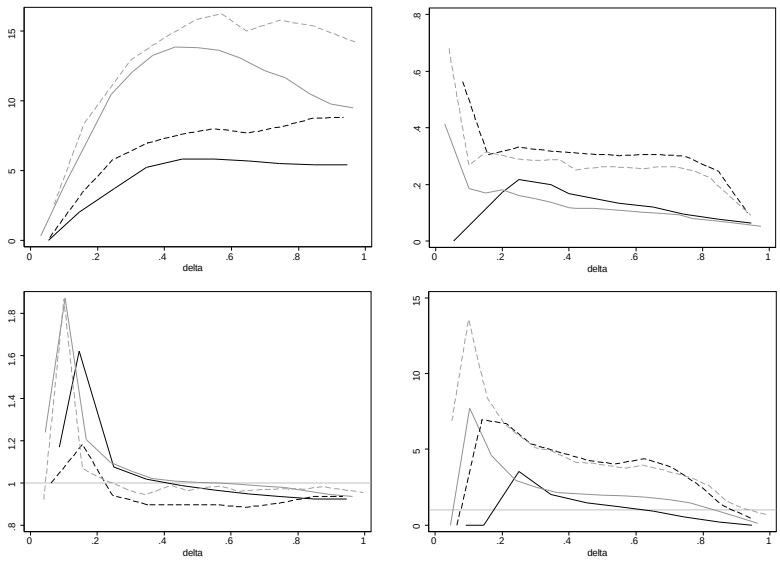

Figure 1. Average degree $\hat{\eta}(g)$ (top left), Gini coefficient $Gi(g)$ (top right), average distance ratio $\widetilde{d}(g)$ (bottom left) and average clustering ratio $\widetilde{c}(g)$ (bottom right) of emergent networks. Black continuous lines correspond to $\left(n^1, S^1, \sigma^1\right)$. Grey continuous lines correspond to $\left(n^2, S^1, \sigma^1\right)$. Black dashed lines correspond to $\left(n^2, S^2, \sigma^2\right)$. Grey dashed correspond to $\left(n^3, S^2, \sigma^2\right)$. Median bands curves computed for 1,000 simulations with randomly drawn values of $\delta \in]0;1[$ for each configuration.

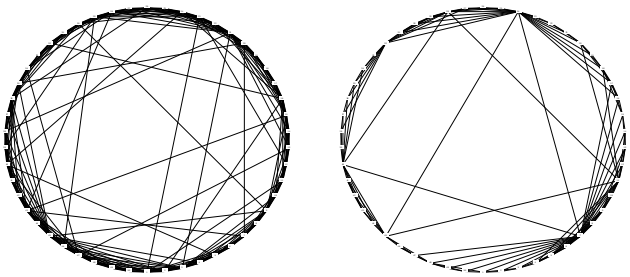

Figure 2. Typical emergent networks generated with $\delta = 0.25$, $n = n^2 = 50$ agents and two configurations of the other parameters : S^1, σ^1 (left network) and S^2, σ^2 (right network).

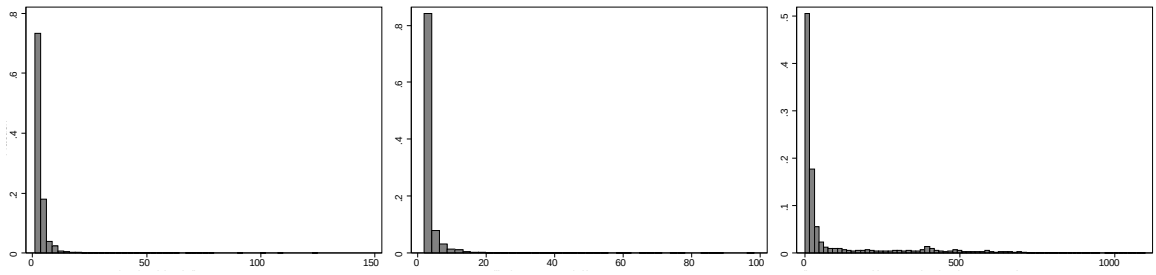

Figure 3. The degree distribution in the whole population of connected inventors (76,612) (left graph), the distribution of components according to their population (central graph) (only components of population below 100 agents are presented since it goes up to more than 33,000), and the distribution of connections according to their geographical distance (right graph) in the whole co-invention network. Agents' location is considered as declared in the patent which evidences the connection. This implies multiple countings of links (given that agents may change locations) which are corrected through a simple weighting procedure.

