Quantifying knowledge exchange in R&D networks: a data-driven model

Mario V. Tomasello[1], Claudio J. Tessone[1,2] and Frank Schweitzer[*,1]
[*]Corresponding author; E-mail: fschweitzer@ethz.ch
[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

Abstract

We develop an agent-based model to reproduce the process of link formation and to understand the effect of knowledge exchange in collaborative inter-firm networks of Research and Development (R&D) alliances. In our model, agents form links based on their previous alliance history and then exchange knowledge with their partners, thus approaching in a knowledge space. We validate our model against real data using a two-step approach. Through an inter-firm alliance dataset, we estimate the model parameters related to the alliance formation, at the same time reproducing the topology of the resulting collaboration network. Subsequently, using a dataset on firm patents, we estimate the parameters related to the process of knowledge exchange. The underlying knowledge space that we consider in our study is defined by real patent classes, allowing for a precise quantification of every firm's knowledge position. We find that real R&D alliances have a duration of around two years, and that the subsequent knowledge exchange occurs at an extremely low rate – a firm's position is rather a determinant than a consequence of its R&D alliances. Finally, we propose an indicator of collaboration performance for the whole network and, remarkably, we find that the empirical R&D network extracted from our data does not maximize such an indicator. However, we find that there exist configurations that can be both realistic and optimized with respect to the collaboration performance. Effective policies, as suggested by our model, would incentivize shorter R&D alliances and higher knowledge exchange rates.

Keywords: Inter-firm network; R&D alliances; Patents; Knowledge exchange; Agent-based model

1 Introduction

The last three decades have been characterized by a growing number of inter-firm alliances, aimed at Research and Development (R&D) purposes. Albeit this phenomenon has especially affected highly technological industries such as IT, Pharmaceuticals or Medical Supplies (Ahuja, 2000; Hagedoorn, 2002), all industrial sectors have simultaneously experienced an increased number of such alliances (Tomasello et al., 2013). Consequently, scholars have investigated the mechanisms behind the formation of R&D alliances (Powell et al., 2005), the complex networks they generate (Rosenkopf and Schilling, 2007; Tomasello et al., 2014), and the way their evolution can be

modeled (König et al., 2012; Garas et al., 2014). From a theoretical point of view, it has been shown that firms engage in alliances for several reasons. They can gain access to more and diverse assets (Liebeskind, 1996; Das and Teng, 2000). Next, alliances foster the exchange of knowledge between firms: by joining their technological resources, firms can actually enlarge their knowledge bases faster than they could do individually (Baum et al., 2000; Mowery et al., 1998; Rosenkopf and Almeida, 2003). Finally, firms 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 of the involved firms, making R&D alliances an important part of every firm's knowledge management strategy.

The focus of the present study is indeed such a learning process, which we model as a mutual exchange of knowledge occurring after the establishment of an alliance between two firms. In particular, we develop an agent-based model to investigate the determinants leading to the formation of inter-firm R&D collaborations and the subsequent emergence of an R&D network. At the same time, we study the effect that such collaborations have on the technological positions of the involved firms, and we estimate the *performance* of such networked systems, in terms of explored technological trajectories.

The approach that we adopt in our study can be defined as *data-driven modeling*. Starting from the empirical evidence, we design a set of realistic and theoretically grounded microscopic interaction rules, which we incorporate in an agent-based model; next, we implement such a model through computer simulations, followed by calibration and validation against empirical data. The fine-tuning of the model parameters gives us not only a deep understanding of the system under examination, but also an indication on how to optimize it. The model that we develop here is based on previous empirical findings (Tomasello et al., 2013; Hanaki et al., 2010; Rosenkopf and Schilling, 2007), and combines two existing agent-based models (Tomasello et al., 2014, 2015), in order to reproduce both the alliance formation and the knowledge exchange process in an R&D network.

1.1 Theoretical foundations: knowledge exchange in inter-firm R&D networks

Our agent-based model follows a number of extant works on bounded confidence and continuous opinion dynamics (Axelrod, 1997; Deffuant et al., 2000; DeGroot, 1974; Hegselmann and Krause, 2002; Groeber et al., 2009), in particular applied to innovation networks (Fischer and Fröhlich, 2001; Baum et al., 2010). In the wake of this previous work, and similar to the model proposed in Tomasello et al. (2015), we assume that the collaborating nodes are characterized by an evolving knowledge basis, that is affected by the set of alliances in which the nodes themselves take part. However, differently from the studies that have been done so far, our model does not focus on the formation of consensus clusters – see Axelrod (1997); Schweitzer and Behera (2009) in the case of social systems, or Fagiolo and Dosi (2003) for technology islands. In addition, our work does

not consider the network of R&D alliances as fixed, but it assumes a dynamically evolving R&D network, whose topology corresponds to those of empirically observed networks (see Tomasello et al., 2013; Gulati et al., 2012).

The knowledge-based view of the firm (Fischer and Fröhlich, 2001) assumes that every company is endowed with a knowledge basis that uniquely identifies its resources and capabilities. In other words, a firm can always be associated with a vector consisting of several components (Sampson, 2007), each of which represents its level of knowledge in a given area. As we explain below, these vectors can in turn be associated with a metric *knowledge space* in which the collaborations occur. Thus, every firm occupies a point in this multi-dimensional space, whose coordinates are given by its knowledge vector. Such an approach is similar to a more general model (Axelrod, 1997), proposed in the broader context of social influence. The concept of a metric knowledge space has already been used in one dimension (Groeber et al., 2009), and in two dimensions (Fagiolo and Dosi, 2003; Baum et al., 2010); we here generalize this approach to metric spaces of arbitrary dimensionality.

On the other hand, R&D alliances have been conceptualized by several studies (Mowery et al., 1998; Owen-Smith and Powell, 2004; Grant and Baden-Fuller, 2004; Gomes-Casseres et al., 2006) as a means to exchange technological knowledge among firms, and such an idea is at the heart of several agent based models (Pyka and Fagiolo, 2007; Gilbert, 2004; Cowan et al., 2007). In these models, the agents' knowledge bases become more similar over time, as a consequence of R&D collaborations. The speed at which the agents approach each other in the knowledge space represents one of the fundamental parameters of this family of models, and our work is no exception. Besides, we rely on the assumption that knowledge spillovers occurring in a R&D alliance cause the partners to exchange knowledge along every dimension of their knowledge bases, not limiting the transfer to a specific R&D project that they have in common (Baum et al., 2010). In other words, we study a scenario in which the two partners approach with respect to all dimensions of the knowledge space.

Finally, we want to study the *performance* of the whole collaboration network as a function of the relevant model parameters. The indicator that we propose to measure such a performance takes into account the global knowledge exploration of the systems. I.e., it actually computes the distance traveled by all agents during the evolution of our simulated R&D network. In our model, we consider that the knowledge exploration itself is represented by the motion in the knowledge space, which is fully captured by such a measure. The underlying assumption is that the exploration of as many locations as possible is beneficial for the R&D 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). Precisely, we make use of an existing performance indicator (Tomasello et al., 2015) and refine it by taking into account the actual number of active collaborations in the system, in order to obtain a more reliable measure.

1.2 Theoretical foundations: formation of inter-firm R&D networks

The extant literature on R&D networks has shown that two types of mechanisms are crucial for the formation of new R&D alliances (Rosenkopf and Padula, 2008): endogenous mechanisms (previous alliances and previous network structures) and exogenous mechanisms (exploratory search of new partners). However, both empirical and theoretical studies have mainly focused only on one of the two mechanisms, also called "network endogeneity" (Walker et al., 1997; Powell et al., 1996; Gulati and Gargiulo, 1999; Garas et al., 2014) and "exogenous partner selection" (Burt, 1992; Rosenkopf and Nerkar, 2001; Cowan et al., 2004) respectively. Typically, the concept of endogenous and exogenous mechanisms has been used in the management literature with respect to the belonging of the firms to the R&D network. We follow such a definition and refer to an alliance as "endogenous" if it involves a partner that is already part of the R&D network. Likewise, an alliance is referred to as "exogenous" if it involves a partner that is not part of the R&D network yet.

