# The effect of R&D collaborations on firms' technological positions

Mario V. Tomasello*,[1], Claudio J. Tessone[1,2], Frank Schweitzer[1]

[1] Chair of Systems Design, ETH Zurich - Department of Management, Technology and Economics, Weinbergstrasse 56/58, CH-8092, Zurich (Switzerland)

[2] URPP Social Networks, Department of Business Administration, Universität Zürich, Andreasstrasse 15, CH-8050, Zurich (Switzerland)

** Corresponding author*

**Structured Abstract**

**Purpose –** We develop an agent-based model to reproduce the processes of link formation and knowledge exchange in a Research and Development (R&D) inter-organizational network.

**Methodology** – In our model, agents form links based on their network features, i.e. their belonging to one of the network's circles of influence and their previous alliance history, and then exchange knowledge with their partners, thus modifying their positions in a metric knowledge space. Furthermore, we validate the model against real data using a two-step approach. Through the Thomson Reuters SDC alliance dataset, we estimate the model parameters related to the link formation, thus reproducing the topology of the resulting R&D network. Subsequently, using the NBER data on firm patents, we estimate the parameters related to the knowledge exchange process, thus evaluating the rate at which firms exchange knowledge and the duration of the R&D alliances themselves.

**Originality** – The underlying knowledge space that we consider in our real example is defined by IPC patent classes, allowing for a precise quantification of every firm's knowledge position. Our novel data-driven approach allows us to unveil the complex interdependencies between the firms' network embeddedness and their technological positions. Through the validation of our model, we find that real R&D alliances have a duration of around two years, and that the subsequent knowledge exchange occurs at a very low rate. Most of the alliances, indeed, have no consequence on the partners' knowledge positions: this suggests that a firm's position – evaluated through its patents – is rather a determinant than a consequence of its R&D alliances. Finally, we propose an indicator of *collaboration performance* for the whole network. We find that the real R&D network does not maximize such an indicator.

**Practical implications** – Our study shows that there exist configurations that can be both *realistic* and *optimized* with respect to the collaboration performance. Effective policies to obtain an optimized collaboration network – as suggested by our model – would

incentivize shorter R&D alliances and higher knowledge exchange rates, for instance including rewards for quick co-patenting by allied firms.

**Keywords –** Agent-Based Modeling; Complex Adaptive Systems; Network Theory; Knowledge Exchange; Technological Trajectory.

**Paper type** – Academic Research Paper

## 1 Introduction

The 1990s have witnessed an unprecedented growth in the number of inter-organizational Research and Development (R&D) alliances. This phenomenon has especially affected industrial sectors such as IT, Pharmaceuticals and other high-technology ones (Ahuja, 2000; Hagedoorn, 2002). Consequently, research has investigated the mechanisms behind the formation of such R&D alliances (Powell et al., 2005), as well as the complex networks they origin (Rosenkopf and Schilling, 2007; Tomasello et al., 2014), or the methodologies to describe, model and forecast their evolution (König et al., 2012; Garas et al., 2014).

A number of theoretical works have shown that firms engage in alliances because they can gain access to different assets more quickly than they could do *in-house* (Liebeskind, 1996; Das and Teng, 2000), or because they can actually enlarge their knowledge basis more than they could do individually, by joining their technological resources (Baum et al., 2000; Mowery et al., 1998; Rosenkopf and Almeida, 2003) or, finally, because they can share the costs and risks of a project, especially when this is expensive or with uncertain outcome (Hagedoorn et al., 2000). All of these aspects result in a learning process by the involved firms, making R&D alliances an important source of knowledge exploration and production.

In this work we investigate such a learning process through an agent-based model, that reproduces the knowledge exchange occurring after the establishment of consecutive alliances between firms in a dynamically evolving R&D network.

The model we propose follows an existing stream of literature in the direction of bounded confidence and continuous opinion dynamic models (Deffuant et al., 2000; DeGroot, 1974; Hegselmann and Krause, 2002), especially applied to innovation networks (Baum et al., 2010). However, differently from the studies that have been carried out so far, our model does not focus on the formation of consensus clusters (see Axelrod, 1997; Groeber et al., 2009, in the case of social systems) or technology islands (see Fagiolo and Dosi, 2003). Instead, we focus on the dynamics that leads the system to the observed final state, with an emphasis on the *exploration* of the knowledge space by the collaborating agents.

We validate the model against real data using a two-step approach, as well as two independent datasets. By means of a dataset reporting inter-organizational R&D alliances (Thomson Reuters SDC), we estimate the model parameters related to the network formation and evolution. Next, using a dataset on firm patents (NBER), we estimate the

parameters related to the knowledge exchange process. The underlying knowledge space we consider in our real example is defined by IPC patent classes, allowing for a precise quantification of every firm's technological position.

## 2 Data and methods

In the present study, an R&D network consists of a set of nodes, or agents (the *firms*) and links (the *alliances*), connecting them in pairs. By R&D alliance (or collaboration), we refer to an event of partnership between two firms, that can span from formal joint ventures to more informal research agreements, specifically aimed at research and development purposes. To detect such events, we use the SDC Platinum database, provided by Thomson Reuters, that reports all publicly announced alliances, from 1984 to 2009, between several kinds of economic actors (including manufacturing firms, investors, banks and universities). In our network representation, we draw an undirected link connecting two nodes every time an alliance between the two corresponding firms is announced in the dataset. When an alliance involves more than two firms, all the involved firms are connected in pairs, resulting into a fully connected clique.

In order to evaluate the position of real firms in a metric knowledge space, we use the Patent Citations Data by the U.S.A. National Bureau of Economic Research (NBER), containing detailed information on patents granted in the U.S.A. and other contracting countries, from 1971 to present. Obviously, we select only the entries that have a match with the SDC alliance dataset, both with respect to assignees and time period, thus obtaining a total of around 1,400,000 listed patents. Every patent is associated with one or more assignees and with an International Patent Classification (IPC) class. Companies are associated with a unique identifier, and a relatively big part of them (5,168 firms, precisely) are matched to the SDC alliance dataset. These firms take part in 7,417 distinct R&D alliances.

