World Scientific
www.worldscientific.com

THE ROLE OF NETWORK EMBEDDEDNESS ON THE SELECTION OF COLLABORATION PARTNERS: AN AGENT-BASED MODEL WITH EMPIRICAL VALIDATION

FRANK SCHWEITZER*, ANTONIOS GARAS, MARIO V. TOMASELLO,
GIACOMO VACCARIO and LUCA VERGINER

*Chair of Systems Design, Department of Management
Technology and Economics, ETH Zürich
Weinbergstrasse 56/58, CH-8092 Zürich, Switzerland
fschweitzer@ethz.ch

We use a data-driven agent-based model to study the core–periphery structure of two collaboration networks, R&D alliances between firms and co-authorship relations between scientists. To characterize the network embeddedness of agents, we introduce a coreness value obtained from a weighted k-core decomposition. We study the change of these coreness values when collaborations with newcomers or established agents are formed. Our agent-based model is able to reproduce the empirical coreness differences of collaboration partners and to explain why we observe a change in partner selection for agents with high network embeddedness.

Keywords: R&D network; agent-based modeling; scientific collaborations; R&D alliances; firm alliances.

1. Introduction

Collaboration is a pervasive phenomenon in the animated world. We find it at various organismic levels [3, 6], ranging from cancer cells to bacteria, from gregarious insects to bats and fish [4, 7, 11, 13, 14, 21]. We observe collaboration also in humans and even between human-generated higher-level structures, e.g., between economic firms [10, 25] or political parties [18]. In abstract terms, a collaborative effort most often leads to better results than the additive outcome of isolated efforts. In economics, phenomena such as the division of labor or the establishment of global supply chains are based on this rationale [23, 24]. In social systems, we find, for instance, that publications are written by a larger group of co-authors or strategic alliances between political actors are formed [19, 31, 32].

*Corresponding author.

Complexity science addresses the question of whether we can detect overarching principles to characterize collaborative systems and, subsequently, mechanisms to establish collaboration. The challenge comes with the abstraction: Instead of focusing on the specificity that distinguishes collaborating firms from, e.g., collaborating scientists, complexity science is primarily interested in the commonalities in the general principles that constitute collaborative systems. This aspiration entails two methodological problems: (i) We have to define quantitative measures that allow us to characterize and compare collaboration structures across systems. (ii) We have to identify dynamic principles that generate collaboration irrespective of the entities, or agents in general, that form the system.

As one successful approach to problem (i), the network representation has been established. Agents as the system elements are represented by nodes and interactions between agents as links. The network structure then allows to define topological measures to characterize the network position of agents and to relate it to their collaboration effort. From this perspective, successful collaboration networks should display similar features and collaborative agents could be identified by their network position.

Such an approach would merely state a relation between observed collaborations and certain network features. It does not explain why agents collaborate and how they establish their collaborations. To address problem (ii), agent-based modeling has been proposed. While this approach is deeply rooted in economics, as well as in computer science, there is no general way of developing agent-based models. We can distinguish at least two different directions [20]. The first one starts from the economic perspective [1, 12, 15–17]. Agents collaborate because they obtain a benefit. This requires defining utility functions, i.e., costs and benefits for agents. Further, one must define how agents evaluate current and expected outcomes, what information they take into account and how they make decisions. This is mostly done in a formal manner that allows us to analyze the mathematical properties of the model, albeit with restrictions for the chosen mathematical expressions. Such an approach can replicate certain topological features of observed collaboration networks, which lends some evidence to the underlying assumptions about utilities and decision rules. At the same time, all results crucially depend on these assumptions. Thus, instead of obtaining a general picture of how collaboration structures evolve, we mostly learn what distinguishes the utilities and decisions of firms from, e.g., scientists.

Therefore, in the second approach to agent-based modeling agents have a set of possible rules, which they follow with a certain probability [28]. This set of rules is neither complete, nor exclusive. It is rather motivated by empirical observations of possible actions that agents can choose in a given situation. Importantly, the probabilities to follow certain rules are obtained from data. We therefore call this a data-driven modeling approach. The rules are in some sense "universal", i.e., they apply to different collaborative systems, while the probabilities reflect the specifics of the system.

Similar to the first perspective, the validity of the modeling approach is measured against its ability to reproduce real-world collaboration structures. Succinctly, if problem (i) is solved by means of a network representation that gives us structural measures to compare different collaboration systems, then solving problem (ii) by means of data-driven agent-based modeling allows to compare mechanisms to establish collaborations across systems.

In our paper, we illustrate the power of our approach by modeling and analyzing two very different collaboration systems: R&D (research and development) collaborations between firms and co-authorship relations between scientists. To show that the same quantitative characterization and the same dynamic assumptions to form collaborations can be applied to systems from different domains, we use an agent-based model, i.e., firms and scientists are abstracted as agents in the following. The details of our agent-based model and its calibration for the two different systems are explained in Sec. 2.

In this paper, we focus on one specific question, namely how the network position of agents affects the selection of collaboration partners. Our study is motivated by the empirical observation that collaboration networks show a pronounced core–periphery structure, where a small, but highly integrated core of agents coexists with a large and sparse periphery. To quantify the network embeddedness of agents, in Sec. 3, we introduce the coreness value to measure the distance from the core. We then study how differences in coreness values evolve if agents start new collaborations.