While the endogenous mechanisms depend on the firms' social capital (describing their position in the network), the exogenous mechanisms are affected by the firms' technological and commercial capital. A firm's social capital can be further explained by two variables (Gulati, 1995; Podolny, 1993): its *prominence* – i.e. the history of its previous alliances – and its *cohesiveness*, defined as the set of its direct and indirect links with other firms in the network. In this regard, some empirical studies (Powell et al., 1996; Rosenkopf and Padula, 2008) found that several firm "clusters" populate the R&D network, thus giving rise to different kinds of alliances depending on the firms' position in the network. In particular, three categories of R&D alliances have been identified: i. within-cluster alliances (the partners belong to the same cluster); ii. semi-distant alliances (the partners form a so-called "shortcut" between two different clusters); iii. distant alliances (at least one of the partners is an isolated node, i.e. a newcomer firm). Obviously, a certain number of R&D alliances is not explained by the partners' social capital – think, for instance, of alliances involving start-up companies or financial institutions that have no previous experience in R&D activities. One rationale to search for this kind of partners, whose technological and commercial capital plays a crucial role, is that they can provide access to new information or unique technical knowledge.

However, neither the network endogeneity nor the exogenous partner selection, taken independently, are able to explain the topology of observed R&D networks. Endogenous mechanisms alone would lead to more and more centralized network structures over time, which we do not observe in reality (Tomasello et al., 2013). On the other hand, exogenous mechanisms alone would lead to more regular networks topologies, which we do not observe either (a prominent example is represented by the "monogamous" networks analyzed in Tomasello et al., 2015). A notable exception is the agent-based model developed by Tomasello et al. (2014), which incorporates both endogenous and exogenous rules of alliance formation and is successfully able to reproduce the structure of a real R&D network.

Inspired by these observations, the agent-based model that we develop in the present study includes all the microscopic rules introduced in Tomasello et al. (2014), and combines them with the knowledge exchange rules briefly discussed above. Our model allows us to modulate the weight of both endogenous and exogenous mechanisms for alliance formation, and to test the outcome against real data. Consistently with Tomasello et al. (2014), we obtain indications that the endogenous mechanisms of alliance formation play a more relevant role than the exogenous ones for the growth of the R&D network.

1.3 Our contribution

As previously mentioned, the model developed here combines two existing agent-based models in a straightforward, yet effective, manner. The model introduced in Tomasello et al. (2015) represents a first attempt to investigate the process of knowledge exchange occurring in a dynamic collaboration network; it has identified a mechanism of volatile alliances to help the collaborating agents better explore a knowledge space, using the approximation of monogamous (i.e. sparse) collaboration networks. On the other hand, the model developed in Tomasello et al. (2014) can realistically reproduce the complex topology of real R&D networks, but without considering the effect of alliances on the firms' knowledge positions.

The agent-based model that we introduce here constitutes an important step toward a general modeling framework that combines two dynamic processes, the formation of alliances and the knowledge exchange in collaboration networks. The microscopic interaction rules of our model, as well as its validation, involve a two-step procedure that can be described as follows. The firms form R&D collaborations based on their network features and their social capital; the model parameters related to these mechanisms are estimated through a comprehensive inter-firm alliance dataset. Next, we assume that the formation of each network link causes a process of knowledge exchange between the involved firms, which consequently approach in the knowledge space; the model parameters related to this mechanism are estimated through a second dataset on firm patents. Remarkably, the underlying knowledge space that we consider in our study is defined by real patent classes, allowing for a precise quantification of every firm's technological position.

Our findings point out a predominance of the endogenous network mechanisms (over the exogenous ones) for the alliance formation; in other words, previous network structures and alliance history matter when selecting new collaboration partners. Next, we find that real R&D alliances have a duration of around two years, and that the subsequent knowledge exchange between the partners occurs at a very low rate. Most of the alliances, indeed, have no consequence on the partners' knowledge position: 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 investigate the performance of such a network in terms of explored knowledge trajectories, and we check whether the real R&D network under examination maximizes our proposed performance indicator. Inter-

estingly, we find that this is not the case: effective policies to obtain an optimized collaboration network – as suggested by our model – would incentivize shorter R& D alliances and higher knowledge exchange rates.

The rest of the paper is organized as follows. Section 2 presents the datasets and the methodology used to build the network, as well as to measure the firms' knowledge positions. Section 3 describes all the microscopic interaction rules defining our agent-based model. Sections 4.1 and 4.2 present the results of our computer simulations and the model validation on the alliance and the patent datasets, respectively. In Section 5, we introduce our performance indicator and study the optimality of the real R&D network under examination. Finally, Section 6 concludes.

2 Data and Methodology

We define an R&D network as a set of nodes, or agents (the firms), and links (the alliances). 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 (Thomson-Reuters, 2013), 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 (consortium), all the involved firms are connected in pairs, resulting into a fully connected clique. This procedure is consistent with a previous empirical study (Tomasello et al., 2013), where there is no conceptual difference between a consortium and a "standard" two-partner alliance, which is only a special case of it (and can be thought of as a fully connected clique of size 2).

A quantity that we measure directly from the data prior to the implementation of our agent-based model itself is the firms' *activity* distribution.[1] The activity expresses the probability that a firm takes part in any alliance event occurring in a given time window. For the validation of the present model, we use the overall firm activity, measured on the entire observation period of the dataset. We define such activity a_i, for a firm i, as the number of alliance events e_i involving firm i divided by the total number of alliance events E involving *any* firm reported in the dataset. We then assign such empirical activities a_i to the agents in our computer simulations (for the empirical activity distribution, see Fig. 1).

The SDC Platinum database (Thomson-Reuters, 2013) reports approximately 672,000 publicly announced alliances in all countries, from 1984 to 2009, with a granularity of 1 day, between several kinds of economic actors (including manufacturing firms, investors, banks and universities).

[1] For a more detailed definition and more empirical examples on agents' activity in collaboration networks see Tomasello et al. (2014).

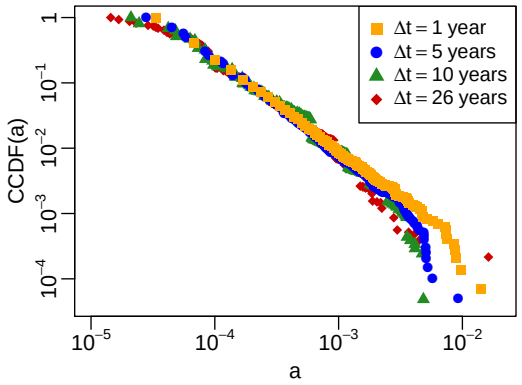

Figure 1: Complementary cumulative distribution function (CCDF) of the empirical firm activities, measured on the SDC dataset with 4 different time windows Δt of 1, 5, 10 and 26 years. When the time window is shorter than 26 years (the entire dataset observation period), we compute the activity by shifting the time window in 1-year increments and then we average the results.

We select all the alliances characterized by the "R&D" flag; after applying this filter, a total of 14,829 alliances, connecting 14,561 firms, are listed in the dataset. However, we keep in our network representation only the firms that have a corresponding entry in the patent dataset, that we utilize to determine their knowledge positions. Such database is the Patent Citations Data by the U.S.A. National Bureau of Economic Research (NBER), that contains 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 first consideration has to be made on the number of classes we take into account, which will correspond to the dimensionality of the knowledge space in which the firms are located. The IPC, introduced in 1971 by the *Strasbourg Agreement*, is a hierarchical system of symbols for the classification of patents according to the different areas of technology to which they pertain.[2] A generic IPC 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". This four-character term is then followed by a group/subgroup indication, represented by additional digits. A typical IPC term can be written as follows: B34H 6/99. The sections identified by the IPC are historically stable and amount to 8, from A (human necessities)

[2]For more information on the International Patent Classification, see `http://www.wipo.int/classifications/ipc`.

to H (electricity). The lower levels are instead subject to more frequent revisions; the eighth and last IPC edition consists of more than 120 classes, 600 subclasses, 7,000 main groups and 60,000 subgroups.

