Newcomers vs. incumbents: How firms select their partners for R&D collaborations

Antonios Garas, Mario V. Tomasello, and Frank Schweitzer
Chair of Systems Design, ETH Zurich, Switzerland
www.sg.ethz.ch

Abstract This paper studies the selection of partners for R&D collaborations of firms both empirically, by analyzing a large data set of R&D alliances over 25 years, and theoretically, by utilizing an agent-based model of alliance formation. We quantify the topological position of a firm in the R&D network by means of the weighted k-core decomposition which assigns a coreness value to each firm. The evolution of these coreness values over time reconstructs the *career path* of individual firms, where lower coreness indicates a better integration of firms in an evolving R&D network. Using a large patent dataset, we demonstrate that coreness values strongly correlate with the number of patents of a firm. Analyzing coreness differences between firms and their partners, we identify a change in selecting partners: less integrated firms choose partners of similar coreness until they reach their best network position. After that, well integrated firms (with low coreness) choose preferably partners with high coreness, either newcomers or firms from the periphery. We use the agent-based model to test whether this change in behavior needs to be explained by means of strategic considerations, i.e. firms switching their strategy in choosing partners dependent on their network position. We find that the observed behavior can be well reproduced without such strategic considerations, this way challenging the role of strategies in explaining macro patterns of collaborations.

1 Introduction

The structure and dynamics of many social and economic systems can be studied by utilizing the network approach. Here, nodes in a network represent the actors, or agents, such as firms in an economic system, while links between nodes represent their interactions, for instance the mutual exchange of knowledge. The *complex network* approach, in particular, builds on the fact that there is a large number of similar agents which allows to express their interactions in terms of statistical rules rather than individual decisions.

Importantly, these complex networks are inherently dynamic, at different scales. *Nodes* can enter or leave the network, which impacts the *size* of the network measured by the number of nodes. But nodes can also start new interactions with other nodes or terminate existing ones, which impacts the *topology* of the network. Eventually, nodes can also strengthen or reduce established interactions which impacts the *weight* of the links. All these dynamic processes feed back on the importance of individual nodes in the network which can be expressed by means of different *centrality* measures [8].

For economic networks, such as the network of research and development (R&D) collaborations between firms discussed in this paper, one generally assumes that the different dynamics do *not* occur at random but are based on the strategic decisions of the firms involved. This leads us to the question addressed in this paper: if we are able to observe the dynamics of the network from large-scale datasets, could we *deduce* (some of) the *strategies* that such firm may follow in entering or leaving the network and in establishing or terminating collaborations? And, more importantly, would we be able to *distinguish* what is deduced as a strategy from a non-strategic behavior that follows only statistical rules?

These questions point to the overarching discussion about the role of *chance* and *choice* in the decision of firms and to the corresponding research strands on economic networks [11, 13, 22]. If the dynamics of firms is driven by *choice*, then the research focus shall be on rational decisions, on utility maximization and on strategic link formation, to explain observed economic networks. Such an approach mainly rooted in economics is indeed able to capture a number of topological features of these networks [1, 3, 12, 14, 16, 17]. Network formation, from this perspective, is modeled as a strategic game where firms form and delete links based on rational strategies using complete or incomplete information. If, on the other hand, the observed dynamics of firms is driven by *chance*, the research focus shall be on stochastic rules, statistical regularities and large-scale structures. This approach is put forward in complex systems science and has been successful in reproducing empirical findings in various economic networks [5, 7, 19–21, 25].

As pointed out in Ref. [22], our aim is to combine these two seemingly different perspectives, i.e. to link economic arguments about strategic link formation of firms to their statistical counterpart of probabilistic rules. For this, we utilize *data driven modeling.* We analyze a large dataset of R&D collaborations described in Sect. 2 to identify, in Sect. 3, strategies that firms possibly use for choosing partners in R&D alliances. As reported in Sect. 3, the observed behavior differs for newcomers and established firms, as well as for firms that have not or have already reached their best network position. The probabilities involved in the link formation with either newcomers or established firms have been deduced from the data by means of statistical calibration [25]. Here, we apply such probabilities in an agent-based model described in Sect. 4 to better understand the observed differences in choosing partners. Since our agent-based model does not build on strategic considerations, it can serve as a *null model* to test to what extent strategic arguments are necessary to explain the observed choice of partners. Sect. 5 draws conclusions with respect to this question.

Along with our investigations, we provide solutions to some methodological issues that can be useful beyond their current application: (a) we quantify the network position of firms by means of the *weighted* k-core decomposition [10], (b) based on this, we reconstruct the time-dependent *career path* of firms in the R&D network, and (c) we link the network position of firms to their success in R&D activities, as measured by their number of patents, to verify that a *topological* measure can indeed be used to proxy *performance*.

2 Network analysis

2.1 Data sets

In this paper, we build on two different data sets. Our first data set is used to *reconstruct* the network of firm interactions. In our network, nodes represent *firms* and links represent *R&D collaborations*. All publicly announced R&D partnerships between firms are available from Thomson Reuters' *SDC Platinum alliances database*. From this database we used in total 21,572 alliance reports involving 13,936 firms between 1984 and 2009. We note that these alliances can involve different kinds of economic actors, e.g. also universities, but we commonly dub them here as *firms*. To classify the industrial sector of a firm's activity, we used its 4-digit Standard Industrial Classification (SIC) code, which allows us to classify universities separately (see Figure 2).

Because the SDC database does not provide a unique identifier for each firm, we used the firm names reported in the dataset. Therefore, we had to correct for the cases where two or more entries with different names corresponded to the same firm, by manually controlling for spelling, legal extensions (e.g. LTD, INC, etc.), and any other recurrent key words (e.g. BIO, TECH, PHARMA, LAB, etc.) that could affect the matching between different entries referring to the same company. We decided to keep those subsidiaries of the same firm that are located in different countries as separated entities.

Our second data set is used to estimate the *success* of each firm, to later relate this to the network position. As a measure of success we use a firm's *knowledge production* as measured by the number of patents in the respective time window. This data is obtained from the *NBER patent database* of the National Bureau of Economic Research. It contains detailed information on about almost three million patents granted in the U.S.A. between 1974 and 2006. Every patent is associated with one or more assignees and is classified according to the International Patent Classification (IPC) system. In general the NBER database is of very high quality, and allowed us to cross-link the firm names involved in alliances in the SDC database with patent information.

2.2 Reconstructing the R&D network

Because the network of firm collaborations is highly dynamic, as pointed out in Sect. 1, there are various ways of studying its evolution over time. In a recent work [24] we focused on the *changing growth pattern* in consecutive 5-year time intervals, to reveal a remarkable *life cycle dynamics* of the R&D network over the whole period of 25 years. Here, our focus is different: we want to find out how a firm's (current) position in the network impacts its (future) strategy of forming alliances with other firms.

This builds on the assumption that *topological measures*, i.e. quantifications of the network position, indeed capture relevant information for strategic decisions. To characterize the information encoded in a network's topology, we distinguish between *local* and *global* information. The former has its focus on the specific firm in relation to its partners: its *degree* characterizes the number of alliance partners, which are direct neighbors in the network at any instance of time. Hence, the *cumulative* degree, aggregated over time, tells us about the variety of partners in the recent history. The *cumulative weight* for each link, on the other hand, tells us how often the same partner was chosen over the available time period.

We argue that cumulative values tell more about a possible strategy because they also reflect the *previous experience* of a firm with respect to diversified and repeated interactions. Hence, in the following we will focus on the cumulative network of R&D collaborations, aggregated over time up to a given year. This implies that the cumulative network further evolves over time, as new annual data is considered. This gives it an advantage over the aggregated network which only considers one static topology. The cumulative network is also preferred over a time-sliced reconstruction of the network which often only results in a large number of disconnected components. Hence, from the cumulative network, we can deduce global information about the position of a firm beyond the relation to its nearest neighbors. For instance, we can characterize its embeddedness in larger *communities*, its belonging to (dis)connected components of the network, its importance in connecting other firms that are not neighbors in the network, etc.