From a dynamic perspective agents entering the network usually do not start from the core, but from the periphery. That means, during the evolution of the network some agents manage to better integrate themselves into the network, but others not. This brings up an important question: do agents follow specific strategies to improve their network embedding? If they do so, then in an agent-based model we should find that their actions could not be captured by the simple probabilistic rules we apply for the "normal" agents. If, on the other hand, they do *not* follow specific strategies to enter the core, then we could argue that their better network integration is the result of *chance* more than of strategic choice. Our detailed discussion in Secs. 3.4 and 4 shows that our agent-based model is able to reveal the feedback mechanisms that lead to the observed practice in partner selection.

The remainder of the paper is divided as follows. In Sec. 2, we present the agent-based model together with an overview of the collaboration data used to calibrate the model parameters. In Sec. 3, we investigate the empirical network embeddedness of firms and of scientists and compare it with the outcome of our agent-based model. This is followed by a discussion of the results in Secs. 3.4 and 4.

2. Modeling the Formation of Collaboration Networks

2.1. *Agent-based model of collaboration networks*

In the following, we utilize a recently proposed agent-based model which was already applied to collaborations between firms [26] and between scientists [29]. We consider

a multi-agent system with N agents. They represent either firms in an R&D network or scientists in a co-authorship network, and links between agents represent collaborations. Collaborating agents form groups of various sizes m, i.e., R&D alliances or co-authorship teams, which appear as fully connected cliques in the collaboration network.

Our model uses two macroscopic features of empirical collaborations as input, the distribution of agents' activities, $P(a)$ (see Sec. 2.2), and the distribution of alliance sizes, $P(m)$. Activity defines the propensity of each agent to initiate a collaboration. From the distribution $P(a)$, we initially sample without replacement an *activity* value a^i for each agent. During the simulations at every time step agent i initiates a collaboration 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 collaboration is formed. This value of m is sampled without replacement from the empirical distribution of collaboration sizes, $P(m)$. By sampling without replacement, the number of created links is exactly equal to the number of links in the empirical network.

The second fundamental attribute of agents is their *label* l^i. 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". For co-authorship teams, labels indicate specific scientific specializations.

Labels do not change over time, but can propagate to other agents. We assume that collaborations allow the transfer of labels to those agents that are not labeled yet. Specifically, at the beginning of a simulation, all agents are *non-labeled*, i.e., they are *newcomers* with a blank membership attribute. Once they received a label, we denote them as established agents, or *incumbents*. A newcomer can obtain its label in two ways: (i) the agent either receives the label from another agent, if the latter initiates an collaboration (label propagation), or (ii) it takes an arbitrary and unique label when it becomes active for the first time (label generation).

This label propagation process is mapped to the formation of collaborations by means of five probabilities for link creation. If the initiator of a collaboration, chosen by its activity, is a newcomer (non-labeled), it links to a labeled agent with probability p_l^{NL}, or to another non-labeled agent with probability p_{nl}^{NL}. If the initiator is an established agent (labeled), 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.

Because the number of collaboration partners, m, is already given from the sampling, the above five probabilities decide how many of the m partners come from each of the three *partner categories*: *same (s)/ different (d)/ no label* (nl). But the link probabilities alone are not sufficient to reproduce the features of empirical collaboration networks. We need an additional dynamic rule to actually *select* the partners *within* the three partner categories. Here, 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 preferential selection affects only incumbents that are already assigned to a partner 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.

Because the preferential attachment rule applies to agents with very different activities, this results in a reinforcement dynamics which is important to understand the emergence of the core–periphery structure of the network. If agents have a high activity, they also have more collaborations over time, and therefore a higher (weighted) degree in the collaboration network. The linear preferential attachment rule implies that within the partner categories "*same/different* label" agents with a higher degree are also 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 the network embeddedness of these agents.

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 consortia or of co-authorship teams. Once an agent has established a collaboration, it will remain in the system until the end of the simulation. It is grounded in the fact that the data set do not contain any information about the duration of the alliances or the exit of agents from the network. Therefore, we could not implement a reasonable exit dynamics into the model.

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

We note that our agent-based model does not make strategic assumptions about collaborations. Instead, the decision of agents in establishing links with newcomers or incumbents are modeled only by means of the mentioned five probabilities, which need to be calibrated for firms and for scientists, separately. Therefore, this is an ideal null model to test whether the observed dynamics of the resulting collaboration network needs strategic agent considerations as an explanation.

2.2. *Calibration of the agent-based model*

In order to calibrate the mentioned probabilities for link formation, we need to use different data sets about collaborations of firms and of scientists. This calibration procedure was carried out and described in detail in previous publications, therefore, we only summarize it here.

Data sets. For firm collaborations we use Thomson Reuters' *SDC Platinum alliances database*. This contains all publicly announced R&D partnerships ("alliances") between 1984 and 2009 with a resolution of 1 year. In total, we have $E = 14,829$ alliance reports, referred to as collaboration events E, involving $N = 14,561$ firms.

For the co-authorship collaborations of scientists, we use the data set from the American Physical Society (APS) about papers published in any APS journal, namely Physical Review Letters, Reviews of Modern Physics, and all Physical Review journals.[a] This data set is quite large, it spans 110 years (1895–2009) and contains $N = 226,724$ unique authors and $E = 1,567.084$ publications, i.e., collaboration events. For the empirical study of network embeddedness we used the full data set. But for the calibration of the agent-based model we only selected papers published between 1984 and 2009 in specific research areas identified by their PACS code. We have restricted ourselves such that the time periods for both data sets are the same and the collaboration networks are of comparable size. In Table 1, we present the example of PACS 42 (Optics), for which we have in total $E = 20,105$ publications involving $N = 27,436$ scientists.