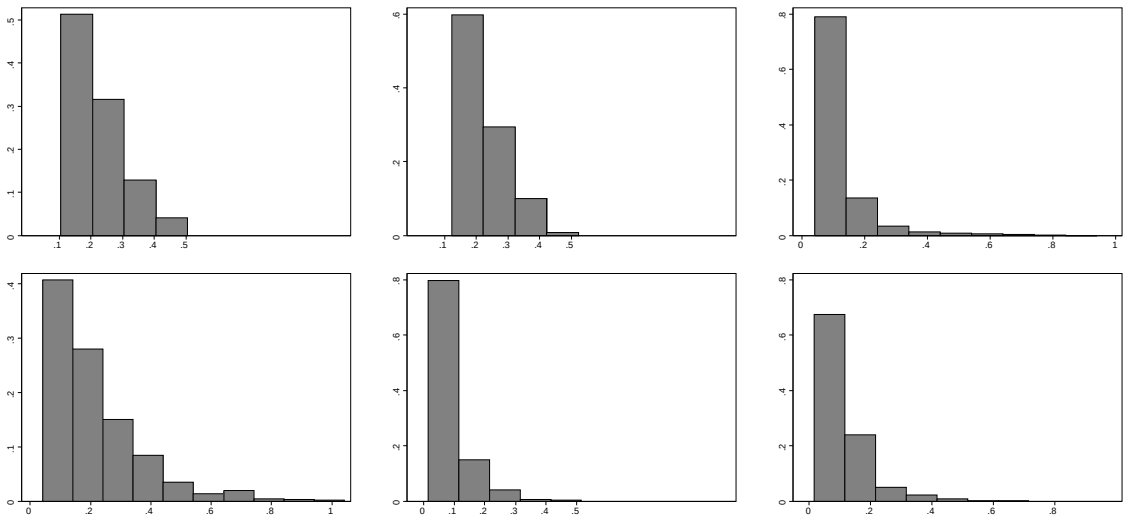

Figure 4. The averaged within-component degree distribution $\rho_C(k)$ computed on theoretical and empirical networks. Top histograms are built with theoretical emergent networks with $\delta = 0.25$ and $\left(n^1, S^1, \sigma^1\right)$ (top left), $\left(n^2, S^1, \sigma^1\right)$ (top central), $\left(n^2, S^2, \sigma^2\right)$ (top right). Bottom histograms are using empirical data on agents belonging to components of population $15 \leq n \leq 25$ (bottom left), on agents belonging to components of both a population $30 \leq n \leq 70$ and a geographical size less than the median $\overline{S}$ (bottom central) and, on agents belonging to components of both a population $30 \leq n \leq 70$ and a geographical size greater than $\overline{S}$ (bottom right).

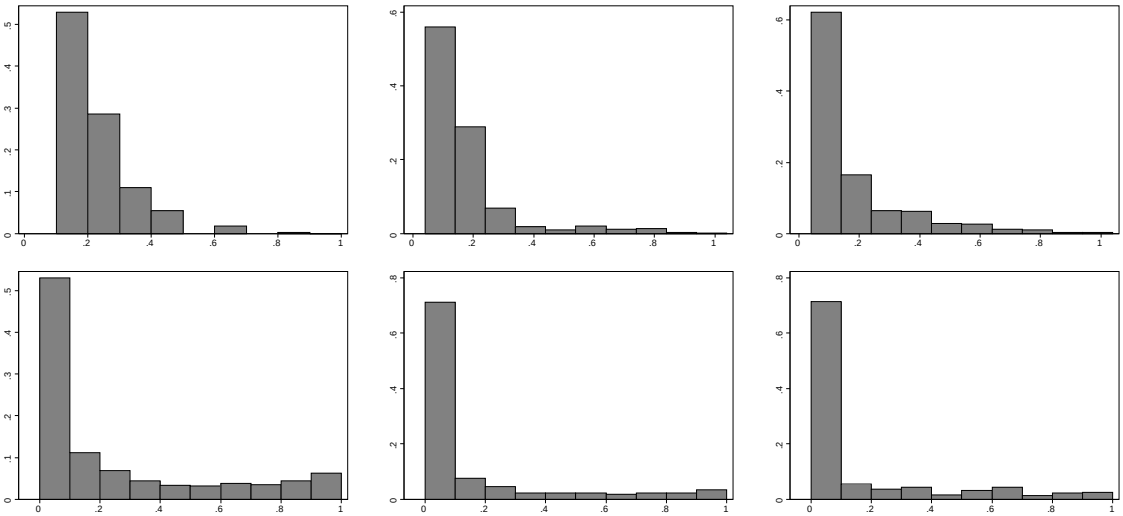

Figure 5. The averaged within-component link relative distance distribution $\phi_C(k)$ computed on theoretical and empirical networks. Top histograms are built with theoretical emergent networks with $\delta = 0.25$ and $\left(n^1, S^1, \sigma^1\right)$ (top left), $\left(n^2, S^1, \sigma^1\right)$ (top central) and $\left(n^2, S^2, \sigma^2\right)$ (top right). Bottom histograms are using empirical data on agents belonging to components of population $15 \leq n \leq 25$ (bottom left), on agents belonging to components of both a population $30 \leq n \leq 70$ and a geographical size less than the median $\overline{S}$ (bottom central) and, on agents belonging to components of both a population $30 \leq n \leq 70$ and a geographical size greater than $\overline{S}$ (bottom right). Multiple countings of links (due to the fact that agents change locations) are corrected through a simple weighting procedure.