We intend to test our model on a broad set of firms, belonging to several industrial sectors, and therefore exhibiting patent activities distributed across all sections, classes and subclasses. Hence, we consider only the section symbol (i.e. the first letter) in our empirical patent classification. Choosing a class- or subclass-level division would result in an excessive patent granularity, meaning a high dimensionality for the corresponding knowledge space. However, for the sake of completeness, we have also tested a division at a class level (i.e. the first letter plus two digits), obtaining a total of 74 classes; we find that the computational burden of operating in a 74-dimensional space does not lead to any significant change in our results.[3]

The titles of the 8 sections, as well as a patent count for each section in our dataset, is reported in Table 1. We find that the number of patents in all sections reflects their technological dynamism (Rosenkopf and Schilling, 2007); indeed, all sections are fairly equally represented, and the two sections exhibiting the lowest patent counts are textiles, paper and fixed constructions, two typical mature industries.

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

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.

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:

$$x_{is} \equiv \frac{N_{is}}{\sum_s N_{is}} \qquad s = A, \ldots, H \tag{1}$$

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, similar to Tomasello et al. (2015). This means that the knowledge distance between two firms i and j reads as:

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

[3] We do not show these results here. The plots are available from the authors upon request.

Preliminary empirical findings. Using the definitions provided in Eqs. 1 and 2, we now compute two empirical measures that will later be used for the validation of our model, namely (i) the knowledge positions of the 5,168 firms listed in our dataset at the beginning of the observation period – i.e. in 1984 – and (ii) the distribution of the knowledge distances between every pair of allied firms, at the moment of alliance formation. When computing the empirical knowledge position of a firm $\mathbf{x}_i$ at a given date t, we consider all the patents for which the firm has applied in a given time window Δt preceding such date t. In order to have a reliable and updated measurement, without losing at the same time too much patent information due to a short time window, we use a length equal to 5 years. 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. 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. In Fig. 2 we report the distribution of the knowledge distances between partner firms at the moment of alliance formation – from now on, the "pre-alliance knowledge distances". The minimum observed value of knowledge distance is 0, while the maximum value of knowledge distance equals $\sqrt{2}$, for normalization reasons. We find that the distribution is peaked around an intermediate distance and left-skewed, i.e. shifted toward small values. In addition, we observe that the counts drop when such preferred distance approaches zero, meaning that firms with the exact same patenting activity tend not to form alliances.

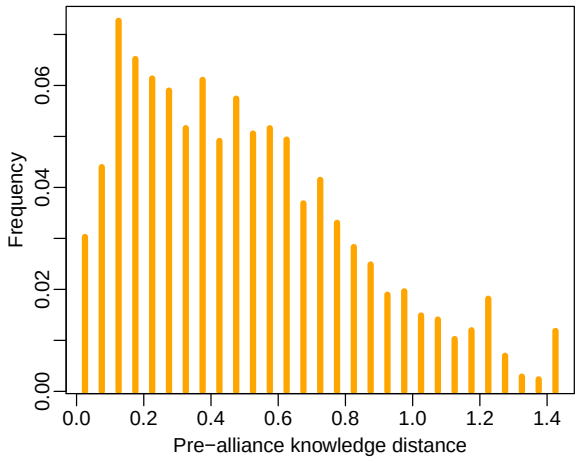

Figure 2: Empirical knowledge distance between every pair of partnered firms, as of the day preceding the alliance formation.

We use the aforementioned distribution to later validate our agent based model. We estimate the value of the knowledge exchange parameters, together with another empirical measure carrying a second, important piece of information: the distribution of the knowledge distances between every pair of allied firms, at the moment of alliance *termination* – from now on, the "post-

alliance knowledge distances". However, the SDC dataset does not report the ending date of any alliance. To overcome this problem, during the validation of the model, we compute the empirical knowledge distance between every pair of linked firms, after a time period equal to the value of the parameter τ (in days) used in the corresponding simulation. The NBER patent dataset has a time-granularity of 1 year, thus forcing us to use a minimum 1-year time window, even when considering τ values smaller than 365 days. Nevertheless, we find that the length of such time window does not affect our results. Precisely, we find that the shape of the knowledge distance distribution appears to have the same shape, irrespectively of the time period following the alliance formation when these distances are computed, even when such time window is reduced to zero. This means that the distribution of post-alliance knowledge distances resembles the one of pre-alliance distances. In Fig. 3 we report the post-alliance knowledge distance distribution for different time windows of length 1, 3, 5 and 10 years.

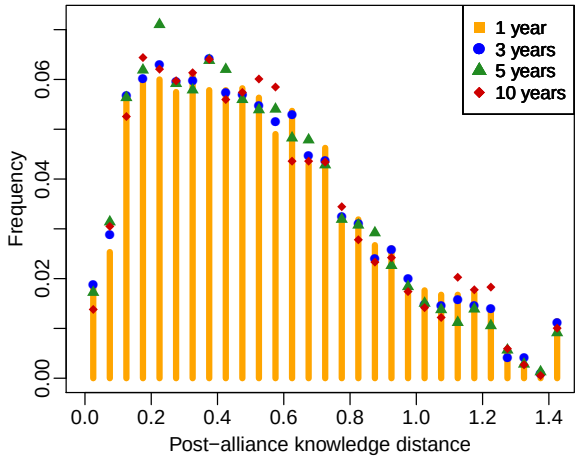

Figure 3: Empirical knowledge distance between every pair of partnered firms, computed 1, 3, 5 and 10 years after the date of the alliance formation.

The fact that the distribution of post-alliance distances differs only slightly from the distribution of pre-alliance distances is in agreement with another, last empirical measure we compute prior to the validation of our model. We calculate the variation of the knowledge distance separating every pair of allied firms between the moments of alliance formation and alliance termination – from now on, the "knowledge distance shift". Again, we report our results in Fig. 4 for four time windows of length equal to 1, 3, 5 and 10 years. We find that the distribution of distance shifts is virtually independent of the chosen time window, as we could already expect from the distribution of the post-alliance knowledge distances. More importantly, the distribution of distance shifts is narrow and centered around zero, confirming our previous finding that post-alliance knowledge distances are subject to an overall weak change as a consequence of R&D alliances.

When looking at the knowledge distance shifts, we do indeed find that most of the R&D alliances

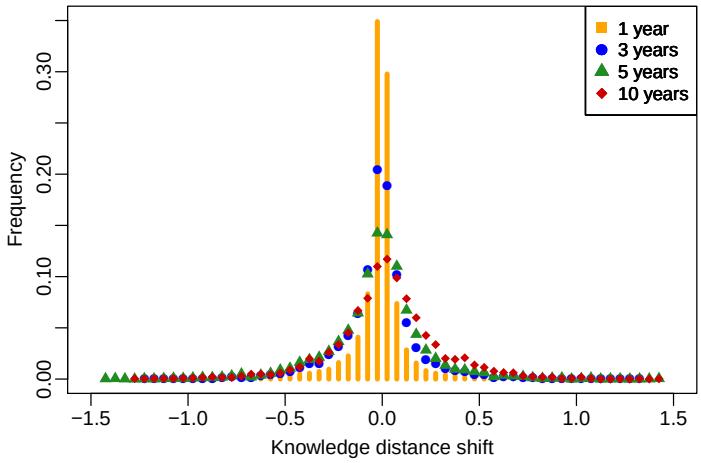

Figure 4: Empirical shift of knowledge distance between every pair of partnered firms, computed 1, 3, 5 and 10 years after the date of the alliance formation.

cause a null change in the knowledge distance between the two partners. However, the 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. As we show through the validation of our agent based model, it can be generated even by microscopic rules considering only an *approach* of the agents in the knowledge space, provided that this is coupled with a complex network dynamics.

3 The model

We now describe the microscopic interaction rules of our agent-based model, divided into two phases. First, the agents form links based on their network features and their social capital; we call such a phase "exploration (link formation)". Second, they exchange knowledge through these links, thus approaching each other in a metric knowledge space; we call this phase "exploitation (knowledge transfer)". In addition, each link can be terminated with a given probability.