The approach we use to determine the knowledge position of a firm is to compute the shares of its patents in a set of different IPC classes. The IPC is a hierarchical system of patent classification. A generic category consists of a letter, the so-called "section symbol", followed by two digits, the so-called "class symbol", and a final letter, the "subclass", plus other additional digits. We intend to test our model on a broad set of firms, exhibiting patent activities distributed across all sections, classes and subclasses. Hence, our choice is to consider only the section symbol (i.e. the first letter) in our empirical patent classification.[1] The titles of the 8 sections, as well as a patent count for each section in our dataset, is reported in Table 1. To ensure a match with our model representation, we define the knowledge position of a firm $\mathbf{x}_i \equiv (x_{iA}, x_{iB}, \ldots, x_{iH})$ as the set of normalized patent counts $x_{is}$ in each section, which in its turn equals:

*1 Choosing a subclass-level division would result in a high dimensionality for the corresponding knowledge space. However, we have also tested a more refined division, obtaining a total of 74 classes, and we have found that the computational burden of operating in a 74-dimensional space does not lead to any significant change in our results.*

$$x_{is} \equiv \frac{N_{is}}{\sum_s N_{is}} \qquad s = A, \ldots, H$$

(1)

where $N_{is}$ is the number of patents that the firm *i* has in a given IPC section *s*. In order to compute knowledge distances between pairs of firms, we use the Euclidean metric. This means that the knowledge distance between two firms *i* and *j* reads as:

$$|\mathbf{x}_i - \mathbf{x}_j| = \sqrt{\sum_{s=A}^{H} (x_{is} - x_{js})^2}$$

(2)

Using the definitions provided in Eqs. 1 and 2, we now compute (i) the knowledge positions of the firms listed in our dataset at the beginning of the observation period and (ii) the distribution of the knowledge distances between every pair of allied firms. It should be noted that, for normalization reasons, such a knowledge distance ranges from 0 to $\sqrt{2}$ . When computing the empirical knowledge position of a firm xi at a given date t, we consider all the patents for which the firm has applied, in a time window$\Delta$ *t* = 5 years preceding such date t.[12] The knowledge positions of the firms at the beginning of the observation period is used as an input for our computer simulations, as we explain below.

Table 1: International Patent Classification (IPC) sections and their description. The last column reports the number of patents registered in our dataset for the corresponding IPC section

| IPC Section | Title | Patents |
|---|---|---|
| A | Human Necessities | 152,974 |
| B | Performing Operations, Transporting | 244,791 |
| C | Chemistry, Metallurgy | 309,675 |
| D | Textiles, Paper | 12,914 |
| E | Fixed Constructions | 17,842 |
| F | Mechanical Engineering, Lighting, Heating, Weapons | 119,581 |
| G | Physics | 508,815 |
| H | Electricity | 476,437 |

.

In Fig. 1 we report the distributions of the knowledge distances between partner firms at the moment of alliance formation – the "pre-alliance knowledge distances" – and at the moment of the alliance termination – the "post-alliance knowledge distances". Given that the SDC alliance dataset does not report the alliance ending dates, we compute such measure for different values of elapsed time after the alliance establishment (1, 3, 5 and 10 years). In addition, Fig. 1 reports the variation of the knowledge distance separating

*1 We have tested different time windows, ranging from 1 to 10 years, and have found that this causes only more missing observations or noise in the distributions, with no effect on our results.*

every pair of allied firms between the moments of alliance formation and alliance termination – the "knowledge distance shift".

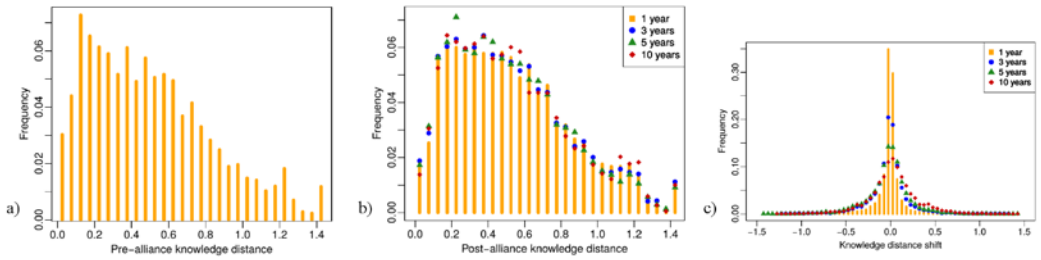


*Figure 1: Histograms of (a) the empirical knowledge distance between every pair of partnered firms, as of the day preceding the alliance formation; (b) 1, 3, 5 and 10 years after the date of the alliance formation; (c) shift of knowledge distance computed 1, 3, 5 and 10 years after the date of the alliance formation..*

We find that the distributions show a peak for intermediate knowledge distance values, meaning that firms with too similar or too different patenting activity tend not to form R&D alliances. We also find that the distribution of post-alliance knowledge distances resembles the one of pre-alliance distances, irrespectively of the selected time window for the alliance termination: this means that most of the R&D alliances cause a null change in the knowledge distance between the two partners, as shown also by the distance shift distribution. At the same time, this last distribution clearly exhibits tails on both sides, meaning that some alliances cause the partners to significantly move closer in the knowledge space, whilst some other alliances cause the partners to significantly move farther away. This is the result of the complex interactions between the collaborating agents, and – as we show later – can be reproduced by our agent-based model.

## 3 The model

The microscopic interaction rules of the present agent-based model are divided into two phases. First, the agents form links based on their network features and their social capital; second, they exchange knowledge through these links, thus approaching each other in a metric knowledge space. In addition, each link can be terminated with a given probability.