Distributions. From these data sets, we calculate the distribution of collaboration sizes, $P(m)$, as well as the activity distribution, $P(a)$, both for the firms and for the scientists. These distributions, which are used as an input for the agent-based simulations, are very broad [29]. In particular, the activity distributions span several orders of magnitude. Here, the empirical activity of a given agent i at time t is the number of collaboration events, $e_{i,t}^{\Delta t}$, involving agent i during a time window Δt (in years) ending at time t divided by the total number of collaboration events, $E_t^{\Delta t}$, involving any agent during the same period of time. We mention that the activity distributions are very stable regardless of the chosen Δt.

Network reconstruction. In a next step, we reconstruct the aggregated collaboration networks for firms and for scientists. We emphasize that these are undirected, but *weighted* networks, where the weight w^{ij} gives the number of collaborations between agents i and j over the whole time. Further, we note that the number of links, L, is not the same as the number of collaboration events, E. A publication co-authored by 4 scientists, for instance, would count as one collaboration event of size $m + 1 = 4$, but it generates 6 links between the 4 involved scientists.

Figure 1 shows these two networks as unweighted networks. There are two important observations: (i) Both collaboration networks have a largest connected component (LCC) and a large number of small disconnected clusters. (ii) The LCC itself shows a prominent *core–periphery* structure, which may be difficult to see because the LCC is quite dense. But for the R&D collaborations, for instance, we note that the inner core contains less than 250 out of 14,000 firms. When we analyze the evolution of network positions of agents in the following, these obviously refer to the core–periphery structure of the LCC only.
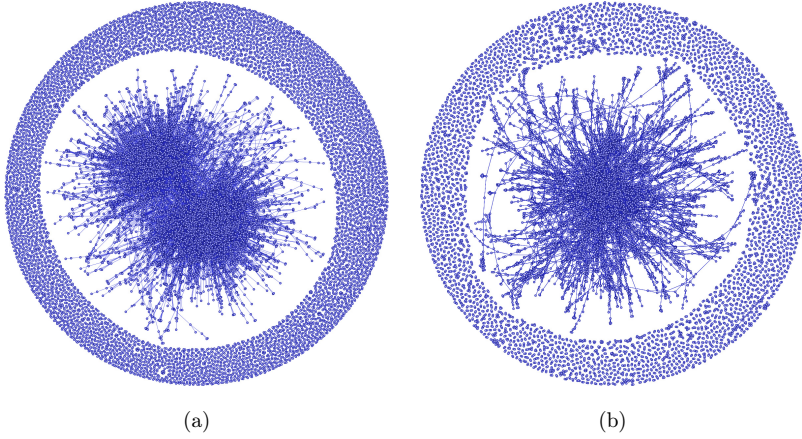
---

[a]http://www.aps.org/.

Fig. 1. Illustration of the collaboration networks of firms (a) and of scientists (b). Data: (a) complete R&D network with about 14,000 nodes and 21,000 links, (b) co-authorship network sampled from the full data set with about 11,000 nodes and 32,000 links.

Calculating average quantities. From the observed aggregated networks, we calculate three mean quantities: (i) Average degree $\langle d \rangle^{\text{obs}} = 2N/L$, where N is the total number of nodes and L is the total number of links. The normalization factor of 2 results from the fact that each link connects two nodes. (ii) Average path length $\langle l \rangle^{\text{obs}}$. A path is formally defined as a sequence of nodes, where any pair of consecutive nodes is connected by a link, i.e., loosely speaking the path length is the number of steps to reach a node over the network from a given starting point. (iii) Average clustering coefficient $\langle c \rangle^{\text{obs}}$. The local clustering coefficient of a node captures the fraction of its neighbors that are directly connected, i.e., loosely speaking it counts the fraction of triangles in a neighborhood. $\langle c \rangle^{\text{obs}}$ is the mean of all local clustering coefficients.

From the R&D network, we further obtain the degree distributions $P^{\text{obs}}(d)$, which give the number of collaboration links of firms, to later compare it with our simulation results in Sec. 3.3.

Calculation of link probabilities. To determine the quantities p_l^{NL}, p_{nl}^{NL}, p_d^{L}, p_s^{L}, p_n^{L}, we run agent-based simulations with all possible combinations of values. We take N and E as input, further we sample agent activities a^i and collaboration sizes m from the respective distributions. From each simulation, we construct the respective network on which we calculate the mean values $\langle d \rangle^{\text{sim}}$, $\langle l \rangle^{\text{sim}}$, $\langle c \rangle^{\text{sim}}$. We can then determine the error ϵ from the differences between the observed and the simulated mean values: $\epsilon_d = \left| \langle d \rangle^{\text{obs}} - \langle d \rangle^{\text{sim}} \right|$, $\epsilon_l = \left| \langle l \rangle^{\text{obs}} - \langle l \rangle^{\text{sim}} \right|$, and $\epsilon_c = \left| \langle c \rangle^{\text{obs}} - \langle c \rangle^{\text{sim}} \right|$. We require that these three errors have to be smaller than a threshold ϵ^0. For all probability combinations we perform 25 simulations and select the combination that gives us the highest fraction of networks that match the criterion $\epsilon < \epsilon^0$.

Table 1. Networks for R&D and for co-authorship collaborations (PACS 42): Number of nodes N, links L, collaboration events E. Optimal sets of link probabilities to simulate the collaboration networks are indicated by $*$. The probabilities of a labeled and an unlabeled agent both sum to 1. These two constraints reduce the number of free parameters from 5 to 3.