3.1 Exploration: link formation

Node activation. We consider a network composed of N nodes; each of them is endowed with two fundamental attributes, an *activity* and a *label*. 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 (see Section 2). The activity defines the propensity of each node to be involved in a collaboration

event. In particular, at every time step, a node i initiates an alliance with probability $p_i = \eta a_i \mathrm{d}t$, and the number of active nodes N_A is:

$$N_A = \eta \langle a \rangle N \mathrm{d}t, \tag{3}$$

where $\langle a \rangle$ is the average node activity and η is a rescaling factor that allows to adjust the activation rates, and consequently the number of active nodes per time step. We find that the model is robust to the choice of η, showing no measurable changes for η 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 $\mathrm{d}t$.

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 number of partners per alliance, directly measured from the SDC Platinum dataset. In other words, we shuffle the sequence of number of partners per alliance and then extract a value every time an activation event occurs; m can be thought of as the number of partners involved in every 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. We remember that the labels model the belonging of the agents to different groups that they implicitly define with their shared practices and/or behaviors. In the example of firms forming R&D alliances, 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. This mechanism can very effectively explain the presence of clusters, or communities, in R&D networks (see Section 1.2), as also testified by the work of Tomasello et al. (2014). 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 (see Fig. 5 for a visual example).

Selection of the partner categories. The presence of labels – as we have already seen in our previous link formation model – induces different types of alliances, that we explicitly distinguish. In particular, if the initiator is a labeled node, this can link to a labeled node having the same

label (with probability p_s^L), or to a node having a different label (p_d^L), or to a node without label (p_n^L). If the initiator is a non-labeled node, i.e. it is a *newcomer* in the collaboration network, this can link to a labeled node (with probability p_l^{NL}), or to another non-labeled node (p_{nl}^{NL}). Similar to our previous model, we define the formation of a link with a labeled node (described by the probabilities p_s^L, p_d^L and p_l^{NL}) as an *endogenous mechanisms*, given that the initiator of the alliance has information about the network position (i.e. social capital) of its potential partners. Likewise, we define the connection with a non-labeled node (events p_n^L and p_{nl}^{NL}) as an *exogenous mechanisms*: in this case, the initiator cannot have any information about the social capital of an agent that is not part of the network yet. As we have done for our previous model, we refer to these mechanisms as endogenous or endogenous with respect to the *network topology* and the *label attributes*. However, the model we now develop includes also rules which are exogenous with respect to the network topology, namely the approach in the knowledge space and the termination of some links.

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 (i.e. number of previous collaborations with distinct partners). 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 (see Fig. 5). 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 transfer

The second group of microscopic rules models a process of knowledge exchange between pairs of collaborating agents, similar to what has been investigated in Tomasello et al. (2015). Basically, we assume that every agent in the network is located in a metric knowledge space and – as a consequence of its collaborations – approaches its partners in this space. In case of multiple partners, the motion of the focal node is determined by the vectorial sum of all the effects due to each of its partners.

Location in a metric knowledge space. Every agent i is a point with coordinates $\mathbf{x_i}$, identified by a vector of D real numbers ranging from 0 to 1. In the case of R&D networks, 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. In order to validate this model against the data, we assign all agents' initial positions by using real patent data, as we explain in more detail in Section 4.

$$\mathbf{x}_i \equiv (x_{i1}, x_{i2}, \ldots, x_{iD}) \qquad i = 1, \ldots, n \tag{4}$$

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. Like in our previous model, we assume that every agent is endowed with a *learning rate* μ. 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 μ 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 μ 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:

$$\dot{\mathbf{x}}_i(t) = \mu \sum_{j \in \mathcal{N}_i(t)} [\mathbf{x}_j(t) - \mathbf{x}_i(t)] \tag{5}$$

where $\mathcal{N}_i(t)$ is the set of partners of the agent i at time t. As we can observe from Equation 5, in the present model there is no proximity condition for the agents' approach in the knowledge space, differently from other existing models (see, for instance, Groeber et al., 2009; Baum et al., 2010; Tomasello et al., 2015). Here, such an effect is fully captured by the network formation mechanisms and its strategies (i.e. the different link probabilities), which are independent of the agents' knowledge positions; every link (i.e. every collaboration) has then the effect to make the involved partners approach in the knowledge space. We then implement the model through computer simulations, using discrete time steps of length $\mathrm{d}t$. The evolution of every agent's position $\mathbf{x_i}$ can be expressed as:

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

Alliance termination. R&D alliances have been proven to have a finite duration (Phelps, 2003; Tomasello et al., 2013). In order to develop a realistic model, we introduce a key parameter, precisely a link characteristic life time τ. We assume that the durations of the alliances are distributed according to a Poisson process with rate $1/\tau$; the mean duration is obviously equal to τ. In our computer simulations, which use discrete time steps of length $\mathrm{d}t$, this translates into the use of a fixed termination probability p_T for any link at any time step, equal to $p_\mathrm{T} = \mathrm{d}t/\tau$.

In order to keep a simplistic set of rules, we assume that the parameter τ is independent of any other feature of the network or the knowledge exchange dynamics.[4]

To sum up, in this section we have described a set of microscopic rules which aim at reproducing the formation of links in a collaboration network, together with the approach of the agents in an underlying knowledge space. We summarize the model microscopic rules by means of a visual example in Fig. 5 and report the nomenclature of all parameters in Table 2.

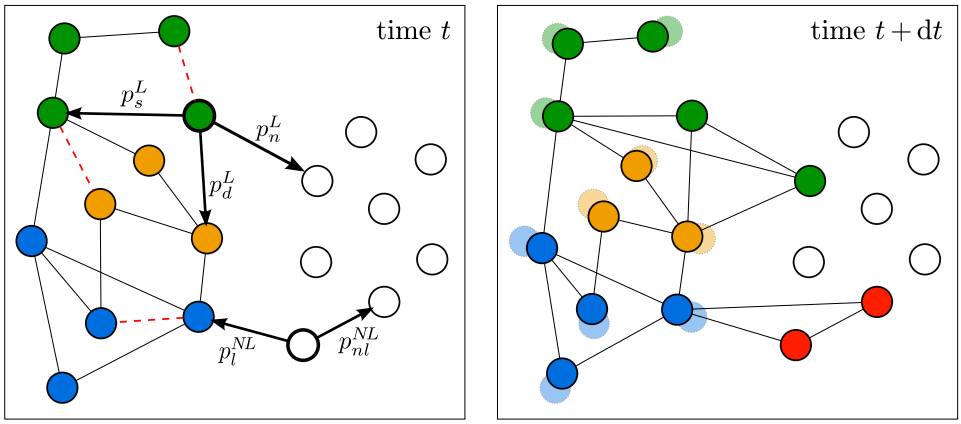

Figure 5: A representative example of network evolution in a bi-dimensional ($D = 2$) knowledge space. The position of the nodes in the plot corresponds to their coordinates in the knowledge space. At time $t + \mathrm{d}t$, all existing links cause the respective agents to approach in the knowledge 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 $t + \mathrm{d}t$, 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.

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
μ	Approaching rate in the knowledge space	Knowledge exchange
τ	Link characteristic life time	Knowledge exchange

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.

[4]One possible extension would be to link τ to the knowledge distance of the two partners, or some other network-related feature.

4 Model validation with a two-step procedure

We now validate our model against the data, in order to estimate the value of its parameters. As already mentioned, we perform our validation procedure in two steps and by using two datasets, R&D alliances and patents.

In the *first step*, we validate the network topology. We fix a set of parameters that we can directly measure from the data (namely, the number of agents and collaborations, the agents' activity distribution and the size of collaboration events). We then estimate the remaining parameters – i.e. p_s^L, p_d^L and p_{nl}^{NL} – by running a set of computer simulations and identifying the simulated collaboration network that matches best 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 – i.e. D, μ and τ – by identifying the simulated collaboration network that best matches with the patent dataset.

4.1 Network formation parameters

The approach that we use to estimate the network parameters p_s^L, p_d^L and p_{nl}^{NL} can be explained as follows. We fix the model parameters that we can directly measure from the data, namely 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, $E = 7,417$. We vary the values of p_s^L, p_d^L and p_{nl}^{NL} in discrete steps spaced by 0.05, in the interval $(0, 1)$. The parameters p_s^L and p_d^L are bounded by the condition $p_n^L = 1 - p_s^L - p_d^L \geq 0$, meaning that their sum has to be smaller or equal to 1. This condition translates into $3,249$ points to explore in the 3-dimensional parameter space, for each of which we run 100 simulations (for a total of $324,900$ runs).