### *3.1 Exploration and link formation*

**Node activation.** We consider a network composed of $N$ nodes; each of them is endowed with a key attribute, its *activity*. The activity is defined as the propensity of a node to be involved in a collaboration event. We assign to each of the $i = 1, \ldots, N$ nodes an activity $a_i$, that is mapped to the empirical activities extracted from the SDC alliance

dataset.[1] In particular, at every time step, a node *i* initiates an alliance with probability $p_i = \eta a_i \, dt$, and the number of active nodes $N_A$ is:

$$N_A = \eta \langle a \rangle N dt, \tag{3}$$

where $\langle a \rangle$ is the average node activity and $\eta$ is a rescaling factor that allows to adjust the number of active nodes per time step. We find that the model is robust to the choice of $\eta$, showing no measurable changes for $\eta$ ranging from $10^{-5}$ to 1; however, we fix $\eta = 0.0115$ to obtain $N_A$ roughly equal to 2, the number of *active firms per day* actually reported in the alliance dataset. More details will follow on the interpretation of the time step duration d*t*.

**Selection of the alliance size.** Upon activation, a node selects the number of partners *m* with whom the alliance is formed. We assume that the value of *m* is totally independent of any characteristic of the active node: we sample it, without replacement, from the empirical distribution of the number of partners per alliance, directly measured from the SDC dataset. The value m can be thought of as the number of partners reported for each alliance event, diminished by 1, because the active node is not counted twice.

**Label propagation.** The second key node attribute is called *label*. This attribute is unique – i.e. every node can have only one label at any time – and fixed – once a node assumes a label, this does not change. The labels model the belonging of the agents to different groups that they implicitly define with their shared practices and/or behaviors. In our network representation, a label symbolizes the membership of the firm in a well defined and recognized "club" or "circle of influence". In addition, we assume that such membership can be transferred to other agents as a consequence of a collaboration, provided that they are not part of any circle of influence yet. In our network representation, every alliance initiator does indeed propagate its label to all of its m partners, if they are non-labeled. At the beginning of every simulation, all nodes are non-labeled, meaning that their membership attribute is blank. There are two ways a non-labeled node can assume its label: (i) the node either receives the label from another node, if the latter initiates an alliance, or (ii) it takes an arbitrary and unique label when it becomes active for the first time.[2]

**Selection of the partner categories.** The presence of labels induces different types of alliances, that we explicitly distinguish. In particular, if the initiator is a labeled node, i.e. an incumbent firm in the R&D network, this can form a link to a labeled node having the same label (with probability $p^L_s$), or to a node having a different label ($p^L_d$), or to a node without label ($p^L_n$); these three probabilities sum up to 1. If the initiator is a non-labeled node, i.e. a newcomer, this can form a link to a labeled node (with probability $p^{NL}_l$), or to another non-labeled node ($p^{NL}_{nl}$); these two probabilities sum up to 1. The

*1 . For more details on the definition of the agents' activity, labels, their theoretical foundation and empirical computation, please refer to Tomasello et al. (2014).*

presence of the two conditions reduces the number of independent parameters; as shown in Table 2, we consider $p^L_s$, $p^L_d$ and $p^{NL}_{nl}$ as the three independent network formation parameters of our model.

**Link formation.** After deciding the category of each of its m partners, we assume that the initiator selects its specific partners within those categories according to their degree (which we define as the number of distinct partners of a firm, and not the number of previous R&D alliances). We use a linear preferential attachment rule, where the probability to attach to a node j linearly scales with its degree $k_j$, meaning that $\Pi(k_j) \sim k_j$. The preferential attachment rule is applied within the pool of all candidate partners, once the selection of the partner category has been made by the alliance initiator. This rule obviously does not apply when the initiator – be it labeled or not – decides to connect to a non-labeled node, which has by definition no previous partners ($k_j = 0$). In this case, the partner is selected among all non-labeled nodes with equal probability. When the selection process is complete, the initiator connects to its m partners. In agreement with our representation of the R&D network, we assume that all the m partners will also link to each other, forming a fully connected clique of size $m$+1.

### *3.2 Exploitation: knowledge exchange*

**Location in a metric knowledge space.** Every agent $i$ is situated in a point with coordinates $\mathbf{x}_i$ , identified by a vector of $D$ real numbers ranging from 0 to 1. The coordinates of every node can be thought of as the ratios of the corresponding firm's expertise along each of the $D$ dimensions of the knowledge space. We assign all agents' initial positions by using real patent data, as explained in Section 2.

$$\mathbf{x}_i \equiv (x_{i1}, x_{i2}, \ldots, x_{iD}) \qquad i = 1, \ldots, n$$

(4)

**Approaching in the metric knowledge space.** We assume that the existence of a link causes the agents at both ends of the link to approach each other in the knowledge space. We assume that every agent is endowed with a *learning rate* $\mu$. This parameter is constant over time and for all nodes in the collaboration network, and can be thought of as the propensity of agents to exchange knowledge with their partners, thus making their knowledge bases more similar over time. It should be noted that the parameter $\mu$ is a rate, not a speed; the actual speed at which the corresponding nodes move in the knowledge space is given by the product of the rate $\mu$ and their distance: therefore, the farther they are in the knowledge space, the faster they approach. When their distance decreases, so does the potential for new learning from the collaboration, and the approaching speed drops consequently. The model dynamics equation can be written as follows:

$$\dot{\mathbf{x}}_i(t) = \mu \sum_{j \in \mathcal{N}_i(t)} [\mathbf{x}_j(t) - \mathbf{x}_i(t)]$$

(5)

where $\mathcal{N}_i(t)$ is the set of partners of the agent *i* at time *t*. We then implement the model through computer simulations, using discrete time steps of length d*t*. The evolution of every agent's position $\mathbf{x}_i$ can be expressed as:

$$\mathbf{x}_i(t+\mathrm{d}t) = \mathbf{x}_i(t) + \mu \sum_{j \in \mathcal{N}_i(t)} [\mathbf{x}_j(t) - \mathbf{x}_i(t)]\,\mathrm{d}t \qquad (6)$$

**Alliance termination.** We then introduce a key parameter to our model: a link characteristic life time $\tau$ . We assume that the collaboration durations are distributed according to a Poisson process with rate $1/\tau$ ; the mean duration is clearly equal to $\tau$ . In our computer simulations, which use discrete time steps of length dt, this translates into the use of a fixed termination probability $p^T$ for any link at any time step, equal to $p^T = \mathrm{d}t/\tau$ . In order to keep a simplistic set of rules, we assume that the parameter $\tau$ is independent of any other feature of the network or the knowledge exchange dynamics.