To reconstruct the cumulative R&D network, we use as time resolution one year and add a new link to the cumulative network every time an alliance of two firms is announced in the dataset in this time window. When an alliance involved more than two firms (*consortium*), all the firms involved are connected in pairs, resulting into a fully connected clique. The *weight* $w^{ij}(t)$ of a link indicates the total number of alliances between firms i and j up to time t. If, during the same time interval, two firms i and j have more than one collaboration on different projects, such multiple links are also considered in the weight.

2.3 Largest connected component and cluster sizes

One prominent feature of the R&D network as shown in Figure 1 is the emergence of a large *connected component*, in which all firms are either directly or indirectly connected. This component is not homogeneous with respect to the economic activities of the firms as the different colors indicate. Instead, it is dominated by the two sectors, `pharmaceuticals` and `computer software`.

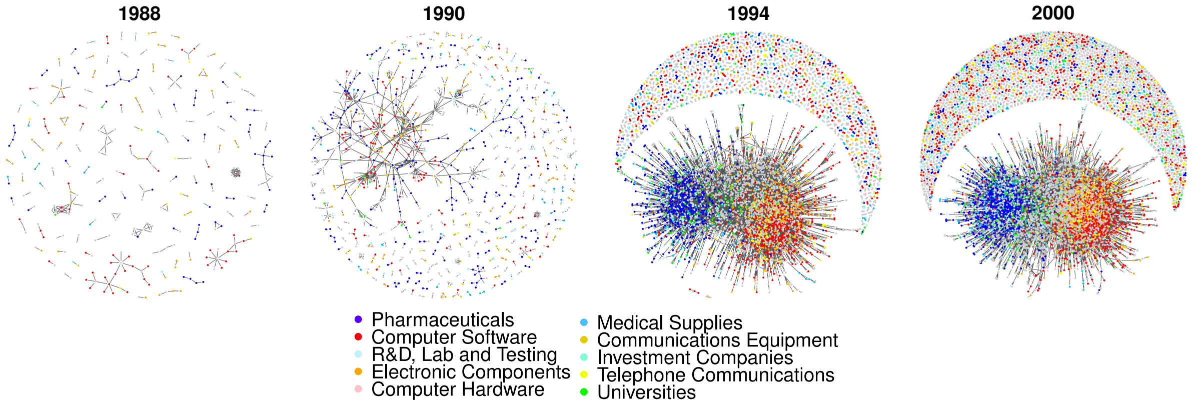

Figure 1: Snapshots of the R&D network showing its evolution and the emergence of a large connected component.

The largest connected component is surrounded by a large number of much smaller components. These indicate isolated clusters of a few firms collaborating. The mixed colors indicate that usually firms from different economic sectors engage in R&D alliances which highlights the importance of their *complementary knowledge bases*.

In order to characterize the distribution of sizes of the different components, or clusters, we use an approach from percolation theory. We compare the *average cluster size* I_{av} of the network *including* the largest connected component with the *reduced* average cluster size I'_{av} of the network *without* the largest connected component. Specifically, these quantities are calculated as:

$$I_{\mathrm{av}} = \sum_{m=2}^{m_{\max}} \frac{h_m m^2}{N^2} \; ; \quad I'_{\mathrm{av}} = \sum_{m=2}^{m_{\max}-1} \frac{h_m m^2}{N^2} = I_{\mathrm{av}} - \frac{m^2_{\max}}{N^2} \tag{1}$$

where $m \geq 2$ are the sizes of the clusters (isolated firms are not considered) and h_m is the number of clusters of size m and N is the total number of firms in the network at a given time.

The evolution of I_{av} and I'_{av} over the time period of 25 years is shown in Figure 2. During early years, the network is still fragmented, i.e many isolated and general small clusters exist. But already at an early stage a larger cluster emerges, which becomes the largest connected component over time. After the years 1992-1993, it dominates the network, which means that *knowledge can potentially diffuse across the whole network.*

2.4 Quantifying network positions

Given that we have reconstructed the cumulative R&D network, we need to characterize the importance of nodes in this network. In a network approach, this importance is quantified by

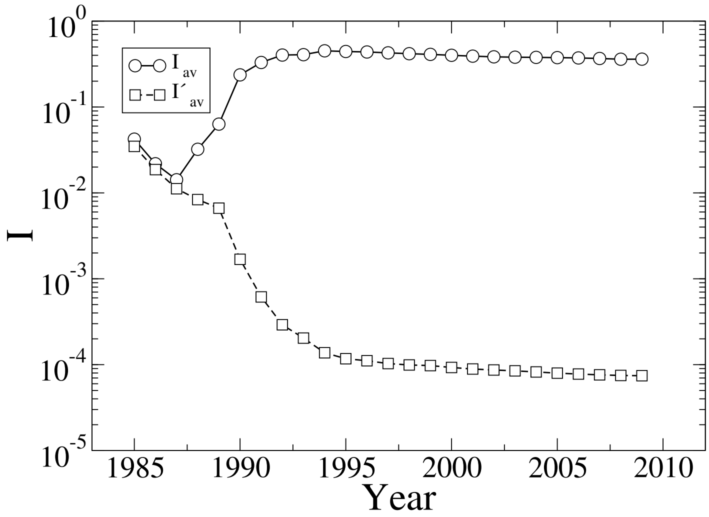

Figure 2: Evolution of the average cluster size (I_{av}), and the reduced average cluster size (I'_{av}) of the R&D network.

purely topological measures, i.e. deduced from the link structure of the network rather than using economic measures (such as market capitalization).

There are various centrality measures to capture the (topological) importance of nodes that are also partially correlated [8]. For example, *degree centrality* simply counts the number of links d^i of node i and thus only captures *local* information about a node and its nearest neighbors. *Betweenness centrality*, on the other hand, quantifies the importance of node i in connecting all other nodes in the system via the shortest path. Such a measure might be of interest in cases where the shortest path has a practical meaning, but it does not allow to capture core-periphery structures.

The main centrality measure used in this paper is called *coreness* C_C^i. It has its roots to social network analysis [4, 23] and aims to measure the importance of a node relative to others. For a given network, a coreness value can be assigned to every node based on a procedure called k-core decomposition. This procedure *recursively* removes all nodes with a degree less than d, i.e., it simulates a *cascade* that consecutively removes nodes with a degree less than d until only nodes with a degree equal or larger than d remain in the network. The procedure starts with $d = 1$, i.e. it removes all nodes that have only *one* neighbor in the networks. The removal may leave these neighboring nodes with one additional neighbor, hence in the second step of the cascade such nodes are also removed. Their removal again may leave other nodes with one remaining neighbor. Thus, in the third step they are also removed and so forth, unless the cascade *stops.* Then, *all* nodes that have been removed during this cascade are assigned a shell number k_s equal to d.

Nodes with a small k_s obviously have been removed very early because they were topologically not well integrated in the network. Consequently, nodes in the shell with the largest $k_s = k_s^{\max}$ are counted as the *the core* of the network. The difference between the k_s^i value of a node i and the value of the core is called *coreness*, $C_C^i = k_s^{\max} - k_s^i$. It can be used to measure the importance of nodes based on the assumption that the most important nodes are part of the core and nodes in the periphery of the network, i.e. at some distance from the core, are less important.

Since the method described uses only information about the node degree and ignores the weight of links, it is called *unweighted k-core* decomposition. This method has been successfully applied to characterize various real-world networks [6, 9]. Kitsak *et al.* [15] showed that the coreness value of a node is a more accurate predictor of its spreading potential as, for example, its degree. This is understandable since coreness also captures information about the second-nearest neighbors of a node (for example, a node A with high degree can have a low coreness if all of its neighbors simply have one neighbor, A)

In this paper, we will use a recent extension of the unweighted k-core decomposition which considers additionally the weights of the links. This *weighted k-core decomposition* [10] uses the same routine to remove nodes as in the unweighted version, but a refined measure for the node degree, called *weighted degree*, d'. For node i it is defined as:

$$d'^i = \left[\left(d^i\right)^{\alpha} \left(\sum_j^{d^i} w^{ij}\right)^{\beta} \right]^{\frac{1}{\alpha+\beta}} \tag{2}$$

d^i is the degree of node i and w^{ij} is the weight of the link between nodes i and j. The summation goes over all neighbors of i. All coreness values reported in this paper are based on the *weighted* k-core decomposition.