The networks that we generate by means of computer simulations are matched to the data with respect to three global indicators: average degree $\langle k \rangle$, average path length $\langle l \rangle$, and global clustering coefficient C,[5] For the empirically observed R&D network, we denote such measures as $\langle k \rangle^{OBS}$, $\langle l \rangle^{OBS}$ and C^{OBS}, respectively, and their values are $\langle k \rangle^{OBS} = 3.45$, $\langle l \rangle^{OBS} = 5.05$ and $C^{OBS} = 0.11$. [6]

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, similar to Tomasello et al. (2014). We

[5] For a rigorous definition of these measures, see Tomasello et al. (2014).

[6] We find that the present network is slightly denser, more clustered, with a shorter average path length than the R&D network analyzed in Tomasello et al. (2014). This happens because we now consider only the firms for which patent data are available, not just any firm reported in the SDC alliance dataset. These firms typically have more alliance partners than average, thus making the resulting network more dense and connected.

do not have a set of observations against which we can validate our model; instead, we only have one empirical point: the real R&D network. In particular, we cannot consider the three measures $\langle k \rangle$, $\langle l \rangle$ and C as independent, therefore the Likelihood function $\mathcal{L}$ reads as:

$$\mathcal{L}(p|net^{OBS}) = f(net^{OBS}|p) \tag{7}$$

where $f(\cdot)$ is the joint density function of all parameter combinations p resulting in a network that is equivalent to the observed one, net^{OBS}. Both p and net^{OBS} are vectors with three components, expressing respectively the three model parameters $p \equiv (p_s^L, p_d^L, p_{nl}^{NL})$ and the three global network measures $net^{OBS} \equiv \left(\langle k \rangle^{OBS}, \langle l \rangle^{OBS}, C^{OBS}\right)$. Therefore, we need to find the parameter combination $(p_s^L, p_d^L, p_{nl}^{NL})$ maximizing the Likelihood $\mathcal{L}(p|net^{OBS})$ to generate a network whose macroscopic properties are *sufficiently similar* to the real network net^{OBS}. By this, we mean that the relative errors from the observed values for the average degree $\varepsilon_{\langle k \rangle}$, the average path length $\varepsilon_{\langle l \rangle}$ and the global clustering coefficient ε_C have to be smaller than a certain threshold ε^0. We empirically compute the Likelihood function $\mathcal{L}$ for each point in the parameter space by counting the fraction of its 100 simulation realizations that fulfill the criteria $\varepsilon_{\langle k \rangle} < \varepsilon^0$; $\varepsilon_{\langle l \rangle} < \varepsilon^0$; $\varepsilon_C < \varepsilon^0$. This way, we obtain values that can range from 0 (no realization of that parameter combination fulfills the criteria) to 1 (all of its realizations fulfill the criteria).

For the choice of the error threshold ε^0, we take a conservative approach and use $\varepsilon^0 = 0.02$, that ensures a good matching with the real R&D network, without cutting out too many points in the parameter space. The corresponding Likelihood scores are reported in Fig. 6 by means of a 3-dimensional color map, where the color scale is representative of the Likelihood. To have a more detailed representation of the likelihood scores, we also show one slice of the parameter space obtained by fixing the parameter p_s^L to 0.45, corresponding to the highest likelihood score region, always using the error threshold $\varepsilon^0 = 0.02$. The 2-dimensional color map reported in Fig. 6 depicts the likelihood score as a function of the other two free parameters p_d^L and p_{nl}^{NL}.

We find that the point with the highest likelihood score has the following coordinates in the parameter space: $p_s^{*L} = 0.45$, $p_d^{*L} = 0.2$ and $p_{nl}^{*NL} = 0.1$. This means that labeled nodes exhibit a fairly balanced alliance strategy, with $p_s^{*L} = 0.45$, $p_d^{*L} = 0.2$, and consequently $p_n^{*L} = 0.35$, while the non-labeled nodes exhibit a very strong tendency to connect to labeled nodes ($p_l^{*NL} = 0.9$), as opposed to a low linking probability with other non-labeled nodes ($p_{nl}^{*NL} = 0.1$). 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.

These results are in line with those we have presented in Tomasello et al. (2014). However, the R&D network with patent data exhibits an even stronger tendency to favor connections with labeled nodes (i.e. incumbent firms) than the pooled R&D network including all firms, irrespectively of their patenting activity. Let us spend a few words on the comparison between these two networks.

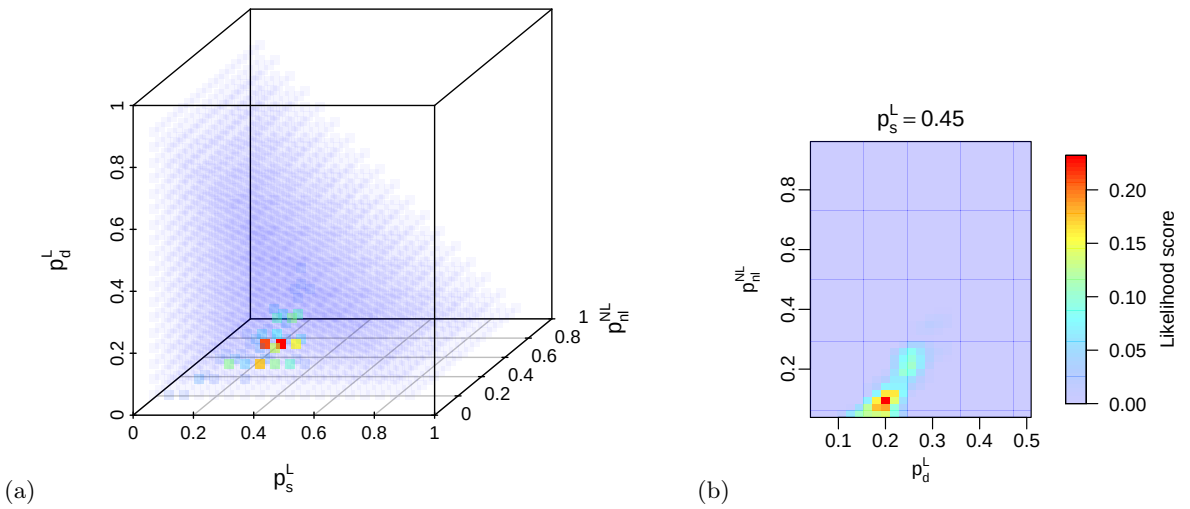

Figure 6: Likelihood scores for all points in the parameter space, for $\varepsilon^0 = 2\%$, represented with a 3-dimensional color map (a). After fixing the value of p_s^L to 0.45 (b), we report the Likelihood score as a function of p_d^L and p_{nl}^{NL}, using the same color scale.

Optimal simulated R&D network				**Real R&D network (with patents)**	
Model parameter	Value	Measure	Value	Measure	Value
p_s^{*L}	0.45	$\langle k \rangle^*$	3.48 ± 0.01	$\langle k \rangle^{OBS}$	3.45
p_d^{*L}	0.2	$\langle l \rangle^*$	5.02 ± 0.08	$\langle l \rangle^{OBS}$	5.05
p_n^{*L}	0.35	C^*	0.111 ± 0.007	C^{OBS}	0.109
p_{nl}^{*NL}	0.1				
p_l^{*NL}	0.9				

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.

Due to the fact that our analysis in now restricted only to firms for which patent data are available, one could expect either an increase in the importance of network endogenous mechanisms – given that we are considering, on the one hand, larger and more active firms – or an increase in the importance of exogenous mechanisms – given that we are considering, on the other hand, firms for which the technological dimension could be more relevant in the alliance formation strategy. Our data confirm the first hypothesis, that is the increase in the relevance of network endogenous mechanisms, which results in higher probabilities for the agents to collaborate with agents that are already part of the network, and therefore already labeled. This behavior is present irrespectively of whether the alliance event is initiated by a labeled or a non-labeled node: precisely, 65% of the collaborations initiated by labeled nodes ($p_s^{*L} + p_d^{*L}$), as well as 90% of the collaborations initiated by non-labeled nodes (p_l^{*NL}), involve a labeled node as a partner.