Collaboration	N	L	E	p_s^{*L}	p_d^{*L}	p_n^{*L}	$p_l^{*\mathrm{NL}}$	$p_{nl}^{*\mathrm{NL}}$
Firms	14,561	21,572	14,829	0.30	0.30	0.40	0.75	0.25
Scientists	27,436	94,961	20,105	0.60	0.05	0.35	0.35	0.65

Optimal simulation values. From our calibration procedure, we have determined for each data set those link probabilities (indicated by $*$) that would generate an *optimal* network in the sense that the expected error between observations and simulations is minimized. We emphasize that this only refers to the *averaged* quantities defined above. We have only used *global information* for the calibration, no specific knowledge about link preferences, etc.

Table 1 summarizes our values obtained, together with some characteristics of the data.

Interpretation of probabilities. We note that the values of the link probabilities allow for an interpretation [29]. For R&D collaborations between firms, for instance, we found that incumbents follow a balanced collaboration strategy. 30% of their collaborations are with agents in the same circle of influence ($p_s^{*L} = 0.3$), 30% with agents in a different circle of influence ($p_d^{*L} = 0.3$) and 40% with newcomers ($p_n^{*L} = 0.4$), represented by non-labeled agents. At the same time, newcomers show a strong tendency to connect to incumbents ($p_l^{*\mathrm{NL}} = 0.75$), as opposed to a low linking probability with other newcomers ($p_{nl}^{*\mathrm{NL}} = 0.25$).

Comparing these values with our findings for co-authorship networks, we note both similar and different tendencies. First, established agents prefer to form links with other established agents ($p_s^{*L} + p_d^{*L} \geq 0.6$). Second, when forming a link with an established agent, the initiator tends to select an agent with the same label ($p_s^{*L} > p_d^{*L}$). This tendency is much more pronounced in co-authorship networks, i.e., to choose a co-author from a different community is less likely than to choose a firm from a different circle of influence. Third, in co-authorship networks newcomers have a stronger tendency to link with other newcomers ($p_{nl}^{*\mathrm{NL}} > p_l^{*\mathrm{NL}}$), while for R&D collaboration the opposite is true: newcomers preferably link to incumbents ($p_l^{*\mathrm{NL}} > p_{nl}^{*\mathrm{NL}}$). This difference could be explained with different entry barriers for publications and patents. Scientists who are newcomers to the publication market can still find opportunities to write papers together, whereas firms that are newcomers to R&D activities may find it more difficult to generate patents.

Validation criteria. We can now use the optimal values for the link probabilities to simulate the evolution of synthetic networks over time. The results, averaged over many runs, are then compared to findings from the empirical network. A mere match between empirics and simulations would not be sufficient to conclude that the

underlying rules of link formation are correct, because we cannot rule out that other sets of rules would lead to equally good results. However, a good agreement lends strong evidence to our agent-based approach, in particular if the *same* model is also able to replicate *different* empirical findings.

In fact, our agent-based model reproduces well the empirical *distributions* of path lengths, clustering coefficients, degrees, component sizes, etc., in two different domains, collaborations between firms and between scientists [29]. Note that the model calibration only uses information about the *mean values* of these distributions, i.e., no complete specification. Reproducing the empirical distributions is remarkable because it means that we are not simply (over)fitting free parameters to available observations. Second, this suggests that the proposed model rules capture the essence of the analyzed collaboration interactions.

3. Dynamics of Network Embeddedness

3.1. *Measuring network embeddedness*

In this paper, we focus on a quantitative measure for network embeddedness, which we take as a benchmark here. To characterize the topological embedding of nodes in a network, various centrality measures have been proposed that are also partially correlated [8]. Because our networks show a clear core–periphery structure, we have introduced a centrality measure for weighted networks, called *coreness* C_C^i [9]. Versions for unweighted networks have been used earlier in social network analysis [2, 22]. In a recent paper [30], we have shown that our coreness measure is well suited to quantify network embeddedness, in particular when compared to other centrality measures. It also strongly correlates with the success of agents, as quantified by non-topological measures such as the number of patents for firms or the number of citations for scientists. Hence, we have argued that our measure of embeddedness can be used to characterize the innovation potential of firms [30].

We do not replicate the argumentation here, but only explain how coreness values for agents are obtained from the reconstructed network. We use the so-called k-core decomposition (see Fig. 2), which *recursively* removes all nodes with a degree less than d from the network, similar to a *cascade*. It 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 are not well integrated in the network, whereas nodes with the largest $k_s = k_s^{\max}$ are in the *core* of the network. The coreness of each node is then defined as $C_C^i = k_s^{\max} - k_s^i$, i.e., it measures the distance from the core. *Low* coreness values characterize the *core*, whereas *high* coreness values characterize the *periphery* of the network.
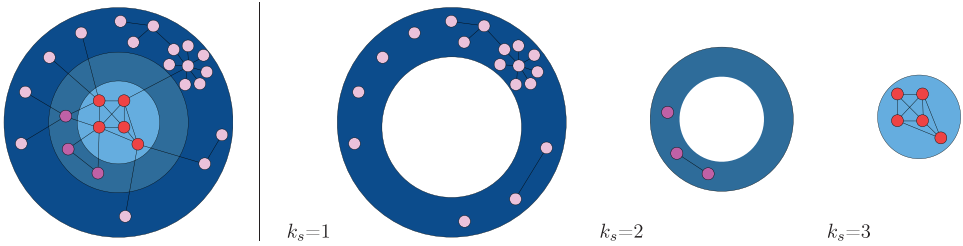
Fig. 2.   Illustration of the weighted *k*-core decomposition.