Eqn. (2) further uses two free parameters α, β to balance the influence of the weights w_{ij} which indicate multiple alliances between the same firms. $\beta = 0$ would return the results of the the unweighted k-core decomposition. In this paper, we use $\alpha = 1$ and $\beta = 0.2$. The choice is justified in Sect. 3.1 where we show that with $\beta = 0.2$ the correlation between network position and performance is maximized. But in Sect. 3 we will also address the question how other values of β impact the list of firms in the core.

2.5 Distribution of coreness values

In the following we apply the k-core decomposition to the cumulative R&D network, to assign coreness values to each firm. Evidently, whenever the network evolves over time most coreness values change. To define a *reference point*, we take the cumulative R&D network at the very

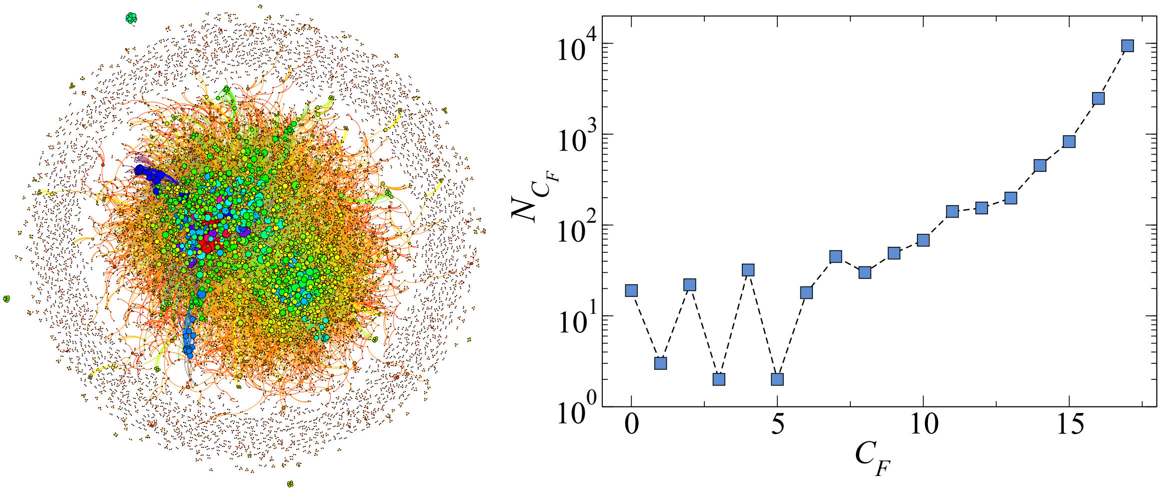

Figure 3: *(left)* Graphical representation of the cumulative R&D network at the end of 2009. This plot is made with Gephi [2] using the OpenOrd layout. The different colors represent different coreness values, with red assigned to the core nodes. *(right)* The network has a strong core-periphery structure [4], i.e. only a small number of nodes having small C_F values, while the majority of the nodes are located in the periphery and have large C_F values.

last time, which is the year 2009, and indicate the coreness values obtained from this maximal network as C_F^i (F stands for final).

Figure 3 (left) shows a network plot of the cumulative R&D network in 2009. The nodes are colored according to their coreness value C_F^i and their size is proportional to their cumulative degree d^i, i.e. the cumulative number of collaboration partners of firms. This makes it obvious that many firms with a large number of collaborations not necessarily belong to the core, but still to the periphery of the R&D network.

Figure 3 (right) shows the normalized *frequency* of coreness values, $f(C_F)$, for the cumulative R&D network in 2009. Taking into account that $N = 13,936$, we see that the total number of firms with small coreness values, $0 \leq C_F \leq 5$, i.e. firms that are part of, or very close to, the core is rather small, but there is a large number of firms in the periphery, $C_F > 5$. I.e. the network exhibits a very distinct *core-periphery structure*.

As an important observation for the real R&D network, we find that only 17 distinct k-shells exist although the maximum degree in the network is $d_{\max} = 336$. It is worth to notice the names of the *firms* that belong to the *core*, i.e. $C_F = 0$, which are listed in Appendix A. There, we also discuss how this list changes if instead of the weighted k-core analysis the unweighted method is used.

3 The career path of firms in the R&D network

3.1 Correlations between network position and success

Before we further discuss how the network position of important firms in the R&D network changes over time, we want to explore to what extent this position is indicative of *success*. As explained in Sect. 2.1, we measure the success of R&D activities of firms by the number of *patents* obtained from the NBER database.

Let us first discuss how this number correlates with the coreness value of firms. The results are shown in Figure 4. According to the distribution of coreness values shown in Figure 3(right), each firm belongs to one out of 17 different classes indicated by the final coreness value C_F. Hence, in Figure 4 we show the distribution of patent numbers for each of these classes, together with the average number of patents and the 95% confidence interval. Although the patent data is rather scattered, there is a very clear trend that the number of patents *increases* with a better network position, i.e. with smaller C_F where $C_F = 0$ indicates the core.

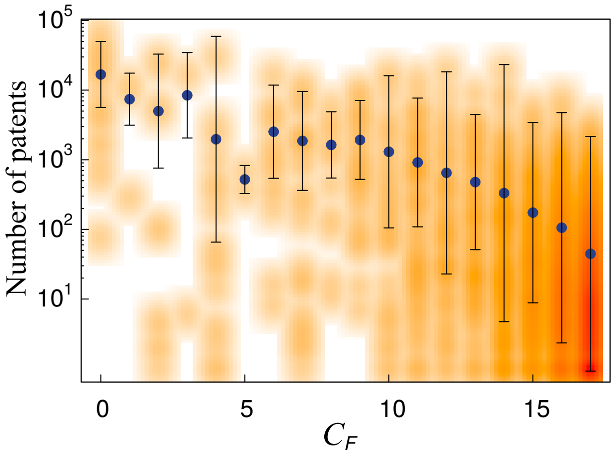

Figure 4: Number of patents against C_F. The heat map is colored according to the number of firms that filled a certain number of patents, the dots represent the average number of patents and the error bars indicate the 95% confidence interval.

More precisely, we calculate the Kendall pairwise correlation between the number of patents $N_p(z)$ in class z and the value of z which quantifies the network position. If z is quantified by *coreness* C_F, as provided by the weighted k-core decomposition method, the correlation implicitly depends on the two free parameters α and β that influence this coreness value. If we take $\alpha = 1$, $\beta = 1$, this correlation reads as $\tau = -0.493$ ($p < 0.001$). However, the correlation could be vastly improved, as we can see if we calculate it for all pairs $(\alpha, \beta) \in [0,1] \times [0,1]$ in steps of $\Delta\alpha =$

$\Delta\beta = 0.1$. This procedure returns that for the parameter values $\alpha = 1$ and $\beta = 0.2$ the correlation between the coreness value and the number of patents is *maximized*, $\tau = -0.84$ ($p < 0.001$).

This means that for $\beta = 0.2$ the weighted coreness of a firm – i.e. a *topological* measure – becomes a very strong indicator of its *success* in R&D activities, as measured by the number of patents. Hence, it establishes a link between *public information* (number of patents) and *network position* (node centrality), which is not directly accessible.

3.2 Evolution of coreness values for individual firms

Now that we have confirmed that the coreness value of a firm is a good indicator of its success in R&D activities, our focus is on the *evolution* of this indicator over time. This helps us to reconstruct a *career path* of individual firms in the R&D landscape. Subsequently, we will address the question how this career path depends on the alliance partners, to identify strategies followed by successful firms.