4.2 Knowledge exchange parameters

We determine the values of the remaining model parameters – i.e. the knowledge exchange parameters – by fixing all the network formation parameters to the values resulting from the first validation step, described in Section 4.1. We then fix the value of one knowledge exchange parameter that we can directly measure from the data, namely the dimensionality D of the knowledge space. We use the eight main sections of the IPC scheme and measure the fractions – not the numbers – of patents in each section, thus giving rise to one bounding condition. Hence, we assume $D = 7$. Consequently, the 7 numbers identifying the knowledge position of every agent are free to vary independently of each other in our simulations; the eighth component of the knowledge position can be inferred from the bounding condition that the patent fractions in every section have to sum up to 1. Obviously, each of the seven x_{is} knowledge components we use in our simulations is bound to be smaller than 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, the agents' approaching rate μ and the characteristic alliance life time τ. We consider the values 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1 and 0.2 for the parameter μ and the values 5, 10, 20, 50, 100, 200, 300, 500, 700, 1000, 2000, 3000 and 5000 for the parameter τ, thus having a total of 104 points to explore in the parameter space. The interpretation of the parameter τ is straightforward: as explained in Section 3.1, we adjust the activation rate of the agents such that the length of a time step dt can be directly interpreted as 1 day. Therefore, the value of τ, which is by design expressed in time steps, can be thought of as the characteristic duration of a real alliance in days.

For each of the 104 parameter combinations, we run 100 simulations, for a total of $10,400$ runs in this second step of our validation procedure. We store the distributions of pre-alliance knowledge distances, post-alliance knowledge distances and knowledge distance shifts in each run. Similar to the first step, we stop every computer simulation when the total number of collaborations equals the number of alliance events reported in the SDC dataset, $E = 7,417$. Finally, we consider each of the collaboration networks resulting from the simulations and compare it to the empirical R&D network, with respect to the first two characteristic distributions, namely the pre-alliance and the post-alliance knowledge distances (see Section 2). We do not use in our validation procedure the third distribution, i.e. the knowledge distance shifts, because it strictly depends on the first two and does not carry any additional information.

While the pre-alliance knowledge distances are unambiguously computable on both the empirical and the simulated networks, the post-alliance knowledge distances are unambiguously computable only on the simulated networks – where every link comes to an end after a definite time. However, alliance ending dates are not available on the real R&D network. To overcome this problem, we compute the empirical knowledge distance between every pair of linked firms after a time period equal to the value of the parameter τ – in days – used in the corresponding simulation.

We use two-sided Kolmogorov-Smirnov (KS) tests to compare each simulated knowledge distance distribution with the corresponding empirical one, and therefore assign a score to every parameter combination. Precisely, we record the value of the resulting D statistics for every KS test we perform; such a value expresses how close two distributions are, and decreases as the two distributions under examination become more similar. We disregard the $p-$value of the KS test, because we are not interested in statistically inferring the provenience of the two distributions from a hypothetical common distribution. Our aim is instead to quantify the similarity between pairs of distributions, a measure that is already fully captured by the $D-$ statistics of a two-sided KS test.

For every simulation, we perform a two-sided KS test on the resulting pre-alliance knowledge distance distribution and the corresponding empirical 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 100 score values for all the simulations in all points of the parameter space. Such goodness scores are presented in Fig. 7, where we make use of a heat-map to summarize our results.

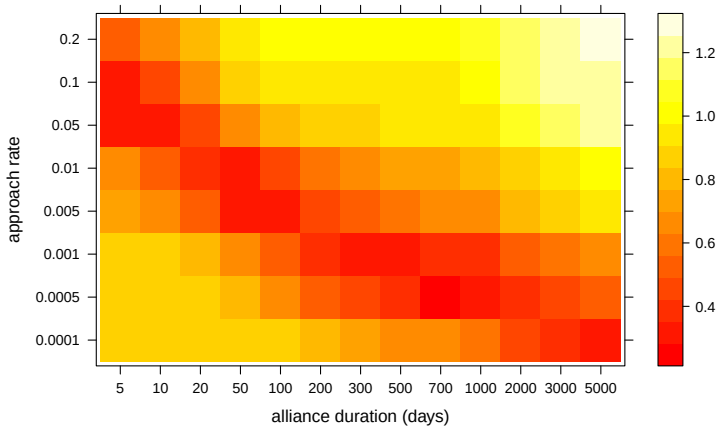

Figure 7: Goodness score for every point in the parameter space, depicted by means of a heat-map. The color scale corresponds to the score value; the lower the score, the closer the simulated R&D network is to the empirical one.

We find that there exists an entire region of the explored bi-dimensional parameter space maximizing the aforementioned goodness score. Such region is identified by low values of the score, corresponding to the red points in Fig. 7. All these points are located in the diagonal of the parameter space connecting the points having large μ values and low τ values, with those having

low μ values and large τ values. This confirms our empirical finding that alliances exert a weak effect on the knowledge positions of firms.

Indeed, the presence of that "optimal" region in the main diagonal of our parameter space clearly indicates that the two parameters are not independent. The product of the two parameters appears to be constant: therefore, only the points with fast approaching rates μ but short alliance life times τ or, on the contrary, with long alliance life times τ but slow approaching rates μ, can generate simulated knowledge distance distributions that correspond to reality. We argue that this is actually an important finding: in real systems, agents do not significantly change their knowledge positions as a consequence of collaborations. They rather use the available information about their mutual knowledge positions in order to establish new collaborations.

Although many parameter combinations exhibit a similar, low goodness score – i.e. they are fairly equally able to reproduce the empirical pre-alliance and post-alliance knowledge distance distributions – the best parameter sets can be ranked quantitatively. We find that the parameter point yielding the best goodness score is identified by the following coordinates: $\mu = 0.0005$ and $\tau = 700$. This means the optimal simulated collaboration network exhibits a low approaching rate, and a characteristic alliance life time slightly shorter than 2 years. This is not only consistent with previous theoretical and empirical observations (Phelps, 2003; Inkpen and Ross, 2001), but it also is surprisingly close to our previous assumption to terminate alliances after 3 years in the empirical network representation that has been used in Tomasello et al. (2013). It is even more surprising if we consider that we have obtained this result by using two different datasets and employing a complex procedure such as the study of the effect of collaborations on knowledge positions through an agent based model.

Additional model tests. The optimal simulated R&D network, as we have shown above, is generated by the set of parameter values $\mu = 0.0005$ and $\tau = 700$. We now want to investigate how well our model, fed with this optimal parameter set, is able to reproduce the knowledge distance distributions of the real R&D network. To this purpose, we report in Fig. 8 and Fig. 9 the distributions of pre-alliance and post-alliance knowledge distances, respectively. In both plots the blue circles correspond to the mean values and the error bars correspond to the standard deviations of all the measures we study on the 200 realizations of the optimal simulated R&D network.

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. We argue that this is due to the fact that our model does not include any self-motion term for the agents in the knowledge space, as our focus is uniquely on the effect of collaborations on the agents' knowledge positions. Therefore, the pre-alliance distance distribution in our simulated network is peaked around a larger value than the real system, and then – as a consequence of the approach in the knowledge space – the post-

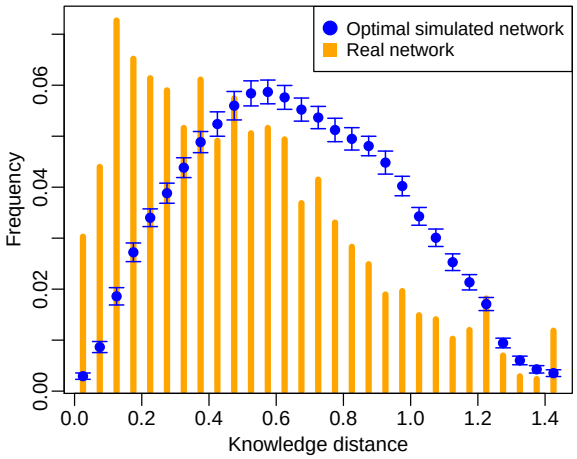

Figure 8: Empirical and simulated distances between firms at the moment of alliance formation.