We summarize the model microscopic rules by means of a visual example in Fig. 2 and report the nomenclature of all parameters in Table 2.

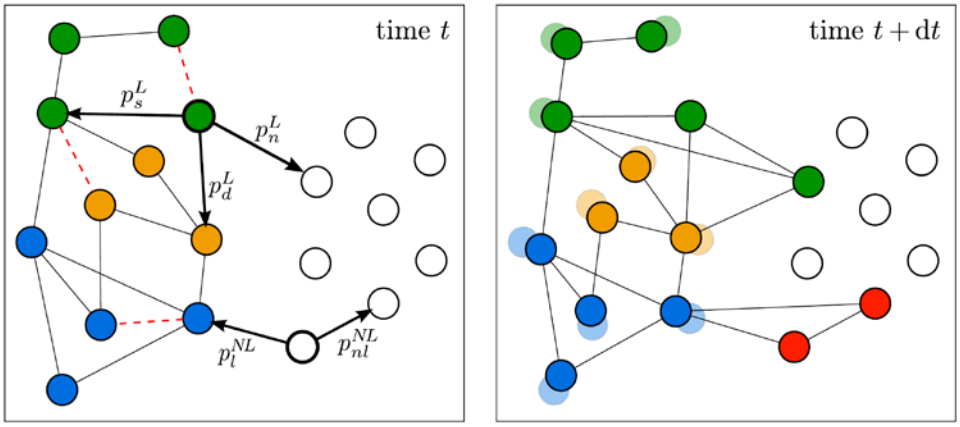


*Figure 2: A representative example of network evolution in a bi-dimensional (D = 2) knowledge space. The position of the nodes in the plot corresponds to their coordinates in the knowledge space. At time t + dt, all existing links cause the respective agents to approach in the knowledge space. Furthermore, we illustrate two collaboration events occurring at time t. The first one is initiated by a labeled node (in green), that has linked to m = 3 new partners, forming a fully connected clique. The second one is initiated by a non-labeled node, that has linked to m = 2 new partners and has taken a new arbitrary label (red). At time t + dt, the alliance initiators propagate their labels (respectively, the green one and the red one) to the partners that were not labeled at time t yet. Finally, we illustrate the termination of 3 links (depicted with red dashed lines) at time t.*

Table 2: Model parameters and their description. The "network formation" parameters are associated with the creation of new links in the collaboration network. The "knowledge exchange" parameters are associated with the approach of the agents in a metric knowledge space, occurring as a consequence of a collaboration

| Parameter | Meaning | Category |
|---|---|---|
| $p_s^L$ | Probability of a labeled node to select a node with the same label | Network formation |
| $p_d^L$ | Probability of a labeled node to select a node with a different label | Network formation |
| $p_{nl}^{NL}$ | Probability of a non-labeled node to select a non-labeled node | Network formation |
| $D$ | Dimensionality of the metric knowledge space | Knowledge exchange |
| $\mu$ | Approaching rate in the knowledge space | Knowledge exchange |
| $\tau$ | Link characteristic life time | Knowledge exchange |

.

## 4 Model validation with a two-step procedure

We perform our validation procedure in two steps and by using two datasets, as already mentioned.

In the *first step*, we validate the network topology. We fix a set of parameters that we can directly measure from the data (the number of agents and R&D alliances, the agents' activity distribution and the size of collaboration events). We then estimate the remaining parameters by running a set of computer simulations and identifying the simulated collaboration network that best matches with the alliance dataset.

In the *second step*, we fix the network formation parameters – using the values obtained in the first step – and run a second set of computer simulations. This time we estimate the knowledge exchange parameters by identifying the simulated collaboration network that best matches with the patent dataset

### *4.1 Alliance dataset and network formation*

The model parameters that we can directly measure from the data are the number of agents ($N$ = 5,168), the distribution of the node activities $a_i$, and the distribution of number of partners $m$ per alliance event. We stop every computer simulation when the total number of formed alliances equals the number of alliance events reported in the SDC dataset, $E$ = 7,417 (considering only the alliances whose partners are also listed in the NBER patent dataset). We vary the values of $p^L{}_s$, $p^L{}_d$ and $p^{NL}{}_{nl}$ in discrete steps spaced by 0.05, in the interval (0, 1). We consider the final aggregated network resulting from each of our computer simulations and we test it against the real data with respect to three properties: average degree $\langle k \rangle$, average path length $\langle l \rangle$ and global clustering coefficient $C$.[14] In order to identify which parameter combination is able to give the best match with the real R&D network, we use a Maximum Likelihood approach. For the sake of brevity, we do not report here all the details of our computations;[25] we summarize our results in Fig. 3, by means of color maps.