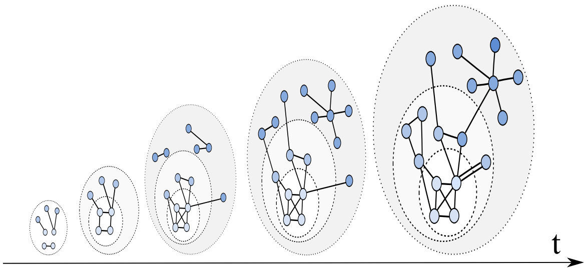

Figure 5: Illustration of the network evolution where new k-shells emerge as new links are formed.

Basically, we need to consider that the coreness value, $C_C^i(t)$, of a firm i can change over time t by means of two different dynamic processes: (a) the growth of the network as a whole, as already demonstrated in Figure 1, and (b) the formation of new alliances involving that particular firm i. Figure 5 sketches a growing network where, similar to real R&D networks, new nodes enter the network by creating new links either with existing nodes or with other newcomers. As the network grows both in size and density, new k-shells emerge and the coreness of all nodes constantly changes with the global connectivity pattern. This implies that the coreness of a particular firm i may change even without any new R&D alliances involving that firm. This contrasts with the second process, where firm i plays an active role in forming new alliances.

We argue that the two dynamic processes can be disentangled in the time evolution of the coreness values of individual firms, as shown in Figure 6. To make coreness values comparable at different times, we define *relative* coreness as $c^i(t) = C_C^i(t)/C_m(t)$, i.e. as the ratio between the

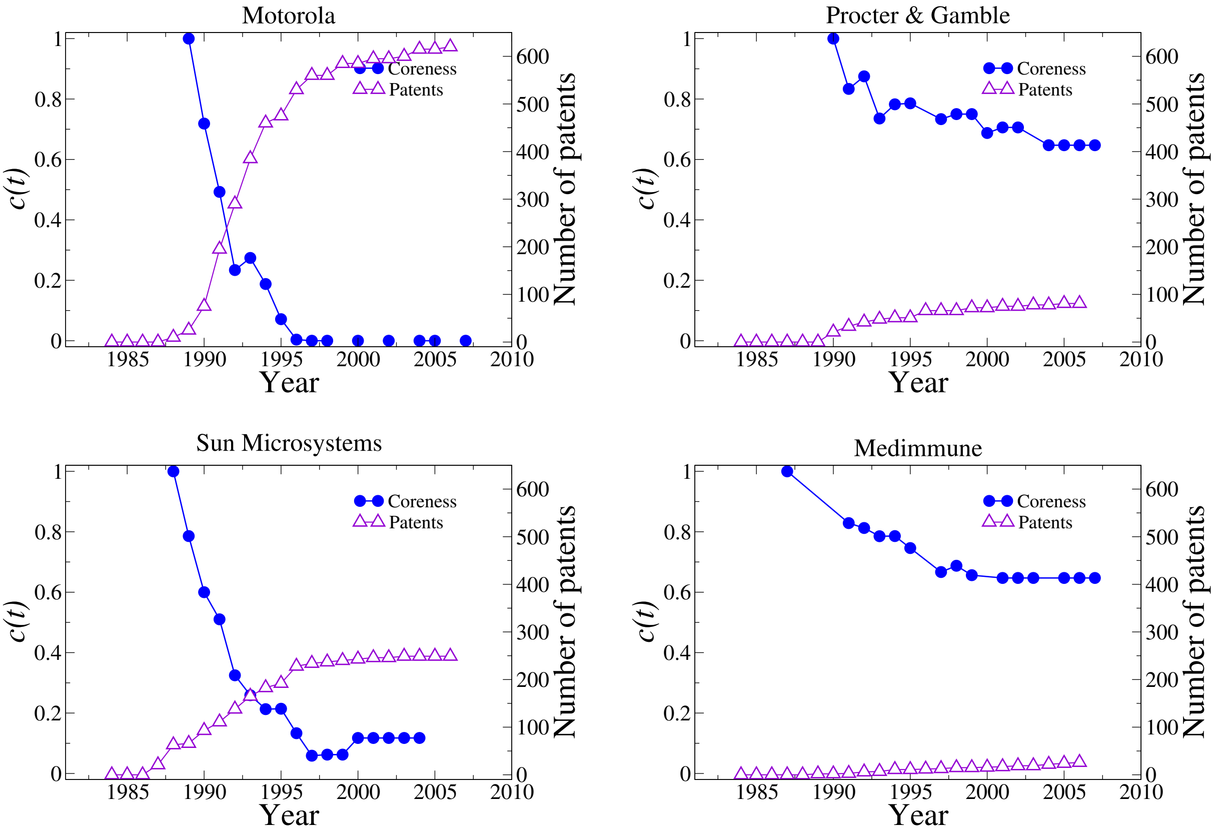

Figure 6: Career path of firms: left successful, right normal

current coreness $C_C^i(t)$ and the maximum coreness $C_m(t)$ at the same time. This returns values between 0 and 1, where 0 always refers to the very *core* and 1 to the outest *periphery.*

In Figure 6, we present two examples of successful firms that made it to the *core* of the R&D network together with two examples of "normal" firms that always belong to the network, but only at the *periphery. All* firms start their career path with high relative coreness values, which is due to the fact that in the beginning no fully developed network existed (cf also Figure 1). The *successful* firms then show a steadily declining relative coreness reaching small values, which means that they manage to get closer to the core as the network evolves. I.e. they *actively* formed alliances such that they stayed away from the periphery, which distinguishes them from the "normal" firms that neither lost or gained in position and thus just stayed in the periphery. Looking at the corresponding patent data, we verify that for the successful firms a better coreness comes along with more patents (as also indicated in Figure 4), whereas for the "normal" firms both the position and the number of patents is rather level (in comparison to the successful firms).

It is important to recognize that even successful firms do not always manage to stay in the very *core* ($c^i = 0$). In most of the cases, there is a minimum c corresponding to the best overall position in the R&D network. In later times, this position can become worse, but successful firms always stay very close to the core.

3.3 Partner selection of newcomers

As already mentioned, the growth of the R&D network occurs by firms selecting partners for R&D alliances. These partners can be *newcomers*, i.e. nodes not yet linked to the network at a given time t, or *incumbents*, i.e. firms already established in the network. This way, the network can grow in *size*, by adding new *nodes*, but also in *density*, by adding new *links*. To disentangle these dynamic processes, we first take the perspective of the *newcomer* that never linked to the network before.

The prevalent *preferential attachment* model of network growth proposes that new nodes choose partners proportional to their degree, i.e. proportional to the number of already established alliances. Such an assumption has several conceptual drawbacks when applying it plainly to the growth of real R&D networks. First, it requires newcomers to have complete information about all the alliances of all firms, to rank these accordingly. Secondly, this neglects any capacity constraints that established firms may have in accepting new alliances. And thirdly, it assumes that a newcomer is attractive enough to be accepted by an established firm as an alliance partner.

Applying the simple preferential attachment rule to growing networks results in a *dissortative network* where nodes with low degree are more likely linked to nodes with high degree. But the real R&D network is *assortative*, i.e nodes with similar degree are more likely to be linked. Specifically, we find for the assortativity coefficient $r = 0.166$. This is in line with positive assortativity coefficients, ranging from 0.12 to 0.363, for various collaboration networks, like scientific co-authorship networks [18] and highlights the existence of degree-degree correlations.

Hence, taking these insights into account, we can expect that newcomers do not follow a preferential attachment rule because they hardly succeed to establish alliances with those firms that are already well integrated in the R&D network. To analyze the real situation, we take again coreness as the most appropriate topological measure of this integration. Figure 7 shows the percentage of all alliances that firms with a given coreness value form with newcomers. Firms in the core only have about 12% of their alliances with newcomers. This share increases with coreness, i.e. with the distance from the core, such that firms in the periphery have almost 40% with newcomers. Taking the perspective of newcomers, this means that the vast majority of partners are firms in the outer parts of the network.