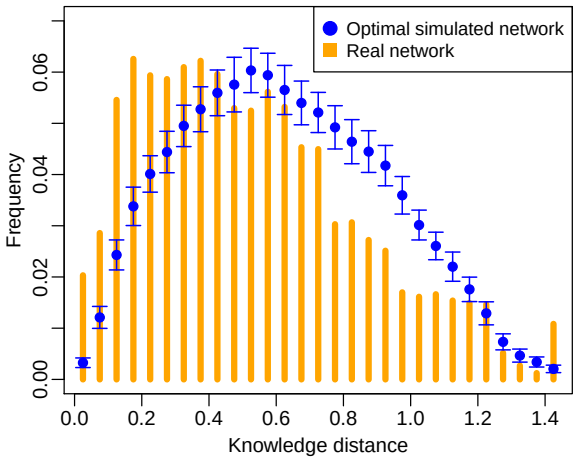

Figure 9: Empirical and simulated distances between firms at the moment of alliance termination.

alliance distance distribution is peaked around a slightly lower value, having a slightly better overlap with the empirical distribution.

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.

Nevertheless, our model is able to reproduce one last empirical distribution – without imposing

it in the validation procedure – i.e. the knowledge distance shifts. This proves that even an approach-only mechanism in a knowledge space is capable to generate positive distance shifts, i.e. increased knowledge distances between two agents as a consequence of a collaboration. We report in Fig. 10 both the empirical and the simulated distribution of the knowledge distance shifts for every pair of connected agents.

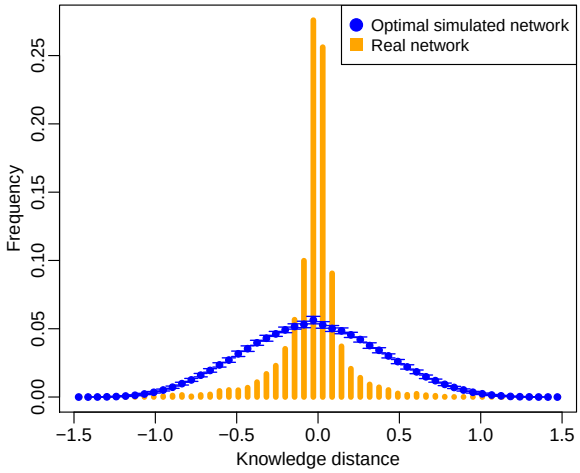

Figure 10: Empirical and simulated distance shifts between all allied firms.

We find that, similar 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 characterizing the system, 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 Network performance

As argued in Section 1, we assume that the exploration of the knowledge space is beneficial for the whole system, and can effectively represent its own performance. In the present Section, we define such a performance indicator for our simulated networks. We do not intend to match this indicator to any possible empirical counterpart, given that we already perform our matching procedure based on empirical knowledge distance distributions. We rather want to investigate whether the empirical R&D network corresponds to a simulated network that is actually optimized with respect to this performance measure.

5.1 Introducing a collaboration performance indicator

We define the path covered by every agent in the knowledge space K_i as the sum of all the distances that the agent travels in every time step of the simulation:

$$K_i = \int_{t=0}^{T_{\max}} |\dot{\mathbf{x}}_i(t)| \, \mathrm{d}t \tag{8}$$

where $T_{\max}$ is the duration of an entire computer simulation. It should be noted that the measure $|\dot{\mathbf{x}}_i(t)| \, \mathrm{d}t$ is a positive scalar and expresses the actual distance traveled by the agent i, differently from its net displacement $\dot{\mathbf{x}}_i(t) \, \mathrm{d}t$, which is a vectorial quantity. The measure K_i is then averaged over all N agents in the network, to obtain the mean knowledge path $\langle K \rangle = N^{-1} \cdot \sum_i K_i$. We hypothesize that this measure can provide a meaningful indication of the macroscopic system performance, because – as already discussed in Section 1.1 with respect to the microscopic level – firms are proven to innovate more when they come in contact with more technological opportunities. Therefore, we assume that a higher value of $\langle K \rangle$, i.e. a better exploration of the knowledge space, corresponds to a higher network performance. We argue that the same reasoning can be as well extended to other types of collaborations that involve learning and/or knowledge exchange processes.

Differently from the model introduced in Tomasello et al. (2015), where the motion of every agent was driven by only one partner at every time step, in the present model the agents are subject to a motion resulting from interactions with multiple partners. Following this reasoning, and considering that our purpose is exactly to quantify the effect of such interactions, we define an indicator that measures how collaborations stimulate the agents' knowledge exploration. We call this indicator the *collaboration performance* $\mathcal{C}$ of the network and define it as:

$$\mathcal{C} = \int_{t=0}^{T_{\max}} \frac{N^{-1} \cdot \sum_{i=1}^{N} |\dot{\mathbf{x}}_i(t)|}{N^{-1} \cdot \sum_{i=1}^{N} k_i^{\mathrm{act}}(t)} \, \mathrm{d}t \; = \; \int_{t=0}^{T_{\max}} \frac{\sum_{i=1}^{N} |\dot{\mathbf{x}}_i(t)|}{\sum_{i=1}^{N} k_i^{\mathrm{act}}(t)} \, \mathrm{d}t. \tag{9}$$

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. In this regard, we remember that not all collaborations are active at a given time t; some are terminated and become inactive, after a characteristic time τ. The quantity $k_i^{\mathrm{act}}(t)$ measures exactly the number of active collaboration in which an agent i is involved at time t. Therefore, the ratio of the two quantities expresses the total distance traveled by the agents in the network per active link, at a given time step t, i.e. a sort of instantaneous collaboration performance of the network. This measure is then integrated over the duration $T_{\max}$ of the simulation, to obtain the overall collaboration performance $\mathcal{C}$ of the network. The quantity at the denominator of Eq.

9 can be thought of as the number of active links in the network at time t, which we indicate with $M^{\mathrm{act}}(t)$,[7] multiplied by a factor 2. By plugging this into Eq. 9, we obtain:

$$\mathcal{C} = \int_{t=0}^{T_{\max}} \frac{\sum_{i=1}^{N} |\dot{\mathbf{x}}_i(t)|}{2 \cdot M^{\mathrm{act}}(t)} \, \mathrm{d}t \tag{10}$$

We use Eq. 10 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. 11, by making use of a heat-map to nicely visualize the average performance $\mathcal{C}$ for every parameter combination.

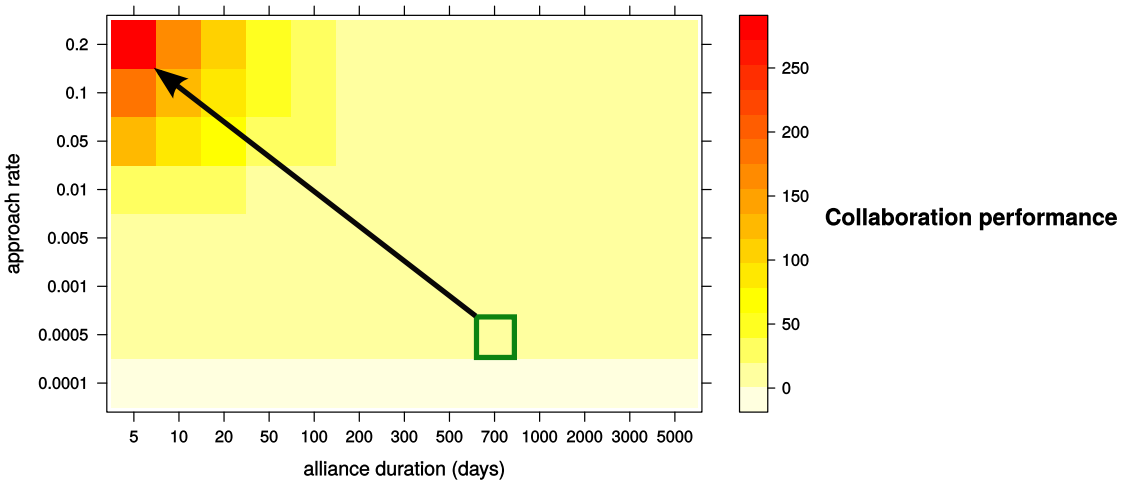

Figure 11: Collaboration performance of the simulated networks, as a function of the characteristic alliance duration and the approach rate. The green square in the parameter space represents the position occupied by the closest simulated networks to the real data.