*1 See Newman (2002) for a rigorous definition of such network measures.*

*2 For more explanations and details, please refer to Tomasello et al. (2014), where the same approach is used for the empirical validation of an agent-based model.*

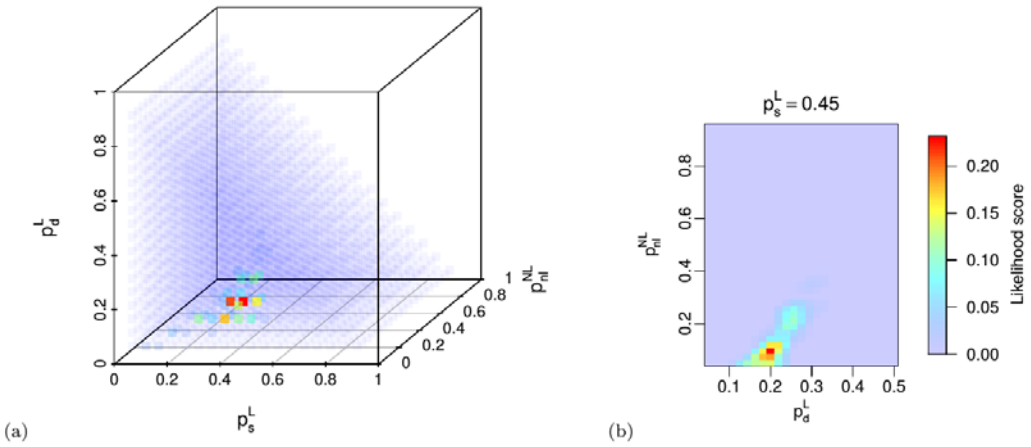


*Figure 3: Likelihood scores for all points in the parameter space, represented with a 3-dimensional color map (a). After fixing the value of $p^L_s$ to 0.45 (b), we report the Likelihood score as a function of $p^L_d$ and $p^{NL}_{nl}$ using the same color scale.*

We find that the point with the highest likelihood score has the following coordinates in the parameter space: $p^{*L}_s = 0.45$, $p^{*L}_d = 0.2$ and $p^{*NL}_{nl} = 0.1$. This means that labeled nodes exhibit a fairly balanced alliance strategy, with $p^{*L}_s = 0.45$, $p^{*L}_d = 0.2$, and consequently $p^{*L}_{nl} = 0.35$, while the non-labeled nodes exhibit a very strong tendency to connect to labeled nodes ($p^{*NL}_l = 0.9$), as opposed to a low linking probability with other non-labeled nodes ($p^{*NL}_{nl} = 0.1$). In other words, we find that he agents have a tendency to collaborate with agents that are already part of the network (i.e. incumbents). We report in Table 3 the set of parameter values maximizing the likelihood score, together with the values of average degree, average path length and global clustering coefficient for the simulated and the real R&D networks.

Table 3: Model parameter set $p^*$ defining the optimal simulated R&D network. The average degree, average path length and global clustering coefficient of the 100 realizations of the optimal R&D network are compared to their analogous empirical values.

| **Optimal simulated R&D network** | | | | **Real R&D network (with patents)** | |
|---|---|---|---|---|---|
| Model parameter | Value | Measure | Value | Measure | Value |
| $p^{*L}_s$ | 0.45 | $\langle k \rangle^*$ | $3.48 \pm 0.01$ | $\langle k \rangle^{OBS}$ | 3.45 |
| $p^{*L}_d$ | 0.2 | $\langle l \rangle^*$ | $5.02 \pm 0.08$ | $\langle l \rangle^{OBS}$ | 5.05 |
| $p^{*L}_n$ | 0.35 | $C^*$ | $0.111 \pm 0.007$ | $C^{OBS}$ | 0.109 |
| $p^{*NL}_{nl}$ | 0.1 | | | | |
| $p^{*NL}_l$ | 0.9 | | | | |

### *4.2 Patent dataset and knowledge exchange*

Prior to our second validation step, we fix all the network formation parameters to the values resulting from the first validation step. We then fix the dimensionality $D$ of our knowledge space: as we use the eight main sections of the IPC scheme, and considering that we measure the fractions of patents in each section, thus giving rise to one bounding condition, we assume $D = 7$. Each of the seven knowledge components is bound to be smaller than or equal to 1. The initial knowledge positions of the agents are assigned from the empirical data (see Section 2).

We then vary the values of the remaining knowledge exchange parameters. Precisely, we consider the values 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1 and 0.2 for the approaching rate $\mu$ and the values 5, 10, 20, 50, 100, 200, 300, 500, 700, 1000, 2000, 3000 and 5000 for the characteristic alliance life time $\tau$. The interpretation of the parameter $\tau$ is straightforward: as explained in Section 3.1, we adjust the activation rate of the agents in such a way that the length of a time step d$t$ can be directly interpreted as 1 day. Therefore, the value of $\tau$ can be thought of as the characteristic duration of a real alliance expressed in days.

For each parameter combination, we run 100 computer simulations and then compare each of the simulated networks to the empirical R&D network, with respect to the pre-alliance and the post-alliance distance distributions. We do not use the distribution of knowledge distance shifts, because it strongly depends on the first two and does not carry any additional information. Given that the alliance ending dates are not reported in the SDC dataset, we compute the empirical knowledge distance between every pair of linked firms after a time period equal to the value of the parameter $\tau$ – in days – used in the corresponding simulation.

For every simulation, we perform a two-sided Kolmogorov-Smirnov (KS) test on the resulting pre-alliance knowledge distance distribution and the corresponding empirical distribution. We use the resulting $D$ statistics, to quantify how close the two distributions are (the lower its value, the more similar they are). We discard the $p$−value of the KS tests, because we are not interested in statistically inferring the provenience of the two distributions from a hypothetical common distribution. We repeat the procedure for the post-alliance knowledge distance distribution, and sum the values of the two resulting $D$−statistics, thus obtaining a goodness score for every simulation. The lower such a score is, the closer the examined simulated R&D network is to the empirical one. We finally average the score values for all the simulations in all points of the parameter space. Such goodness scores are presented in Fig. 4, where we make use of a heatmap to summarize our results.