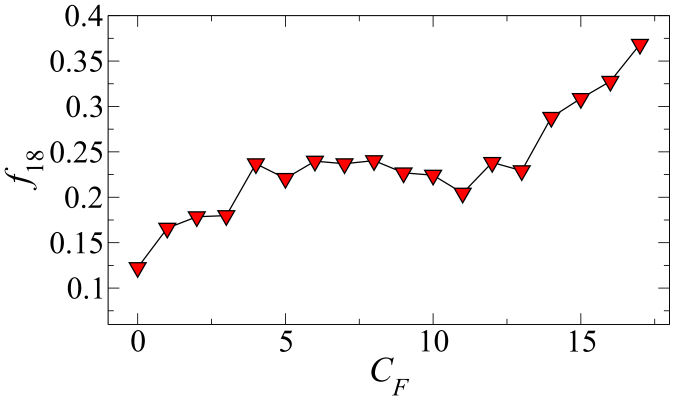

Figure 7: Histogram of the fraction f_{18} i.e. the number of new entries (nodes that were previously not part of the network, assigned to $k_s = 18$) that partner with nodes having coreness C_F divided with the total number of partners of these nodes.

3.4 Partner selection of incumbents

We now look at the formation of alliances from the perspective of *incumbents*, i.e. firms that are already part of the network. Such a firm could form alliances with *newcomers*, with firms that are integrated at a level *comparable* to itself, or with firms that are *much better* integrated and may even belong to the core. The choice between such alternatives of course depend on the *attractiveness* of the firm itself. Again, we consider coreness as the most appropriate measure to quantify both the embeddedness in the network and the success of R&D activities, i.e. the attractiveness of a firm. That means we should be interested in the coreness value *differences* between firms at the time of alliance formation.

Figure 8 shows, for two successful firms, their relative coreness together with the relative coreness of their alliance partners. We observe that a successful firm was able to steadily improve its coreness in the course of time. It does not always reach the very core, $c = 0$, but it is very close to it, indicated by the small values of c. As Figure 8(left) shows, the network position can also worsen at later times, in particular because *other* firms managed to become better integrated into the core. We recall that the network position is quantified by a relative measure that takes the core-periphery structure of the whole network into account.

As discussed for newcomers, in the early period when firms are rather new to the network and thus part of the periphery, they have a strong tendency to choose alliance partners that are also part of the periphery. When these firms become better integrated in the network, as indicated by a decreasing coreness value, they choose partners of *comparable* coreness that are also better integrated. The decrease in coreness, both of firm i and its partners, continues until the firm reaches a state of minimal coreness, the time of which, t_c^i, is indicated by a red line in Figure 8.

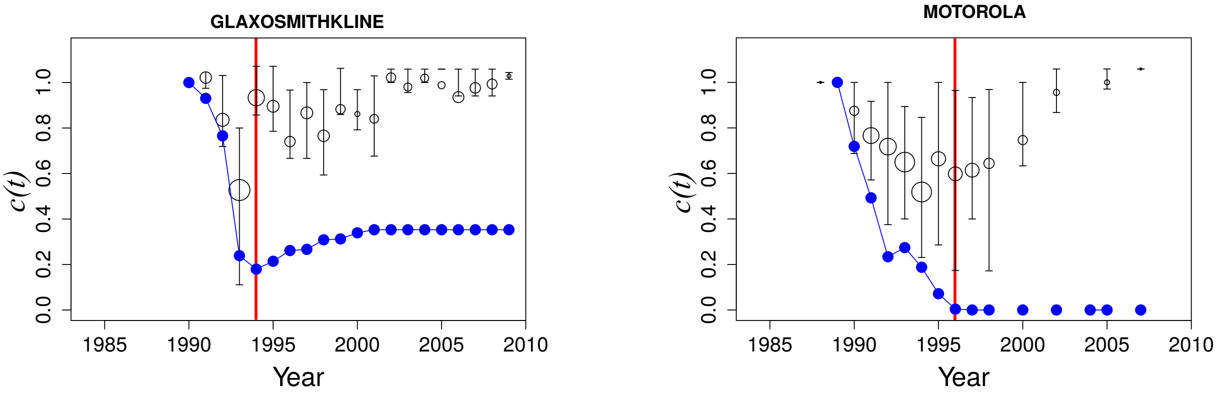

Figure 8: examples of the normalized coreness evolution of firms with different C_F values (blue circles), and the average normalized coreness evolution of their partners (open circles). The size of the open circle is proportional to the fraction of collaborations involving the particular firm that happened in a given year over the total number of collaborations of this firm. With a red vertical line we mark the t_c, to visually clarify the change in collaboration behavior before and after this point. It is interesting to note that the firms are more active during the first phase, when they try to maximize their centrality.

Hence, t_c^i gives the time of maximum integration of firm i in the network and the corresponding relative coreness value $c^i(t_c^i)$ quantifies the best network position obtained. If we compare the relative coreness values of a firm and its alliance partners with respect to t_c^i, we observe a change: While for times $t < t_c^i$, firm i chooses partners of *similar coreness*, for times $t > t_c^i$ the firm chooses partners of *high coreness*, i.e. newcomers or firms from the periphery.

This raises the question whether this observation can be interpreted as a change in the *strategy* of a firm in selecting its partners. It could also be that firms in the periphery or newcomers act differently from core firms simply because they have different options to choose from. This would then probably not count as a change in their strategy, but as a direct reflection of their network position. We will address precisely this question in Sect. 4 by means of an agent-based model that allows us to disentangle strategic behavior from restrictions of choice. But before, we verify the above finding by taking into account all alliances of all firms over time.

For each firm i, we calculate the relative coreness $c^i(t)$ for every year t and the time t_c^i of minimal coreness. We further calculate the number of alliances with each of their partners j, i.e. $w^{ij}(t)$, and the total number of alliances $A^i(t) = \sum_j w^{ij}(t)$ in the given year. Eventually, we calculate the relative coreness $c^j(t)$ of each of their partners j. Combining all these information, we obtain the weighted average of *coreness differences*:

$$\left\langle dc^i(t)\right\rangle = \frac{1}{A^i(t)} \sum_j w^{ij}(t) \left[c^i(t) - c^j(t)\right] \tag{3}$$

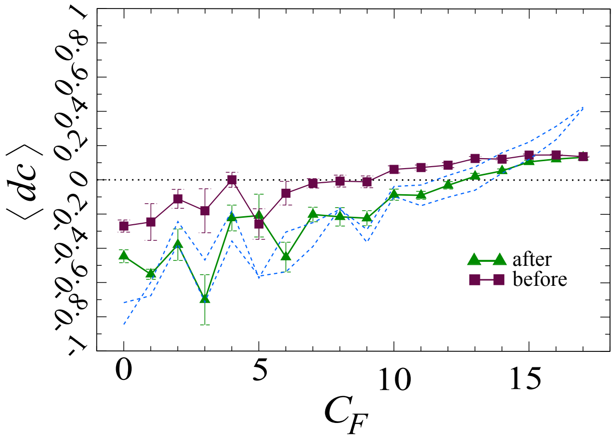

Figure 9: Average partner coreness deviation. Plot of the average normalized partner coreness deviation $\langle dc \rangle$ against C_F before and after t_c.

After calculating $\left\langle dc^i(t) \right\rangle$ for all times t between $t_{\text{start}} = 1984$ and $t_{\text{end}} = 2009$, we divide the values according to two time periods, `before` and `after` t_c^i and average for each of these periods separately:

$$\left\langle dc^i_{\text{before}} \right\rangle = \frac{1}{t^i_c - t_{\text{start}}} \sum_{t=t_{\text{start}}}^{t^i_c} \left\langle dc^i(t) \right\rangle \ ; \quad \left\langle dc^i_{\text{after}} \right\rangle = \frac{1}{t_{\text{end}} - t^i_c} \sum_{t=t^i_c}^{t_{\text{end}}} \left\langle dc^i(t) \right\rangle \tag{4}$$

For each firm i, these two values are related to the final coreness value of that firm, C_F at t_{end}, i.e., $\left\langle dc^i_{\text{before}} \right\rangle (C_F)$ and $\left\langle dc^i_{\text{after}} \right\rangle (C_F)$. Then, for each value of C_F (between 0 and 17), we average the $\left\langle dc^i \right\rangle$ with the same C_F separately `before` and `after` t_c.