Instead of the unweighted *k*-core decomposition, in this paper we apply the *weighted k-core decomposition* [9]. It uses the *weighted degree*, d', defined as follows:

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

Here, 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. The free parameters α and β can balance the influence of the weights w^{ij}. Following [30], we set $\alpha = 1$ and $\beta = 0.2$.

3.2. *Empirical dynamics of network embeddedness*

In the following, our focus is on the *evolution* of network embeddedness, as quantified by the coreness values $C_C^i(t)$ of individual agents i. These values change over time either because new collaborations are established that involve agent i or because the network as a whole grows. The latter means that new agents enter the network and form new links to incumbents, i.e., established agents, or to other newcomers. As the result of network growth, new *k*-shells appear, which instantaneously affect the coreness values of all agents, even if they not establish new collaborations.

To compare coreness values at different times, we introduce the *relative* coreness $c^i(t) = C_C^i(t)/C_m(t)$, i.e., the ratio between the current coreness $C_C^i(t)$ and the maximum coreness $C_m(t)$ at the same time. $c^i(t)$ can have values between 0, which indicates the very *core*, and 1, which indicates the outermost *periphery*.

Figure 3 shows two examples, one for a firm, the other one for a scientist, of how these relative coreness values change over time in the respective R&D or co-authorship network. We see that in both cases agents start with high relative coreness values because the collaboration network still has to be established. In the two examples, for both agents the relative coreness values then decline over time, indicating that they become part of the core as the network evolves.

We note that agents do not always keep their network position in the very core ($c^i = 0$), as can be seen for the firm GlaxoSmithKline. It has entered the R&D network in 1990 and reached the minimal distance to the core in 1994. Then, it slowly
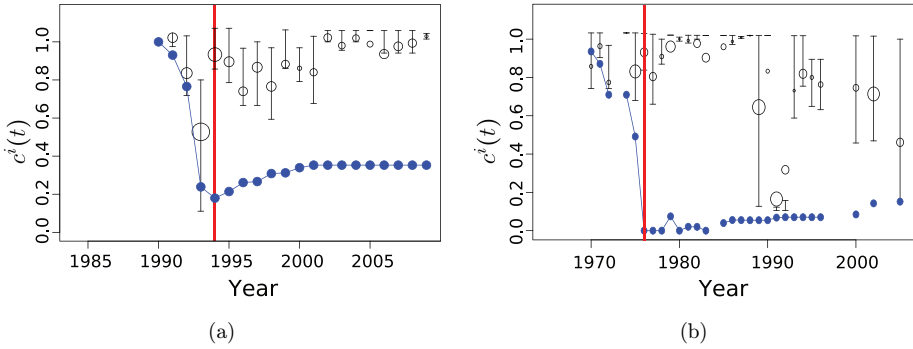
Fig. 3. (Color online) Evolution of relative coreness in blue for (a) Glaxo and (b) Feldmann. The year in which the minimum relative coreness value is marked with a red vertical line. The average relative coreness of the partners of the two selected agents is plotted using black circles. The size of the circle is proportional to the number of partners in that year.

moved away from the core, in particular because *other* firms managed to become better integrated into the core. Only in 2001, the firm reached a stable network position. Hence, for each agent we can identify a minimum relative coreness value $c^i(t^i_c)$ corresponding to the best overall position in the collaboration network and a time t^i_c at which the maximum embeddedness of agent i in the network is obtained. This is indicated by a red line in Fig. 3.

As an additional information, Fig. 3 also shows for the two selected agents their relative coreness together with the relative coreness of their collaboration partners. This allows an interesting observation. In the early period when the two agents are rather new to the network and are thus still part of the periphery, they have a strong tendency to collaborate with partners that have a *comparable* coreness value. This continues until the agents reach their state of minimal coreness, $c^i(t^i_c)$. For times $t > t^i_c$ we observe a change: once agents have reached the core, i.e., are well embedded in the network, they collaborate with partners of *high coreness*, i.e., new-comers or agents from the periphery.

We verify this finding by taking into account all collaborations of firms and of scientists over time. For each agent i, we calculate the relative coreness $c^i(t)$ for every year t and the time t^i_c of minimal coreness. We further calculate the number of collaborations with each of their partners j, i.e., $w^{ij}(t)$, and the total number of collaborations $A^i(t) = \sum_j w^{ij}(t)$ 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*:

$$\langle dc^i(t) \rangle = \frac{1}{A^i(t)} \sum_j w^{ij}(t)[c^i(t) - c^j(t)]. \tag{2}$$

After calculating $\langle dc^i(t) \rangle$ for all times t between t_{start} and t_{end}, i.e., between 1984 and 2009, we divide the values according to two time periods, before and after t^i_c and

average for each of these periods separately:

$$\left\langle dc_{\text{before}}^i \right\rangle = \frac{1}{t_c^i - t_{\text{start}}} \sum_{t=t_{\text{start}}}^{t_c^i} \left\langle dc^i(t) \right\rangle; \quad \left\langle dc_{\text{after}}^i \right\rangle = \frac{1}{t_{\text{end}} - t_c^i} \sum_{t=t_c^i}^{t_{\text{end}}} \left\langle dc^i(t) \right\rangle. \quad (3)$$

For each agent i, these two values are related to the final coreness value of that agent, C_F at t_{end}, i.e., $\left\langle dc_{\text{before}}^i \right\rangle(C_F)$ and $\left\langle dc_{\text{after}}^i \right\rangle(C_F)$. Then, for each value of C_F, e.g., between 0 and 17 for the case of firms, 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 Fig. 4. 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 agent has, on average, a higher coreness than its chosen partners. This applies for initiators with high coreness, i.e., newcomers or agents in the periphery that strive to get a better network position by choosing better integrated partners. Negative values mean that this relation switches: initiating agents 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 agents closer to the core have more collaborations with newcomers or agents 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 agents that have already reached their best network position. This means that established core agents choose even more partners with high coreness (newcomers, periphery) than agents in the period `before` t_c that are still striving for a better network position.