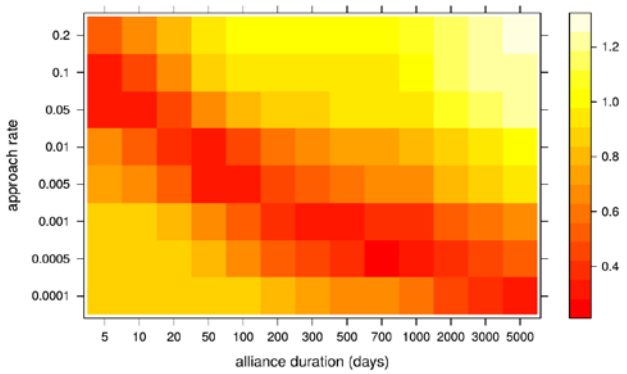


*Figure 4: Goodness score for every point in the parameter space, depicted by means of a heatmap. The color scale corresponds to the score value; the lower the score, the closer the simulated R&D network is to the empirical one.*

We find that there exists an entire region of the explored bi-dimensional parameter space maximizing the aforementioned goodness score, corresponding to the red points in Fig. 4. Indeed, the presence of such a region indicates that the two parameters are not independent, and that their product appears to be constant. Indeed, only the points with fast approaching rates $\mu$ but short alliance life times $\tau$ or – on the contrary – with long alliance life times $\tau$ but slow approaching rates $\mu$, can generate knowledge distance distributions that correspond to reality. Based on this finding, which is consistent with our previous empirical analysis (see Section 2), we argue that real companies do not significantly change their knowledge positions as a consequence of R&D collaborations. They rather use the available information about their mutual knowledge positions in order to establish new alliances.

The parameter point yielding the absolute best goodness score is identified by the following coordinates: $\mu$ = 0.0005 and $\tau$ = 700. This means that the optimal simulated collaboration network exhibits a low approaching rate, and a characteristic alliance life time slightly shorter than 2 years. This is consistent with previous theoretical and empirical observations (Phelps, 2003; Inkpen and Ross, 2001), and it is remarkable to consider that we have obtained this result by using two different datasets and employing a complex procedure, such as an agent-based model reproducing the effect of collaborations on knowledge positions.

As additional test, we show in Fig. 5 the distributions of pre-alliance distances, post-alliance distances and knowledge distance shifts, generated by the model fed with the optimal parameter set (i.e. $\mu$ = 0.0005 and $\tau$ = 700). As we have imposed an equivalence criterion through the KS test, we expect that the empirical and the simulated distributions are fairly similar, which is what we find from our analysis. However, the post-alliance distance distribution generated by our model performs slightly better than the pre-alliance distance distribution. Given that our model includes only an approach mechanism, and not a self-motion nor a drift, the post-alliance distance distribution is peaked around a

slightly lower value than the pre-alliance distribution, having a slightly better overlap with the empirical distribution.

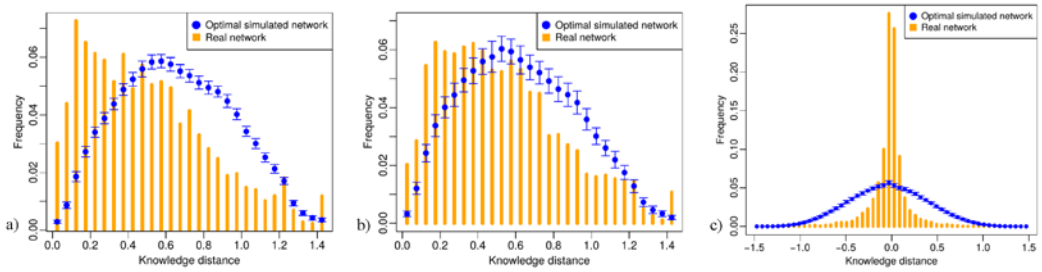


*Figure 5: Distributions of empirical and simulated (a) pre-alliance knowledge distances, (b) post-alliance knowledge distances and (c) knowledge distance shifts.*

Obviously, in every collaboration network, the agents produce knowledge on their own and explore new trajectories in the knowledge space even without being involved in collaborations or alliances. However, we intentionally do not include this behavior in our agent based model, in order not to over-complicate the microscopic rules and isolate the effects of collaboration formation on the positions of the agents.

Our model is nevertheless able to reproduce also the distribution of knowledge distance shifts, as we report in Fig. 5. Similarly to the real system, the simulated distance shift distribution is peaked around zero. For the reasons explained above, the collaborations in our model have an overall null (or very weak) effect on the knowledge distances between agents. However, given the complex network structure, we also find a number of cases in which the two partners find themselves farther away in the knowledge space than they were at the moment of the collaboration establishment. Remarkably, our model can retrieve this positive right-tail of the knowledge distance shift distribution, even if the microscopic rules do not include any drift, nor self-motion, nor distancing mechanisms for the agents.

## 5 Introducing a performance indicator

We now want to study the performance of the whole collaboration network. To this purpose, we propose an indicator taking into account the global knowledge exploration of the systems, i.e. it quantifies the distance traveled by all agents during the evolution of our simulated R&D network. The underlying assumption is that the exploration of as many locations as possible is beneficial for the collaboration network, in that it allows the agents to come in contact with many technological opportunities, potentially leading to more frequent innovations (Fagiolo and Dosi, 2003). We call our indicator the *collaboration performance* $\mathcal{C}$ of the network and define it as:

$$\mathcal{C} = \int_{t=0}^{T_{\max}} \frac{N^{-1} \cdot \sum_{i=1}^{N} |\dot{\mathbf{x}}_i(t)|}{N^{-1} \cdot \sum_{i=1}^{N} k_i^{\mathrm{act}}(t)} \, \mathrm{d}t = \int_{t=0}^{T_{\max}} \frac{\sum_{i=1}^{N} |\dot{\mathbf{x}}_i(t)|}{\sum_{i=1}^{N} k_i^{\mathrm{act}}(t)} \, \mathrm{d}t$$

(6)