The results are shown in Figure 9. We observe that, for the two periods `before` and `after` t_c, the averaged coreness differences *decrease* monotonously with final coreness C_F and even become negative. Positive values mean that the initiating firm has, on average, a higher coreness than its chosen partners. This applies for initiators with high coreness, i.e. newcomers or firms in the periphery that strive to get a better network position by choosing better integrated partners. Negative values mean that this relation switches: initiating firms have on average a lower coreness, i.e. they are better integrated than their partners. This applies for initiators with low coreness that made it to the core and confirms the previous discussion that firms closer to the core have more alliances with newcomers or firms with high coreness.

Looking particularly at differences between the two time periods, we find that this shift from positive to negative coreness differences becomes much stronger in the period `after` t_c, i.e., for firms that have already reached their best network position. This means that established core firms choose even more partners with high coreness (newcomers, periphery) than firms in the period `before` t_c that are still striving for a better position.

To test the robustness of this finding, we perform a random reshuffling of the alliance links while preserving the degree sequence of the empirical network. The dashed curves in Figure 9 show the averaged coreness differences `before` and `after` t_c for the reshuffled network. We see that the trend is the same as for the empirical network. However, the *differences* between the two curves are much larger for the *empirical* network than for the reshuffled network. This means that the *observed* change `before` and `after` t_c is not random. We preformed a two-sided Kolmogorov-Smirnov test to the distributions of the $\langle dc \rangle$ for the empirical and the reshuffled network, and we can reject that they are the same with $p = 0.056$.

As discussed above, it needs further investigation to decide whether our observation that firms change their selection criteria for partners can be interpreted as a change in *strategy*. In the next section, we will use an agent-based model that allows to test to what extent the observed pattern can be reproduced *without* stragetic considerations.

4 Modeling the alliance formation

4.1 Agent-based model

Now that we have identified the dynamics of firms forming R&D alliances dependent on their success, it remains to *reproduce* this behavior by means of an agent-based model. Such a modeling approach can be regarded as successful if it is able to reproduce the pattern in alliance formation observed in Figure 9, which we take as a benchmark here. At the same time, a *correct* agent-based model should be able to additionally reproduce other empirical observations, such as the empirical degree distribution of the R&D network or the coreness distribution. Only if different dimensions of this rather complex phenomenon are reproduced by the same model, we can claim that the model captures the essence of the *dynamic interactions* that lead to R&D alliances, rather than simply (over)fitting free parameters to available observations.

In the following, we utilize a recently proposed agent-based model of alliance formation in R&D networks [25]. Agents represent firms in an R&D network and links between agents represent R&D collaborations. The model uses two macroscopic features of empirical R&D collaborations as input, the distribution of agents activities and the distribution of alliance sizes (see Fig 10). But it does not make strategic assumptions about alliance formation. Instead, the decision of agents in establishing alliances with newcomers or incumbents are modeled by means of five probabilities, which need to be calibrated. Therefore, this is an ideal null model to test whether the observed change in firms strategies need strategic agent considerations as an explanation. Hereafter, we briefly sketch the outline of the model, further explanations and mathematical details are given in [25].

Assuming a multi-agent system with N agents, each agent is assigned two fundamental attributes: an *activity* and a *label*. The activity defines the propensity of each agent to be involved in a collaboration event, and the label is a unique attribute that once set for an agent and does not change afterwards.

Initially we assign to each agent an *activity* value a_i, which is sampled without replacement from the distribution of empirical activities shown in Fig. 10(left). At every time step agent i initiates an alliance with probability $p_i \propto a_i \mathrm{d}t$. Thus, at each time step the number of active agents is $N_A \propto \langle a \rangle N \mathrm{d}t$, where $\langle a \rangle$ is the average agent activity. Upon activation, an agent becomes an *initiator*, i.e. selects the number of partners, m, with whom the alliance is formed. This value of m is sampled without replacement from the empirical distribution of alliance sizes shown in Fig. 10(right).

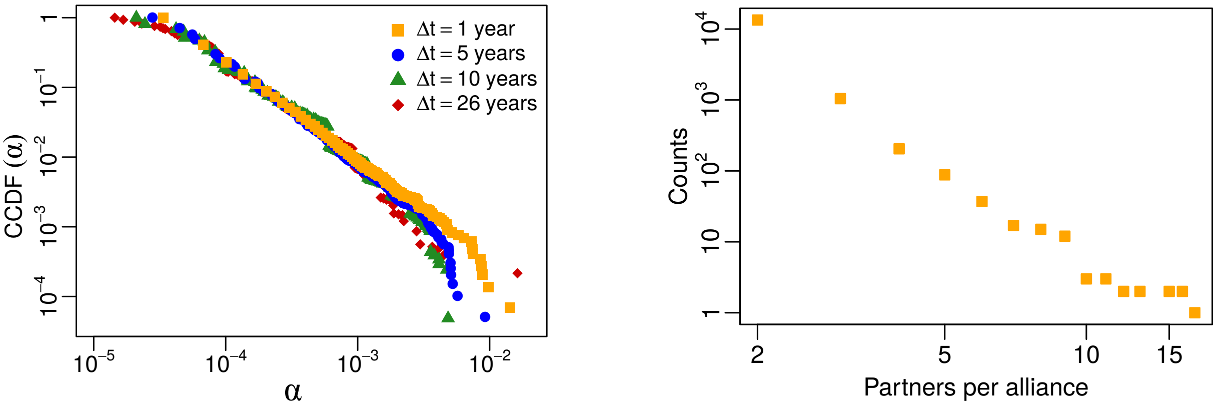

Figure 10: *(left)* Complementary cumulative distribution function (CCDF) of the empirical firm activities, measured from the SDC dataset for 4 different time windows. *(right)* Distribution of alliance sizes, as measured from the SDC alliance dataset. Both figures are adapted from Ref. [25].

The *label* attribute is used to model the participation of an agent in different groups with shared practices and/or behaviors. For the case of firms forming R&D alliances, labels translate to membership, "clubs" or "circles of influence". In our model we assume that collaborations allow the transfer of such membership to other agents that are not part of any circle of influence yet. At the beginning of a simulation, all agents are *non-labeled*, meaning that their membership attribute is blank. There are two ways a *non-labeled* agent can obtain its label: (i) the agent either receives the label from another agent, if the latter initiates an alliance, or (ii) it takes an arbitrary and unique label when it becomes active for the first time. Hence, if the agent that initiates the alliance is a non-labeled one (*newcomer*), it links to a labeled agent with probability p_l^{NL}, or to another non-labeled agent with probability p_{nl}^{NL}. If the agent that initiates the alliance is already labeled (*incumbent*), it has three options to form a link. It can i) link to an agent with the *same* label with probability p_s^L, ii) link to an agent with a *different* label with probability p_d^L, or iii) link to an agent without a label with probability p_n^L.

As the number of alliance partners, m, is already given, the above five probabilities decide how many of the m partners come from each available partner category (*same/different/no label*). To actually *select* the partners *within* these categories, we use a linear preferential attachment rule, where the probability to attach to a node j linearly scales with its degree d_j, i.e. $\Pi(d_j) \propto d_j$. This rule applies only for incumbents that are already assigned to a category, as by definition newcomers are non-labeled and have no previous partners ($d_j = 0$). Therefore, if the initiator connects to a newcomer, the partner is selected among all non-labeled nodes with equal probability. When the partner selection process is complete, all m partners are mutually connected, forming a fully connected clique of size $m + 1$. This reflects the meaning of R&D alliances in a consortium.