To test the robustness of this finding, we performed a random reshuffling of the collaboration links of firms while preserving the degree sequence of the empirical R&D network. The dashed curves in Fig. 4(a) show the averaged coreness differences
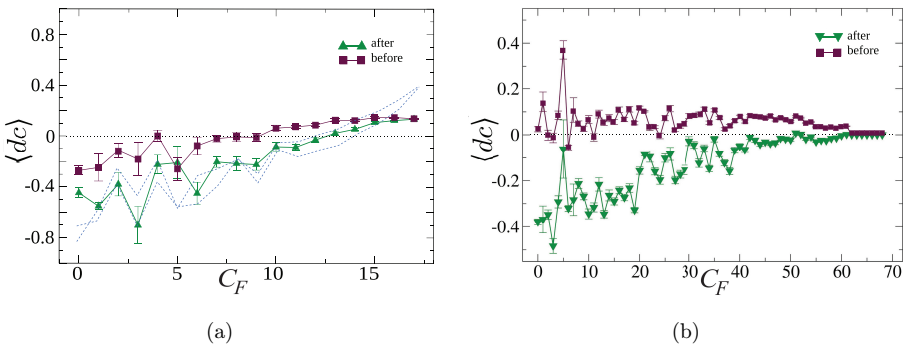

Fig. 4. (Color online) Average partner coreness deviation. Plot of the average normalized partner coreness deviation $\langle dc \rangle$ against C_F before and after t_c. Firms (a) and Scientists (b). The dashed blue lines in panel (a) are obtained by randomly reshuffling the collaboration links of firms while preserving the degree sequence of the empirical R&D network.
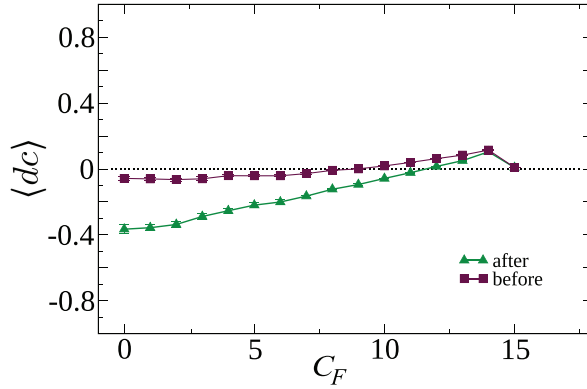
Fig. 5.   Average normalized partner coreness deviation $\langle dc \rangle$ before and after t_c obtained from the agent-based model, using the calibration from the R&D network.

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 performed 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 $p = 0.056$.

3.3. *Validation of the agent-based model*

To validate our agent-based model on the macro level, we have to verify that the dynamics observed in Fig. 4 for firms and for scientists can indeed be replicated by
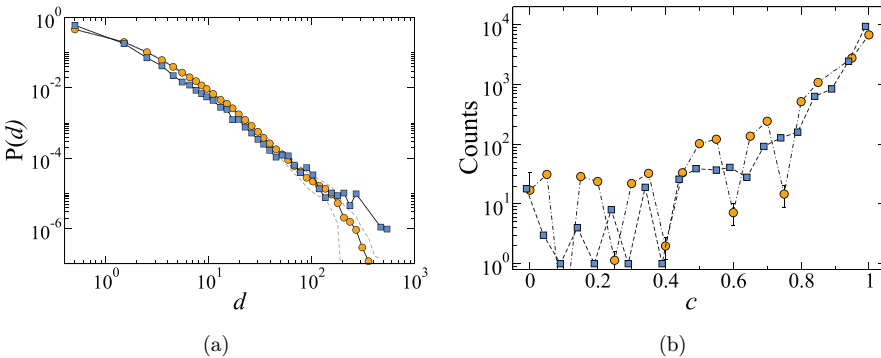

(a)	(b)

Fig. 6.   (Color online) Model results and validation for R&D collaborations of firms: (a) Degree distribution of the R&D network obtained from the agent-based model (blue line) and from the empirical network (circles). (b) Comparison of the coreness distribution for the R&D network obtained from the agent-based model (orange) and from the empirical network (blue). The results are averaged over 100 model realizations and the error bars (when visible) indicate standard errors.

our model. Our main result is presented in Fig. 5 which should be compared to the empirical findings shown in Fig. 4. It demonstrates that our agent-based model generates the same pattern in partner selection that was observed empirically, namely after reaching their lowest coreness value, agents establish collaborations with more peripheral agents.

To demonstrate that the agent-based model is also able to reproduce other empirical findings without overfitting, we plot in Fig. 6(a) the degree distribution of the network ensembles obtained from the 100 realizations, alongside of the empirical degree distribution for the R&D network, both for the final time. The excellent match of the two distributions should be noted. We further plot in Fig. 6(b) the distribution of the coreness values both from the empirical data and from the computer simulations of the R&D network. 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 values, we note that the core–periphery structure of the network is well captured by the model — without any additional assumptions.

3.4. *Improving network embeddedness: Chance or choice?*

Our observations about the impact of network embedding on the partner selection raise the question whether this impact should be interpreted as a change in the *strategy* of an agent in selecting its partners.