The quantity at the numerator $\sum_i |\dot{\mathbf{x}}_i(t)|$ represents the total distance traveled by all agents in the network at time *t*. The measure $k_i^{\mathrm{act}}(t)$ is defined as the number of active links incident on an agent *I*, or – in other words – the number of active collaboration in which an agent *i* is involved at time *t*. In this regard, we remember that not all collaborations are active at a given time t; some are terminated and become inactive. Therefore, the ratio inside the integral in Eq. 6 expresses the total distance traveled by the agents in the network per active link at time *t* (a sort of instantaneous collaboration performance). This measure is then integrated over the duration $T^{\max}$ of the simulation, to obtain the collaboration performance $\mathcal{C}$. The quantity at the denominator of Eq. 6 can be thought of as the number of active links in the network at time t, which we indicate with $M^{\mathrm{act}}(t)$,[1] multiplied by a factor 2. By plugging this into Eq. 6, we obtain:

$$\mathcal{C} = \int_{t=0}^{T_{\max}} \frac{\sum_{i=1}^{N} |\dot{\mathbf{x}}_i(t)|}{2 \cdot M^{\mathrm{act}}(t)} \, \mathrm{d}t$$

(7)

We use Eq. 7 to compute the collaboration performance $\mathcal{C}$ in every network we generate through the exploration of our parameter space. We report our results in Fig. 10, by making use of a heatmap to nicely visualize the average performance for every parameter combination.

We find that the configurations having the highest collaboration performance are located in one region of the parameter space, exhibiting high approach rates and short characteristic alliance life times. This means that a network with the highest collaboration performance exhibits links with (i) a short characteristic life time and (ii) allowing for a fast knowledge transfer between the involved partners. While the dependence of the performance on the approach rate $\mu$ is easily predictable, the effect of the collaboration life time $\tau$ is not trivial, given all the complex interdependencies between the network dynamics and the motion of the agents in the knowledge space.

*1 From network theory, we know that at any given time t, the sum of all node degrees $k_i$ equals the number of network links M multiplied by two.*

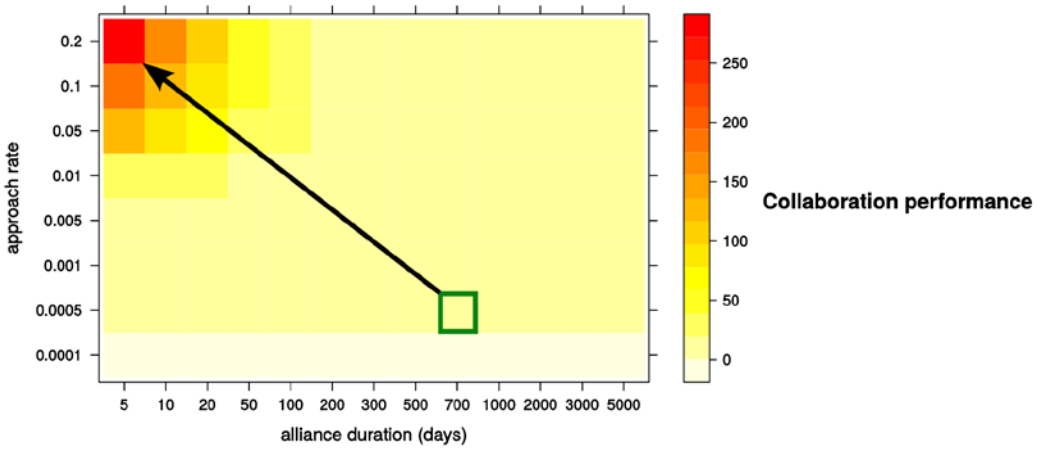


*Figure 6: Collaboration performance of the simulated networks, as a function of the characteristic alliance duration and the approach rate. The green square in the parameter space represents the position occupied by the closest simulated networks to the real data.*

We argue that a short collaboration life time is beneficial for the performance of the collaboration network, because a reduced number of collaborations allows an agent to move efficiently along one or a few directions in the knowledge space. When the characteristic life time $\tau$ increases, more links are active at the same time, thus forcing the agents to cope with the effect of multiple partnerships; this results in a reduced exploration of the knowledge space. In other words, the density of the collaboration network increases with $\tau$ and, after a certain threshold, the addition of a new link has a negative marginal effect on the overall exploration of the knowledge space.

Indeed, we have found that there exist configurations that can be both *realistic* and *optimized* with respect to the collaboration performance at the same time, as can be seen by comparing Fig. 5 and Fig. 6 (note the red points on the main diagonal). Therefore, effective policies to obtain an improved collaboration network would incentivize shorter R&D alliances and higher knowledge exchange rates, for instance including rewards for quick co-patenting by allied firms.

## 6 Conclusions

We have developed an agent-based model that is able to reproduce both the link formation and the knowledge exchange process in a collaboration network. We have used a novel approach, by combining previous results on knowledge exchange and collaboration network growth. In this new modelling framework, agents form links based on their network features and then exchange knowledge with their partners. Our agents are endowed with three key attributes: an activity (representing their propensity to engage in new alliances), a label (representing their membership in a given circle of influence), and a position in a metric knowledge space defined by a vector.

The validation of our model against real data has been performed through a novel two-step approach. By means of the SDC alliance dataset, we have estimated the network

formation parameters, thus reproducing the topology of the resulting collaboration network. Subsequently, through the NBER dataset (on firm patents), we have estimated the knowledge exchange parameters, thus evaluating the rate at which firms exchange knowledge and the duration of the R&D alliances themselves.

We have found that the agents in our model exhibit a strong tendency to connect to network incumbents: precisely, 65% of the collaborations initiated by labeled nodes (i.e. incumbents), as well as a surprising 90% of the collaborations initiated by non-labeled nodes (i.e. newcomers), are addressed to a labeled (incumbent) partner node. In this regard, the validation of our model brings additional support to the theory of the importance of existing network structures in the formation of new R&D collaborations.