4.2 Model results

In order to run agent-based computer simulations, we need to determine the five probabilities. This was done in Ref. [25] using the R&D data set described in Sect. 2.1. Specifically, we found that incumbent firms follow a balanced alliance strategy, forming 30% of their alliances with firms in the same circle of influence ($p_s^{*L} = 0.3$), 30% of their alliances with firms in a different circle of influence ($p_d^{*L} = 0.3$) and 40% of their alliances with newcomer firms ($p_n^{*L} = 0.4$), represented by non-labeled nodes. At the same time, the newcomer firms show a strong tendency to connect to incumbent firms ($p_l^{*NL} = 0.75$), as opposed to a low linking probability with other newcomers ($p_{nl}^{*NL} = 0.25$). We use these results here, to simulate the evolution of 100 synthetic networks over time. Their (time dependent) topology is analyzed with respect to the weighted k-core decomposition, to allow a comparison with the empirical R&D network.

Our main result is presented in Figure 11 which should be compared to the empirical finding shown in Figure 9. It demonstrates that the agent-based model we used is indeed able to *reproduce* the change of strategies in firms after they have reached the core. This is remarkable as our simulations do not use assumptions about the strategic behavior of firms.

To verify that the agent-based model is also able to reproduce other empirical findings without overfitting, we plot in Figure12(left) the degree distribution of the network ensembles obtained from the 100 realizations, alongside of the empirical degree distribution, both for the final time. The excellent match of the two distributions should be noted. We further plot in Figure12(right) the distribution of the coreness values both from the empirical data (also shown in Figure 3 (right) and from the computer simulations. Here, we use the normalized coreness C' instead of C_F. While one could argue about some deviations between the two in the range of small coreness

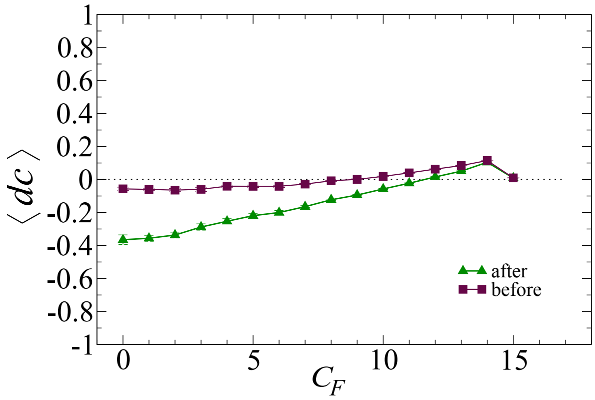

Figure 11: Plot of the average normalized partner coreness deviation $\langle dC' \rangle$ before and after t_c for the networks obtained using our model.

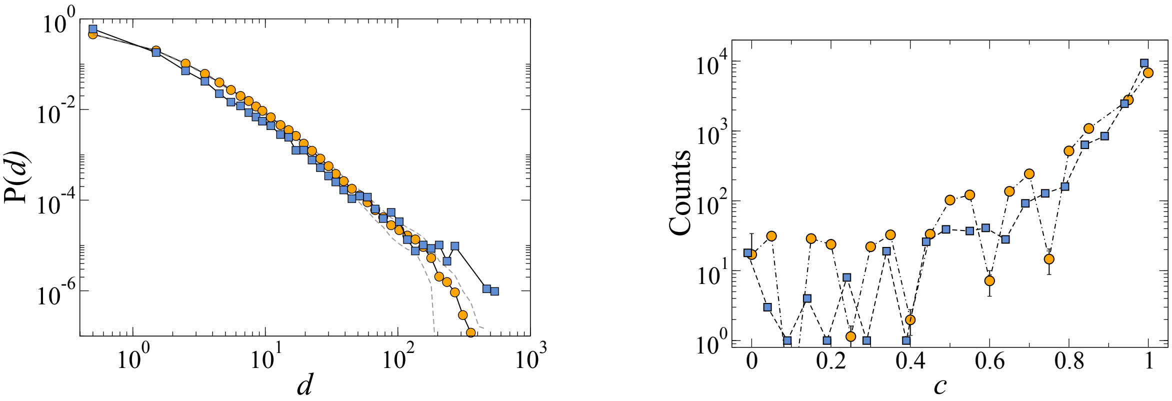

Figure 12: Model results and validation: *(left)* Degree distribution of the network obtained using our model (blue line), alongside the degree distribution of the full empirical network (circles). *(right)* Comparison of the coreness distribution obtained from the model (orange) versus the distribution of the empirical cumulative network (blue). The results are averaged over 100 model realizations and the error bars (when visible) indicate standard errors.

values, we note that the core-periphery structure of the network is well captured by the model – without any additional assumptions.

5 Discussion

In this paper we address an important question for the evolution of real networks, in particular R&D networks, namely which partners firms select for their collaborations. Our approach can be best described as *data-driven modeling*, i.e. we first *analyze* large-scale time-resolved data about R&D collaborations of firms. The aggregated outcome of this analysis is then used to *calibrate* an *agent-based model* that captures the interactions of firms in forming R&D alliances. This agent-based model is then *validated* by its ability to reproduce the *dynamics* of the R&D network at the macro level, in addition to *topological* features observed in the network. In this particular application, we want to reproduce the observed *change* of partners by successful firms, from partners with similar coreness to partners with high coreness.

Coreness and success In order to *quantify* the network position of individual firms, we propose a method that is rather new compared to established centrality measures, the *weighted k-core decomposition.* This assumes a sequence of cascades to prune the network and assigns a coreness value to each node that indicates its distance from the core.

As the network evolves both by adding new nodes and new links, individual coreness values change to reflect the relative position of nodes. We observe the emergence of a clear *core-periphery structure* characterized by a dense core with a smaller number of firms and a sparse periphery with the majority of less integrated firms.

Analyzing the coreness of firms and their number of patents as a proxy of their success in R&D activities, we find a strong correlation ($\tau = -0.84$ for $\alpha = 1$, $\beta = 0.2$ for the weighted k-core decomposition). That means, we can use a *topological* measure, coreness, that quantifies the *network position* as an indicator of successful R&D activities and, hence, as a measure of the *attractiveness* of a firm as potential collaboration partner.

Selection of partners Monitoring a firm's coreness over time allows us to determine its *career path.* The most successful firms move from the *periphery* of the R&D network (close) to the *core* in the course of time, i.e. from high to low coreness values. From the data, we obtain a time t_c^i for each firm when the minimal coreness, i.e. the best network integration, is reached. At about t_c^i we observe a *change* in partner selection, from partners of *similar* coreness to partners of *high* coreness.

Such an observed behavior may have a rational basis. Firms new to the network may have little chances to get connected to core firms. Therefore, in the absence of better alternatives, they may eventually team up with other newcomers or firms from the periphery with comparable coreness. Together with their partners, they then try to improve their network position. However, at the time of maximum network integration, the competition with other firms of similar or lower coreness can become more important than the opportunity to further increase their (already optimal) position. So, while previous partners may have become competitors, successful firms more likely search for, and to team up with, new start-up companies with fresh ideas.

Strategic behavior vs. changed opportunities The question is whether this observed change in partner selection indeed follows a *strategy*, i.e. a deliberative process, or whether the "strategic" behavior is still the same but the opportunities for firms have changed.

To decide between these two alternative explanations, we propose an agent-based model that assigns fixed probabilities to firms for choosing their partners, irrespective of any network position. We demonstrate that this agent-based model is able to reproduce the observed change in partner selection together with other topological features, such as degree distribution and coreness distribution. This indicates that the model indeed captures the essence of forming R&D alliances between firms.

From the results, we can conclude that the change in choosing partners can be reproduced *without* assuming changes in the selection rules for partners and *without* assuming any dependence of these rules on the network position. This does *not* allow to conclude that firms do not follow strategies in selecting their partners, or change these strategies dependent on the network position. It just demonstrates that firms do not need to *change strategies* in order to display a "behavior" that is observed in their career path.

Agent-based model This leaves us with the concluding question how we obtain from selection rules that do *not depend* on the network position (but on five constant probabilities) strategies in selecting partners that *depend* on the network position.

Our agent-based model is an *activity driven model*, i.e. from the empirical distribution of activities firms get assigned a (fixed) activity a_i to form alliances. Obviously, in a stochastic simulation firms with a higher activity are on average chosen *earlier* and *more often*. This generates a first mover advantage because such firms can increase their degree early on. In the beginning, they also get a higher chance to propagate their *label* to other (unlabeled) firms.