Such a change of strategy could indeed have a rational explanation, as follows. Agents new to the network may have little chances to get connected to core agents. Therefore, in the absence of better alternatives, they may eventually team up with other newcomers or agents 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 agents 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 agents more likely search for, and to team up with, newcomers with fresh ideas.

The question is whether the observed change in partner selection indeed follows a *strategy*, i.e., a deliberative process to become more successful, or whether the "strategy" is still the same but opportunities have changed. Then, contrasting the above explanation, one could argue that differences in partner selection are caused by different opportunities to be involved in a collaboration. To decide between these two alternative explanations, we can use our agent-based model because it allows to disentangle strategic behavior from probabilistic actions. Precisely, in our model agents are assigned constant probabilities for linking to newcomers or incumbents. Hence, differences in observed actions are not expressed by the link probabilities, which are the same for all agents, but by the process to become initiators of collaborations, and to be selected within a partner category "same/different/no label".

Our results have demonstrated that this model is able to reproduce the observed change in partner selection together with other topological features, such as degree distribution and coreness distribution. Therefore, we can conclude that the change in choosing partners can be reproduced *without* assuming changes in the selection rules. This does *not* allow to conclude that agents do not follow strategies in selecting their partners, or change these strategies dependent on the network position. But it demonstrates that agents do not need to *change strategies* to act in a way that is observed in their evolution of network embeddedness.

In fact, the abundance of collaboration opportunities is one of the driving forces behind the process of improving network embeddedness. While it is true that agents with a high activity are more often selected, it is also true that the number of newcomers or peripheral agents is much larger than the number of core agents. Whenever newcomers or peripheral agents are activated to become initiators, they follow the preferential selection rule, i.e., they tend to select partner agents with a larger degree. These are most likely agents from the core, which are well embedded in the network. Because these agents are more often selected for collaboration, their degree increase over time which further increases their probability to be selected, next time. A higher degree, on the other hand, relates to a lower coreness value as outlined above, albeit not in a linear manner. It is this feedback process that eventually helps some agents to further improve their network embeddedness, whereas the majority of agents still stay in the periphery.

4. Discussion

In this paper we have focused on the dynamics of collaboration networks, using two data sets from different domains, about R&D collaborations between firms and about co-authorship collaborations between scientists. To answer our initial question, whether these different collaboration networks can be characterized and modeled from a unifying perspective, we made two contributions.

First, we proposed a new measure for network embeddedness, the (relative) coreness value $c^i(t)$, to compare the topological positions of individual agents. We would like to emphasize that our measure of network embeddedness is also a good predictor of success of individual agents. Looking at the R&D collaboration network of firms and their corresponding patent data [30], we verified that for the successful firms a better coreness comes along with more patents whereas for the "normal" firms both the position and the number of patents is rather level (in comparison to the successful firms).

Monitoring the change of individual coreness values, we found in both data sets the same dynamics (shown in Fig. 4). Some agents over time have improved their network embeddedness by moving from the *periphery* of the collaboration network close to the *core*, i.e., from high to low coreness values. From the data, we obtained a time t_c^i for each agent when the minimal coreness, i.e., the best network

embeddedness, is reached. At about t_c^i we observed a *change* in partner selection, from partners of *similar* coreness to partners of *different* (i.e., high) coreness.

To explain this seemingly strategic behavior was one aim of our agent-based model, which is our second contribution. We utilized a stochastic label propagation model with five probabilities to form collaboration links between newcomers and established agents with similar or different labels. These probabilities could be calibrated using aggregated quantities from the respective empirical collaboration networks. As the result, we obtained a *data-driven* agent-based model, where domain specific information is included in interpretable link probabilities. Additionally, we implemented a hypothesis of how agents choose collaboration partners *within* the three partner categories same/different/no label, namely by preferentially choosing agents with a higher degree.

Our model assumes that link formation is unilateral, i.e., mutual consensus is not modeled. This shortcoming is discussed in [27]. Further, our model does not include the termination of collaborations or the exit of agents because the data sets do not contain the respective information. As a consequence, the model may overestimate the number of active agents. To assess the impact on the maximum k-core value, let us assume that the number of active partners in the R&D collaboration network is proportional to the size of a firm. It was observed [5] that firms with a high degree less likely disappear. Hence, we expect that $\langle dc_{\text{after}} \rangle$ is less negative, as links to firms with low degree (which implies low coreness) are more likely to disappear. In other words, the distance between $\langle dc_{\text{before}} \rangle$ and $\langle dc_{\text{after}} \rangle$ could be smaller than estimated. A similar argument applies to the model of scientific collaborations.

Despite these simplifying modeling assumptions we could demonstrate that the model is able to explain the observed impact of network embeddedness on the selection of collaboration partners. In a nutshell, there exists a feedback between an agent's activity on the one hand and its ability to increase the degree by establishing new collaborations and to propagate the own label to newcomers, on the other hand. More collaborations not only lead to higher degrees, but also to lower coreness values, i.e., agents become embedded in the core of the network. The chance to propagate the own label later increases the chance to be selected for new collaborations, because empirics has shown that firms or scientists prefer to collaborate with partners with the same label.

These combined effects eventually explain the observation that core agents tend to collaborate preferably with newcomers or agents from the periphery. In fact, newcomers and peripheral agents choose these core agents with larger probability, once they managed to establish their strong network embeddedness. Thus, in conclusion, what seems to be a deliberative strategy of successful agents, namely to switch their rules of partner selection, can be basically explained without strategic considerations. These cannot be excluded, but the model suggests that the empirical observations do not already imply such considerations. Hence, the emergence of realistic core–periphery structures in collaboration networks can be successfully modeled without deliberative agents.