5.2 Optimality of the real R&D network

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 an optimized network, in terms of collaboration performance $\mathcal{C}$, exhibits links with (i) a short characteristic life time and (ii) allowing for a fast knowledge transfer between the involved partners – and thus a fast approach in the knowledge space. While the dependence of the performance $\mathcal{C}$ on the approach rate μ is easily predictable, the effect of the collaboration

[7] From network theory, we know that at any given time t, the sum of all node degrees k_i equals the number of links M multiplied by two, i.e. $\sum_i k_i(t) = 2 \cdot M(t)$

life time τ is not trivial, given all the complex interdependencies between the network dynamics and the motion of the agents in the knowledge space.

We argue that a short collaboration life time is beneficial for the performance $\mathcal{C}$ 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 τ 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 motion – i.e. a reduced exploration – in the knowledge space. In other words, the density of the collaboration network increases with τ and, after a certain threshold, the addition of a new link has a negative marginal effect on the overall exploration of the knowledge space. Such non-trivial effect, which we could detect only through the implementation and development of our agent-based model, has several implications for policies aimed at optimizing real systems.

Indeed, we have found that the empirical configuration of the real R&D network is generated by parameter sets $[\mu; \tau]$ which are located along the main diagonal of the parameter space, as we show in Fig 7. This means that it is possible to obtain a configuration that is both *realistic* and *optimized* with respect to the collaboration performance. 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 Discussion and conclusions

We have developed an agent-based model that is able to reproduce both the link formation and the knowledge exchange process in a real R&D alliance network. We have used a novel approach, by combining previous results on knowledge exchange and collaboration network growth. In this new modeling framework, agents form links based on their network features and then exchange knowledge with their partners. 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 two-step procedure. By means of an alliance dataset, we have estimated a set of network formation parameters, thus reproducing the topology of the resulting collaboration network. Subsequently, through a second dataset on firm patents, we have estimated a set of parameters expressing the rate at which firms exchange knowledge and the duration of the R&D alliances themselves. 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.

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) 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 (see Podolny, 1993; Raub and Weesie, 1990).

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). The finding of a typical life time τ of around 2 years is consistent with our previous theoretical assumptions and a number of previous studies (see Inkpen and Ross, 2001; Phelps, 2003).

In addition, both our agent-based model and our empirical analysis have allowed us to test the effect of R&D collaborations on the firms' positions in the knowledge space. Our results show that, *overall*, such collaborations exert a null or weak effect on the partners' positions. However, we have also observed and reproduced some examples of extreme shifts: some alliances bring the partners much closer, while some others push them farther from each other in the knowledge space. Such a weak effect of R&D collaborations 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: therefore, 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 traveled by all agents per active link. We have found that the configuration exhibiting the highest performance is characterized by the shortest possible alliance duration, and the largest possible approach rate. In the case of R&D collaborations, obviously, it would be impossible to directly require short alliance durations or enforce a fast learning rate between real companies. However, in principle, it is possible to obtain a configuration that is both *realistic* and *optimized* with respect to the collaboration performance. 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 would be to stimulate 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. In this way, we aim at offering a complete and straightforward interpretation of the effects of knowledge exchange in a dynamically evolving collaboration network.

Compliance with Ethical Standards

M. V. T. acknowledges financial support from the Seed Project – Risk Center (SP-RC) 01-15 "Performance and resilience of collaboration networks", granted by the ETH Zurich Risk Center. The authors declare that they have no conflict of interest.

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., and Jonard, N. (2010). Network-independent partner selection and the evolution of innovation networks. *Management Science*, 56(11):2094–2110.

Baum, J. A., Calabrese, T., and Silverman, B. S. (2000). Don't go it alone: Alliance network composition and startups' performance in Canadian biotechnology. *Strategic management journal*, 21(3):267–294.

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

Cowan, R., Jonard, N., and Ozman, M. (2004). Knowledge Dynamics in a Network Industry. *Technological Forecasting and Social Change*, 71(5):469–484.

Cowan, R., Jonard, N., and Zimmermann, J. (2007). Bilateral Collaboration and the Emergence of Innovation Networks. *Management Science*, 53(7):1051–1067.

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., and 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. and 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.

Fischer, M. M. and Fröhlich, J. (2001). *Knowledge, complexity and innovation systems*. Springer Science & Business Media.

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.

Gilbert, N. (2004). Agent-based social simulation: dealing with complexity. Technical report, Center for Research on Social Simulation ,University of Surrey, Guildford, UK.

Gomes-Casseres, B., Hagedoorn, J., and Jaffe, A. (2006). Do alliances promote knowledge flows? *Journal of Financial Economics*, 80(1):5–33.

Grant, R. and Baden-Fuller, C. (2004). A knowledge accessing theory of strategic alliances. *Journal of Management Studies*, 41(1):61–84.

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

Gulati, R. (1995). Social structure and alliance formation patterns: A longitudinal analysis. *Administrative science quarterly*, 40(4):619–652.

Gulati, R. and Gargiulo, M. (1999). Where do interorganizational networks come from? *The American Journal of Sociology*, 104(5):1398–1438.

Gulati, R., Sytch, M., and Tatarynowicz, A. (2012). The Rise and Fall of Small Worlds: Exploring the Dynamics of Social Structure. *Organization Science*, 23(2):449–471.

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

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

Hanaki, N., Nakajima, R., and Ogura, Y. (2010). The dynamics of R&D network in the IT industry. *Research policy*, 39(3):386–399.

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

Inkpen, A. C. and 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., and 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 Managment 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.

Owen-Smith, J. and Powell, W. W. (2004). Knowledge networks as channels and conduits: The effects of spillovers in the Boston biotechnology community. *Organization science*, 15(1):5–21.

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

Podolny, J. M. (1993). A status-based model of market competition. *American journal of sociology*, 98(4):829–872.

Powell, W., Koput, K., and Smith-Doerr, L. (1996). Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology. *Administrative science quarterly*, 41(1):116–145.

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

Pyka, A. and Fagiolo, G. (2007). *Agent-based modelling: a methodology for neo-Schumpeterian economics*. Elgar companion to neo-schumpeterian economics.

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

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

Rosenkopf, L. and Nerkar, A. (2001). Beyond local search: boundary-spanning, exploration, and impact in the optical disk industry. *Strategic Management Journal*, 22(4):287–306.

Rosenkopf, L. and Padula, G. (2008). Investigating the Microstructure of Network Evolution: Alliance Formation in the Mobile Communications Industry. *Organization Science*, 19(5):669.

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

Sampson, R. C. (2007). R&D Alliances and Firm Performance: the Impact of Technological Diversity and Alliance Organization on Innovation. *Academy of Management Journal*, 50(2):364–386.

Schweitzer, F. and Behera, L. (2009). Nonlinear voter models: the transition from invasion to coexistence. *The European Physical Journal B-Condensed Matter and Complex Systems*, 67(3):301–318.

Thomson-Reuters (2013). SDC Platinum dataset, `http://thomsonreuters.com/sdc-platinum/`.

Tomasello, M. V., Napoletano, M., Garas, A., and Schweitzer, F. (2013). The Rise and Fall of R&D Networks. *ArXiv preprint,* arXiv:1304.3623.

Tomasello, M. V., Perra, N., Tessone, C. J., Karsai, M., and 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., and Schweitzer, F. (2015). A model of dynamic rewiring and knowledge exchange in R&D networks. *ArXiv preprint,* arXiv:1505.01756.

Walker, G., Kogut, B., and Shan, W. (1997). Social Capital, Structural Holes and the Formation of an Industry Network. *Organization Science*, 8(2):109–125.