Firms with high activity also have more alliances over time, and therefore a higher (weighed) degree in the cumulative R&D network. This becomes important when the initiator of an alliance has to select partners *within* the two categories "*same/different* label". There, we use a linear preferential attachment rule, i.e. within each category firms with a higher degree are chosen with higher probability. This further increases their degree or at least the weight of the link in case of repeated collaborations, which eventually improves their coreness value.

This preferential selection rule applies to both newcomers and incumbents. Because the network is dominated by newcomers and low-degree/high-coreness incumbent nodes, all these firms offer to establish alliances with more central firms. If this offer is accepted, the likelihood that a firm with high degree is chosen becomes even higher over time because the differentiation between firms with higher and lower degree increases. It is less pronounced in the beginning, but with an established core-periphery structure the degree distribution becomes rather skew (see also Figure 12 left).

These combined reinforcement effects eventually lead to the observation that core firms tend to form alliances preferably with newcomers. It is the newcomers that, with larger probability, choose these firms after they managed to have a low coreness/high degree (in comparison to others).

Thus, in conclusion, what appears as a deliberative strategy of successful firms, namely to switch from partners of comparable coreness to partners of high coreness, can be basically explained by the choice of newcomers. This does not exclude other strategic considerations, it just tells that the observation does not already imply such considerations.

Which of the firms eventually end up in the core of the R&D network basically depends on their individual activity, which is very skew distributed (see Figure 10). More active firms have an early mover advantage, they can accumulate more experience in collaborations over time, which increases their attractiveness as partners. More collaborations also lead to better integration in the network, as quantified by the decreasing coreness, and to more success in R&D activities, as quantified by the number of patents.

Acknowledgments

The authors acknowledge financial support from the EU-FET project MULTIPLEX 317532. The calculations of network measures, and the statistical treatment of data were performed using the R software for statistical analysis v2.15.2, and the igraph library v0.6.2.

Author Contributions

A.G. and F.S. designed the research, A.G. and M.V.T. performed the research, and analyzed the data. F.S. and A.G. wrote the manuscript.

References

[1] Bala, V.; Goyal, S. (2000). A noncooperative model of network formation. *Econometrica* **68(5)**, 1181–1229.

[2] Bastian, M.; Heymann, S.; Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks.

[3] Billand, P.; Bravard, C.; Durieu, J.; Sarangi, S. (2015). *Uncertain Innovation and R&D Network Formation. Post-print*, HAL.

[4] Borgatti, S. P.; Everett, M. G. (2000). Models of core/periphery structures. *Social Networks* **21(4)**, 375 – 395.

[5] Boss, M.; Elsinger, H.; Summer, M.; Thurner 4, S. (2004). Network topology of the interbank market. *Quantitative Finance* **4(6)**, 677–684.

[6] Carmi, S.; Havlin, S.; Kirkpatrick, S.; Shavitt, Y.; Shir, E. (2007). A model of Internet topology using k-shell decomposition. *Proceedings of the National Academy of Sciences of the United States of America* **104(27)**, 11150–4.

[7] Fagiolo, G.; Reyes, J.; Schiavo, S. (2009). World-trade web: Topological properties, dynamics, and evolution. *Physical Review E* **79(3)**, 036115.

[8] Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social Networks* **1(3)**, 215–239.

[9] Garas, A.; Argyrakis, P.; Rozenblat, C.; Tomassini, M.; Havlin, S. (2010). Worldwide spreading of economic crisis. *New Journal of Physics* **12(11)**, 113043.

[10] Garas, A.; Schweitzer, F.; Havlin, S. (2012). A k-shell decomposition method for weighted networks. *New Journal of Physics* **14(8)**, 083030.

[11] Goyal, S. (2007). *Connections: an Introduction to the Economics of Networks.* Princeton University Press.

[12] Goyal, S.; Moraga-González, J. L. (2001). R&D Networks. *The RAND Journal of Economics* **32(4)**, 686–707.

[13] Jackson, M. O. (2010). *Social and Economic Networks.* Princeton University Press.

[14] Jackson, M. O.; Wolinsky, A. (1996). A strategic model of social and economic networks. *Journal of economic theory* **71(1)**, 44–74.

[15] Kitsak, M.; Gallos, L. K.; Havlin, S.; Liljeros, F.; Muchnik, L.; Stanley, H. E.; Makse, H. A. (2010). Identification of influential spreaders in complex networks. *Nature Physics* **6(11)**, 888–893.

[16] Koenig, M. D.; Battiston, S.; Napoletano, M.; Schweitzer, F. (2011). Recombinant knowledge and the evolution of innovation networks. *Journal of Economic Behavior & Organization* **79(3)**, 145–164.

[17] Koenig, M. D.; Tessone, C. J.; Zenou, Y. (2014). Nestedness in networks: A theoretical model and some applications. *Theoretical Economics* **9(3)**, 695–752.

[18] Newman, M. E. J. (2002). Assortative Mixing in Networks. *Phys. Rev. Lett.* **89**, 208701.

[19] Orsenigo, L.; Pammolli, F.; Riccaboni, M.; Bonaccorsi, A.; Turchetti, G. (1997). The evolution of knowledge and the dynamics of an industry network. *Journal of Management and Governance* **1(2)**, 147–175.

[20] Powell, W. W.; White, D. R.; Koput, K. W.; Owen-Smith, J. (2005). Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences 1. *American Journal of Sociology* **110(4)**, 1132–1205.

[21] Scholl, T.; Garas, A.; Schweitzer, F. (2015). The spatial component of R&D networks. *arXiv:1509.08291 (Journal of Evolutionary Economics – revised and resubmitted)* .

[22] Schweitzer, F.; Fagiolo, G.; Sornette, D.; Redondo, F. V.; White, D. R. (2009). Economic Networks: What do we know and What do we need to know? *Advances in Complex Systems* **12(04)**, 407– 422.

[23] Seidman, S. B. (1983). Network structure and minimum degree. *Social Networks* **5(3)**, 269–287.

[24] Tomasello, M. V.; Napoletano, M.; Garas, A.; Schweitzer, F. (2016). The Rise and Fall of R&D Networks. *Industrial and Corporate Change* **dtw041**.

[25] 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**.

Appendix

A Identification of core firms

In order to better illustrate the advantage of the *weighted* k-core decomposition we present the list of firms in the *core* ($C_F = 0$) for $\alpha = 1$ and varying β, where $\beta = 0$ corresponds to the *unweighted* k-core decomposition. The analysis was done on the cumulative R&D network for the year 2009. As shown in Table 1 the list of firms identified in the core strongly depend on the value of $\beta \in \{0, 1\}$. The list provided by the weighted method matches better our economic intuition as it contains big and well known firms. These are mayor international players in research and development activities. Note that the results for $\beta = 0.2$ are much closer to $\beta = 1.0$ than to $\beta = 0$. I.e., we argue that the core firms are considerably well captured as long as the *weighted* method is used.

$\beta = 1$	$\beta = 0.2$	$\beta = 0$
Apple	Apple	Aiscorp
AT&T	AT&T	Arbortext
Fujitsu	France Telecom	Avalanche Dev.
Hewlett Packard	Fujitsu	Broadvision
Hitachi	Hewlett Packard	Computer Task
IBM	Hitachi	Database Publishing Sys.
Intel	IBM	EBT
Matsushita Electric	Matsushita Electric	Furlcrum
Microsoft	Microsoft	Information Design
Mitsubishi Electric	Mitsubishi Electric	Information Dimensions
Motorola	Motorola	Intergraph
NEC	NEC	Object Design
Nippon Telegraph & Telephone	Nippon Telegraph & Telephone	OfficeSmith CTMG
OKI Electric Ind.	Nortel Networks	Open Text
Sanyo Electric	Oki Electric Ind.	Oracle Sys.
Sharp	Philips Electronics	SoftQuad
Sony	Sanyo Electric	XSoft
Toshiba	Sony	
	Toshiba	

Table 1: Firms identified as the core of the R&D network using the weighed k-core decomposition method with $\alpha = 1$ and different values of β.