Acknowledgments

Author Contributions

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

References

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

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

[3] Couzin, I. D., Krause, J. *et al.*, Self-organization and collective behavior in vertebrates, *Adv. Study Behav.* **32**(1) (2003) 1–75.

[4] Deisboeck, T. S. and Couzin, I. D., Collective behavior in cancer cell populations, *Bioessays* **31**(2) (2009) 190–197.

[5] Dunne, P. and Hughes, A., Age, size, growth and survival: Uk companies in the 1980s. *J. Ind. Econ.* **42**(2) (1994) 115–140.

[6] Ebeling, W. and Schweitzer, F., Self-organization, active brownian dynamics, and biological applications, *Nova Acta Leopold.* **88**(332) (2003) 169–188.

[7] Erban, R. and Othmer, H. G., From individual to collective behavior in bacterial chemotaxis, *SIAM J. Appl. Math.* **65**(2) (2004) 361–391.

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

[9] Garas, A., Schweitzer, F. and Havlin, S., A k-shell decomposition method for weighted networks, *New J. Phys.* **14**(8) (2012) 083030.

[10] Goyal, S. and Moraga-González, J. L., R&d networks, *RAND J. Econ.* **32**(4) (2001) 686–707.

[11] Hecht, I., El, Y. B., Balmer, F., Natan, S., Tsarfaty, I., Schweitzer, F. and Jacob, E. B., Tumor invasion optimization by mesenchymal-amoeboid heterogeneity, *Sci. Rep.* **5** (2015) 10622.

[12] Jackson, M. O. and Wolinsky, A., A strategic model of social and economic networks, *J. Econ. Theory* **71**(1) (1996) 44–74.

[13] Jeanson, R. and Deneubourg, J.-L., Conspecific attraction and shelter selection in gregarious insects, *Am. Naturalist* **170**(1) (2007) 47–58.

[14] Kerth, G., Perony, N. and Schweitzer, F., Bats are able to maintain long-term social relationships despite the high fission-fusion dynamics of their groups, *Proc. R. Soc. B: Biological Sci.* **278**(1719) (2011) 2761–2767.

[15] Koenig, M. D., Battiston, S., Napoletano, M. and Schweitzer, F., Recombinant knowledge and the evolution of innovation network, *J. Econ. Behav. Organ.* **79**(3) (2011) 145–164.

[16] Koenig, M. D., Battiston, S., Napoletano, M. and Schweitzer, F., The efficiency and stability of r&d networks, *Games Econ. Behav.* **75**(2) (2012) 694–713.

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

[18] Otjes, S. and Rasmussen, A., The collaboration between interest groups and political parties in multi-party democracies: Party system dynamics and the effect of power and ideology, *Party Polit.* **23**(2) (2017) 96–109.

[19] Sarigol, E., Pfitzner, R., Scholtes, I., Garas, A. and Schweitzer, F., Predicting scientific success based on coauthorship networks, *EPJ Data Sci.* **3** (2014) 9.

[20] Schweitzer, F., Fagiolo, G., Sornette, D., Redondo, F. V. and White, D. R., Economic networks: What do we know and what do we need to know? *Adv. Complex Syst.* **12**(4) (2009) 407–422.

[21] Schweitzer, F., Lao, K. and Family, F., Active random walkers simulate trunk trail formation by ants, *Biosystems* **41**(3) (1997) 153–166.

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

[23] Suzuki, N., Kamiya, T., Umata, I., Ito, S. and Iwasawa, S., Analyzing the structure of the emergent division of labor in multiparty collaboration, in *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (2012), pp. 1233–1236.

[24] Takeishi, A., Knowledge partitioning in the interfirm division of labor: The case of automotive product development, *Organ. Sci.* **13**(3) (2002) 321–338.

[25] Tomasello, M. V., Napoletano, M., Garas, A. and Schweitzer, F., The rise and fall of R&D networks, *Ind. Corporate Change* **26**(4) (2016) 617–646.

[26] Tomasello, M. V., Perra, N., Tessone, C. J., Karsai, M. and Schweitzer, F., The role of endogenous and exogenous mechanisms in the formation of r&d networks, *Sci. Rep.* **4** (2014) 5679.

[27] Tomasello, M. V., Burkholz, R. and Schweitzer, F., Modeling the formation of r&d alliances: An agent-based model with empirical validation, Economics Discussion Papers 2017-107, Kiel Institute for the World Economy (IfW Kiel) (2017).

[28] Tomasello, M. V., Tessone, C. J. and Schweitzer, F., A model of dynamic rewiring and knowledge exchange in r&d networks, *Adv. Complex Syst.* **19**(1–2) (2016) 1650004.

[29] Tomasello, M. V., Vaccario, G. and Schweitzer, F., Data-driven modeling of collaboration networks: A cross-domain analysis, *EPJ Data Sci.* **6** (2017) 22.

[30] Vaccario, G., Verginer, L., Garas, A., Tomasello, M. V. and Schweitzer, F., Network embeddedness of firms predicts their innovation potential, submitted for publication (2022).

[31] Van Dyke, N. and McCammon, H. J., *Strategic Alliances: Coalition Building and Social Movements* (University of Minnesota Press, 2010).

[32] Zingg, C., Nanumyan, V. and Schweitzer, F., Citations driven by social connections? A multi-layer representation of coauthorship networks, *Quant. Sci. Stud.* **1**(4) (2020) 1493–1509.