As for the knowledge exchange parameters, we find that the real R&D network is best reproduced by a configuration exhibiting a relatively low approach rate and a characteristic duration of around two years (700 days). Both our agent-based model and our empirical analysis, indeed, show that collaborations exert an overall null or weak effect on the partners' knowledge position. However, despite such an effect, some collaborations can cause extreme shifts: some bring the partners closer, while some others push them farther in the metric knowledge space.

This suggests that real firms do not significantly change their knowledge positions as a consequence of their collaborations. They rather use the available information about their mutual knowledge positions in order to establish new collaborations: this means that a firm's position, evaluated through its patents, is more a determinant than a consequence of its R&D alliances.

Finally, we have investigated the outcome of our generated networks with respect to a novel performance indicator, which we define as the distance travelled by all agents per active link. We find that the configuration exhibiting the highest performance is characterized by the shortest possible alliance duration, and the largest possible approach rate. Indeed, we have found that it is possible to obtain a configuration that is both realistic and optimized with respect to the collaboration performance. In the case of R&D alliances, obviously, it would be impossible to directly require short alliance durations or enforce a fast learning rate between real companies. However, effective policies could include, for instance, rewards for co-patenting activities from partner companies, when these are carried out as early as possible after the establishment of an R&D alliance. The goal is to push companies to always explore new knowledge positions with new partners, although limiting the duration of a single alliance, and avoiding having too many active collaborations at the same time.

In conclusion, we argue that our model can successfully reproduce both network-related and knowledge-related features of a real inter-organizational R&D network, while providing at the same time a unique methodology to estimate the network performance. In addition, we argue that our model is extendable to other collaboration systems, beyond the domain of R&D networks, provided that the agents can be unequivocally positioned in a knowledge space. This way, we can offer a complete and straightforward interpretation of the effects of knowledge exchange in a dynamically evolving collaboration network.

## References


Ahuja, G. (2000). Collaboration networks, structural holes, and innovation: A longitudinal study. Administrative science quarterly 45(3), 425–455.

Axelrod, R. (1997). The dissemination of culture. Journal of conflict resolution 41(2), 203–226.

Baum, J.; Cowan, R.; Jonard, N. (2010). Network-independent partner selection and the evolution of innovation networks. Management Science 56(11), 2094–2110.

Das, T. and Teng, B. (2000). A resource-based theory of strategic alliances. Journal of management, 26(1):31.

Deffuant, G.; Neau, D.; Amblard, F.; Weisbuch, G. (2000). Mixing beliefs among interacting agents. Advances in Complex Systems 3(4), 87–98.

DeGroot, M. H. (1974). Reaching a consensus. Journal of the American Statistical Association 69(345), 118–121.

Fagiolo, G.; Dosi, G. (2003). Exploitation, exploration and innovation in a model of endogenous growth with locally interacting agents. Structural Change and Economic Dynamics 14(3), 237-273.

Garas, A., Tomasello, M. V., and Schweitzer, F. (2014). Selection rules in alliance formation: strategic decisions or abundance of choice? ArXiv preprint, arXiv:1403.3298.

Groeber, P.; Schweitzer, F.; Press, K. (2009). How Groups Can Foster Consensus: The Case of Local Cultures. Journal of Artificial Societies and Social Simulation 12(2), 4.

Hagedoorn, J., Link, A. N., and Vonortas, N. S. (2000). Research partnerships. Research Policy, 29(4-5):567–586.

Hagedoorn, J. (2002). Inter-firm R&D partnerships: an overview of major trends and patterns since 1960. Research policy 31(4), 477–492.

Hegselmann, R.; Krause, U. (2002). Opinion dynamics and bounded confidence: models, analysis and simulation. Journal of Artificial Societies and Social Simulation 5(3).

Inkpen, A. C.; Ross, J. (2001). Why do some strategic alliances persist beyond their useful life? California Management Review 44(1), 132–148.

König, M. D.; Battiston, S.; Napoletano, M.; Schweitzer, F. (2012). The efficiency and stability of R&D networks. Games and Economic Behavior 75(2), 694–713.

Liebeskind, J. P. (1996). Knowledge, Strategy, and the Theory of the Firm. Strategic Management Journal, 17:93–109.

Mowery, D., Oxley, J., and Silverman, B. (1998). Technological overlap and interfirm cooperation: implications for the resource-based view of the firm. Research Policy, 27(5):507–523.

Newman, M. (2010). Networks: an introduction. Oxford University Press.

Phelps, C. (2003). Technological exploration: A longitudinal study of the role of recombinatory search and social capital in alliance networks. Ph.D. thesis, New York University, Graduate School of Business Administration.

Powell, W., White, D., Koput, K., and Owen-Smith, J. (2005). Network Dynamics and Field Evolution: The Growth of Interorganizational Collaboration in the Life Sciences1. American journal of sociology, 110(4):1132–1205.

Rosenkopf, L. and Almeida, P. (2003). Overcoming local search through alliances and mobility. Management science, 49(6):751–766.

Rosenkopf, L. and Schilling, M. (2007). Comparing alliance network structure across industries: observations and explanations. Strategic Entrepreneurship Journal, 1(3-4):191–209.

Tomasello, M. V; Napoletano, M.; Garas, A.; Schweitzer, F. (2013). The Rise and Fall of R&D Networks. Arxiv:1304.3623. Submitted to: Industrial and Corporate Change; *current state: under revision.*

Tomasello, M. V.; Perra, N.; Tessone, C. J.; Karsai, M.; Schweitzer, F. (2014). The Role of Endogenous and Exogenous Mechanisms in the Formation of R&D Networks. Scientific Reports 4, 5679.

Tomasello, M. V.; Tessone, C. J.; Schweitzer, F. (2015). Quantifying Knowledge Exchange in R&D Networks. *Unpublished working paper.*